Biometrika (2014), **101**, 1, *pp.* 103–120
Printed in Great Britain

doi: 10.1093/biomet/ast059
Advance Access publication 12 February 2014

Sparse precision matrix estimation via lasso penalized D-trace loss

By TENG ZHANG

Department of Mathematics, Princeton University, Fine Hall, Washington Rd, Princeton, New Jersey 08544, U.S.A.
tengz@princeton.edu

and HUI ZOU

School of Statistics, University of Minnesota, 224 Church St SE, Minneapolis, Minnesota 55455, U.S.A.
zouxx019@umn.edu

Summary

We introduce a constrained empirical loss minimization framework for estimating high-dimensional sparse precision matrices and propose a new loss function, called the D-trace loss, for that purpose. A novel sparse precision matrix estimator is defined as the minimizer of the lasso penalized D-trace loss under a positive-definiteness constraint. Under a new irrepresentability condition, the lasso penalized D-trace estimator is shown to have the sparse recovery property. Examples demonstrate that the new condition can hold in situations where the irrepresentability condition for the lasso penalized Gaussian likelihood estimator fails. We establish rates of convergence for the new estimator in the elementwise maximum, Frobenius and operator norms. We develop a very efficient algorithm based on alternating direction methods for computing the proposed estimator. Simulated and real data are used to demonstrate the computational efficiency of our algorithm and the finite-sample performance of the new estimator. The lasso penalized D-trace estimator is found to compare favourably with the lasso penalized Gaussian likelihood estimator.

Some key words: Constrained minimization; D-trace loss; Graphical lasso; Graphical model selection; Precision matrix; Rate of convergence.

1. Introduction

Assume that we have n independent and identically distributed p-dimensional random variables. Let Σ^* be the population covariance matrix and let $\Theta^* = (\Sigma^*)^{-1}$ denote the corresponding precision matrix. Massive high-dimensional data frequently arise in computational biology, medical imaging, genomics, climate studies, finance and other fields, and it is of both theoretical and practical importance to estimate high-dimensional covariance or precision matrices. In this paper we focus on estimating a sparse precision matrix Θ^* when the dimension is large. Sparsity in Θ^* is interesting because each nonzero entry of Θ^* corresponds to an edge in a Gaussian graphical model for describing the conditional dependence structure of the observed variables (Whittaker, 1990). Specifically, if $x \sim N_p(\mu, \Sigma^*)$, then $\Theta^*_{ij} = 0$ if and only if $x_i \perp\!\!\!\perp x_j \mid \{x_k : k \neq i, j\}$. The construction of Gaussian graphical models has applications in a wide range of fields, including genomics, image analysis and macroeconomics (Li & Gui, 2006; Wille & Bühlmann, 2006; Dobra et al., 2009; Li, 2009). Meinshausen & Bühlmann (2006) proposed a neighbourhood selection scheme in which one can sequentially estimate the support of each row of Θ^* by fitting an ℓ_1 or lasso penalized least squares regression model (Tibshirani, 1996). Yuan (2010) used the Dantzig selector (Candès & Tao, 2007) to replace the lasso penalized least squares in the neighbourhood selection scheme. Peng et al. (2009) proposed a joint neighbourhood estimator using the lasso penalization. Cai et al. (2011) proposed a constrained ℓ_1 minimization estimator for estimating sparse precision matrices and established its convergence rates under the elementwise ℓ_∞ norm and Frobenius norm. A common drawback of the methods mentioned above is that they do not always guarantee that the final estimator is positive definite. One can also use Cholesky decomposition to estimate the precision or covariance matrix, as in Huang et al. (2006). With this approach, a sparse regularized estimator of the Cholesky factor is first derived and then the estimated Cholesky factor is used to construct the final estimator of Θ^*. The regularized Cholesky decomposition approach always gives a positive-semidefinite matrix but does not necessarily produce a sparse estimator of Θ^*.

To the best of our knowledge, the only existing method for deriving a positive-definite sparse precision matrix is via the lasso or ℓ_1 penalized Gaussian likelihood estimator or its variants. Yuan & Lin (2007) proposed the lasso penalized likelihood criterion and suggested using the maxdet algorithm to compute the estimator. Motivated by Banerjee et al. (2008), Friedman et al. (2008) developed a blockwise coordinate descent algorithm, called the graphical lasso, for solving the lasso penalized Gaussian likelihood estimator. Witten et al. (2011) presented some computational tricks to further boost the efficiency of the graphical lasso. In the literature, the graphical lasso is often used as an alternative name for the lasso penalized Gaussian likelihood estimator. Convergence rates for the graphical lasso have been established by Rothman et al. (2008) and Ravikumar et al. (2011).

The graphical lasso estimator is outside the penalized maximum likelihood estimation paradigm, as it works for non-Gaussian data (Ravikumar et al., 2011). To gain a better understanding, we propose a constrained convex optimization framework for estimating large precision matrices, within which the graphical lasso can be viewed as a special case. We further introduce a new loss function, the D-trace loss, which is convex and minimized at Θ^{-1}. We define a novel estimator as the minimizer of the lasso penalized D-trace loss under the constraint that the solution be positive definite. The D-trace loss is much simpler than the graphical lasso loss, thus permitting a more direct theoretical analysis and offering significant computational advantages. Under a new irrepresentability condition, we prove the sparse recovery property of the new estimator and show through examples that our irrepresentability condition is satisfied while that for the graphical lasso fails. Asymptotically, the new estimator and the graphical lasso have comparable rates of convergence in the elementwise maximum, Frobenius and operator norms. Through simulation, we show that in finite samples the new estimator outperforms the graphical lasso, even when the data are generated from Gaussian distributions.

2. METHODOLOGY

2·1. *An empirical loss minimization framework*

We begin with some notation and definitions. For a $p \times p$ matrix $X = (X_{i,j}) \in \mathbb{R}^{p\times p}$, its Frobenius norm is $\|X\|_{\rm F} = (\sum_{i,j} X_{i,j}^2)^{1/2}$. We also use $\|X\|_{1,\rm off}$ to denote the off-diagonal ℓ_1 norm: $\|X\|_{1,\rm off} = \sum_{i \neq j} |X_{i,j}|$. Let $\mathcal{S}(p)$ denote the space of all $p \times p$ positive-definite matrices. For any two symmetric matrices $X, Y \in \mathbb{R}^{p\times p}$, we write $X \succcurlyeq Y$ when $X - Y$ is positive semidefinite. We use $\text{vec}(X)$ to denote the p^2-vector formed by stacking the columns of X, and $\langle X, Y\rangle$ means $\text{tr}(XY^{\mathrm{T}})$ throughout the paper.

Suppose that we want to use a Θ from $\mathcal{S}(p)$ to estimate $(\Sigma_0)^{-1}$. We use a loss function $L(\Theta, \Sigma_0)$ for this estimation problem, and we require it to satisfy the following two conditions.

Condition 1. The loss function $L(\Theta, \Sigma_0)$ is a smooth convex function of Θ.

Condition 2. The unique minimizer of $L(\Theta, \Sigma_0)$ occurs at $(\Sigma_0)^{-1}$.

Condition 1 is required for computational reasons, and Condition 2 is needed so that we get the desired precision matrix when the loss function $L(\Theta, \Sigma_0)$ is minimized. It is also important that $L(\Theta, \Sigma_0)$ be constructed directly through Σ_0, not $(\Sigma_0)^{-1}$, because in practice we need to use its empirical version $L(\Theta, \hat{\Sigma}_0)$, where $\hat{\Sigma}_0$ is an estimate of Σ_0, to compute the estimator of $(\Sigma_0)^{-1}$. With such a loss function in hand, we can construct a sparse estimator of $(\Sigma_0)^{-1}$ via the convex program

$$\mathop{\arg\min}_{\Theta\in\mathcal{S}(p)} L(\Theta, \hat{\Sigma}) + \lambda_n\|\Theta\|_{1,\text{off}}, \tag{1}$$

where $\hat{\Sigma}$ denotes the sample covariance matrix and $\lambda_n > 0$ is the ℓ_1 penalization parameter.

The graphical lasso can be seen as an application of the empirical loss minimization framework, defined as

$$\mathop{\arg\min}_{\Theta\in\mathcal{S}(p)} \langle\Theta, \hat{\Sigma}\rangle - \log\det(\Theta) + \lambda_n\|\Theta\|_{1,\text{off}}. \tag{2}$$

Yuan & Lin (2007) proposed this estimator by following the penalized maximum likelihood estimation paradigm: $\langle\Theta, \hat{\Sigma}\rangle - \log\det(\Theta)$ corresponds to the negative loglikelihood function of the multivariate Gaussian model. Comparing (2) to (1), we see that the graphical lasso is an empirical loss minimizer where the loss function is $L_{\mathrm{G}}(\Theta, \Sigma_0) = \langle\Theta, \Sigma_0\rangle - \log\det(\Theta)$. One can verify that $L_{\mathrm{G}}(\Theta, \Sigma_0)$ satisfies Conditions 1 and 2. Although $L_{\mathrm{G}}(\Theta, \Sigma_0)$ has dual interpretations, it has been shown that the graphical lasso provides a consistent estimator even when the data do not follow a multivariate Gaussian distribution (Ravikumar et al., 2011). Thus, the empirical loss minimization view of the graphical lasso is more fundamental and can better explain its broader successes with non-Gaussian data.

2·2. *A new estimator*

From the empirical loss minimization viewpoint, L_{G} is not the most natural and convenient loss function for precision matrix estimation because of the log-determinant term. We show in this paper that there is a much simpler loss function than L_{G} for estimating sparse precision matrices. The new loss function is

$$L_{\mathrm{D}}(\Theta, \Sigma_0) = \frac{1}{2}\langle\Theta^2, \Sigma_0\rangle - \text{tr}(\Theta). \tag{3}$$

As L_{D} is expressed as the difference of two trace operators, we call it the D-trace loss. We first verify that L_{D} satisfies the two conditions above. To check Condition 1, observe that

$$L_{\mathrm{D}}(\Theta_1, \Sigma_0) + L_{\mathrm{D}}(\Theta_2, \Sigma_0) - 2L_{\mathrm{D}}\left\{\frac{1}{2}(\Theta_1+\Theta_2),\ \Sigma_0\right\} = 2\left\langle\left(\frac{\Theta_1-\Theta_2}{2}\right)^2,\ \Sigma_0\right\rangle \geqslant 0.$$

To check Condition 2, we show that the derivative of (3) is $(\Theta\Sigma_0 + \Sigma_0\Theta)/2 - I$ and that the Hessian of L_{D} can be expressed as $(\Sigma_0 \otimes I + I \otimes \Sigma_0)/2$, where $\otimes$ denotes the Kronecker product. Since Σ_0 is positive definite, the Hessian has only positive eigenvalues (see, e.g., Pease, 1965, § XIV.7) and so is positive definite. It is then easy to see that $\Theta = \Sigma_0^{-1}$ is the unique minimizer of $L_{\mathrm{D}}(\Theta_1, \Sigma_0)$ as a function of Θ.

We have verified that the D-trace loss is a valid loss function to be used in the empirical loss minimization framework. The corresponding estimator is then defined according to (1):

$$\hat{\Theta} = \arg\min_{\Theta \in \mathcal{S}(p)} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_{1,\mathrm{off}}, \tag{4}$$

where λ_n is a nonnegative penalization parameter. One can also use the ℓ_1 norm, $\|\Theta\|_1 = \sum_{i \neq j} |\Theta_{i,j}|$, in (4). In many applications we know a priori that the smallest eigenvalue of the true precision matrix is at least ϵ, where ϵ is a certain threshold. We can easily incorporate this into the estimator by considering

$$\hat{\Theta} = \arg\min_{\Theta \succeq \epsilon I} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_{1,\mathrm{off}}. \tag{5}$$

In § 3 we derive an efficient algorithm for solving (5), setting $\epsilon = 10^{-8}$ as the default.

From a computational point of view, L_{D} is more convenient than L_{G}. We can view the D-trace loss as an analogue of the least squares loss, used in regression, for precision matrix estimation. It is difficult to come up with a simpler loss function than L_{D} that satisfies Conditions 1 and 2. One might argue that L_{G} should be the optimal loss function at least for estimating Θ^* for Gaussian distributions, owing to its likelihood interpretation. However, the conventional wisdom does not necessarily hold true in the empirical loss minimization framework for precision matrix estimation. For simplicity, let $\lambda_n = 0$ and compare the minimizer of the empirical loss with the maximum likelihood estimator when Σ^{-1} exists. Then we see that if the loss function satisfies Conditions 1 and 2, the solution in (1) is always Σ^{-1}, regardless of the actual form of the loss function. This is different from what happens in conventional regression problems, where unpenalized loss functions produce different estimates, such as in the case of Huber's regression versus least squares. In the rest of the paper we study the theoretical and numerical properties of the lasso penalized D-trace loss estimator for estimating sparse precision matrices. We have found that the new estimator enjoys theoretical and empirical advantages over the lasso penalized Gaussian likelihood estimator.

Our estimator has an interesting connection to the constrained ℓ_1 minimization estimator (Cai et al., 2011) defined through

$$\text{minimize} \sum_{i,j}^{p} |\Theta_{ij}| \quad \text{subject to} \quad \max_{i,j} |\hat{\Sigma}\Theta - I| \leqslant \lambda_n. \tag{6}$$

Cai et al. (2011) regularized the diagonal elements of Θ. To simplify the discussion, we can do the same for our estimator by using $\|\Theta\|_1$ in (4); then the penalized L_{D} estimator is

$$\arg\min_{\Theta \in \mathcal{S}(p)} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_1. \tag{7}$$

The solution of (7) satisfies

$$\frac{1}{2}(\Theta\hat{\Sigma} + \hat{\Sigma}\Theta) - I = \lambda_n \hat{Z}, \tag{8}$$

where $\hat{Z}$ represents the subgradient, taking values in $[-1, 1]$. Therefore, following the derivation of the Dantzig selector (Candès & Tao, 2007), we can relax (8) and drop the positive-definiteness constraint to define a constrained minimization estimator through

$$\text{minimize} \sum_{i,j}^{p} |\Theta_{ij}| \quad \text{subject to} \quad \max_{i,j} \left| \frac{1}{2}(\Theta\hat{\Sigma} + \hat{\Sigma}\Theta) - I \right| \leqslant \lambda_n. \tag{9}$$

Comparing (9) and (6), we see that the Dantzig version of the penalized L_{D} estimator is very similar to the estimator of Cai et al. (2011). An important difference between (9) and (6) is that the solution of (9) is guaranteed to be symmetric, which is not the case for (6).

A referee called our attention to an unpublished manuscript by Liu and Luo, available at http://arxiv.org/abs/1203.3896. Let θ_k be the kth column vector of Θ, and let e_k denote a p-dimensional vector with 1 in the kth coordinate and 0 in all other coordinates. Liu and Luo's estimator is motivated by the fact that the constrained ℓ_1 minimization estimator in (6) has the following equivalent formulation:

$$\text{minimize } |\theta_k|_1 \quad \text{subject to} \quad |\hat{\Sigma}\theta_k - e_k|_\infty \leqslant \lambda_n \quad (k = 1, \ldots, p). \tag{10}$$

See Lemma 1 of Cai et al. (2011). Liu and Luo's estimator of θ_k is defined by

$$\arg\min_{\theta_k} \frac{1}{2}\theta_k^{\mathrm{T}}\hat{\Sigma}\theta_k - e_k^{\mathrm{T}}\theta_k + \lambda_n|\theta_k|_1. \tag{11}$$

Liu and Luo used a reverse Dantzig selector step to get (11) from (10). A major advantage of doing so is that solving (11) can be computationally more efficient than solving (10). On the other hand, the penalized L_{D} estimator in (7) can be rewritten as

$$\mathop{\arg\min}_{\Theta=[\theta_1,\ldots,\theta_p]\in\mathcal{S}(p)} \sum_{k=1}^{p} \left(\frac{1}{2}\theta_k^{\mathrm{T}}\hat{\Sigma}\theta_k - e_k^{\mathrm{T}}\theta_k + \lambda_n|\theta_k|_1 \right). \tag{12}$$

Therefore, if we drop the positive-definiteness constraint, (12) reduces to solving (11) for $k = 1, \ldots, p$. The fundamental difference between our estimator and Liu and Luo's estimator is that our method respects the positive-definite nature of a precision matrix, while Liu and Luo's method treats a precision matrix estimation problem as p separate vector estimation problems; their estimator is not even guaranteed to be symmetric.

3. Algorithm

3·1. *Architecture of the algorithm based on the alternating direction method*

In this section we develop an efficient algorithm for solving the constrained optimization problem in (5), based on the alternating direction method. Before delving into the technical details, it is interesting to first review the efforts that have been devoted to solving the lasso penalized Gaussian likelihood estimator. Yuan & Lin (2007) used the maxdet algorithm to compute the lasso penalized Gaussian likelihood estimator, but that algorithm is very slow for high-dimensional data. Banerjee et al. (2008) and Friedman et al. (2008) developed blockwise descent algorithms. Duchi et al. (2008) proposed a projected gradient method, and Lu (2009) proposed a method that involves applying Nesterov's smooth optimization technique; in both these papers the authors showed that their algorithms perform faster than blockwise descent algorithms. More recently, Scheinberg et al. (2010) developed an alternating direction method for solving the lasso penalized Gaussian likelihood estimator and showed that their method is faster than the projected gradient method (Duchi et al., 2008) as well as Nesterov's smooth optimization method (Lu, 2009). Based on previous work, the alternating direction method is the state-of-the-art algorithm for solving the lasso penalized Gaussian likelihood estimator. In order to compare the D-trace loss and Gaussian likelihood function in computational terms, we derive an alternating direction method for solving the lasso penalized D-trace estimator and compare the computational efficiency of the lasso penalized D-trace loss with that of the Gaussian likelihood estimators, showing that the new estimator is faster.

We introduce two new matrices, Θ_0 and Θ_1, and rewrite (4) as

$$\mathop{\arg\min}_{\Theta_1 \succeq \epsilon I} \frac{1}{2}\langle \Theta^2, \hat{\Sigma}\rangle - \operatorname{tr}(\Theta) + \lambda_n \|\Theta_0\|_{1,\mathrm{off}} \quad \text{subject to } [\Theta, \Theta] = [\Theta_0, \Theta_1]. \tag{13}$$

From (13), we consider the augmented Lagrangian

$$\begin{aligned} L(\Theta, \Theta_0, \Theta_1, \Lambda_0, \Lambda_1) = {} & \frac{1}{2}\big\langle \Theta^2, \hat{\Sigma}\big\rangle - \operatorname{tr}(\Theta) + \lambda_n \|\Theta_0\|_{1,\mathrm{off}} + h(\Theta_1 \succeq \epsilon I) \\ & + \langle \Lambda_0, \Theta - \Theta_0\rangle + \langle \Lambda_1, \Theta - \Theta_1\rangle \\ & + (\rho/2)\|\Theta - \Theta_0\|_{\mathrm{F}}^2 + (\rho/2)\|\Theta - \Theta_1\|_{\mathrm{F}}^2, \end{aligned}$$

where $h(\Theta_1 \succeq \epsilon I)$ is an indicator function defined by

$$h(\Theta_1 \succeq \epsilon I) = \begin{cases} 0, & \Theta_1 \succeq \epsilon I; \\ \infty, & \text{otherwise.} \end{cases}$$

Let $(\Theta^k, \Theta_0^k, \Theta_1^k, \Lambda_0^k, \Lambda_1^k)$ be the solution at step k, for $k = 0, 1, 2, \ldots$. We update $(\Theta, \Theta_0, \Lambda)$ according to

$$\Theta^{k+1} = \mathop{\arg\min}_{\Theta = \Theta^{\mathrm{T}}} L(\Theta, \Theta_0^k, \Theta_1^k, \Lambda_0^k, \Lambda_1^k), \tag{14}$$

$$[\Theta_0^{k+1}, \Theta_1^{k+1}] = \mathop{\arg\min}_{\Theta_0 = \Theta_0^{\mathrm{T}}, \Theta_1 \succeq \epsilon I} L(\Theta^{k+1}, \Theta_0, \Theta_1, \Lambda_0^k, \Lambda_1^k), \tag{15}$$

$$[\Lambda_0^{k+1}, \Lambda_1^{k+1}] = [\Lambda_0^k, \Lambda_1^k] + \rho[\Theta^{k+1} - \Theta_0^{k+1}, \Theta^{k+1} - \Theta_1^{k+1}]. \tag{16}$$

Step (16) is trivial. For (14), we can write

$$\Theta^{k+1} = \mathop{\arg\min}_{\Theta = \Theta^{\mathrm{T}}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} + 2\rho I\rangle - \langle \Theta, I + \rho\Theta_0^k + \rho\Theta_1^k - \Lambda_0^k - \Lambda_1^k\rangle.$$

Let $G(A, B)$ denote the solution to the optimization problem

$$\mathop{\arg\min}_{\Theta = \Theta^{\mathrm{T}}} \frac{\rho}{2}\langle \Theta^2, A\rangle - \langle \Theta, B\rangle, \qquad A > 0. \tag{17}$$

Then we can write

$$\Theta^{k+1} = G(\hat{\Sigma} + 2\rho I, I + \rho\Theta_0^k + \rho\Theta_1^k - \Lambda_0^k - \Lambda_1^k). \tag{18}$$

The explicit solution to (17) is given in the following theorem.

THEOREM 1. *Given any p-dimensional symmetric positive-definite matrix A and any p-dimensional matrix B, let $A = U_A \Sigma_A U_A^{\mathrm{T}}$ be the eigenvalue decomposition of A, with ordered eigenvalues $\sigma_1 \geqslant \cdots \geqslant \sigma_p$. Define*

$$G(A, B) = \underset{\Theta=\Theta^{\mathrm{T}}}{\arg\min} \frac{1}{2}\langle \Theta^2, A\rangle - \langle B, \Theta\rangle.$$

Then

$$G(A, B) = U_A \left\{ \left(U_A^{\mathrm{T}} B U_A \right) \circ C \right\} U_A^{\mathrm{T}}, \tag{19}$$

where $\circ$ denotes the Hadamard product of matrices and $C_{i,j} = 2/(\sigma_i + \sigma_j)$.

To update Θ_0^{k+1}, from (15) we write

$$\Theta_0^{k+1} = \underset{\Theta_0=\Theta_0^{\mathrm{T}}}{\arg\min} \frac{\rho}{2}\langle \Theta_0^2, I\rangle - \langle \Theta_0, \rho\Theta^{k+1} + \Lambda_0^k\rangle + \lambda_n \|\Theta_0\|_{1,\mathrm{off}}.$$

Let $\mathrm{S}(A, \lambda)$ denote the solution to the optimization problem

$$\underset{\Theta_0=\Theta_0^{\mathrm{T}}}{\arg\min} \frac{1}{2}\langle \Theta_0^2, I\rangle - \langle \Theta_0, A\rangle + \Lambda_0 \|\Theta_0\|_{1,\mathrm{off}}.$$

Then we can write

$$\Theta_0^{k+1} = \mathrm{S}\left(\Theta^{k+1} + \frac{1}{\rho}\Lambda_0^k, \frac{\lambda_n}{\rho} \right), \tag{20}$$

where the operator S is defined by

$$\mathrm{S}(A, \lambda)_{i,j} = \begin{cases} A_{i,j}, & i = j, \\ A_{i,j} - \lambda, & i \neq j,\ A_{i,j} > \lambda, \\ A_{i,j} + \lambda, & i \neq j,\ A_{i,j} < -\lambda, \\ 0, & i \neq j,\ -\lambda \leqslant A_{i,j} \leqslant \lambda. \end{cases}$$

To update Θ_1^{k+1}, we write

$$\Theta_1^{k+1} = \underset{\Theta_1 \succeq \epsilon I}{\arg\min} \frac{\rho}{2}\langle \Theta_1^2, I\rangle - \langle \Theta_1, \rho\Theta^{k+1} + \Lambda_1^k\rangle. \tag{21}$$

For a symmetric matrix X we define the matrix operator $[X]_+$ as follows: let the eigenvalue decomposition of X be $U_X \operatorname{diag}(\lambda_1, \ldots, \lambda_p) U_X^{\mathrm{T}}$; then

$$[X]_+ = U_X \operatorname{diag}\left\{ \max(\lambda_1, \epsilon), \ldots, \max(\lambda_p, \epsilon) \right\} U_X^{\mathrm{T}}.$$

The solution to (21) is then

$$\Theta_1^{k+1} = \left[\Theta^{k+1} + \frac{\Lambda_1^k}{\rho}\right]_+ .$$

We now have all the pieces needed to carry out the alternating direction method for solving (5). Algorithm 1 summarizes the details.

Algorithm 1. Alternating direction method for solving (5).

Step 1. Initialization: $k = 0,\ \Lambda^0,\ \Theta_0^0 = \Theta_1^0$.

Step 2. Repeat (a)–(d) until convergence:

(a) $k = k + 1$;
(b) use Theorem 1 to compute $G(\hat{\Sigma} + 2\rho I, I + \rho\Theta_0^k + \rho\Theta_1^k - \Lambda_0^k - \Lambda_1^k)$, and set

$$\Theta^{k+1} = G(\hat{\Sigma} + 2\rho I, I + \rho\Theta_0^k + \rho\Theta_1^k - \Lambda_0^k - \Lambda_1^k);$$

(c) let $\Theta_0^{k+1} = \mathrm{S}(\Theta^{k+1} + \Lambda_0^k/\rho, \lambda_n/\rho)$ and $\Theta_1^{k+1} = [\Theta^{k+1} + \Lambda_1^k/\rho]_+$;
(d) let $\Lambda_0^{k+1} = \Lambda_0^k + \rho(\Theta^{k+1} - \Theta_0^{k+1})$ and $\Lambda_1^{k+1} = \Lambda_1^k + \rho(\Theta^{k+1} - \Theta_1^{k+1})$.

3·2. *Implementation*

Here we discuss the implementation details for Algorithm 1. The most computationally expensive part is the update of Θ_1^{k+1}, owing to the eigenvalue constraint. If we drop that constraint and consider

$$\breve{\Theta} = \mathop{\arg\min}_{\Theta \in \mathbb{R}^{p\times p}, \Theta = \Theta^{\mathrm{T}}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma}\rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_{1,\mathrm{off}}, \tag{22}$$

then we can derive a much simpler alternating direction method for computing $\breve{\Theta}$. If $\breve{\Theta} \succeq \epsilon I$, we must have $\hat{\Theta} = \breve{\Theta}$. If we find that $\breve{\Theta}$ has an eigenvalue less than ϵ, then we can always use Algorithm 1 to find $\hat{\Theta} = \breve{\Theta}$, in which $\breve{\Theta}$ can be taken as the initial value of Θ_0. This implementation strategy could save a lot of computational time.

We now work out the simplified alternating direction method for computing $\breve{\Theta}$. Following the same steps as in § 3·1, we consider the augmented Lagrangian

$$L(\Theta, \Theta_0, \Lambda) = \frac{1}{2}\langle \Theta^2, \hat{\Sigma}\rangle - \mathrm{tr}(\Theta) + \lambda_n\|\Theta_0\|_1 + \langle \Lambda, \Theta - \Theta_0\rangle + (\rho/2)\|\Theta - \Theta_0\|_{\mathrm{F}}^2.$$

We update $(\Theta, \Theta_0, \Lambda)$ according to the following three steps:

$$\Theta^{k+1} = \mathop{\arg\min}_{\Theta = \Theta^{\mathrm{T}}} L(\Theta, \Theta_0^k, \Lambda^k), \tag{23}$$

$$\Theta_0^{k+1} = \mathop{\arg\min}_{\Theta_0} L(\Theta^{k+1}, \Theta_0, \Lambda^k), \tag{24}$$

$$\Lambda^{k+1} = \Lambda^k + \rho(\Theta^{k+1} - \Theta_0^{k+1}).$$

The solutions to (23) and (24) are given in (18) and (20), respectively. Algorithm 2 summarizes the details for computing $\breve{\Theta}$ and the final estimator $\hat{\Theta}$.

Algorithm 2. Alternating direction method implementation for our estimator.

Step 1. Initialization: $k=0$, Λ^0, $\Theta_0^0=\{\mathrm{diag}(\hat{\Sigma})\}^{-1}$ where $\mathrm{diag}(\hat{\Sigma})$ is a diagonal matrix which keeps the diagonal elements of $\hat{\Sigma}$.

Step 2. Repeat (a)–(d) until convergence:

(a) $k=k+1$;
(b) $\Theta^{k+1}=G(\hat{\Sigma}+\rho I, I+\rho\Theta_0^k-\Lambda^k)$;
(c) $\Theta_0^{k+1}=\mathrm{S}(\Theta^{k+1}+\rho^{-1}\Lambda^k, \rho^{-1}\lambda_n)$;
(d) $\Lambda^{k+1}=\Lambda^k+\rho(\Theta^{k+1}-\Theta_0^{k+1})$.

Step 3. Report the converged Θ^k as the solution to $\breve{\Theta}$ defined in (22).

Step 4. If $\lambda_{\min}(\breve{\Theta})>\epsilon$, report $\breve{\Theta}$ as $\hat{\Theta}$.

Step 5. Otherwise, use Algorithm 1 to calculate $\hat{\Theta}$, and in its Step 1 use $\breve{\Theta}$ as the initial value for Θ_0^0 and Θ_1^0.

We have implemented Algorithm 2 in Matlab. In our implementation, we take $\rho=1$ and stop the algorithm when both of the following criteria are satisfied:

$$\frac{\|\Theta^{k+1}-\Theta^k\|_{\mathrm{F}}}{\max(1, \|\Theta^k\|_{\mathrm{F}}, \|\Theta^{k+1}\|_{\mathrm{F}})}<10^{-7}, \qquad \frac{\|\Theta_0^{k+1}-\Theta_0^k\|_{\mathrm{F}}}{\max(1, \|\Theta_0^k\|_{\mathrm{F}}, \|\Theta_0^{k+1}\|_{\mathrm{F}})}<10^{-7}.$$

4. Numerical results

Among existing methods, the lasso penalized Gaussian likelihood estimator is the only popular precision matrix estimator that can simultaneously retain sparsity and positive definiteness. To show the virtue of the D-trace loss, we use simulations to compare the performance of our estimator with that of the lasso penalized Gaussian likelihood estimator.

In the simulation study, data were generated from $N(0, \Sigma^*)$. The following three forms of Σ^* were considered.

Model 1: $\Theta^*_{i,i}=1$, $\Theta^*_{i,j}=0{\cdot}2$ for $1\leqslant|i-j|\leqslant 2$ and $\Theta^*_{i,j}=0$ otherwise.

Model 2: $\Theta^*_{i,i}=1$, $\Theta^*_{i,j}=0{\cdot}2$ for $1\leqslant|i-j|\leqslant 4$ and $\Theta^*_{i,j}=0$ otherwise.

Model 3: $\Theta^*_{i,i}=1$, $\Theta^*_{i,i+1}=0{\cdot}2$ for $\mathrm{mod}(i, p^{1/2})\neq 0$, $\Theta^*_{i,i+p^{1/2}}=0{\cdot}2$ and $\Theta^*_{i,j}=0$ otherwise; this is the grid model in Ravikumar et al. (2011) and requires $p^{1/2}$ to be an integer.

The sample size was taken to be $n=400$ in all three models. We let $p=500$ in Models 1 and 2, and $p=484$ in Model 3. Each estimator was tuned by five-fold crossvalidation. Simulation results based on 100 independent replications are reported in Table 1, where we compare the two estimators in terms of five quantities: the Frobenius risk $E(\|\hat{\Theta}-\Theta^*\|_{\mathrm{F}})$, the operator risk $E(\|\hat{\Theta}-\Theta^*\|_2)$, the matrix $\ell_{1,\infty}$ risk $E(\|\hat{\Theta}-\Theta^*\|_{1,\infty})$, and the percentages of correctly estimated nonzeros and zeros. Table 1 shows that our estimator performs better than the lasso penalized Gaussian likelihood estimator, even though the data are Gaussian. We also recorded the running time of each estimator by fixing the parameter λ_n at the value chosen by crossvalidation. We computed the lasso penalized Gaussian likelihood estimator by using the alternating direction

Table 1. *Results of simulation study: comparison of our estimator with the lasso penalized Gaussian likelihood estimator, i.e., graphical lasso, in terms of three different matrix norms and the percentages of correctly estimated nonzeros and zeros. Reported numbers are averages over* 100 *independent runs, with standard errors given in parentheses. In the first three columns smaller numbers are better; in the last two columns larger numbers are better*

	Frobenius	Operator	$\ell_{1,\infty}$	TP	TN
			Model 1		
Our estimator	7·19 (0·06)	0·77 (0·02)	1·06 (0·04)	88·80 (0·86)	98·77 (0·03)
Graphical lasso	7·49 (0·19)	0·78 (0·02)	1·26 (0·09)	88·12 (2·82)	97·65 (0·71)
			Model 2		
Our estimator	11·70 (0·09)	1·59 (0·01)	1·92 (0·03)	63·47 (1·57)	98·66 (0·20)
Graphical lasso	11·88 (0·03)	1·61 (0·01)	2·11 (0·05)	64·88 (0·69)	97·40 (0·06)
			Model 3		
Our estimator	5·07 (0·06)	0·56 (0·02)	0·91 (0·04)	99·41 (0·22)	98·57 (0·04)
Graphical lasso	5·26 (0·06)	0·58 (0·02)	1·06 (0·06)	99·76 (0·13)	97·48 (0·07)

TP, percentage of correctly estimated nonzeros; TN, percentage of correctly estimated zeros.

method as implemented by Scheinberg et al. (2010). The average running time for our estimator was 1·2 seconds, whereas that for the lasso penalized Gaussian likelihood estimator was 2 seconds.

5. Theoretical results

5·1. *Notation*

In this section we study the theoretical properties of the proposed estimator in the ultrahigh-dimensional setting. Under suitable regularity conditions, the proposed estimator is consistent under various matrix norms and has a sparse recovery property with high probability. In particular, when the x_i are sampled from a sub-Gaussian distribution, consistency holds if $\log(p)$ is small compared to n.

We assume that the true precision matrix Θ^* is sparse. Let $S = \{(i, j) : \Theta^*_{i,j} \neq 0\}$ denote the support of Θ^* and S^{c} the complement of S. Let d be the maximum node degree in Θ^*, and denote by s the number of edges in the graph corresponding to Θ^*. We introduce some additional notation to facilitate the presentation. For a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the ℓ_1 norm $\sum |x_i|$ is written as $|x|_1$, and the ℓ_2 norm $(\sum_{i=1}^{p} x_i^2)^{1/2}$ is written as $\|x\|$. For a matrix X, the elementwise matrix norm $\max_{i,j} |X_{i,j}|$ is written as $\|X\|_\infty$, the ℓ_1 norm $\sum_{i,j} |X_{i,j}|$ is written as $\|X\|_1$, the $\ell_{1,\infty}$ norm $\max_i(\sum_j |X_{i,j}|)$ is written as $\|X\|_{1,\infty}$, and the operator norm $\max_{\|x\|=1} \|Xx\|$ is written as $\|X\|$. For any subset T of $\{1, \ldots, p\} \times \{1, \ldots, p\}$, we denote by $\mathrm{vec}(X)_T$ the subvector of $\mathrm{vec}(X)$ indexed by T. For any two subsets T_1 and T_2 of $\{1, \ldots, p\} \times \{1, \ldots, p\}$, we denote by $X_{T_1 T_2}$ the submatrix of X with rows and columns indexed by T_1 and T_2, respectively. We use $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ to denote the largest and smallest eigenvalues of a symmetric matrix X. We write $\theta_{\min} = \min_{i,j \in S} |\Theta^*_{i,j}|$, $\alpha = 1 - \max_{e \in S^{\mathrm{c}}} \|\Gamma^*_{e,S}(\Gamma^*_{S,S})^{-1}\|_1$, $\Delta_\Gamma = \hat{\Gamma}_{S,S} - \Gamma^*_{S,S}$, $\Delta_\Sigma = \hat{\Sigma} - \Sigma^*$, $\varepsilon = \|\Delta_\Sigma\|_\infty$, $\kappa_\Gamma = \|\Gamma^{*-1}_{S,S}\|_{1,\infty}$ and $\kappa_\Sigma = \|\Sigma^*\|_{1,\infty}$.

5·2. *The irrepresentability condition*

We first present the irrepresentability condition for establishing the model selection consistency of our estimator. An irrepresentability condition is also required for the lasso penalized Gaussian likelihood estimator for estimating sparse precision matrices (Ravikumar et al., 2011).

Denoting the Kronecker matrix sum by $\oplus$ and the Kronecker matrix product by $\otimes$, our irrepresentability condition involves the function

$$\Gamma(\Sigma) = \frac{1}{2}(\Sigma \oplus \Sigma) = \frac{1}{2}(\Sigma \otimes I + I \otimes \Sigma).$$

Upon using the definition of the Kronecker matrix sum, we see that $\Gamma(\Sigma)$ is a $p^2 \times p^2$ matrix indexed by vertex pairs and that

$$\Gamma(\Sigma)_{(j,k),(l,m)} = \Sigma_{k,m}\delta(j,l) + \Sigma_{j,l}\delta(k,m), \tag{25}$$

where $\delta(j,l) = 1$ if $j = l$ and $\delta(j,l) = 0$ if $j \neq l$. For simplicity, we write $\Gamma^* = \Gamma(\Sigma^*)$ and $\hat{\Gamma} = \Gamma(\hat{\Sigma})$. In our theoretical analysis, the following irrepresentability condition is assumed:

$$\max_{e \in S^{\mathrm{c}}} \|\Gamma^*_{e,S}(\Gamma^*_{S,S})^{-1}\|_1 < 1. \tag{26}$$

It is interesting to compare (26) with the irrepresentability condition for the lasso penalized Gaussian likelihood estimator (Ravikumar et al., 2011, Assumption 1), which is

$$\max_{e \in S^{\mathrm{c}}} \|(\Sigma^* \otimes \Sigma^*)_{e,S}\{(\Sigma^* \otimes \Sigma^*)_{S,S}\}^{-1}\|_1 < 1. \tag{27}$$

Notice that (26) involves the Kronecker sum $\Sigma^* \oplus \Sigma^*$ while (27) uses the Kronecker product $\Sigma^* \otimes \Sigma^*$.

It is difficult to compare (26) and (27) in general. Here we compare them on a specific example used by Meinshausen (2008) and Ravikumar et al. (2011), with $\Theta^* \in \mathbb{R}^{4\times 4}$, $\Theta^*_{i,i} = 1$, $\Theta^*_{2,3} = \Theta^*_{3,2} = 0$, $\Theta^*_{1,4} = \Theta^*_{4,1} = 2c^2$ and $\Theta^*_{i,j} = c$ otherwise, where we assume $c \in [-2^{-1/2}, 2^{-1/2}]$ so that Θ^* is positive definite. For this example, we can verify numerically that (26) holds for $|c| \leqslant 0{\cdot}31$ while (27) requires that $|c| < 0{\cdot}2017$ (Ravikumar et al., 2011, § 3.1.1). Thus, when $|c| \in [0{\cdot}2017, 0{\cdot}31]$, (26) holds while (27) fails.

We also compared (26) and (27) on two autoregressive models of orders 1 and 3. In the first, we let $\Theta^* \in \mathbb{R}^{p\times p}$, $\Theta^*_{i,i} = 1$, $\Theta^*_{i,j} = c$ for $|i - j| = 1$ and $\Theta^*_{i,j} = 0$ otherwise. In the second, we let $\Theta^* \in \mathbb{R}^{p\times p}$, $\Theta^*_{i,i} = 1$, $\Theta^*_{i,j} = c$ for $1 \leqslant |i - j| \leqslant 3$ and $\Theta^*_{i,j} = 0$ otherwise. The condition (26) was less restrictive than (27) for all values of p that we tested. For example, consider $p = 30$. For the autoregressive model of order 1, (26) holds for $|c| < 0{\cdot}41$ and (27) holds only for $|c| < 0{\cdot}35$; for the autoregressive model of order 3, (26) holds for $|c| < 0{\cdot}22$ while (27) holds only for $|c| < 0{\cdot}14$.

5·3. *Rates of convergence*

We establish rates of convergence and the model selection consistency of the penalized D-trace estimator under the assumption that $x_1, \ldots, x_n$ are independent and identically sampled from a sub-Gaussian distribution with covariance Σ^* such that all the $X_i/\Sigma^{*\,1/2}_{i,i}$ are sub-Gaussian with parameter σ. Here X_i is the ith coordinate of the random vector X, so we assume that

$$E[\exp\{tX_i(\Sigma^*_{i,i})^{-1/2}\}] \leqslant \exp(\sigma^2 t^2/2) \quad (t \in \mathbb{R}). \tag{28}$$

Theorem 2. *Under* (28) *and the irrepresentability condition* (26), *choose*

$$\lambda_n = 12\alpha^{-1}(\kappa_\Sigma \kappa_\Gamma^2 + \kappa_\Gamma)\left\{128(1 + 4\sigma^2)^2 \max_i(\Sigma^*_{i,i})^2(\eta \log p + \log 4)/n\right\}^{1/2}$$

for some $\eta > 2$ and

$$n > C_1 \max\Big[\lambda_{\min}(\Theta^*)^{-1}\min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\},$$
$$12d\kappa_\Gamma,\, 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma),\, \{\max_i \Sigma_{i,i}^* 8(1+4\sigma^2)\}^{-1}\Big]^2(\eta\log p + \log 4),$$

where $C_1 = 128(1+4\sigma^2)^2 \max_i(\Sigma_{i,i}^*)^2$. Then, with probability greater than $1 - p^{\eta-2}$, we have

$$\|\hat\Theta - \Theta^*\|_\infty \leqslant \{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}\{C_1(\eta\log p + \log 4)/n\}^{1/2},$$
$$\|\hat\Theta - \Theta^*\|_{\rm F} \leqslant (s+p)^{1/2}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}\{C_1(\eta\log p + \log 4)/n\}^{1/2},$$
$$\|\hat\Theta - \Theta^*\|_2 \leqslant \min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}\{C_1(\eta\log p + \log 4)/n\}^{1/2}.$$

In addition, $\hat\Theta$ recovers all zeros in Θ^*. Moreover, if

$$n > C_1\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}^2(\eta\log p + \log 4)/\theta_{\min}^2,$$

then $\hat\Theta$ recovers all zeros and nonzeros in Θ^*.

Next, we establish rates of convergence and model selection consistency of the penalized D-trace estimator under a weaker polynomial tail assumption. Assume that $x_1, \ldots, x_n$ are independent and identically sampled from a distribution with polynomial tails having covariance Σ^* such that $(\Sigma_{i,i}^*)^{-1/2}X_i$ has finite $4m$th moments, i.e., there exist m and $K_m \in \mathbb{R}$ such that

$$E\{(\Sigma_{i,i}^*)^{-1/2}X_i\}^{4m} \leqslant K_m \quad (i = 1, \ldots, p). \tag{29}$$

THEOREM 3. *Under (29) and the irrepresentability condition (26), choose*

$$\lambda_n = 24n^{-1/2}\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)(\max_i \Sigma_{i,i}^*)(K_m+1)^{1/(2m)}p^{\eta/(2m)}$$

for some $\eta > 2$ and

$$n > C_2 p^{\eta/m}\max\Big[\lambda_{\min}(\Theta^*)^{-1}\min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\},$$
$$12d\kappa_\Gamma,\, 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\Big]^2$$

where $C_2 = \{2^{2m}(\max_i \Sigma_{i,i}^*)^{2m}(K_m+1)\}^{1/m}$. Then, with probability $1 - p^{\eta-2}$, we have

$$\|\hat\Theta - \Theta^*\|_\infty \leqslant \{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}C_2^{1/2}p^{\eta/(2m)}n^{-1/2},$$
$$\|\hat\Theta - \Theta^*\|_{\rm F} \leqslant (s+p)^{1/2}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}C_2^{1/2}p^{\eta/(2m)}n^{-1/2},$$
$$\|\hat\Theta - \Theta^*\|_2 \leqslant \min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}C_2^{1/2}p^{\eta/(2m)}n^{-1/2}.$$

In addition, $\hat\Theta$ recovers all zeros in Θ^*. Moreover, if

$$n > C_2 p^{\eta/m}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2+\kappa_\Gamma)\}^2/\theta_{\min}^2,$$

then $\hat\Theta$ recovers all zeros and nonzeros in Θ^*.

These rate-of-convergence results look similar to those of Ravikumar et al. (2011). However, our technical analysis is different from theirs. The key component in their analysis is Brouwer's fixed-point theorem, but we can use a more direct approach to analyse the penalized D-trace estimator, thanks to its simple expression.

6. Discussion

In the empirical loss minimization framework, the D-trace loss is much simpler than the Gaussian likelihood loss, which is basically a quadratic function of the precision matrix. Its simplicity leads to theoretical and computational advantages. We have provided theoretical and empirical evidence to support the D-trace loss and the lasso penalized D-trace estimator. On the other hand, our results do not imply that the D-trace loss estimator is superior to the graphical lasso. Conceptually, the D-trace loss is to the Gaussian likelihood as the hinge loss underlying the support vector machine is to the binomial likelihood for logistic regression. Each has its own merits, and neither dominates the other. An open question remains concerning the irrepresentability condition: we can neither prove nor disprove that (26) is always weaker than (27). This technical problem will be studied in another paper.

Acknowledgement

We thank Dr Shiqian Ma for sharing the Matlab code that implements his alternating direction method for computing the lasso penalized Gaussian likelihood estimator. We thank the editor, associate editor and referees for their suggestions, as well as Professors Peter Bühlmann and Tony Cai for helpful discussions. Zou's research was supported in part by the U.S. National Science Foundation and the Office of Naval Research.

Appendix: technical proofs

Proof of Theorem 1

With a positive definite A, $\langle A, \Theta^2\rangle/2 - \langle B, \Theta\rangle$ is a strictly convex function over Θ. Therefore, we only need to check that its derivative is zero at $G(A, B)$, i.e., $2^{-1}\{AG(A, B) + G(A, B)A\} - B = 0$. Equivalently, we need to check that

$$\mathrm{diag}(\sigma_1, \ldots, \sigma_p)\{U_A^{\mathrm{T}} G(A, B) U_A\} + \{U_A^{\mathrm{T}} G(A, B) U_A\}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_p) = U_A^{\mathrm{T}} B U_A.$$

The above equation can be verified by calculation for $G(A, B)$ defined in (19), and so Theorem 1 is proved.

Proofs of Theorems 2 and 3

We prove these two theorems simultaneously. For clarity of presentation, we first sketch the proof and then fill in the details of the technical lemmas and their proofs.

Following Definition 1 in Ravikumar et al. (2011), we assume that there exists a constant $v_* > 0$ and a function f such that

$$\mathrm{pr}(|\hat{\Sigma}_{i,j} - \Sigma^*_{i,j}| \geqslant \delta) \leqslant 1/f(n, \delta) \quad (1 \leqslant i, j \leqslant p;\ 0 < \delta < 1/v_*). \tag{A1}$$

We also define

$$n_f(\delta, r) = \arg\max\{n : f(n, \delta) \leqslant r\}, \quad \delta_f(n, r) = \arg\max\{\delta : f(n, \delta) \leqslant r\}.$$

The tail assumption (A1) holds for a large class of random vectors. Two special cases, sub-Gaussian tails and polynomial tails, are defined in (28) and (29). When (28) holds, we have $v_* = \{\max_i \Sigma^*_{i,i} 8(1 + 4\sigma^2)\}^{-1}$ and $f(n, \delta) = \exp(c_* n\delta^2)/4$, where $c_* = \{128(1 + 4\sigma^2)^2 \max_i(\Sigma^*_{i,i})^2\}^{-1}$ (Ravikumar et al., 2011, § 2.3.1). Straightforward calculation gives $\delta_f(n, p^\eta) = \{128(1 + 4\sigma^2)^2 \max_i(\Sigma^*_{i,i})^2(\eta \log p + \log 4)/n\}^{1/2}$

and $n_f(\delta, p^\eta) = 128(1+4\sigma^2)^2 \max_i(\Sigma^*_{i,i})^2(\eta \log p + \log 4)/\delta^2$. When (29) holds, we have $v_* = 0$ and $f(n, \delta) = c_* n^m \delta^{2m}$, where $c_* = 2^{-2m}(\max_i \Sigma^*_{i,i})^{-2m}(K_m + 1)^{-1}$ (Ravikumar et al., 2011, § 2.3.2). Thus $\delta_f(n, p^\eta) = p^{\eta/(2m)} c_*^{-1/(2m)} n^{-1/2}$ and $n_f(\delta, p^\eta) = p^{\eta/m} c_*^{-1/m} \delta^{-2}$. With these preparations in place, Theorems 2 and 3 can be proved using the following technical lemma.

LEMMA A1. *Define $\breve{\Theta}$ by*

$$\breve{\Theta} = \mathop{\arg\min}_{\Theta \in \mathbb{R}^{p\times p}, \Theta = \Theta^{\mathrm{T}}} \frac{1}{2}\langle \Theta^2, \hat{\Sigma} \rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_{1,\mathrm{off}}. \tag{A2}$$

Then the following hold:

(a) $\mathrm{vec}(\breve{\Theta})_{S^c} = 0$ *if*

$$\max_{e \in S^c} \left| \hat{\Gamma}_{e,S} \hat{\Gamma}^{-1}_{S,S} \mathrm{vec}(I)_S \right| < \alpha\lambda_n/2, \quad \max_{e \in S^c} \left\| \hat{\Gamma}_{e,S}(\hat{\Gamma}_{S,S})^{-1} \right\|_1 \leqslant 1 - \alpha/2; \tag{A3}$$

(b) $\mathrm{vec}(\breve{\Theta})_{S^c} = 0$ *if*

$$\varepsilon < \frac{1}{12 d \kappa_\Gamma}, \tag{A4}$$

$$6\varepsilon(\kappa_\Sigma \kappa_\Gamma^2 + \kappa_\Gamma) \leqslant 0{\cdot}5\,\alpha \min(\lambda_n, 1); \tag{A5}$$

(c) *assuming the conditions in part* (b), *we also have*

$$\|\breve{\Theta} - \Theta^*\|_\infty < \lambda_n \kappa_\Gamma + \frac{5}{2} d(1 + \lambda_n)\varepsilon\kappa_\Gamma^2. \tag{A6}$$

The proof of Lemma A1 is based on the following auxiliary lemma, which is used to control $\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*}_{S,S}{}^{-1}\|_\infty$ and $\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*}_{S,S}{}^{-1}\|_{1,\infty}$ by $\varepsilon = \|\Delta_\Sigma\|_\infty$. For convenience we present it here.

LEMMA A2. *Assuming* (A4)*, we have*

$$\|R(\Delta_\Gamma)\|_{1,\infty} \leqslant 6d^2\varepsilon^2\kappa_\Gamma^3, \quad \|R(\Delta_\Gamma)\|_\infty \leqslant 12 d\varepsilon^2\kappa_\Gamma^3, \tag{A7}$$

where $R(\Delta_\Gamma) = \{\Gamma^*_{S,S} + (\Delta_\Gamma)_{S,S}\}^{-1} - \Gamma^{*-1}_{S,S} + \Gamma^{*-1}_{S,S}(\Delta_\Gamma)_{S,S}\Gamma^{*-1}_{S,S}$. *Moreover, we have*

$$\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_{1,\infty} \leqslant 6d^2\varepsilon^2\kappa_\Gamma^3 + 2d\varepsilon\kappa_\Gamma^2, \tag{A8}$$

$$\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_\infty \leqslant 12 d\varepsilon^2\kappa_\Gamma^3 + 2\varepsilon\kappa_\Gamma^2. \tag{A9}$$

In this proof we assume the general choices of n and λ_n:

$$\begin{aligned} n > n_f \Big\{ 1/\max\Big(& \sigma_{\min}^{-1}\big[\min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}\big], \\ & \theta_{\min}^{-1}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}, 12d\kappa_\Gamma, 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma), v_*\Big),\ p^\eta \Big\} \end{aligned} \tag{A10}$$

and $\lambda_n = 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\delta_f(n, p^\eta)$ for some $\eta > 2$.

(a) By the definition of n_f, with probability at least $1 - 1/p^{\eta-2}$ we have

$$\varepsilon = \|\hat{\Sigma} - \Sigma^*\|_\infty \leqslant \delta_f(n, p^\eta) < 1/\max\big\{12d\kappa_\Gamma, 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma), v_*\big\}. \tag{A11}$$

Now we verify the two assumptions in Lemma A1(b). Assumption (A4) is easy to verify using (A11). From (A11) and the definition of λ_n we also have $\lambda_n \leqslant 1$, and (A5) follows from the definition of λ_n and the fact that $\lambda_n \leqslant 1$.

The convergence rate of $\|\breve{\Theta} - \Theta^*\|_\infty$ then follows from (A6), (A4), the control of ε by $\delta_f(n, p^\eta)$ in (A11), the definition of λ_n, and the fact that $\lambda_n \leqslant 1$:

$$\begin{aligned}\|\breve{\Theta} - \Theta^*\|_\infty &< \lambda_n \kappa_\Gamma + \frac{5}{2} d(1+\lambda_n)\varepsilon\kappa_\Gamma^2 \leqslant \lambda_n\kappa_\Gamma + 5d\varepsilon\kappa_\Gamma^2 \\ &\leqslant \{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}\delta_f(n, p^\eta). \qquad \text{(A12)}\end{aligned}$$

The estimation of $\|\hat{\Theta} - \Theta^*\|$ follows from (A12), the fact that $\hat{\Theta} = \breve{\Theta}$, which will be shown at the end of the proof of Lemma A1, and the estimation of v_*, $f(n, p^\eta)$ and $\delta_f(n, p^\eta)$.

(b) Combining the bound on $\|\breve{\Theta} - \Theta^*\|_\infty$ with the fact that there are at most $s + p$ nonzero elements in $\breve{\Theta}$ and that the nonzeros of $\breve{\Theta}$ form a subset of Θ^*, we obtain

$$\|\breve{\Theta} - \Theta^*\|_{\mathrm{F}} \leqslant (s+p)^{1/2}\|\breve{\Theta} - \Theta^*\|_\infty \leqslant (s+p)^{1/2}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}\delta_f(n, p^\eta) \qquad \text{(A13)}$$

and

$$\begin{aligned}\|\breve{\Theta} - \Theta^*\|_2 &\leqslant \min\{(s+p)^{1/2}, d\}\|\breve{\Theta} - \Theta^*\|_\infty \\ &\leqslant \min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}\delta_f(n, p^\eta). \qquad \text{(A14)}\end{aligned}$$

The estimation of $\|\hat{\Theta} - \Theta^*\|_2$ and $\|\hat{\Theta} - \Theta^*\|_{\mathrm{F}}$ follows from (A13), (A14), the equality $\hat{\Theta} = \breve{\Theta}$, and the estimation of v_*, $f(n, p^\eta)$ and $\delta_f(n, p^\eta)$.

(c) By part (b) of Lemma A1, $\breve{\Theta}$ specifies all zeros in Θ^*. When (A10) holds, with probability at least $1 - 1/p^{\eta-2}$ we have that $\delta_f(n, p^\eta) \leqslant \{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}/\theta_{\min}$. By combining this with (A12), $\breve{\Theta}$ recovers all zeros and nonzeros in Θ^*. Finally, we show that $\hat{\Theta} = \breve{\Theta}$. Using the fact that with probability at least $1 - 1/p^{\eta-2}$, $\delta_f(n, p^\eta) \leqslant \lambda_{\min}(\Theta^*)/[\min\{(s+p)^{1/2}, d\}\{5d\kappa_\Gamma^2 + 12\alpha^{-1}(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma)\}]$, together with (A14), we deduce that $\lambda_{\min}(\breve{\Theta}) > 0$ and therefore $\hat{\Theta} = \breve{\Theta}$. This completes the proof of Theorems 2 and 3.

Proof of Lemma A1

(a) First, we define $\tilde{\Theta}$ as the solution to the hypothetical problem

$$\tilde{\Theta} = \mathop{\arg\min}_{\Theta = \Theta^{\mathrm{T}}, \Theta_{S^{\mathrm{c}}} = 0} \frac{1}{2}\langle \Theta^2, \hat{\Sigma}\rangle - \mathrm{tr}(\Theta) + \lambda_n \|\Theta\|_{1,\mathrm{off}}. \qquad \text{(A15)}$$

From its directional derivative, we obtain the equality

$$\{(\tilde{\Theta}\hat{\Sigma} + \hat{\Sigma}\tilde{\Theta})/2 - I + Z\}_S = 0$$

where

$$Z_{i,j}\begin{cases} = 0, & (i,j) \in S^{\mathrm{c}} \text{ or } i = j, \\ = \mathrm{sign}(\tilde{\Theta}_{i,j}), & (i,j) \in S,\ i \neq j,\ \tilde{\Theta}_{i,j} \neq 0, \\ \in [-1, 1], & (i,j) \in S,\ i \neq j,\ \tilde{\Theta}_{i,j} = 0. \end{cases}$$

Applying the definition of $\hat{\Gamma} = \Gamma(\hat{\Sigma})$ in (25), this can be rewritten as

$$\{\hat{\Gamma}\mathrm{vec}(\tilde{\Theta}) - \mathrm{vec}(I) + \lambda_n \mathrm{vec}(Z)\}_S = 0. \qquad \text{(A16)}$$

Recall that $\tilde{\Theta}_{S^{\mathrm{c}}} = 0$, (A16) is equivalent to $\hat{\Gamma}_{S,S}\mathrm{vec}(\tilde{\Theta})_S - \mathrm{vec}(I)_S + \lambda_n\mathrm{vec}(Z)_S = 0$, and the explicit solution to (A15) is

$$\mathrm{vec}(\tilde{\Theta})_S = \hat{\Gamma}_{S,S}^{-1}\{\mathrm{vec}(I)_S - \lambda_n \mathrm{vec}(Z)_S\}. \qquad \text{(A17)}$$

Now we verify that $\tilde{\Theta}$ is also the solution to (A2). Since the objective function in (A2) is convex, we only need to verify that its derivative at $\Theta = \tilde{\Theta}$ is zero; that is,

$$\left|\frac{1}{2}(\hat{\Sigma}\tilde{\Theta} + \tilde{\Theta}\hat{\Sigma}) - I\right|_{i,j} \leqslant \lambda_n \quad (1 \leqslant i \neq j \leqslant p),$$
$$\left|\frac{1}{2}(\hat{\Sigma}\tilde{\Theta} + \tilde{\Theta}\hat{\Sigma}) - I\right|_{i,i} = 0 \quad (i = 1, \ldots, p). \tag{A18}$$

Applying (A16), we have that (A18) holds when $(i, j) \in S$. Therefore we need only verify (A18) for $(i, j) \in S^{\mathrm{c}}$. As $\mathrm{vec}(I)_{S^{\mathrm{c}}} = 0$, to prove (A18) it is sufficient to prove that for $e \in S^{\mathrm{c}}$,

$$|\hat{\Gamma}_{e,S}\mathrm{vec}(\tilde{\Theta})_S| \leqslant \lambda_n. \tag{A19}$$

Upon combining (A17) with (A19), it suffices to prove that for $e \in S^{\mathrm{c}}$,

$$|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(I)_S - \lambda_n\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S| \leqslant \lambda_n. \tag{A20}$$

Since $\|\mathrm{vec}(Z)_S\|_\infty \leqslant 1$, we have $|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S| \leqslant \|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\|_1$, Combining this upper bound of $|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S|$ with the assumptions in (A3), we prove (A20) as follows:

$$\begin{aligned}|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(I)_S - \lambda_n\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S| &\leqslant |\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(I)_S| + \lambda_n|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S| \\ &\leqslant |\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(I)_S| + \lambda_n\|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\|_1 \\ &\leqslant \alpha\lambda_n/2 + \lambda_n(1 - \alpha/2) = \lambda_n.\end{aligned}$$

Since (A20) implies (A18), we have shown that $\tilde{\Theta}$ is also the solution $\breve{\Theta}$ in (A2). By the definition of $\tilde{\Theta}$, we obtain $\mathrm{vec}(\breve{\Theta})_S = 0$.

(b) We prove this part in two steps. First, we show that (A21) implies the two conditions in (A3):

$$\max_{e \in S^{\mathrm{c}}} \|\hat{\Gamma}_{e,S}(\hat{\Gamma}_{S,S})^{-1} - \Gamma^*_{e,S}(\Gamma^*_{S,S})^{-1}\|_1 \leqslant 0{\cdot}5\,\alpha \min(\lambda_n, 1). \tag{A21}$$

Then we prove (A21). Therefore we get $\mathrm{vec}(\breve{\Theta})_S = 0$ upon applying the result of part (a).

Combining $\alpha = 1 - \max_{e \in S^{\mathrm{c}}} \|\Gamma^*_{e,S}(\Gamma^*_{S,S})^{-1}\|_1$ with the triangle inequality, we obtain the second assumption in (A3) from (A21). Using the fact that

$$\Gamma^{*\,-1}_{S,S}\mathrm{vec}(I)_S = \mathrm{vec}(\Theta^*)_S, \tag{A22}$$

we have $\Gamma^*_{S^{\mathrm{c}},S}\{\Gamma^{*\,-1}_{S,S}\mathrm{vec}(I)_S\} = \Gamma^*_{S^{\mathrm{c}},S}\{\mathrm{vec}(\Theta^*)_S\} = \mathrm{vec}\{(\Sigma^*\Theta^* + \Theta^*\Sigma^*)/2\}_{S^{\mathrm{c}}} = 0$, and the first condition in (A3) can be verified as follows:

$$\begin{aligned}|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S}\,\mathrm{vec}(I)_S| &= |(\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S} - \Gamma^*_{e,S}\Gamma^{*\,-1}_{S,S})\,\mathrm{vec}(I)_S| + |\Gamma^*_{e,S}\Gamma^{*\,-1}_{S,S}\,\mathrm{vec}(I)_S| \\ &\leqslant \|\hat{\Gamma}_{e,S}\hat{\Gamma}^{-1}_{S,S} - \Gamma^*_{e,S}\Gamma^{*\,-1}_{S,S}\|_1 + 0 \\ &\leqslant \alpha\lambda_n/2.\end{aligned}$$

We now prove (A21). Since the right-hand side of (A21) is equivalent to the right-hand side of (A5), we need only prove that the left-hand side of (A21) is smaller than the left-hand side of (A5). Note that $\|\Gamma^*\|_\infty \leqslant 2\|\Sigma^*\|_\infty$, $\|\Gamma^*\|_{1,\infty} \leqslant 2\|\Sigma^*\|_{1,\infty}$, and the left-hand side of (A21) can be controlled as follows, by applying (A9), (A26) and (A27): for any $e \in S^{\mathrm{c}}$,

$$\begin{aligned}
&\|\hat{\Gamma}_{e,S}(\hat{\Gamma}_{S,S})^{-1} - \Gamma^*_{e,S}(\Gamma^*_{S,S})^{-1}\|_1\\
&\quad = \|(\hat{\Gamma}_{e,S} - \Gamma^*_{e,S})(\Gamma^*_{S,S})^{-1} + \Gamma^*_{e,S}(\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}) + (\hat{\Gamma}_{e,S} - \Gamma^*_{e,S})(\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S})\|_1\\
&\quad \leqslant \|(\hat{\Gamma}_{e,S} - \Gamma^*_{e,S})\Gamma^{*-1}_{S,S}\|_1 + \|\Gamma^*_{e,S}(\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S})\|_1 + \|(\hat{\Gamma}_{e,S} - \Gamma^*_{e,S})(\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S})\|_1\\
&\quad \leqslant \|\hat{\Gamma}_{e,S} - \Gamma^*_{e,S}\|_\infty \|\Gamma^{*-1}_{S,S}\|_{1,\infty} + 2\|\Sigma^*\|_{1,\infty}\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_\infty + 2d\varepsilon\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_\infty\\
&\quad \leqslant 2\varepsilon\kappa_\Gamma + (2\kappa_\Sigma + 2d\varepsilon)(12d\varepsilon^2\kappa_\Gamma^3 + 2\varepsilon\kappa_\Gamma^2).
\end{aligned} \tag{A23}$$

Inserting (A4) into the right-hand side of (A23) yields the simplification

$$\begin{aligned}
2\varepsilon\kappa_\Gamma + (2\kappa_\Sigma + 2d\varepsilon)(12d\varepsilon^2\kappa_\Gamma^3 + 2\varepsilon\kappa_\Gamma^2) &\leqslant 2\varepsilon\kappa_\Gamma + \left(2\kappa_\Sigma + \frac{1}{6\kappa_\Gamma}\right)(\varepsilon\kappa_\Gamma^2 + 2\varepsilon\kappa_\Gamma^2)\\
&= \varepsilon\left(2\kappa_\Gamma + 6\kappa_\Sigma\kappa_\Gamma^2 + \frac{1}{2}\kappa_\Gamma\right)\\
&< 6\varepsilon(\kappa_\Sigma\kappa_\Gamma^2 + \kappa_\Gamma).
\end{aligned} \tag{A24}$$

Upon combining (A5), (A23) and (A24), we obtain (A21).

(c) By using the fact that $\breve{\Theta} = \tilde{\Theta}$, along with (A17) and (A22), we obtain

$$\begin{aligned}
\|\breve{\Theta} - \Theta^*\|_\infty &= \|\mathrm{vec}(\tilde{\Theta})_S - \mathrm{vec}(\Theta^*)_S\|_\infty\\
&= \|(\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S})\mathrm{vec}(I)_S - \lambda_n\hat{\Gamma}^{-1}_{S,S}\mathrm{vec}(Z)_S\|_\infty\\
&\leqslant \|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_{1,\infty} + \lambda_n\|\hat{\Gamma}^{-1}_{S,S}\|_{1,\infty}\\
&\leqslant (1+\lambda_n)\|\hat{\Gamma}^{-1}_{S,S} - \Gamma^{*-1}_{S,S}\|_{1,\infty} + \lambda_n\|\Gamma^{*-1}_{S,S}\|_{1,\infty}.
\end{aligned} \tag{A25}$$

Then we prove (A6) by applying (A4) and (A8) to the right-hand side of (A25):

$$\|\breve{\Theta} - \Theta^*\|_\infty \leqslant \lambda_n\kappa_\Gamma + (1+\lambda_n)(6d^2\varepsilon^2\kappa_\Gamma^3 + 2d\varepsilon\kappa_\Gamma^2) < \lambda_n\kappa_\Gamma + \frac{5}{2}(1+\lambda_n)d\varepsilon\kappa_\Gamma^2.$$

This completes the proof of Lemma A1.

Proof of Lemma A2

Using the definition of Γ^* and $\hat{\Gamma}$, we have

$$\|(\Delta_\Gamma)_{S,S}\|_{1,\infty} \leqslant 2d\varepsilon, \tag{A26}$$

and then (A4) implies that $\|\Gamma^{*-1}_{S,S}\|_{1,\infty}\|(\Delta_\Gamma)_{S,S}\|_{1,\infty} < 1/3$. Following the proof of Ravikumar et al. (2011, Appendix B), we obtain that $\|R(\Delta_\Gamma)\|_\infty \leqslant 3\|(\Delta_\Gamma)_{S,S}\|_\infty\|(\Delta_\Gamma)_{S,S}\|_{1,\infty}\kappa_\Gamma^3/2$ and $\|R(\Delta_\Gamma)\|_{1,\infty} \leqslant 3\|(\Delta_\Gamma)_{S,S}\|^2_{1,\infty}\kappa_\Gamma^3/2$. Then we prove (A7) by combining (A26) with the fact that

$$\|(\Delta_\Gamma)_{S,S}\|_\infty \leqslant \|\hat{\Gamma} - \Gamma^*\|_\infty \leqslant 2\|\hat{\Sigma} - \Sigma^*\|_\infty = 2\varepsilon. \tag{A27}$$

Moreover,

$$\|\Gamma^{*-1}_{S,S}(\Delta_\Gamma)_{S,S}\Gamma^{*-1}_{S,S}\|_{1,\infty} \leqslant \|(\Delta_\Gamma)_{S,S}\|_{1,\infty}\|\Gamma^{*-1}_{S,S}\|^2_{1,\infty} \leqslant 2d\varepsilon\kappa_\Gamma^2, \tag{A28}$$

$$\|\Gamma^{*-1}_{S,S}(\Delta_\Gamma)_{S,S}\Gamma^{*-1}_{S,S}\|_\infty \leqslant \|(\Delta_\Gamma)_{S,S}\|_\infty\|\Gamma^{*-1}_{S,S}\|^2_{1,\infty} = \|(\Delta_\Gamma)_{S,S}\|_\infty\kappa_\Gamma^2 \leqslant 2\varepsilon\kappa_\Gamma^2. \tag{A29}$$

Then (A8) and (A9) are obtained by combining (A7), (A28), (A29) and the definition of $R(\Delta_\Gamma)$. This completes the proof of Lemma A2.

References

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

Dobra, A., Eicher, T. & Lenkoski, A. (2009). Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. *Statist. Methodol.* **7**, 292–306.

Duchi, J., Gould, S. & Koller, D. (2008). Projected subgradient methods for learning sparse Gaussians. In *Proc. 24th Annual Conf. Uncertainty Artif. Intel. (UAI 2008)*. Corvallis, Oregon: AUAI Press, pp. 145–52.

Friedman, J. H., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Huang, J., Liu, N., Pourahmadi, M. & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

Li, H. & Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7**, 302–17.

Li, S. (2009). *Markov Random Field Modeling in Image Analysis*. New York: Springer.

Lu, Z. (2009). Smooth optimization approach for sparse covariance selection. *SIAM J. Optimiz.* **19**, 1807–27.

Meinshausen, N. (2008). A note on the lasso for Gaussian graphical model selection. *Statist. Prob. Lett.* **78**, 880–4.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

Pease, M. (1965). *Methods of Matrix Algebra*. London: Academic Press.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Assoc.* **104**, 735–46.

Ravikumar, P., Wainwright, M., Raskutti, G. & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–80.

Rothman, A., Bickel, P., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

Scheinberg, K., Shiqian Ma, S. & Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems 23*, J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta, eds. New York: Curran Associates, pp. 2101–9.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wille, A. & Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statist. Appl. Genet. Molec. Biol.* **5**, Issue 1, Article 1.

Witten, D., Friedman, J. H. & Simon, N. (2011). New insights and faster computations for the graphical lasso. *J. Comp. Graph. Statist.* **20**, 892–900.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach Learn. Res.* **11**, 2261–86.

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

[*Received May* 2012. *Revised October* 2013]