Generalized Pseudolikelihood Methods for Inverse Covariance Estimation

Alnur Ali
Machine Learning Dept.
Carnegie Mellon University
alnurali@cmu.edu

Kshitij Khare
Dept. of Statistics
University of Florida
kdkhare@stat.ufl.edu

Sang-Yun Oh
Dept. of Stats. and Applied Prob.
UC Santa Barbara
syoh@pstat.ucsb.edu

Bala Rajaratnam
Dept. of Statistics
UC Davis
brajaratnam01@gmail.com

This document contains proofs and supplementary details for the paper "Generalized Pseudolikelihood Methods for Inverse Covariance Estimation". All section, equation, table, and figure numbers in this supplementary document are preceded by the letter S (all section, equation, table, and figure numbers without an S refer to the main paper).

S.1 COMPUTATIONAL ASPECTS OF THE PseudoNet ESTIMATOR

S.1.1 Proximal gradient method for computing the PseudoNet estimate

In Algorithm 1 below, we fully specify our proximal gradient method for computing the PseudoNet estimate (it is straightforward give an accelerated proximal gradient method as well). Assuming the iterates are sparse, the computational cost of each iteration of Algorithm 1 is dominated by computing the soft-thresholding operator, and therefore costs $O(p^2)$.

Algorithm 1 Proximal gradient method for computing the PseudoNet estimate

```
Input: data matrix $X \in \mathbf{R}^{n \times p}$, tuning parameters $\lambda_1, \lambda_2 > 0$
Output: estimate $\hat{\Omega}^{\text{net}}$
initialize starting point $\Omega \in \mathbf{S}^p_{++}$ (the space of $p \times p$ positive definite matrices); optimization tolerance $\epsilon > 0$;
line search parameters $\tau_{\text{init}}, \beta \in (0, 1)$
repeat
  compute $\nabla g(\Omega)$, using Equation 4
  choose $\tau$ via backtracking line search as follows
      set $\tau \leftarrow \tau_{\text{init}}$
      set $\tilde{\Omega} \leftarrow \mathbf{prox}_{(\lambda_1 \tau) h}(\Omega - \tau \nabla g(\Omega))$ using Equation 3
      while $g(\tilde{\Omega}) \geq g(\Omega) - \mathbf{Tr}\left( (\nabla g(\Omega))^T (\Omega - \tilde{\Omega}) \right) + \frac{1}{2\tau} \|\Omega - \tilde{\Omega}\|_F^2$ do
        update $\tilde{\Omega} \leftarrow \mathbf{prox}_{(\lambda_1 \tau) h}(\Omega - \tau \nabla g(\Omega))$
        update $\tau \leftarrow \beta \tau_{\text{init}}$
      end while
      output $\tau$
  update $\Omega \leftarrow \mathbf{prox}_{(\lambda_1 \tau) h}(\Omega - \tau \nabla g(\Omega))$
until stopping criterion is satisfied, i.e., $\|\nabla g(\Omega) + z\|_F / \|\Omega\|_F \leq \epsilon$ ($z$ is any subgradient of $h$ evaluated at $\Omega$)
output $\hat{\Omega}^{\text{net}} = \Omega$
```

S.1.2 Choice of tuning parameters

Here, we give a way to choose the tuning parameters λ_1 and λ_2 in the PseudoNet optimization problem (1). We propose choosing these parameters by selecting the (λ_1, λ_2) pair that minimizes the following Bayesian information criterion-like score over a grid of tuning parameter values:

$$\mathbf{Bic}(\lambda_1, \lambda_2) = \sum_{j=1}^{p} \mathbf{Bic}(\lambda_1, \lambda_2, j), \tag{S.1}$$

where

$$\mathbf{Bic}(\lambda_1, \lambda_2, j) = n \log \mathbf{rss}(\lambda_1, \lambda_2, j) + \log n \times \left| \left\{ \ell : \ell \in \{1, \dots, p\},\ \ell \neq j,\ \hat{\Omega}^{\text{net}}_{j\ell}(\lambda_1, \lambda_2) \neq 0 \right\} \right|,$$
$$\mathbf{rss}(\lambda_1, \lambda_2, j) = \sum_{i=1}^{n} \left(X_{ij} - \sum_{k \neq j}^{p} \frac{\hat{\Omega}^{\text{net}}_{jk}(\lambda_1, \lambda_2)}{\hat{\Omega}^{\text{net}}_{jj}(\lambda_1, \lambda_2)} X_{ik} \right)^2,$$

and $\hat{\Omega}^{\text{net}}(\lambda_1, \lambda_2)$ is the solution of the PseudoNet optimization problem (1) for a particular λ_1 and λ_2. This method is simple to implement and computationally inexpensive, especially when combined with the screening rules that we described in Section S.1.4, and which we elaborate on below.

S.1.3 Sequential strong screening rules for PseudoNet

Lemma 2.1, presented in the main paper, forms the foundation for our screening rules; its proof is given below. We also provide an algorithmic specification of our screening rules in Algorithm 2.

Sequential strong rules build on the work of Banerjee et al. (2008, Theorem 4), who first observed that variables can be dropped from their particular optimization problem by arguing from their dual problem and block coordinate descent procedure. Mazumder and Hastie (2012) also derive screening rules for the GLasso by arguing from the GLasso's optimality conditions. Although all of these rules are *safe*, *i.e.*, they do not commit violations, we unfortunately do not use block coordinate descent to compute the PseudoNet estimate, and a careful inspection of PseudoNet's optimality conditions reveals that these conditions are not separable in the entries of $\hat{\Omega}^{\text{net}}$, making the framework of Tibshirani et al. (2012) more appropriate here.

Algorithm 2 Sequential strong screening rules for PseudoNet

```
Input: data matrix $X \in \mathbf{R}^{n \times p}$; nonincreasing sequences of tuning parameters $(\lambda_1^{(k)})_{k=1}^{r}, (\lambda_2^{(\ell)})_{\ell=1}^{s}$
Output: estimates $\hat{\Omega}^{\text{net}}(\lambda_1^{(k)}, \lambda_2^{(\ell)}),\ k = 1, \dots, r,\ \ell = 1, \dots, s$
for $\ell = 1, \dots, s$ do
  compute $\hat{\Omega}^{\text{net}}(\lambda_1^{(1)}, \lambda_2^{(\ell)})$ using Equation 1 with $\lambda_1^{(1)}, \lambda_2^{(\ell)}$
  for $k = 2, \dots, r$ do
    compute $N$, the set of nonzero variables, using Equation 6 with $\hat{\Omega}^{\text{net}}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}), \lambda_1^{(k-1)}, \lambda_2^{(\ell)}$
    repeat
      compute $\hat{\Omega}^{\text{net}}(\lambda_1^{(k)}, \lambda_2^{(\ell)})$ using Equation 1 with $N, \lambda_1^{(k)}, \lambda_2^{(\ell)}$
      check (all variables) for violations using the optimality conditions for (1) (see Equation S.2)
      add any violating variables back into $N$
    until there are no violations
    output $\hat{\Omega}^{\text{net}}(\lambda_1^{(k)}, \lambda_2^{(\ell)})$
  end for
end for
```

S.1.4 Proof of Lemma 2.1

Proof. By considering the gradient of the smooth term in the objective of the PseudoNet optimization problem (1), given by (4), in a componentwise fashion, we can express the optimality conditions for (1), evaluated at the off-diagonal entries of $\hat{\Omega}^{\text{net}}_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)})$, as

$$\begin{array}{llll} \left|C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)})\right| & \leq \lambda_1^{(k)} & \text{if} \;\; \hat{\Omega}^{\text{net}}_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) & = 0 \\ C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) & = \lambda_1^{(k)} & \text{if} \;\; \hat{\Omega}^{\text{net}}_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) & > 0 \\ C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) & = -\lambda_1^{(k)} & \text{if} \;\; \hat{\Omega}^{\text{net}}_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) & < 0. \end{array} \tag{S.2}$$

Now assume that $\left|C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)})\right| < 2\lambda_1^{(k)} - \lambda_1^{(k-1)}$. Then we have that

$$\begin{aligned} \left|C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)})\right| &\leq \left|C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) - C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)})\right| + \left|C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)})\right| \\ &< |\lambda_1^{(k)} - \lambda_1^{(k-1)}| + 2\lambda_1^{(k)} - \lambda_1^{(k-1)} \\ &= \lambda_1^{(k)}, \end{aligned}$$

with the first inequality following by the triangle inequality; the second by the assumptions that the C_{ij} are nonexpansive and nonincreasing, as well as by the assumption in the statement of the lemma; and the third because we assumed that $\lambda_1^{(k-1)} \geq \lambda_1^{(k)}$. By checking (S.2), this implies that $\hat{\Omega}^{\text{net}}_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) = 0$ is a solution. □

S.2 ADDITIONAL NUMERICAL RESULTS FOR THE SYNTHETIC EXAMPLES

We present the variable selection accuracies and estimation errors for PseudoNet and CONCORD on the synthetic data described in Section 3.1, with $p = 1000$, In Table S.1. We also present the percentages of variables that our screening rules suggest dropping, as well as the percentages of violations, on this same data (*i.e.*, with $p = 1000$ again) in Figure S.1. (We presented these results with $p = 3000$ in Section 3.1 of the main paper.)

Our experimental settings correspond to ultimately running PseudoNet and CONCORD 145,200 and 6,600 times and estimating $p(p+1)/2 = 500,500$ and $4,501,500$ parameters, respectively. Computing the mean across λ_1, λ_2 gave similar results.

		$n = 200$		$n = 400$		$n = 800$	
		PseudoNet	CONCORD	PseudoNet	CONCORD	PseudoNet	CONCORD
AUC	Median	**0.68**	0.65	**0.81**	0.73	**0.91**	0.86
	IQR	0.02	0.01	0.01	0.01	0.01	0.01
Squared Frobenius norm	Median	**6391.48**	20150.68	**5722.84**	18805.59	**4205.49**	14990.35
	IQR	84.70	513.99	26.18	245.65	18.22	192.78
ℓ_2 operator norm	Median	**2.51**	5.17	**2.41**	5.07	**2.56**	5.84
	IQR	0.01	0.06	0.01	0.03	0.01	0.03
Elementwise ℓ_1 norm	Median	**17480.45**	35959.79	**21640.10**	46951.74	**21749.16**	51526.46
	IQR	65.71	323.46	35.01	240.09	26.38	276.32
Elementwise ℓ_∞ norm	Median	**1.34**	2.93	**1.06**	2.32	**0.67**	1.38
	IQR	0.01	0.04	0.01	0.02	0.01	0.03
Wallclock time (secs.)	Median	**73.72**	103.23	**40.76**	71.02	**14.60**	20.46
	IQR	3.23	41.53	1.76	29.54	0.70	7.08

Table S.1: Median and interquartile range for PseudoNet and CONCORD's areas under the curves (AUCs), estimation errors in several matrix norms, and wallclock times ($p = 1000$). Higher median AUC is better, lower median estimation error and wallclock time is better; best in **bold**. PseudoNet outperforms CONCORD across all sample sizes and metrics.

S.3 ADDITIONAL NUMERICAL RESULTS FOR THE MINIMUM VARIANCE PORTFOLIO OPTIMIZATION EXAMPLE

In addition to the numerical results given in the main paper, we consider here the realized risk and Sharpe ratios for various estimators and estimation horizons, after accounting for borrowing costs (at an 8% annual percentage rate) and transaction costs (at 0.5% of the principal); Tables S.2 and S.3 present the results, and we generally see the same trends as in the main paper. PseudoNet achieves the lowest risk when the estimation horizon is small, and otherwise is within 5% of the lowest risk. PseudoNet also achieves the highest Sharpe ratio four (out of eight) times, and is otherwise within 5% of the highest Sharpe ratio.

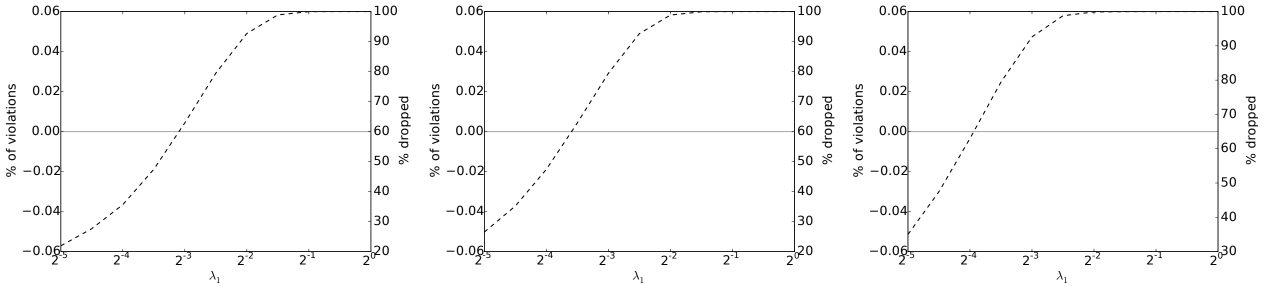

Figure S.1: Percentages of dropped variables excluding diagonal entries (dashed line, right vertical axes) and violations (solid line, left vertical axes) for PseudoNet's screening rules ($\lambda_2 = 1$, $p = 1000$); first column is $n = 0.2p$, second is $n = 0.4p$, third is $n = 0.8p$. The rules never commit a violation.

H	PseudoNet	CONCORD	Sample	GLasso	CondReg	Ledoit
35	**14.98**	16.75	33.70	16.29	17.61	15.32
40	**14.79**	16.73	26.46	16.27	17.54	15.21
45	**14.98**	16.75	23.13	16.28	17.43	15.21
50	**14.77**	16.73	20.87	16.10	17.39	15.15
75	**14.82**	16.76	17.25	15.38	16.98	14.91
150	14.81	16.80	15.18	14.74	16.17	**14.45**
225	14.85	16.81	14.77	14.64	15.85	**14.29**
300	14.96	16.86	14.74	14.73	15.88	**14.29**

Table S.2: Realized risk for various estimators and estimation horizons H, after accounting for borrowing and transaction costs, in the portfolio optimization example. Lower is better; best in **bold**. PseudoNet is best 5/8 times.

Qualitatively, we find that, although PseudoNet does provide sparse estimates, these estimates are usually somewhat denser than those provided by CONCORD (as expected); Figure S.2 plots these estimates (from a randomly chosen investment horizon and trading period). Thus, owing to its (comparatively) denser and better estimates, PseudoNet can reduce risk by hedging, for example, by taking a short position in a stock whose returns are negatively correlated with another stock that it also takes a long position in. To this end, we consider the *size of the short side* of a portfolio $x \in \mathbf{R}^p$, which is defined as the ratio of the magnitude of all the short positions in the portfolio to the magnitude of the portfolio, expressed as a percentage, *i.e.*,

$$100 \times \left(\sum_{i=1}^{p} \min\{x_i, 0\} \right) / \left(\sum_{i=1}^{p} |x_i| \right).$$

Table S.4 presents the size of the short side, averaged over all trading periods, for various estimators and estimation horizons, and we indeed see that the size of PseudoNet's short side is larger than CONCORD's, GLasso's, and CondReg's.

S.4 SUSTAINABLE ENERGY APPLICATION

Here, we present an evaluation of PseudoNet on the task of recovering the conditional independencies between several wind farms on the basis of historical wind power measurements at these farms; wind power is naturally intermittent (as are many renewable resources), and thus understanding the relationships between wind farms can help operators forecast, plan, and dispatch. We obtained hourly wind power measurements from July 1, 2009 through September 14, 2010 (440 days) at seven wind farms from `http://www.kaggle.com/c/GEF2012-wind-forecasting`; see Hong et al. (2014) for further details, as well as a summary of a recent Kaggle competition based on this data. Each group of 48 columns in the data set corresponds to two days (*i.e.*, 48 hours) of hourly wind power measurements at a particular farm; to model the nonlinear relationship between wind power at different locations, we consider five radial basis function kernels spread evenly and evaluated at each hourly measurement (see, for example, Wytock and Kolter (2013); Ali et al. (2016) for a similar approach). Thus, $p = 7 \times 48 \times 5 = 1680$. Each row in the data set considers wind power measurements starting 12 hours after the (start of the) previous row; for example, the first row considers wind power measurements from 1:00 pm on July 1, 2009 through 12:00 pm on July 3, 2009, the second row from 1:00

H	PseudoNet	CONCORD	Sample	GLasso	CondReg	Ledoit
35	**0.47**	0.42	0.35	0.42	0.42	0.44
40	0.46	0.42	**0.50**	0.43	0.43	0.41
45	0.40	**0.41**	0.30	0.40	0.41	0.36
50	**0.43**	0.42	0.23	0.40	0.41	0.38
75	**0.41**	0.40	0.36	0.34	0.40	0.33
150	0.42	0.42	0.27	0.33	**0.43**	0.36
225	0.46	0.45	0.33	0.33	**0.48**	0.38
300	**0.49**	0.45	0.32	0.32	0.44	0.37

Table S.3: Sharpe ratios for various estimators and estimation horizons H, after accounting for borrowing and transaction costs, in the portfolio optimization example. Higher is better; best in **bold**. PseudoNet is best 4/8 times.

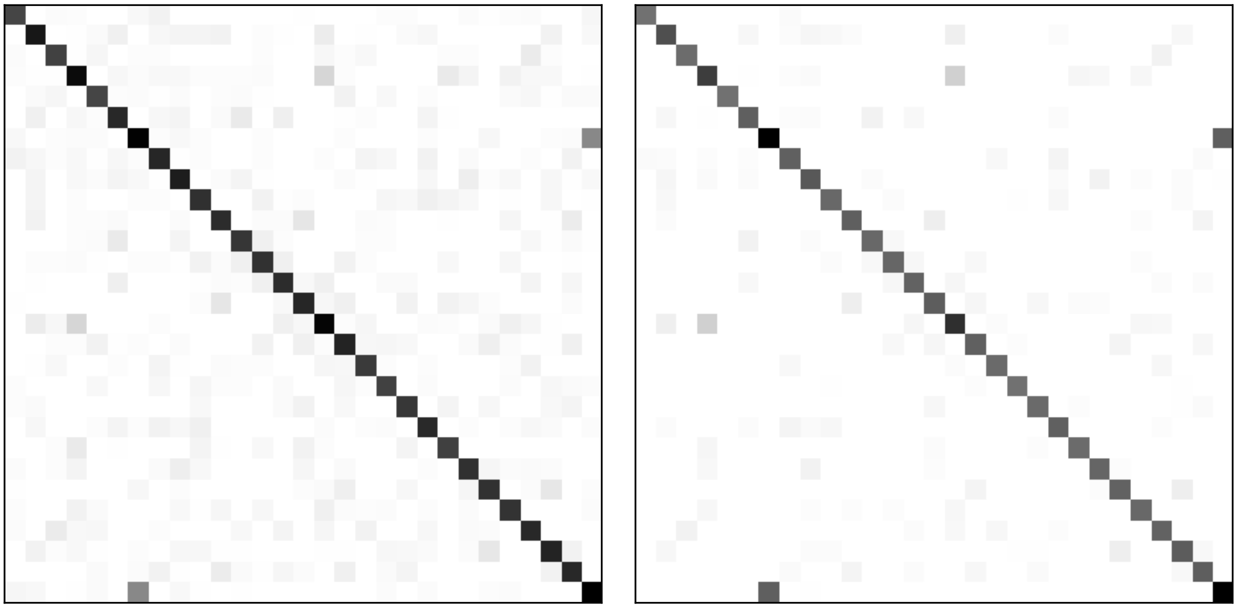

Figure S.2: Estimates provided by PseudoNet (left) and CONCORD (right); darker means larger in magnitude.

am on July 2, 2009 through 12:00 am on July 4, 2009, and the last row from 1:00 am on September 12, 2010 through 12:00 am on September 14, 2010. Thus, $n = 877$. Computing the PseudoNet estimate here therefore corresponds to learning the structure of a spatiotemporal graphical model.

The left panel of Figure S.3 presents the PseudoNet estimate's sparsity pattern. The nonzero super- and sub-diagonal entries suggest that at any wind farm the previous hour's wind power (naturally) influences the next hour's, while the nonzero off-diagonal entries, for example, in the (4,6) block, uncover farms that may influence one another: for example, farms 4 and 6 may be nearby, or (perhaps more interestingly) they may not be nearby[1]. Wytock and Kolter (2013), whose method placed fifth in the Kaggle competition, as well as Ali et al. (2016) report similar findings (see the left panel of Figure 7 as well as Figure S.3, respectively, in these papers). The right panel of Figure S.3 evaluates PseudoNet's screening rules on this data set: the rules never commit a violation.

S.5 PROOF OF LEMMA 4.1

We prove this result by first establishing, in the following lemma, that the gradient of the smooth term in the objective in the PseudoNet optimization problem (1), ∇g, is Lipschitz continuous. The squared Frobenius norm penalty in the PseudoNet optimization problem (1) makes doing this much cleaner, letting us move completely away from the strategy used in Oh et al. (2014, Theorem 3.1).

Lemma S.5.1. *Suppose $(\Omega^{(i)})_{i=0}^{k}$ is a sequence of PseudoNet iterates with nonincreasing objective value. Let Ω be any of the iterates here. Also, let $L = 1/\ell^2 + \|S\|_2 + \lambda_2$, with $\|\cdot\|_2$ denoting the ℓ_2 operator norm*

[1] The true wind farm locations are censored in the data set.

H	PseudoNet	CONCORD	Sample	GLasso	CondReg	Ledoit
35	6.91	0.06	41.13	0.63	1.77	20.50
40	6.80	0.06	38.64	0.67	1.91	20.45
45	6.64	0.05	36.89	0.83	2.21	20.31
50	6.60	0.04	35.46	1.36	2.43	20.33
75	5.93	0.04	30.89	8.60	4.11	20.13
150	5.74	0.02	25.65	23.34	7.58	19.60
225	5.59	0.01	23.68	23.35	9.34	19.26
300	5.22	0.00	22.45	22.43	9.41	18.85

Table S.4: Average size of the short side for various estimators and estimation horizons H in the portfolio optimization example.

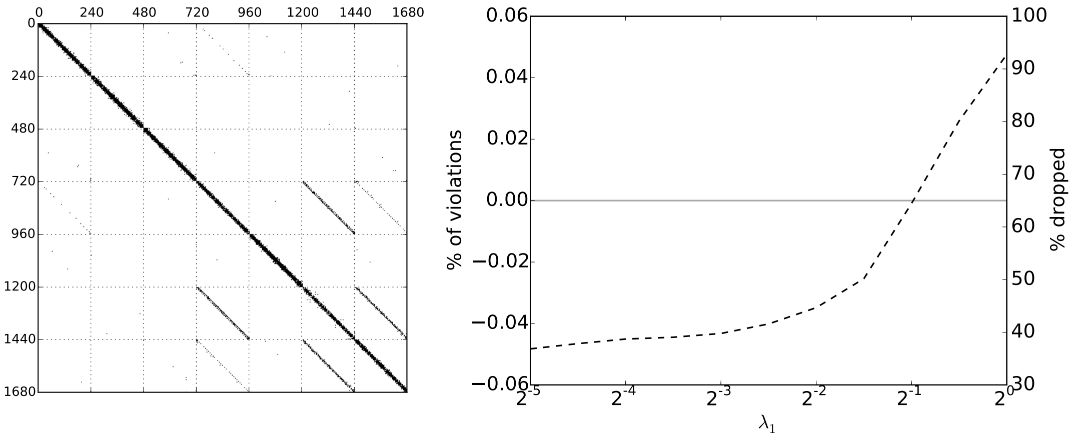

Figure S.3: Left: sparsity pattern for the PseudoNet estimate (black means nonzero, and each block corresponds to a wind farm). Right: percentages of dropped variables excluding diagonal entries (dashed line, right vertical axes) and violations (solid line, left vertical axes) for PseudoNet's screening rules ($\lambda_2 = 1$); the rules never commit a violation.

(maximum singular value), and let ℓ be a constant that uniformly lower bounds Ω_{ii}, $i = 1, \dots, p$. Then we get that $\nabla^2 g(\Omega) \preceq L I_{p^2 \times p^2}$.

Proof. Let J_{net} be the objective in the PseudoNet optimization problem (1). Then we have that

$$-\sum_{i=1}^{p} \log \Omega_{ii} + (\lambda_2/2) \sum_{i=1}^{p} \Omega_{ii}^2 \leq J_{\text{net}}(\Omega^{(0)}),$$

since the ℓ_1 term in the objective in (1) is nonnegative, and the trace term can be expressed as a nonnegative quadratic form. The lefthand side here approaches ∞ as either $\Omega_{ii} \to \infty$ or $\Omega_{ii} \to 0$, *i.e.*, Ω_{ii} must be uniformly bounded away from ∞ and 0 by some u and ℓ, respectively, for $i = 1, \dots, p$, owing to the righthand side of the expression. Thus, we can upper bound the eigenvalues of (5) with

$$1/\ell^2 + \|S\|_2^2 + \lambda_2,$$

as claimed. □

Obtaining linear convergence is now immediate. As g is smooth, the conclusion in Lemma S.5.1 is equivalent to $\nabla^2 g(\Omega) \preceq L I_{p^2 \times p^2} \iff \|\nabla g(\Omega) - \nabla g(\tilde{\Omega})\|_F \leq L \|\Omega - \tilde{\Omega}\|_F$, where $\tilde{\Omega} \in \mathbf{S}_{++}^{p}$, and $L = 1/\ell^2 + \|S\|_2 + \lambda_2$. Now, since g is also λ_2-strongly convex, the claim follows by Schmidt et al. (2011, Proposition 3). □

S.6 SATURATION RESULTS

The statement and proof of Theorem 4.4 make use of a matrix $A \in \mathbf{R}^{np \times p(p-1)/2}$ containing the columns of the data matrix X arranged in a particular fashion; this matrix is defined as

$$A = -\begin{bmatrix} X_2 & X_3 & X_4 & \cdots & X_{p-1} & X_p & 0 & & & & & & \cdots & & & & & & & & & 0 \\ X_1 & 0 & & \cdots & & 0 & X_3 & X_4 & X_5 & \cdots & X_{p-1} & X_p & 0 & & & & \cdots & & & & & 0 \\ 0 & X_1 & 0 & & \cdots & 0 & X_2 & 0 & & \cdots & & 0 & X_4 & X_5 & X_6 & \cdots & X_{p-1} & X_p & 0 & \cdots & & 0 \\ & & & & & & & & & & \vdots & & & & & & & & & & & \\ 0 & & \cdots & & 0 & X_1 & 0 & & \cdots & & 0 & X_2 & 0 & & \cdots & & 0 & X_3 & 0 & \cdots & 0 & X_{p-1} \end{bmatrix}.$$

In order to make the statement and proof of Corollary 4.5 clearer, we also describe the SPLICE and SPACE estimators in more detail here.

We can obtain a SPLICE estimate by first minimizing the following objective, alternately over the variables $D \in \mathbf{R}^{p \times p}$ and $B \in \mathbf{R}^{p \times p}$, where D is a diagonal matrix and the diagonal entries of the matrix B are set to zero,

$$-\frac{1}{2} \log \det D + \frac{1}{2} \sum_{i=1}^{p} \frac{1}{D_{ii}^2} \|X_i - X_{-i}(B_{i\cdot})^T\|_2^2 + \lambda_1 \|B\|_1, \tag{S.3}$$

where X_{-i} denotes the data matrix X after removing the ith column, and $B_{i\cdot}$ here means the ith row of B after removing the entry B_{ii}; then, for any iteration i, we compute the estimate

$$\hat{\Omega}^{\text{spl},(i)} = (\hat{D}^{(i-1)})^{-2}(I - \hat{B}^{(i)}), \tag{S.4}$$

with $\hat{\Omega}^{\text{spl},(i)}$ referring to the estimate at the end of the ith iteration ($\hat{D}^{(i-1)}$ and $\hat{B}^{(i)}$ are interpreted similarly).

Turning to SPACE, we can compute a SPACE estimate by minimizing the following objective, alternately over the variables Ω_{diag} and Ω_{off},

$$-(1/2) \log \det \Omega_{\text{diag}} + \lambda_1 \|\Omega_{\text{off}}\|_1 + (1/2) \sum_{i=1}^{p} \Omega_{\text{diag},ii} \left\| X_i - \sum_{j \neq i}^{p} \Omega_{\text{off},ij} \sqrt{\frac{\Omega_{\text{diag},jj}}{\Omega_{\text{diag},\,ii}}} X_j \right\|_2^2, \tag{S.5}$$

where $\Omega_{\text{diag},ii}$ refers to the (i,i)th entry of Ω_{diag} ($\Omega_{\text{off},ij}$ is interpreted similarly). As a reminder, $\Omega_{\text{diag}} \in \mathbf{R}^{p \times p}$ is a matrix of the diagonal entries of Ω, with its off-diagonal entries set to zero; $\Omega_{\text{off}} \in \mathbf{R}^{p \times p}$ is a matrix of the off-diagonal entries of Ω, with its diagonal entries set to zero; and we form the SPACE estimate, for any iteration i, as $\hat{\Omega}^{\text{spc},(i)} = \hat{\Omega}_{\text{diag}}^{(i)} + \hat{\Omega}_{\text{off}}^{(i)}$. To be clear, the superscripts involving i here are interpreted just as with SPLICE above (also, we note that in the optimization problem (S.5), we have set the "weights" for each regression subproblem i to $\Omega_{\text{diag},ii}$, as recommended by Peng et al. (2009)).

Since the objectives in the defining optimization problems for many pseudolikelihood-based estimators include terms that go beyond pure lasso regressions, it is perhaps not clear that pseudolikelihood-based estimators can also saturate.

S.6.1 Proof of Theorem 4.4

Proof. We proceed by first showing that there exists a CONCORD estimate that saturates; then we show that the PseudoNet estimate does not saturate.

A CONCORD estimate is defined as a solution to the following (convex) optimization problem:

$$\underset{\Omega \in \mathbf{R}^{p \times p}}{\text{minimize}} \quad -(1/2) \log \det(\Omega_{\text{diag}}^2) + (n/2)\, \mathbf{Tr}\, S\Omega^2 + \lambda_1 \|\Omega_{\text{off}}\|_1, \tag{S.6}$$

where, as a reminder, $\Omega_{\text{diag}} \in \mathbf{R}^{p \times p}$ is a matrix of the diagonal entries of Ω, with its off-diagonal entries set to zero; $S \in \mathbf{R}^{p \times p}$ is the sample covariance matrix, *i.e.*, $S = (1/n)X^T X$, and $X \in \mathbf{R}^{n \times p}$ is a data matrix; $\Omega_{\text{off}} \in \mathbf{R}^{p \times p}$ is a matrix of the off-diagonal entries of Ω, with its diagonal entries set to zero; λ_1 is a tuning parameter; and $\| \cdot \|_1$ is the elementwise ℓ_1 norm.

Letting

$$\tilde{J}_{\text{con}}(\Omega_{\text{diag}}) = \inf_{\Omega_{\text{off}}} (1/2) \sum_{i=1}^{p} \left\| \sum_{j=1}^{p} \Omega_{ij} X_j \right\|_2^2 + \lambda_1 \|\Omega_{\text{off}}\|_1, \tag{S.7}$$

we see that the optimization problem (S.6) above is equivalent to

$$\underset{\Omega_{\text{diag}}}{\text{minimize}} \quad -(1/2) \log\det(\Omega_{\text{diag}}^2) + \tilde{J}_{\text{con}}(\Omega_{\text{diag}}).$$

Next, define

$$b = \begin{bmatrix} \Omega_{11} X_1 \\ \Omega_{22} X_2 \\ \Omega_{33} X_3 \\ \vdots \\ \Omega_{pp} X_p \end{bmatrix}, \quad \omega = \begin{bmatrix} \Omega_{12} \\ \Omega_{13} \\ \vdots \\ \Omega_{1p} \\ \Omega_{23} \\ \Omega_{24} \\ \vdots \\ \Omega_{2p} \\ \Omega_{34} \\ \Omega_{35} \\ \vdots \\ \Omega_{3p} \\ \vdots \\ \Omega_{p-1,p} \end{bmatrix},$$

i.e., $b \in \mathbf{R}^{np}$ and $\omega = \mathbf{vech}\,\Omega \in \mathbf{R}^{p(p-1)/2}$.

Then we can express (S.7) as

$$\inf_{\omega} (1/2)\|b - A\omega\|_2^2 + \lambda_1 \|\omega\|_1, \tag{S.8}$$

which is evidently a lasso problem with variable ω.

Then, by Tibshirani (2013, Lemma 14), for any b, A, and $\lambda_1 > 0$, there exists a solution $\hat{\omega}(\Omega_{\text{diag}})$ of (S.8) (note that we have written here the solution $\hat{\omega}$ as a function of Ω_{diag} to emphasize the dependence on Ω_{diag}) that will have at most $\min\{np, p(p-1)/2\}$ nonzero entries for any value of Ω_{diag}; thus, when $p \gg n$, $\mathbf{card}\,\hat{\omega}(\Omega_{\text{diag}}) \leq np$, as claimed. The final claim in the statement of the result follows by invoking Tibshirani (2013, Lemma 3).

Now, turning to the PseudoNet optimization problem (1), we have that the trace plus the squared Frobenius norm penalty in the objective in (1) can be expressed as

$$\begin{aligned} (n/2)\,\mathbf{Tr}\,S\Omega^2 + (\lambda_2/2) \sum_{i,j=1}^{p} \Omega_{ij}^2 &= (1/2) \sum_{i=1}^{p} \Omega_i^T X^T X \Omega_i + (\lambda_2/2) \sum_{i=1}^{p} \Omega_i^T \Omega_i \\ &= (1/2) \sum_{i=1}^{p} \Omega_i^T \left(X^T X + \lambda_2 I \right) \Omega_i \\ &= (1/2) \sum_{i=1}^{p} \left\| \sum_{j=1}^{p} \Omega_{ij} \begin{bmatrix} X_j \\ \sqrt{\lambda_2} e_j \end{bmatrix} \right\|_2^2, \end{aligned} \tag{S.9}$$

where, as a reminder, e_i is the ith standard basis vector in $\mathbf{R}^p$.

Thus, following a similar argument as above, we can express (1) as a lasso problem with variable $\omega \in \mathbf{R}^{p(p-1)/2}$, $A \in \mathbf{R}^{p(n+p) \times p(p-1)/2}$, and $b \in \mathbf{R}^{p(n+p)}$; however, in this case, the solution $\hat{\omega}(\Omega_{\text{diag}})$ can have $p(p-1)/2$ nonzeros, as claimed. □

S.6.2 Proof of Corollary 4.5

Proof. We prove these results by following a strategy similar to the one we used in the proof of Theorem 4.4. Note that, at the end of some iteration $i-1$, we can consider the variables D (for SPLICE) and Ω_{diag} (for SPACE) fixed, and then optimize over B (for SPLICE) and Ω_{off} (for SPACE). Accordingly, we let (for SPLICE)

$$b_{\text{spl}}^{(i-1)} = \begin{bmatrix} (1/\hat{D}_{11}^{(i-1)})X_1 \\ (1/\hat{D}_{22}^{(i-1)})X_2 \\ (1/\hat{D}_{33}^{(i-1)})X_3 \\ \vdots \\ (1/\hat{D}_{pp}^{(i-1)})X_p \end{bmatrix}, \quad \omega_{\text{spl}} = \begin{bmatrix} (B_{1\cdot})^T \\ (B_{2\cdot})^T \\ (B_{3\cdot})^T \\ \vdots \\ (B_{p\cdot})^T \end{bmatrix},$$

$$A_{\text{spl}}^{(i-1)} = \begin{bmatrix} (1/\hat{D}_{11}^{(i-1)})X_{-1} & 0 & 0 & 0 & \cdots & 0 \\ 0 & (1/\hat{D}_{22}^{(i-1)})X_{-2} & 0 & 0 & \cdots & 0 \\ 0 & 0 & (1/\hat{D}_{33}^{(i-1)})X_{-3} & 0 & \cdots & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & (1/\hat{D}_{pp}^{(i-1)})X_{-p} \end{bmatrix},$$

i.e., $b_{\text{spl}}^{(i-1)} \in \mathbf{R}^{np}$, $\omega_{\text{spl}} \in \mathbf{R}^{p(p-1)}$, and $A_{\text{spl}}^{(i-1)} \in \mathbf{R}^{np \times p(p-1)}$. We also let (for SPACE)

$$b_{\text{spc}}^{(i-1)} = \begin{bmatrix} \sqrt{\hat{\Omega}_{11}^{(i-1)}}X_1 \\ \sqrt{\hat{\Omega}_{22}^{(i-1)}}X_2 \\ \sqrt{\hat{\Omega}_{33}^{(i-1)}}X_3 \\ \vdots \\ \sqrt{\hat{\Omega}_{pp}^{(i-1)}}X_p \end{bmatrix}, \quad \omega_{\text{spc}} = \begin{bmatrix} \Omega_{12} \\ \Omega_{13} \\ \vdots \\ \Omega_{1p} \\ \Omega_{23} \\ \Omega_{24} \\ \vdots \\ \Omega_{2p} \\ \Omega_{34} \\ \Omega_{35} \\ \vdots \\ \Omega_{3p} \\ \vdots \\ \Omega_{p-1,p} \end{bmatrix},$$

$$A_{\text{spc}}^{(i-1)} = \begin{bmatrix} \tilde{X}2 & \tilde{X}3 & \tilde{X}4 & \cdots & \tilde{X}p-1 & \tilde{X}p & 0 & \cdots & 0 \\ \tilde{X}1 & 0 & \cdots & 0 & \tilde{X}3 & \tilde{X}4 & \tilde{X}5 & \cdots & \tilde{X}p-1 & \tilde{X}p & 0 & \cdots & 0 \\ 0 & \tilde{X}1 & 0 & \cdots & 0 & \tilde{X}2 & 0 & \cdots & 0 & \tilde{X}4 & \tilde{X}5 & \tilde{X}6 & \cdots & \tilde{X}p-1 & \tilde{X}p & 0 & \cdots & 0 \\ & & & & & & & \vdots & & & & & & & & & & \\ 0 & \cdots & 0 & \tilde{X}1 & 0 & \cdots & 0 & \tilde{X}2 & 0 & \cdots & 0 & \tilde{X}3 & 0 & \cdots & 0 & \tilde{X}p-1 \end{bmatrix},$$

where we write $\tilde{X}_j = \sqrt{\hat{\Omega}_{jj}^{(i-1)}}X_j$; so, $b_{\text{spc}}^{(i-1)} \in \mathbf{R}^{np}$, $\omega_{\text{spc}} \in \mathbf{R}^{p(p-1)/2}$, and $A_{\text{spc}}^{(i-1)} \in \mathbf{R}^{np \times p(p-1)/2}$. Applying Tibshirani (2013, Lemma 14) as before, and noting that applying (S.4) does not affect the sparsity pattern of $\hat{B}^{(i)}$ for SPLICE, gives the required results. □

S.7 STATEMENT OF REGULARITY CONDITIONS FOR AND PROOF OF THEOREM 4.2

S.7.1 Statement of regularity conditions for Theorem 4.2

Below, we state the regularity conditions required to establish the consistency of PseudoNet. The conditions are similar those required in Khare et al. (2015), which are in turn similar to those in Peng et al. (2009), except that here we must additionally control how the new tuning parameter λ_2 grows with n. When it is particularly helpful to emphasize the dependence of the tuning parameters λ_1 and λ_2 on n, we write $\lambda_{1,n} = \lambda_1$ and $\lambda_{2,n} = \lambda_2$.

i. *Sub-Gaussian rows.* We require that the rows of the data matrix X are i.i.d. sub-Gaussian random vectors, *i.e.*, there exists a constant $c \geq 0$ such that, for all $t \in \mathbf{R}^p$, we have that $\mathbf{E}\exp(t^T X_{i\cdot}) \leq \exp((c^2/2)t^T t),\ i = 1,\ldots,n$, where, as a reminder, $X_{i\cdot}$ is the ith row of X.

ii. *Correlation restrictions.* For all n, we require that the minimum and maximum eigenvalues of the underlying covariance matrix Σ^0, *i.e.*, $\lambda_{\min}(\Sigma^0)$ and $\lambda_{\max}(\Sigma^0)$, are uniformly bounded away from zero and infinity (note that we omit the notational dependence of Σ^0, as well as some related quantities, on n, for simplicity).

iii. *Incoherence.* We require that there exists a constant $\alpha < 1$ such that, for all $(i,j) \in \mathcal{A}_n^c$, where $\mathcal{A}_n$ here is the support of the off-diagonal entries of the underlying inverse covariance matrix Ω^0_{off}, *i.e.*,

$$\mathcal{A}_n = \left\{(i,j) : 1 \leq i < j \leq p,\ \Omega^0_{ij} \neq 0\right\},$$

we have that

$$\left|\bar{L}''_{ij,\mathcal{A}_n}(\omega^0_{\text{off}},\omega^0_{\text{diag}})(\bar{L}''_{\mathcal{A}_n\mathcal{A}_n}(\omega^0_{\text{off}},\omega^0_{\text{diag}}))^{-1}\,\mathbf{sign}\,\omega^0_{\mathcal{A}_n}\right| \leq \alpha. \tag{S.10}$$

Here, the **sign** here is interpreted elementwise; ω^0_{off} and ω^0_{diag} are the vectorizations of the off-diagonal and diagonal entries, respectively, of the underlying inverse covariance matrix Ω^0, *i.e.*,

$$\omega^0_{\text{off}} = \mathbf{vec}\,\Omega^0_{\text{off}}, \quad \omega^0_{\text{diag}} = \mathbf{vec}\,\Omega^0_{\text{diag}};$$

$L(\omega^0_{\text{off}},\omega^0_{\text{diag}})$ equals the $\log\det$ plus trace terms in (1) evaluated at $(\omega^0_{\text{off}},\omega^0_{\text{diag}})$, *i.e.*,

$$L(\omega^0_{\text{off}},\omega^0_{\text{diag}}) = -(1/2)\log\det((\Omega^0_{\text{diag}})^2) + (n/2)\,\mathbf{Tr}(S(\Omega^0)^2);$$

and $\bar{L}''_{ij,k\ell}(\omega^0_{\text{off}},\omega^0_{\text{diag}})$ is an element of the negative $(p^2 \times p^2)$-dimensional Fisher information matrix at $(\omega^0_{\text{off}},\omega^0_{\text{diag}})$, *i.e.*,

$$\bar{L}''_{ij,k\ell}(\omega^0_{\text{off}},\omega^0_{\text{diag}}) = \mathbf{E}\,\frac{\partial^2 L(\omega^0_{\text{off}},\omega^0_{\text{diag}})}{\partial\omega^0_{\text{off},ij}\omega^0_{\text{off},k\ell}}, \quad i,j,k,\ell = 1,\ldots,p$$

(we abuse notation somewhat and write $\omega_{ij} = \Omega_{ij}$).

iv. *Accurate diagonal estimates.* We require the existence of accurate diagonal estimates $\hat{\omega}_{\text{diag}}$ such that

$$\|\hat{\omega}_{\text{diag}} - \omega^0_{\text{diag}}\|_\infty = O_P(\sqrt{(\log n)/n}).$$

v. *Support size and tuning parameter restrictions.* As $n \to \infty$, we let $q_n = o(\sqrt{n/\log n})$, $\lambda_{1,n}\sqrt{q_n} \to 0$, $\lambda_{1,n}\sqrt{n/\log n} \to \infty$, and $\lambda_{2,n} = o(\lambda_{1,n})$, where $q_n = |\mathcal{A}_n|$ (note that we make explicit here the notational dependence of the tuning parameters on n).

vi. *Signal restrictions.* As $n \to \infty$, we require that $s_n/(\lambda_{1,n}\sqrt{q_n}) \to \infty$, where $s_n = \max_{(i,j)\in\mathcal{A}_n} |\omega^0_{\text{off},ij}|$.

Condition (iii) can be interpreted as requiring bounded correlation between the rows of $\bar{L}''_{\mathcal{A}_n^c\mathcal{A}_n}(\omega^0_{\text{off}},\omega^0_{\text{diag}})$ and the columns of $(\bar{L}''_{\mathcal{A}_n\mathcal{A}_n}(\omega^0_{\text{off}},\omega^0_{\text{diag}}))^{-1}$. Khare et al. (2015) as well as Peng et al. (2009) also use this condition; see Khare et al. (2015) for examples that satisfy this condition.

S.7.2 Proof of Theorem 4.2

Proof. Define $w_i = \hat{\Omega}^2_{ii},\ i = 1,\ldots,p$, where, as a reminder, the $\hat{\Omega}_{ii}$ are estimates of the diagonal entries of Ω^0 that are assumed in condition (iv) (see the statement of Theorem 4.2), and consider the change of variables for the off-diagonal entries of Ω

$$\omega_{ij} = -\theta_{ij}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}, \quad i,j = 1,\ldots,p,\ i \neq j,$$

where $\theta \in \mathbf{R}^{p(p-1)}$ and again $\omega = \mathbf{vec}\,\Omega$; then we can express the trace term in the objective in the PseudoNet optimization problem (1) as

$$
\begin{aligned}
n\,\mathbf{Tr}\,S\Omega^2 &= \sum_{i=1}^{p} (w_i/\hat{\Omega}_{ii}^2)\Omega_i^T X^T X \Omega_i \\
&= \sum_{i=1}^{p} (w_i/\hat{\Omega}_{ii}^2) \left\| \sum_{j=1}^{p} \omega_{ij} X_j \right\|_2^2 \\
&= \sum_{i=1}^{p} w_i \left\| (1/\hat{\Omega}_{ii}) \left(\hat{\Omega}_{ii} X_i + \sum_{j \neq i}^{p} \omega_{ij} X_j \right) \right\|_2^2 \\
&= \sum_{i=1}^{p} w_i \left\| X_i + \sum_{j \neq i}^{p} (\omega_{ij}/\hat{\Omega}_{ii}) X_j \right\|_2^2 \\
&= \sum_{i=1}^{p} w_i \left\| X_i - \sum_{j \neq i}^{p} \theta_{ij} \left(\hat{\Omega}_{jj}/\hat{\Omega}_{ii} \right)^{1/2} X_j \right\|_2^2 . \qquad \text{(S.11)}
\end{aligned}
$$

Equation S.11 is equal to the objective of the SPACE optimization problem (*cf.* Peng et al. (2009, Equation 10) and/or the trace term in Khare et al. (2015, Equation 12)), up to constants and for fixed diagonal entries; thus, the log det term (which is only a function of diagonal entries) plus the trace term in the objective in (1) are also equivalent to the corresponding terms in the SPACE's objective. This implies that properties A1–A4 and B0–B3 in the supplement for Peng et al. (2009) also apply to the log det plus trace terms in the objective in (1).

Now, let $L(\theta)$ denote the log det plus trace terms in the objective in (1) (with variable off-diagonal entries $\theta \in \mathbf{R}^{p(p-1)}$ and fixed diagonal entries $\hat{\omega}_{\text{diag}}$), and let $B_{c_1}(\theta^0_{\text{off}}, c_1 q_n^{1/2} \lambda_{1,n})$ be a ball of radius $c_1 q_n^{1/2} \lambda_{1,n}$, for a constant $c_1 > 0$, with center θ^0_{off}, *i.e.*, $B_{c_1} = \{\theta : \|\theta - \theta^0_{\text{off}}\|_2 \leq c_1 q_n^{1/2} \lambda_{1,n}\}$, where θ^0_{off} is the application of the same (strictly monotone) transformation in (S.7.2) to the underlying off-diagonal entries ω^0_{off}.

First, we show that the unique, global solution (owing to the strong convexity of (S.12)) of the following "restricted" optimization problem lies in B_{c_1} with probability tending to one as $n \to \infty$:

$$
\underset{\theta : \theta_{\mathcal{A}_n^c} = 0}{\text{minimize}} \quad L(\theta) + \lambda_{1,n} \textstyle\sum_{i \neq j}^{p} \left| (\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} \theta_{ij} \right| + (\lambda_{2,n}/2) \textstyle\sum_{i \neq j}^{p} \hat{\Omega}_{ii}\hat{\Omega}_{jj} \theta_{ij}^2 . \qquad \text{(S.12)}
$$

Let $\alpha_n = q_n^{1/2} \lambda_{1,n}$, and let $u \in \mathbf{R}^{p(p-1)}$ with $u_{\mathcal{A}_n^c} = 0$ and $\|u\|_2 = c$, for a constant $c > 0$. Fix $\theta \in B_{c_1}$ to be equal to $\theta^0_{\text{off}} + \alpha_n u$. Then we have that

$$
\begin{aligned}
&\lambda_{1,n} \left(\sum_{i \neq j}^{p} \left| (\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} \theta^0_{\text{off},ij} \right| - \sum_{i \neq j}^{p} \left| (\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} \theta_{ij} \right| \right) \\
&\quad \leq \lambda_{1,n} \sum_{i \neq j}^{p} \left| (\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} (\theta^0_{\text{off},ij} - \theta_{ij}) \right| \\
&\quad = \lambda_{1,n} \alpha_n \sum_{i \neq j}^{p} \left| (\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} u_{ij} \right| \\
&\quad = O(\lambda_{1,n} \alpha_n q_n^{1/2} \|u\|_2) \\
&\quad = O(\alpha_n^2), \qquad \text{(S.13)}
\end{aligned}
$$

with probability at least $1 - O(n^{-\beta})$, as the diagonal estimates $\hat{\Omega}_{ii}$ are uniformly bounded with high probability; the second line here follows by the triangle inequality, the third by the choice of θ, the fourth by the Cauchy-Schwarz inequality and the definition of u, and the fifth by the definition $\alpha_n = q_n^{1/2} \lambda_{1,n}$.

We also have that

$$
\begin{aligned}
&(\lambda_{2,n}/2)\left(\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}(\theta^0_{\text{off},ij})^2 - \sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}^2\right) \qquad \text{(S.14)}\\
&\quad = (\lambda_{2,n}/2)\left(\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}(\theta^0_{\text{off},ij})^2 - \sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}(\theta^0_{\text{off},ij}+\alpha_n u_{ij})^2\right)\\
&\quad = -\lambda_{2,n}\alpha_n\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta^0_{\text{off},ij}u_{ij} - (\lambda_{2,n}/2)\alpha_n^2\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}u_{ij}^2. \qquad \text{(S.15)}
\end{aligned}
$$

We get for the first term in (S.15) that

$$
\begin{aligned}
-\lambda_{2,n}\alpha_n\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta^0_{\text{off},ij}u_{ij} &\leq O(\lambda_{2,n}\alpha_n q_n^{1/2})\|u\|_2\\
&= o(\alpha_n^2)\|u\|_2, \qquad \text{(S.16)}
\end{aligned}
$$

with probability at least $1-O(n^{-\beta})$; the first line here follows by the Cauchy-Schwarz inequality, and the second by the assumption that $\lambda_{2,n} = o(\lambda_{1,n})$.

Similarly, we get for the second term in (S.15)

$$
-(\lambda_{2,n}/2)\alpha_n^2\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}u_{ij}^2 \leq o(\alpha_n^2)\|u\|_2^2, \qquad \text{(S.17)}
$$

with probability at least $1 - O(n^{-\beta})$.

Putting (S.16) and (S.17) together, we get for (S.14) that

$$
(\lambda_{2,n}/2)\left(\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}(\theta^0_{\text{off},ij})^2 - \sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}^2\right) \leq o(\alpha_n^2)\left(\|u\|_2 + \|u\|_2^2\right) \qquad \text{(S.18)}
$$

with probability at least $1 - O(n^{-\beta})$.

Next, let $J_{\text{net}}(\theta)$ equal the objective in (1) (with fixed diagonal entries $\hat{\omega}_{\text{diag}}$); combining (S.13) and (S.18), we get

$$
\begin{aligned}
J_{\text{net}}(\theta) - J_{\text{net}}(\theta^0_{\text{off}}) \geq\ & L(\theta) - L(\theta^0_{\text{off}})\\
&- \lambda_{1,n}\left(\sum_{i\neq j}^{p}\left|(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\theta^0_{\text{off},ij}\right| - \sum_{i\neq j}^{p}\left|(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\theta_{ij}\right|\right)\\
&- (\lambda_{2,n}/2)\left(\sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}(\theta^0_{\text{off},ij})^2 - \sum_{i\neq j}^{p}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}^2\right)\\
\geq\ & L(\theta) - L(\theta^0_{\text{off}}) - O(\alpha_n^2) - o(\alpha_n^2)\\
=\ & L(\theta) - L(\theta^0_{\text{off}}) - O(\alpha_n^2).
\end{aligned}
$$

By the same arguments in the proof of Lemma S-3 in the supplement for Peng et al. (2009), it follows that the (unique, global) solution to the restricted problem (S.12) lies in B_{c_1}, with probability at least $1 - O(n^{-\beta})$; this also implies (by a simple contradiction argument) that the event $\mathbf{sign}\,\hat{\theta}_{\mathcal{A}_n} = \mathbf{sign}\,\theta^0_{\mathcal{A}_n}$ occurs with high probability.

By construction, the solution $\hat{\theta}$ to the restricted optimization problem (S.12) satisfies the support "block" of the optimality conditions for the unrestricted optimization problem (1). Next, we show that $\hat{\theta}$ satisfies the non-support (the complement of the support) block of the optimality conditions for the unrestricted optimization problem (1).

The optimality conditions for the unrestricted optimization problem (1) are

$$\begin{array}{lll} L_{ij}^{'}(\theta) + \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij} & = -\lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\,\mathbf{sign}\,\theta_{ij} & \text{if } \theta_{ij} \neq 0 \\ |L_{ij}^{'}(\theta) + \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}| & \leq \lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2} & \text{if } \theta_{ij} = 0, \end{array} \tag{S.19}$$

where $L_{ij}^{'}(\theta) = \partial L(\theta)/\partial\theta_{ij}$; this establishes the analog of Lemma S-1 in the supplement for Peng et al. (2009), and also implies that Lemma S-2 there applies to the unrestricted optimization problem (1) here. We wish to show that (with high probability)

$$\max_{(i,j)\in\mathcal{A}_n^c} |L_{ij}^{'}(\hat{\theta}) + \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\hat{\theta}_{ij}| < \lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}.$$

We begin by taking an exact (since $L_{\mathcal{A}_n}^{'}$ is affine) first-order Taylor expansion of $L_{\mathcal{A}_n}^{'}(\hat{\theta})$ around θ^0, *i.e.*,

$$\begin{aligned} L_{\mathcal{A}_n}^{'}(\hat{\theta}) &= L_{\mathcal{A}_n}^{'}(\theta^0) + L_{\mathcal{A}_n\mathcal{A}_n}^{''}\underbrace{(\hat{\theta}-\theta^0)}_{v} \\ &= L_{\mathcal{A}_n}^{'}(\theta^0) + \underbrace{(L_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0) - \bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0))}_{\Delta_{\mathcal{A}_n\mathcal{A}_n}}v + \bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)v. \end{aligned} \tag{S.20}$$

However, we also have that, with probability at least $1 - O(n^{-\beta})$,

$$L_{\mathcal{A}_n}^{'}(\hat{\theta}) = -\lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\,\mathbf{sign}\,\theta_{\mathcal{A}_n}^0. \tag{S.21}$$

Equating (S.20) and (S.21) and rearranging, we get

$$v = -\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}\left(\lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\,\mathbf{sign}\,\theta_{\mathcal{A}_n}^0 + L_{\mathcal{A}_n}^{'}(\theta^0) + \Delta_{\mathcal{A}_n\mathcal{A}_n}v\right). \tag{S.22}$$

Repeating a similar analysis for any $(i,j) \in \mathcal{A}_n^c$, we get

$$L_{ij}^{'}(\hat{\theta}) = L_{ij}^{'}(\theta^0) + \Delta_{ij,\mathcal{A}_n}(\theta^0)v + \bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)v. \tag{S.23}$$

Now, plugging (S.22) into the third term on the righthand side of (S.23), we get

$$\begin{aligned} L_{ij}^{'}(\hat{\theta}) = L_{ij}^{'}(\theta^0) &+ \Delta_{ij,\mathcal{A}_n}(\theta^0)v \\ &- \lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}\mathbf{sign}\,\theta_{\mathcal{A}_n}^0 \\ &- \bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}L_{\mathcal{A}_n}^{'}(\theta^0) \\ &- \bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}\Delta_{\mathcal{A}_n\mathcal{A}_n}v. \end{aligned}$$

Applying the triangle inequality and rearranging yields

$$\begin{aligned} |L_{ij}^{'}(\hat{\theta})| \leq &\left|\lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}\bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}\mathbf{sign}\,\theta_{\mathcal{A}_n}^0\right| \\ &+ \left|\left(\Delta_{ij,\mathcal{A}_n}(\theta^0) - \bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}\Delta_{\mathcal{A}_n\mathcal{A}_n}\right)v\right| \\ &+ \left|\bar{L}_{ij,\mathcal{A}_n}^{''}(\theta^0)\left(\bar{L}_{\mathcal{A}_n\mathcal{A}_n}^{''}(\theta^0)\right)^{-1}L_{\mathcal{A}_n}^{'}(\theta^0)\right| \\ &+ |L_{ij}^{'}(\theta^0)|. \end{aligned}$$

The first term here is (strictly) less than $\lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}/2$ by condition (iii), and the remaining terms are $o(\lambda_{1,n})$, with probability at least $1 - O(n^{-\beta})$, by the same arguments in the proof of Peng et al. (2009, Theorem 2).

Now, let $R_{ij}^{'}(\theta) = \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}$; repeating a similar analysis as above, we get

$$\begin{aligned} R_{ij}^{'}(\hat{\theta}) &= R_{ij}^{'}(\theta^0) + \left(R_{ij,\mathcal{A}_n}^{''}(\theta^0) - \bar{R}_{ij,\mathcal{A}_n}^{''}(\theta^0)\right)v + \bar{R}_{ij,\mathcal{A}_n}^{''}(\theta^0)v \\ &= R_{ij}^{'}(\theta^0) + \bar{R}_{ij,\mathcal{A}_n}^{''}(\theta^0)v \\ &= \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}\theta_{ij}^0 + \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}v_{ij} \\ &\leq o(\lambda_{1,n}) + \lambda_{2,n}\hat{\Omega}_{ii}\hat{\Omega}_{jj}c_1 q_n^{1/2}\lambda_{1,n} \\ &= o(\lambda_{1,n}), \end{aligned}$$

where the penultimate line follows since $\|v\|_2 = \|\hat{\theta} - \theta^0\|_2 \leq c_1 q_n^{1/2}\lambda_{1,n} \implies v_{ij} \leq c_1 q_n^{1/2}\lambda_{1,n}$, and the last line since $q_n^{1/2}\lambda_{1,n} \to 0$ by condition (v).

Putting these findings together, we get, with probability at least $1 - O(n^{-\beta})$,

$$\max_{(i,j)\in\mathcal{A}_n^c} |L_{ij}^{'}(\hat{\theta}) + R_{ij}^{'}(\hat{\theta})| < \lambda_{1,n}(\hat{\Omega}_{ii}\hat{\Omega}_{jj})^{1/2}/2 + o(\lambda_{1,n}),$$

as required.

Thus, since the (unique, global) solution to the restricted optimization problem (S.12) satisfies the optimality conditions for the unrestricted optimization problem (1) (which also admits a unique, global solution), and since the restricted solution lies in B_{c_1}, we obtain the required results. □

S.8 STATEMENT OF REGULARITY CONDITIONS FOR AND PROOF OF THEOREM 4.3

The assumption that sufficiently accurate estimates of the diagonal entries of the underlying inverse covariance matrix $\Omega_{ii}^0,\ i = 1,\ldots,p$, are available is critical in establishing that the SPACE, CONCORD, and PseudoNet estimates are consistent, even in a high-dimensional setup. An obvious estimate for the underlying diagonal entries is $1/S_{ii},\ i = 1,\ldots,p$, where S here is the sample covariance matrix. However, S is not invertible when $p > n$, and so these estimates are not defined; an alternative approach is to consider the entries of a generalized inverse of S, but this turns out to be difficult, as well. Theorem 4.3 below instead provides a two-step method to obtain these estimates and rigorously establishes their accuracy, resolving an important gap in the literature, and may also be of independent interest.

S.8.1 Statement of regularity conditions for Theorem 4.3

Assume conditions (i), (ii), (v), and (vi) that were stated above for Theorem 4.2. Assume further that there exists a constant $\delta < 1$ such that

$$\left|\Sigma_{i,\mathcal{A}_n^j}^0\left(\Sigma_{\mathcal{A}_n^j,\mathcal{A}_n^j}^0\right)^{-1}\mathbf{sign}\,\Omega_{\mathcal{A}_n^j,j}^0\right| \leq \delta, \quad i \notin \mathcal{A}_n^j,\ j = 1,\ldots,p, \tag{S.24}$$

where

$$\begin{aligned} d_n &= \max_{k=1,\ldots,p}\left|\left\{\ell : \ell \in \{1,\ldots,p\},\ \ell \neq k,\ \Omega_{k\ell}^0 \neq 0\right\}\right|, \\ \mathcal{A}_n^j &= \left\{k : k \in \{1,\ldots,p\},\ k \neq j,\ \Omega_{jk}^0 \neq 0\right\}, \end{aligned}$$

and the **sign** in (S.24) is interpreted elementwise. As a reminder, d_n denotes the maximum number of nonzero entries in any row of Ω^0; when d_n is bounded in n, this theorem yields estimates satisfying condition (iv) above, even when $p > n$. We note that (S.24) is similar but not equivalent to condition (iii) above.

S.8.2 Proof of Theorem 4.3

We start by considering the estimation of the pth diagonal entry for ease of exposition. As discussed later, the argument below (all the way to Equation (S.47)) can be repeated verbatim for estimation of the ith diagonal entry with obvious notational changes.

Note that, since $d_n = O(q_n)$, conditions (i), (ii), (v), and (vi) imply that $d_n^{1/2}\lambda_{1,n} \to 0$, $d_n(\log n/n)^{1/2} \to 0$, and $(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$.

Let $(\eta^T, 1) = \Omega_{p\cdot}/\Omega_{pp}$, *i.e.*, η is the pth (off-diagonal) row of Ω divided by the pth diagonal entry. Let S again denote the sample covariance matrix. Consider the function

$$J_p(\eta) = (\eta^T, 1)S(\eta^T, 1)^T + \lambda_{1,n}\sum_{i=1}^{p-1}|\eta_i|,$$

where again $\lambda_{1,n}$ is the tuning parameter. This a convex function, and any global minimizer of this function will be sparse in η. This will immediately lead to an estimate of the sparsity in the pth row of Ω. The function J_p is the same objective function used by Meinshausen and Bühlmann (2006) in their neighborhood selection procedure (up to a simple transformation of the parameter η). Note that Meinshausen and Bühlmann (2006) provide a consistency proof for the sparsity pattern obtained by minimizing J_p under a set of regularity assumptions (for example, Gaussianity).[2] We provide a proof of sparsity selection consistency for J_p below under a set of related but different assumptions from those in Meinshausen and Bühlmann (2006) (for example, under a general sub-Gaussian tail setting).

Let η^0 denote the true value of the parameter η. Also, for ease of exposition, we use $\eta_p = \eta_p^0 = 1$ below, but the vector η will always refer to the $(p-1)$-dimensional parameter defined above. We now obtain the required result through a sequence of lemmas.

Lemma S.8.1. *For any $\gamma > 0$, there exists a constant $C_\gamma > 0$ such that, with probability at least $1 - O(n^{-\gamma})$,*

$$\max_{1\le i,j,\le p}|S_{ij} - \Sigma_{ij}^0| \le C_\gamma\sqrt{\frac{\log n}{n}},$$

for large enough n.

Proof. Fix $1 \le i, j \le p$. Let $\mu_+ = \mathbf{E}_{\Sigma_n^0}\left[(X_{1i} + X_{1j})^2\right]$ and $\mu_- = \mathbf{E}_{\Sigma_n^0}\left[(X_{1i} - X_{1j})^2\right]$. It follows that

$$\begin{aligned}
&\mathbf{Pr}(|S_{ij} - \Sigma_{ij}^0| > t)\\
&= \mathbf{Pr}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n}(X_{\ell i} + X_{\ell j})^2 - (X_{\ell i} - X_{\ell j})^2 - (\mu_+ - \mu_-)\right| > 4t\right)\\
&\le \mathbf{Pr}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n}(X_{\ell i} + X_{\ell j})^2 - \mu_+\right| > 2t\right) + \mathbf{Pr}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n}(X_{\ell i} - X_{\ell j})^2 - \mu_-\right| > 2t\right).
\end{aligned} \tag{S.25}$$

Note that $X_{\ell i} + X_{\ell j}$ are sub-Gaussian random variables (by condition (i)), and their variances are uniformly bounded in i, j, and n (by condition (ii)). For any $c_3 > 0$, it follows, by (S.25) and Rudelson and Vershynin (2013, Theorem 1.1), that there exist constants K_1 and K_2 independent of i, j, and n such that

$$\mathbf{Pr}\left(|S_{ij} - \Sigma_{ij}^0| > C\sqrt{\frac{\log n}{n}}\right) \le K_1 e^{-K_2 n\left(c_3\sqrt{\frac{\log n}{n}}\right)^2} = K_1 e^{-K_2 C^2 \log n},$$

for large enough n. Using the union bound and the fact that $p = O(n^\kappa)$, for some $\kappa > 0$, gives us the required result. □

Next, let

$$\tilde{L}(\eta) = (\eta^T, 1)S(\eta^T, 1)^T,$$

and let

$$d_i(\eta) = 2\sum_{j=1}^{p}\eta_j S_{ij}, \tag{S.26}$$

for $1 \le i \le p-1$, denote the elements of the gradient of $\tilde{L}$. Then we obtain the following results.

[2] Note that, by combining the sparsity patterns for all the rows of Ω using the neighborhood selection procedure, one can obtain an estimate for the sparsity pattern in Ω^0. However, a drawback is that the resulting pattern is not necessarily symmetric. On the other hand, our goal in this section is to show consistency of a procedure, which uses the sparsity pattern for neighborhood selection solely for estimating the diagonal entries of Ω^0.

Lemma S.8.2 (Optimality conditions). *η minimizes J_p if and only if*

$$\begin{array}{lll} d_i(\eta) & = -\lambda_{1,n}\,\mathbf{sign}\,\eta_i & \textit{if } \eta_i \neq 0,\ 1 \leq i \leq p-1 \\ |d_i(\eta)| & \leq \lambda_{1,n} & \textit{if } \eta_i = 0, 1 \leq i \leq p-1. \end{array} \tag{S.27}$$

Also, if $|d_i(\hat{\eta})| < \lambda_{1,n}$, for any minimizer $\hat{\eta}$, then by the continuity of d_i and the convexity of J_p, it follows that $\tilde{\eta}_i = 0$, for every minimizer $\tilde{\eta}$ of J_p.

Lemma S.8.3. *For every $1 \leq i \leq p-1$,*

$$\mathbf{E}_{\Sigma_n^0}\left[d_i(\eta^0)\right] = 0.$$

Proof. Let Σ_r^0 denote the submatrix of Σ^0 formed by using the first r rows and columns. It follows, by the definition of η^0, that, for every $1 \leq i < p$,

$$\mathbf{E}_{\Sigma^0}\left[d_i(\eta^0)\right] = 2\sum_{j=1}^{p} \eta_j^0 \Sigma_{ij}^0 = \frac{2}{\Omega_{pp}^0}\sum_{j=1}^{p} (\Sigma^0)^{-1}_{pj}\Sigma_{ij}^0 = 0.$$

□

Lemma S.8.4. *For any $\gamma > 0$, there exists a constant $C_{1,\gamma} > 0$ such that, with probability at least $1 - O(n^{-\gamma})$,*

$$\max_{1 \leq i \leq p} |d_i(\eta^0)| \leq C_{1,\gamma}\sqrt{\frac{\log n}{n}}.$$

Proof. It follows, by Lemma S.8.2, that

$$d_i(\eta^0) = \frac{2}{n}\sum_{\ell=1}^{n} X_{\ell i}\left(\sum_{j=1}^{p} \eta_j^0 X_{\ell j}\right)$$

is the difference between the sample covariance and population covariance of X_i and $\sum_{j=1}^p \eta_j^0 X_j$. It follows, by condition (ii) and the definition of η^0, that the variance of $\sum_{j=1}^p \eta_j^0 X_j$, given by $\left((\eta^0)^T, 1\right)\Sigma^0\left((\eta^0)^T, 1\right)^T$, is uniformly bounded over n. The proof now follows along the same lines as the proof of Lemma S.8.1. □

Note that $\mathcal{A}_n^p$ is the set of indices corresponding to the nonzero entries of η_n^0. Also note that $|\mathcal{A}_n^p| \leq d_n$. Next, we establish properties for the following "restricted" minimization problem:

$$\underset{\eta:\eta_j=0,\ j\notin\mathcal{A}_n^p}{\text{minimize}} \quad J_p(\eta). \tag{S.28}$$

Lemma S.8.5. *There exists $C > 0$ such that, for any $\gamma > 0$, a global minimum of the restricted minimization problem (S.28) exists within the ball $\{\eta : \|\eta - \eta^0\|_2 < C\sqrt{d_n}\lambda_{1,n}\}$, with probability at least $1 - O(n^{-\gamma})$ for sufficiently large n.*

Proof. Let $\tilde{\alpha}_n = \sqrt{d_n}\lambda_{1,n}$. Then, for any constant $C > 0$ and any $u \in \mathbf{R}^{p-1}$ satisfying $u_j = 0$ for every $j \notin \mathcal{A}_n^p$ and $\|u\|_2 = C$, we get by the triangle inequality that

$$\sum_{j=1}^{p-1}|\eta_j^0| - \sum_{j=1}^{p-1}|\eta_j^0 + \tilde{\alpha}_n u_j| \leq \tilde{\alpha}_n\sum_{j=1}^{p-1}|u_j| \leq C\tilde{\alpha}_n\sqrt{d_n}. \tag{S.29}$$

Again, let

$$\tilde{L}(\eta) = (\eta^T, 1)^T S(\eta^T, 1)^T.$$

By (S.29) and a second-order Taylor series expansion around η^0, we get

$$\begin{aligned}
&J_p(\eta^0 + \tilde{\alpha}_n u) - J_p(\eta^0) \\
&= \tilde{L}(\eta^0 + \tilde{\alpha}_n u) - \tilde{L}(\eta^0) - \lambda_{1,n} \left(\sum_{j=1}^{p-1} |\eta_j^0| - \sum_{j=1}^{p-1} |\eta_j^0 + \tilde{\alpha}_n u_j| \right) \\
&\geq \tilde{\alpha}_n \sum_{j \in \mathcal{A}_n^p} u_j d_j(\eta^0) + \tilde{\alpha}_n^2 \sum_{j \in \mathcal{A}_n^p} \sum_{k \in \mathcal{A}_n^p} u_j u_k S_{jk} - C \tilde{\alpha}_n \sqrt{d_n} \lambda_{1,n} \\
&\geq \tilde{\alpha}_n \sum_{j \in \mathcal{A}_n^p} u_j d_j(\eta^0) + \tilde{\alpha}_n^2 \sum_{j \in \mathcal{A}_n^p} \sum_{k \in \mathcal{A}_n^p} u_j u_k (S_{jk} - \Sigma_{jk}^0) + \tilde{\alpha}_n^2 \sum_{j \in \mathcal{A}_n^p} \sum_{k \in \mathcal{A}_n^p} u_j u_k \Sigma_{jk}^0 - C \tilde{\alpha}_n^2. \qquad \text{(S.30)}
\end{aligned}$$

Note that $\lambda_{1,n}\sqrt{\frac{n}{\log n}} \to \infty$ and $d_n\sqrt{\frac{\log n}{n}} \to 0$ as $n \to \infty$, since $(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$ and $d_n^{1/2}\lambda_{1,n} \to 0$. It follows, by the Cauchy-Schwarz inequality, Lemma S.8.1, and Lemma S.8.4, that for any $\gamma > 0$ there exist constants C_γ and $C_{1,\gamma} > 0$ such that, with probability at least $1 - O(n^{-\gamma})$,

$$\tilde{\alpha}_n \sum_{j \in \mathcal{A}_n^p} u_j d_j(\eta^0) \leq C C_{1,\gamma} \sqrt{\frac{d_n \log n}{n}} \tilde{\alpha}_n = o(\tilde{\alpha}_n^2) \qquad \text{(S.31)}$$

and

$$\frac{\tilde{\alpha}_n^2}{2} \left| \sum_{j \in \mathcal{A}_n^p} \sum_{k \in \mathcal{A}_n^p} u_j u_k (S_{jk} - \Sigma_{jk}^0) \right| \leq C_\gamma C^2 d_n \sqrt{\frac{\log n}{n}} = o(\tilde{\alpha}_n^2). \qquad \text{(S.32)}$$

Also, by condition (ii), it follows that

$$\sum_{j \in \mathcal{A}_n^p} \sum_{k \in \mathcal{A}_n^p} u_j u_k \Sigma_{jk}^0 \geq \frac{C^2 \tilde{\alpha}_n^2}{2\lambda_{\max}(\Omega^0)}. \qquad \text{(S.33)}$$

Combining (S.30), (S.31), (S.32), and (S.33), we get that

$$J_p(\eta^0 + \tilde{\alpha}_n u) - J_p(\eta^0) > \frac{C^2 \tilde{\alpha}_n^2}{2\lambda_{\max}(\Omega^0)} - 2C\tilde{\alpha}_n^2,$$

with probability at least $1 - O(n^{-\gamma})$, for large enough n.

Choosing $C = 4\lambda_{\max}(\Omega^0) + 1$, we obtain that

$$\inf_{u: u_{(\mathcal{A}_n^p)^c} = 0,\ \|u\|_2 = C} J_p(\eta^0 + \tilde{\alpha}_n u) > J_p(\eta^0),$$

with probability at least $1 - O(n^{-\gamma})$, for large enough n. Hence, for every $\eta > 0$, a local minimum (in fact a global minimum due to convexity) of the restricted minimization problem (S.28) exists within the ball $\{\eta : \|\eta - \eta^0\|_2 < C\sqrt{d_n}\lambda_{1,n}\}$, with probability at least $1 - O(n^{-\eta})$, for sufficiently large n. □

Lemma S.8.6. *There exists a constant $C_1 > 0$ such that, for any $\gamma > 0$, the following holds with probability at least $1 - O(n^{-\gamma})$.*

For any η in the set

$$S = \{\eta : \|\eta - \eta^0\|_2 \geq C_1 \sqrt{d_n}\lambda_{1,n},\ \eta_j = 0\ \forall j \notin \mathcal{A}_n^p\},$$

we have $\left\|d_{\mathcal{A}_n^p}(\eta)\right\|_2 > \sqrt{d_n}\lambda_{1,n}$, where $d_{\mathcal{A}_n^p}(\eta) = (d_j(\eta))_{j \in \mathcal{A}_n^p}$.

Proof. Recall that $\tilde{\alpha}_n = \sqrt{d_n}\lambda_{1,n}$. Choose $\eta \in S$ arbitrarily. Let $u = \eta - \eta^0/\tilde{\alpha}_n$. It follows that $u_j = 0$, for every $j \notin \mathcal{A}_n^p$ and $\|u\| \geq C_1$. By a first-order Taylor series expansion of $d_{\mathcal{A}_n^p}$, it follows that

$$\begin{aligned}
d_{\mathcal{A}_n^p}(\eta) &= d_{\mathcal{A}_n^p}(\eta^0) + 2\tilde{\alpha}_n S_{\mathcal{A}_n^p \mathcal{A}_n^p} u_{\mathcal{A}_n^p} \\
&= d_{\mathcal{A}_n^p}(\eta^0) + 2\tilde{\alpha}_n \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p} u_{\mathcal{A}_n^p} + 2\tilde{\alpha}_n \left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right) u_{\mathcal{A}_n^p}. \qquad \text{(S.34)}
\end{aligned}$$

By Lemma S.8.1 and Lemma S.8.4, it follows that, for any $\gamma > 0$, there exist constants $C_{2,\gamma}$ and $C_{3,\gamma}$ such that

$$\begin{aligned}
&\|d_{\mathcal{A}_n^p}(\eta)\|_2 \\
&\quad \geq 2\tilde{\alpha}_n \left\|\Sigma^0_{\mathcal{A}_n^p\mathcal{A}_n^p} u_{\mathcal{A}_n^p}\right\|_2 - C_{2,\gamma}\sqrt{\frac{d_n \log n}{n}} - C_{3,\gamma}\|u\|_2 \frac{\tilde{\alpha}_n d_n \sqrt{\log n}}{\sqrt{n}} \\
&\quad \geq \frac{\tilde{\alpha}_n}{\lambda_{\max}(\Omega^0)}\|u\|_2 \\
&\quad = \sqrt{d_n}\lambda_{1,n}\frac{C_1}{\lambda_{\max}(\Omega^0)},
\end{aligned}$$

with probability at least $1 - O(n^{-\gamma})$ for large enough n. The last inequality follows by condition (iii) and since $d_n(\log n/n)^{1/2} \to 0$.

Choosing $C_1 = \lambda_{\max}(\Omega^0) + 1$ leads to the required result. □

The next lemma establishes estimation and model selection (sign) consistency for the restricted minimization problem (S.28).

Lemma S.8.7. *There exists $C_2 > 0$ such that, for any $\gamma > 0$, the following holds with probability at least $1 - O(n^{-\gamma})$ for large enough n:*

a. there exists a solution to the restricted minimization problem (S.28)

b. (estimation consistency) any global minimum of the restricted minimization problem (S.28) lies within the ball $\{\eta : \|\eta - \eta^0\|_2 < C_2\sqrt{d_n}\lambda_{1,n}\}$

c. (sign consistency) for any solution $\hat{\eta}$ of the restricted minimization problem (S.28), $\mathbf{sign}\,\hat{\eta}_j = \mathbf{sign}\,\eta_j^0$, for every $1 \leq j \leq r$.

Proof. The existence of a solution follows from Lemma S.8.6.

By the optimality conditions for the restricted minimization problem (S.28) (along the lines of Lemma S.8.2), it follows that, for any solution $\hat{\eta}$ of (S.28), $|d_j(\hat{\eta})| \leq \lambda_{1,n}$, for every $j \in \mathcal{A}_n^p$. It follows that $\left\|d_{\mathcal{A}_n^p}(\hat{\eta})\right\|_2 \leq \sqrt{d_n}\lambda_{1,n}$. Estimation consistency now follows from Lemma S.8.7.

Note that, by condition (vi) and the fact that $d_n \leq q_n$,

$$\eta_j^0 \geq \frac{s_n}{\lambda_{\max}(\Omega^0)} > 2C_2\sqrt{d_n}\lambda_{1,n},$$

for every $j \in \mathcal{A}_n^p$ and for sufficiently large n. Sign consistency now follows by combining this fact with $\|\eta - \eta^0\|_2 < C_2\sqrt{d_n}\lambda_{1,n}$. □

The next lemma will be instrumental in showing that the solution set of the restricted minimization problem (S.28) is the same as the solution set of the unrestricted minimization problem for J_p with high probability.

Lemma S.8.8. *For any $\gamma > 0$, any solution $\hat{\eta}$ of (S.28) satisfies*

$$\max_{j \notin \mathcal{A}_n^p} |d_j(\hat{\eta})| < \lambda_{1,n},$$

with probability at least $1 - O(n^{-\gamma})$ for large enough n.

Proof. Let $\gamma > 0$ be given, and let $\hat{\eta}$ be a solution of (S.28). If $C_n = \{\mathbf{sign}\,\hat{\eta} = \mathbf{sign}\,\eta^0\}$, then $\mathbf{Pr}(C_n) \geq 1 - O(n^{-\gamma-\kappa})$ for large enough n (by Lemma S.8.7). Now, on C_n, it follows by a first-order expansion of $d_{\mathcal{A}_n^p}$ around η^0 and the optimality conditions for (S.28), that

$$\begin{aligned}
-\lambda_{1,n}\,\mathbf{sign}\,\eta^0_{\mathcal{A}_n^p} &= d_{\mathcal{A}_n^p}(\hat{\eta}) \\
&= d_{\mathcal{A}_n^p}(\eta^0) + 2S_{\mathcal{A}_n^p\mathcal{A}_n^p}\hat{u}_n \\
&= H_n\hat{u}_n + d_{\mathcal{A}_n^p}(\eta^0) + 2\left(S_{\mathcal{A}_n^p\mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p\mathcal{A}_n^p}\right)\hat{u}_n, \qquad \text{(S.35)}
\end{aligned}$$

where $\hat{u}_n = \hat{\eta} - \eta^0$, and $H_n = 2\Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}$.

Hence,

$$\hat{u}_n = -\lambda_{1,n} H_n^{-1} \,\mathbf{sign}\, \eta^0_{\mathcal{A}_n^p} - H_n^{-1} d_{\mathcal{A}_n^p}(\eta^0) - 2H_n^{-1}\left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right)\hat{u}_n. \tag{S.36}$$

Now, let us fix $j \notin \mathcal{A}_n^p$. By a first-order Taylor series expansion of d_j, it follows that

$$d_j(\hat{\eta}) = d_j(\eta^0) + 2S^T_{i,\mathcal{A}_n^p}\hat{u}_n.$$

Using (S.36), we get that

$$\begin{aligned} d_j(\hat{\eta}) &= d_j(\eta^0) + 2(S_{j,\mathcal{A}_n^p} - \Sigma^0_{j,\mathcal{A}_n^p})^T \hat{u}_n + 2(\Sigma^0_{j,\mathcal{A}_n^p})^T \hat{u}_n \\ &= -2\lambda_{1,n}(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1} \,\mathbf{sign}\, \eta^0_{\mathcal{A}_n^p} + d_j(\eta^0) - 2(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1} d_{\mathcal{A}_n^p}(\eta^0) + \\ &\quad -4(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1}\left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right)\hat{u}_n + 2(S_{i,\mathcal{A}_n^p} - \Sigma^0_{i,\mathcal{A}_n^p})^T \hat{u}_n. \end{aligned} \tag{S.37}$$

We now individually analyze all the terms in (S.37).

It follows, by (S.24), that the first term satisfies

$$\left| -2\lambda_{1,n}(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1} \,\mathbf{sign}\, \eta^0_{\mathcal{A}_n^p} \right| \le \delta\lambda_{1,n} < \lambda_{1,n}. \tag{S.38}$$

It follows, by Lemma S.8.4 and since $(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$ and $d_n^{1/2}\lambda_{1,n} \to 0$, that the second term $d_j(\eta^0)$ is $o(\lambda_{1,n})$ with probability at least $1 - O(n^{-\gamma-\kappa})$ for large enough n.

Also, by condition (ii) and the definition of H_n, we get that

$$\left\| 2(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1} \right\|_2 \le \left\| \Sigma^0_{j,\mathcal{A}_n^p} \right\|_2 \|2H_n^{-1}\|_2 \le \frac{1}{\lambda_{\min}(\Omega^0)} \left\| \left(\Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right)^{-1} \right\|_2 \le \frac{\lambda_{\max}(\Omega^0)}{\lambda_{\min}(\Omega^0)}, \tag{S.39}$$

where $\|\cdot\|_2$ here denotes the ℓ_2 operator norm (maximum singular value). It follows, by Lemma S.8.4 and since $(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$ and $d_n^{1/2}\lambda_{1,n} \to 0$, that the third term in (S.37) satisfies

$$\left| 2(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1} d_{\mathcal{A}_n^p}(\eta^0) \right| \le \frac{\lambda_{\max}(\Omega^0)}{\lambda_{\min}(\Omega^0)} \sqrt{d_n} \max_{j \in \mathcal{A}_n^p} |d_j(\eta^0)| = o(\lambda_{1,n}). \tag{S.40}$$

Let $b = 2H_n^{-1}\Sigma_{j,\mathcal{A}_n^p}$. Note that, by (S.39), the norm of b is uniformly bounded in n and r. Also note that the jth element of the vector $\left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right) b$ is the difference between the sample and the population covariance of X_j and $\sum_{k \in \mathcal{A}_n^p} b_k X_k$. Using the same line of arguments as in the proof of Lemma S.8.4, it follows that there exists a constant $C_{4,\gamma} > 0$ such that

$$\max_{j \in \mathcal{A}_n^p} \left| \left(\left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right) b \right)_j \right| \le C_{4,\gamma}\sqrt{\frac{\log n}{n}}, \tag{S.41}$$

with probability at least $1 - O(n^{-\gamma-\kappa})$ for large enough n. By (S.39), (S.41), claim (b) in Lemma S.8.7, and since $(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$ and $d_n^{1/2}\lambda_{1,n} \to 0$, we have that the fourth term in (S.37) satisfies

$$\begin{align} \left| 4(\Sigma^0_{j,\mathcal{A}_n^p})^T H_n^{-1}\left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right)\hat{u}_n \right| &\le 2\left\| \left(S_{\mathcal{A}_n^p \mathcal{A}_n^p} - \Sigma^0_{\mathcal{A}_n^p \mathcal{A}_n^p}\right) b \right\|_2 \|\hat{u}_n\|_2 \notag \\ &= O\left(\sqrt{\frac{d_n \log n}{n}} \sqrt{d_n}\lambda_{1,n} \right) \tag{S.42} \\ &= o(\lambda_{1,n}), \tag{S.43} \end{align}$$

with probability at least $1 - O(n^{-\gamma-\kappa})$ for large enough n.

By Lemma S.8.1, claim (b) in Lemma S.8.7, and condition (ii), the fifth term in (S.37) satisfies

$$\left|2(S_{i,\mathcal{A}_n^p} - \Sigma^0_{i,\mathcal{A}_n^p})^T \hat{u}_n\right| \leq 2\left\|S_{i,\mathcal{A}_n^p} - \Sigma^0_{i,\mathcal{A}_n^p}\right\|_2 \|\hat{u}_n\|_2 = O\left(\sqrt{\frac{d_n \log n}{n}}\sqrt{d_n}\lambda_{1,n}\right) = o(\lambda_{1,n}). \tag{S.44}$$

It follows, by (S.37), (S.38), (S.40), and (S.42)-(S.44), that, for any $j \notin \mathcal{A}_n^p$,

$$|d_j(\hat{\eta})| < \lambda_{1,n},$$

with probability at least $1 - O(n^{-\gamma-\kappa})$ for large enough n. The result now follows by the union bound, and from the fact that $p = O(n^\kappa)$. □

Let $\gamma > 0$ be chosen arbitrarily. Let $C_{p,n}$ denote the event on which Lemma S.8.7 and Lemma S.8.8 hold. It follows that $\mathbf{Pr}(C_{p,n}) \geq 1 - O(n^{-\gamma-\kappa})$, for large enough n. Now, on $C_{p,n}$, any solution of the restricted problem (S.28) is also a global minimizer of J_p (by Lemma S.8.2). Hence, there is at least one global minimizer of J_p for which the components corresponding to $(\mathcal{A}_n^p)^c$ are zero. It again follows, by Lemma S.8.2, that these components are zero for all global minimizers of J_p. Hence, the solution set of the restricted minimization problem (S.28) is the same as the solution set for the unrestricted problem (*i.e.*, the set of global minimizers of J_p). Hence, on $C_{p,n}$, the assertions of Lemma S.8.7 hold for the solutions of the unrestricted minimization problem for J_p.

Now, let $\mathcal{B}_n^p = \mathcal{A}_n^p \cup \{p\}$. Using the sparsity in Ω^0 it can be shown that Ω^0_{pp} is also the diagonal entry corresponding to the index p in $\left(\Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}$. Let $\hat{\mathcal{A}}_n^p$ be the set of indices corresponding to the nonzero entries of any minimizer $\hat{\eta}$ of J_p, let $\hat{\Omega}_{pp}$ be the diagonal entry corresponding to the index p for $\left(S_{\hat{\mathcal{B}}_n^p\hat{\mathcal{B}}_n^p}\right)^{-1}$, and let $\hat{\mathcal{B}}_n^p = \hat{\mathcal{A}}_n^p \cup \{p\}$. It follows that $\hat{\mathcal{B}}_n^p = \mathcal{B}_n^p$ on $C_{p,n}$, and that

$$\begin{aligned}
|\hat{\Omega}_{pp} - \Omega^0_{pp}| &\leq \left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1} - \left(\Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \\
&\leq \left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \left\|S_{\mathcal{B}_n^p\mathcal{B}_n^p} - \Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right\|_2 \left\|\left(\Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \\
&\leq \lambda_{\max}(\Omega^0)\left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \left\|S_{\mathcal{B}_n^p\mathcal{B}_n^p} - \Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right\|_2 \\
&\leq d_n\lambda_{\max}(\Omega^0)\left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \max_{1\leq i,j\leq p}|S_{ij} - \Sigma^0_{ij}|. \qquad (S.45)
\end{aligned}$$

Note that, by Lemma S.8.1, there exists a constant $C_{\gamma+\kappa}$ such that

$$\|S - \Sigma^0_n\|_{\max} = \max_{1\leq i,j\leq p}|S_{ij} - \Sigma^0_{ij}| \leq C_{\gamma+\kappa}\sqrt{\frac{\log n}{n}},$$

with probability at least $1 - O(n^{-\gamma-\kappa})$ for large enough n. Let D_n denote the event on which the above inequality holds. Hence, on D_n, we get

$$\begin{aligned}
\left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 &\leq \left\|\left(\Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 + \left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1} - \left(\Sigma^0_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \\
&\leq \lambda_{\max}(\Omega^0) + d_n\lambda_{\max}(\Omega^0)\left\|\left(S_{\mathcal{B}_n^p\mathcal{B}_n^p}\right)^{-1}\right\|_2 \max_{1\leq i,j\leq p}|S_{ij} - \Sigma^0_{ij}| \\
&\leq \lambda_{\max}(\Omega^0) + \lambda_{\max}(\Omega^0)C_{\gamma+\kappa}d_n\sqrt{\frac{\log n}{n}} \qquad (S.46)
\end{aligned}$$

for large enough n. It follows, by (S.45), (S.46), and since $d_n(\log n/n)^{1/2} \to 0$, that on $C_{p,n} \cap D_n$

$$|\hat{\Omega}_{pp} - \Omega^0_{pp}| \leq 2\lambda^2_{\max}(\Omega^0)C_{\gamma+\kappa}d_n\sqrt{\frac{\log n}{n}} \tag{S.47}$$

for large enough n.

For every $1 \leq i \leq p$, the above argument can be repeated verbatim by considering η to be the ith (off-diagonal) row of Ω^0 normalized by the corresponding entry, and constructing the J_i, $\mathcal{A}_n^i$, *etc.* accordingly. Then, by maximizing J_i, we can obtain $\hat{\mathcal{A}}_n^i$ such that there exists a set $C_{i,n}$ with $\mathbf{Pr}(C_{i,n}) = 1 - O(n^{-\gamma-\kappa})$ for large enough n, and $\hat{\mathcal{A}}_n^i = \mathcal{A}_n^i$ on $C_{i,n}$. Again, it can be shown in exactly the same way as above (for the case of the pth row), that if $\hat{\Omega}_{ii}$ is the diagonal entry corresponding to the index i for $\left(S_{\hat{\mathcal{B}}_n^i \hat{\mathcal{B}}_n^i}\right)^{-1}$, then on $C_{i,n} \cap D_n$

$$|\hat{\Omega}_{ii} - \Omega_{ii}^0| \leq 2\lambda_{\max}(\Omega^0)^2 C_{\gamma+\kappa} d_n \sqrt{\frac{\log n}{n}}. \tag{S.48}$$

It follows, by (S.47) and (S.48), that on $(\cap_{i=1}^p C_{i,n}) \cap D_n$

$$\max_{1 \leq i \leq p} |\hat{\Omega}_{ii} - \Omega_{ii}^0| \leq 2\lambda_{\max}^2(\Omega^0) C_{\gamma+\kappa} d_n \sqrt{\frac{\log n}{n}}. \tag{S.49}$$

Since

$$\mathbf{Pr}\left((\cap_{i=1}^p C_{i,n}) \cap D_n\right) \geq 1 - (p+1)O(n^{-\gamma-\kappa}) = 1 - O(n^{-\gamma})$$

for large enough n, we have achieved our goal.

Note that the estimation accuracy in Lemma S.8.7 is $\sqrt{d_n}\lambda_{1,n}$. Hence, an estimate of Ω_{pp} based on $\hat{\eta}$ has estimation accuracy larger than or equal to $\sqrt{d_n}\lambda_{1,n}$. Since

$$d_n \sqrt{\frac{\log n}{n}} = \sqrt{d_n} \sqrt{\frac{d_n \log n}{n}} = o(\sqrt{d_n}\lambda_{1,n}),$$

$(1/\lambda_{1,n})((d_n/n)\log n)^{1/2} \to 0$, and $d_n^{1/2}\lambda_{1,n} \to 0$, it follows that a two-step procedure gives a provably better estimation accuracy than direct lasso based estimates of the diagonal entries of Ω^0.

References

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. The multiple quantile graphical model. In *Advances in Neural Information Processing Systems*, 2016. Available at `http://arxiv.org/pdf/1607.00515.pdf`.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9: 485–516, 2008.

Tao Hong, Pierre Pinson, and Shu Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30:357–363, 2014.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B*, 77(4):803–825, 2015.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13:781–794, 2012.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Sang-Yun Oh, Onkar Dalal, Kshitij Khare, and Bala Rajaratnam. Optimization methods for sparse pseudolikelihood graphical model selection. In *Advances in Neural Information Processing Systems*, pages 667–675. 2014.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

Mark Schmidt, Nicolas Roux, and Francis Bach. Convergence rates of inexact proximal gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 1458–1466, 2011.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B*, 74(2):245–266, 2012.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Matt Wytock and J. Zico Kolter. Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1265–1273, 2013.