Asymptotic Normality of Graphical LASSO(s)

Torben Staud

January, 2024

Contents

Introduction

Consider the following setting. Let $\mathcal{G}$ be an undirected graph with vertices $\mathcal{V} = \{1, \dots, p\}$, $p \in \mathbb{N}$. Let the distribution of the vector $X^0 = (X_1, \dots, X_p)$ be determined by the graphical model of $\mathcal{G}$, where a vertex between X_i and X_j denotes conditional dependence given $X^{-i,j}$.

In the case that X is Gaussian with covariance matrix Σ and precision matrix $\Theta = \Sigma^{-1}$ we already heard often, that conditional independence of X_i and X_j given the rest of X is equivalent to $\Theta_{i,j} = 0$. Hence, in order to make inference about the graphical dependence structure, estimating the precision matrix Θ is well motivated. Even though this correspondence is generally not valid for other distributions, in the literature estimating the precision matrix is still an active research field (e.g. for Subgaussian families).

In the paper(s) they consider Subgaussian random variables (random variables which have a tail decay which is at least as fast as a Gaussian random variable, in order to obtain concentration inequalities).

The papers do not focus on the traditional situation where n is big and the number of coordinates p is fixed and small compared to n. Instead they investigate the situation of $p = p_n$ and allow even for $p = o(n), n = o(p)$. These settings are called *large dimensional* and *high dimensional*, respectively. In those cases it is known that the sample covariance does that perform well (it is singular with probability tending to 1). This among other has sparked interest in researching different types of estimators for the precision matrix in high dimensional statistics. Note, that if $p >> n$ and the model is not sparse (meaning only few parameters are $\equiv 0$) or has any other specification, which decreases the dimension in a suitable sense, it is to my understanding, simply not possible to obtain good estimators.

In the papers [Janková and van de Geer, 2019], [Rothman, 2008] presented here **Graphical LASSO**-type estimators were investigated, which are penalized maximum likelihood based estimators, which induce sparsity via a ℓ_1 penalty (L(east)A(bsolute)S(hrinkage and)S(election)O(perator)). It is called like that because firstly, norm penalization implies for growing penalty λ a (norm) shrinking optimal parameter. And selection because the choice of the ℓ_1 norm leads to sparsity. As Holger said, in certain models, it shrinks exactly to zero. Rothman et. al. called the LASSO Sparce permutation invariant co-

variance estimator. Now there are two categories of LASSO-type estimation: The first being *global* methods, which typically estimate the whole precision matrix (by a ℓ_1 regularized log-likelihood). The second being *nodewise* methods, which estimate columns (or smaller parts) of the precision matrix individually one by one.

1 Asymptotic Normality in the low dimensional setting

It is instructive to investigate asymptotic normality for the model of p-dimensional centered multivariate normal distributions with regular covariance matrix $\Sigma_0 \in \mathbb{R}^{p\times p}$ and $\Theta_0 = \Sigma^{-1} \in \mathbb{R}^{p\times p}$, where p is **fixed**. Let $\hat{\Sigma} := X^tX/n$ denote the sample covariance matrix, where $X := (X^1, \dots, X^n)^t$ denotes the design matrix of the samples $X^i \in \mathbb{R}^p; i = 1, \dots, n$. $\hat{\Sigma}$ is the MLE for Σ (Marius sketched the proof last time) and by the functional invariance of the MLE we know that $\hat{\Theta} := \hat{\Sigma}^{-1}$ (which exists with probability 1 for absolutely continuous random variables by the Spectral Theorem) is the MLE for $\Theta = \Sigma^{-1}$. We then have

$$\hat{\Theta} - \Theta_0 = -\Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 + \text{rem}_0, \tag{1.1}$$

where $\text{rem}_0 = -\Theta_0(\hat{\Sigma} - \Sigma_0)(\hat{\Theta} - \Theta_0)$, since

$$\begin{aligned}
&\Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 - \Theta_0(\hat{\Sigma} - \Sigma_0)(\hat{\Theta} - \Theta_0) \\
&= -\Theta_0(\hat{\Sigma} - \Sigma_0)(\Theta_0 + \hat{\Theta} - \Theta_0) \\
&= (I - \Theta_0\hat{\Sigma})\hat{\Theta} \\
&= \hat{\Theta} - \Theta_0.
\end{aligned}$$

Now by the multivariate CLT (we will see why in a later proof) $\hat{\Sigma} - \Sigma_0 = O_{\mathbb{P}}(1/\sqrt{n})$, hence $\hat{\Theta} - \Theta_0 = o_{\mathbb{P}}(1)$ and thus $\text{rem}_0 = o_{\mathbb{P}}(1/\sqrt{n})$. The asymptotic normality now follows from (1.1). Problems in the high-dimensional settings?

- The sample covariance becomes singular
- Since $p = p_n$ we have no multivariate CLT at hand.

A workaround is a non-sparsity penalized MLE called **Graphical LASSO (GLASSO)** or **SPICE**:

$$\hat{\Theta} := \text{argmin}_{\Theta=\Theta^t,\Theta>0}\, \text{tr}(\hat{\Sigma}\Theta) - \log\det\Theta + \lambda\|\Theta\|_{\text{off},1},$$

where $\|\Theta\|_{\text{off},1} := \sum_{i\neq j}|\Theta_{i,j}|$ is the induced ℓ^1-norm without the diagonal as we do not want to punish variances. In 2000 Knight and Fu [Knight and Fu, 2000] proved (in the linear regression model) that among other the limiting distribution of the LASSO estimator may have positive mass at 0 if 0 is the true parameter, this was complemented by a paper of Pötscher and Leeb in 2014 [Pötscher and Leeb, 2009] Also the limiting distribution (already in the low dimensional setting) depends in a difficult way on the true parameter. This was a major difficulty in constructing confidence intervals. In 2014 van de Geer et. al. [van de Geer et al., 2014] introduced a method to remove the sparsity in a certain sense, result in asymptotic normality:

2 De-biasing the GLASSO

We will start by mentioning *One Step Estimators*, whose motivation and some properties can be found in Van der Vaart [van der Vaart, 1998]. The idea is based on having a prelimary or initial estimator and to improve upon in a certain way, when the estimator is defined as a root of a problem (Z-estimator, often MLE).

Assume R_n is a real valued risk function and regular in a certain sense. The ordinary MLE may then be translated to the root finding problem

$$\dot{R}_n(\Theta) = 0. \tag{2.1}$$

Now introducing the penalty leads (often) to the root problem

$$\dot{R}_n(\Theta) + \xi(\Theta) = 0, \tag{2.2}$$

where ξ is a subgradient (generalization of the differential to convex functions) corresponding to the penalty (in our case $\|\cdot\|_1$ is convex). The idea is, to use the one step estimator for (2.1) on the estimator obtained by (2.2) in a first order (linear) Taylor expansion of $R_n(\Theta)$ around $R_n(\tilde{\Theta})$, where the latter is an initial estimator:

$$\begin{aligned} 0 = R_n(\Theta) &\approx R_n(\tilde{\Theta}) + \dot{R}_n(\tilde{\Theta})(\Theta - \tilde{\Theta}) \\ \iff \Theta &= \tilde{\Theta} - \dot{R}_n^{-1}(\tilde{\Theta}) R_n(\tilde{\Theta}). \end{aligned}$$

In that way we "remove" the bias in the one step estimator procedure but keep the sparsity in a sense, as the initial estimator was constructed as a penalized estimator.

Lets move on to the specific GLASSO case. We know that without penalization (2.1) corresponds to

$$\hat{\Sigma}\Theta = I, \tag{2.3}$$

(that's the MLE and it's functional invariance). But as mentioned (2.3) is with high probability not solvable as $\hat{\Sigma}$ becomes singular. Instead of the linear approximation we require of our initial estimator $\tilde{\Theta}$ that

$$\hat{\Sigma}\tilde{\Theta} - I = -\eta(\tilde{\Theta}), \tag{2.4}$$

where $\eta(\tilde{\Theta})$ is small. One can now show by calculation that by (2.4)

$$\tilde{\Theta} + \tilde{\Theta} \cdot \eta(\tilde{\Theta}) - \Theta_0 = -\Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 + \text{rem}_0 + \text{rem}_{\text{reg}}, \tag{2.5}$$

where $\text{rem}_{\text{reg}} := (\hat{\Theta} - \Theta_0)\eta(\hat{\Theta})$. Using (2.4) yields $\eta(\tilde{\Theta}) = -(\hat{\Sigma}\tilde{\Theta} - I)$ and thus by (2.5) we define the de-sparsified estimator:

$$\hat{\Theta}_D := 2\hat{\Theta} - \hat{\Theta}\hat{\Sigma}\hat{\Theta}, \tag{2.6}$$

where $\hat{\Theta}$ is the GLASSO as the approximate inverse of $\hat{\Sigma}$. Note, that compared to (1.1) the rem_{reg} term is new and stems from the regularization. We will find that indeed the remainder terms are small enough (essentially because of the Oracle inequalities Marius talked about last time).

3 Conditions

There are three central conditions under which asymptotic normality may be shown.

Condition 3.1 (Bounded Eigenvalues). The precision matrix $\Theta = \Sigma^{-1}$ exists and there is a universal constant $L \geq 1$ such that

$$\frac{1}{L} \leq \Lambda_{\min}(\Theta_0) \leq \Lambda_{\max}(\Theta_0) \leq L,$$

where $\Lambda_{\min}(\cdot), \Lambda_{\max}(\cdot)$ denote the minimal and maximal eigenvalue of a matrix, respectively.

In the low dimensional case the former condition is more of a notation, as invertibility of the Covariance Matrix is included in the model anyways. But in the large- and high-dimensional setting $p = p_n \to \infty$ and thus both $\Lambda_{\min}(\Theta_0) \to 0$ and $\Lambda_{\max}(\Theta_0) \to \infty$ are possible.

Condition 3.2 (Sub-Gaussian entries). The design matrix X has indepent rows $X^1, \ldots, X^n \in \mathbb{R}^p$ which have zero mean and each entry (of a vector) is Sub-Gaussian with a universal parameter $K > 0$, meaning:

$$\mathbb{P}(|X_j^i| \geq t) \leq 2 \exp\left(-\frac{t^2}{K^2}\right), \quad (t \geq 0), i = 1, \ldots, n, j = 1, \ldots, p.$$

It is important, that the parameter is universal as there are p_n many. There are 200 equivalent definitions of Sub-Gaussianity like bounds for exponential moments, Laplace Transformation bounds etc; just keep in mind that this means fast tail decay which leads to good concentration inequalities.

Recall $\mathcal{V} = \{1, \ldots, p\}$. To encode that the precision matrix is sparse, meaning it thas a lot of zero entries, we denote the following unknown constants of Θ_0 for $j \in \mathcal{V}$:

$$D_j := \{(i,j) \colon i \in \mathcal{V}, i \neq j, \Theta^0_{i,j} \neq 0\}, \quad d_j := |D_j|, \quad d := \max_{j=1,\ldots,p} d_j.$$

The number d_j is the degree of the node j and describes, how much information the i-th coordinate has. Define

$$S := \bigcup_{j=1}^{p} D_j, \quad s := \sum_{j=1}^{p} d_j,$$

where $s \leq p^2$ measures sparsity (if s is small Θ_0 is sparse in our sense). In the sparse settings we need s to increase slower with regards to n.

4 Asymptotic Normality of the Desparsified Global LASSOs

Theorem 4.1 (Asymptotic Normality for High Dimensions). *Assume Conditions 3.1, 3.2 and the sparsity Condition* $(p+s)\sqrt{d} = o(\sqrt{n}/\log p)$. *Then, for* $\lambda \asymp \sqrt{\log p/n}$,

$$\| \text{rem} \|_\infty = o_{\mathbb{P}}(n^{-1/2}). \tag{4.1}$$

Furthermore, for $i, j = 1, \dots, p$,

$$\sqrt{n} \frac{(\hat{\Theta}_D - \Theta_0)_{i,j}}{\sigma_{i,j}} \rightsquigarrow \mathcal{N}(0,1). \tag{4.2}$$

Before we give a proof of the Theorem, we briefly mention two things: In order to obtain confidence intervals, we need a consistent estimator for the asymptotic variance $\sigma_{i,j}^2$. For the Gaussian case there are estimators at hand, which were proven to be consistent see [Janková and Van de Geer, 2017] Section 3.1. Secondly, recall the **Weighted GLASSO**, which Marius introduced in the last talk:

$$\hat{\Theta}_W := \text{argmin}_{\Theta=\Theta^t, \Theta>0} \, \text{tr}(\hat{\Sigma}\Theta) - \log\det\Theta + \sum_{i \neq j} \hat{W}_{i,i}\hat{W}_{j,j}|\Theta_{i,j}|,$$

where $\hat{W}^2 := \text{diag}(\hat{\Sigma})$. The weighting results in better oracle bounds (last talk) and this allows for the weaker sparsity condition $s\sqrt{d} = o(\sqrt{n}/\log p)$:

Theorem 4.2. *Assume Conditions 3.1, 3.2 and the sparsity Condition* $s\sqrt{d} = o(\sqrt{n}/\log p)$. *Then, for* $\lambda \asymp \sqrt{\log p/n}$,

$$\| \text{rem} \|_\infty = O_{\mathbb{P}}(n^{-1/2}).$$

Furthermore, for $i, j = 1, \dots, p$,

$$\sqrt{n} \frac{(\hat{\Theta}_W - \Theta_0)_{i,j}}{\sigma_{i,j}} \rightsquigarrow \mathcal{N}(0,1).$$

Since $n = o(p)$ in the high-dimensional case, the latter Theorem might be understood as a formulation for the $p >> n$ regime, as there is no restriction on the growth of p anymore.

Proof of Theorem 4.1. We will start by proving (4.1). For that we will need a result from convex optimization applied to $\hat{\Theta}$. The Karush-Kuhn-Tucker Conditions (generalization of First Order conditions in multivariate minimization problems to convex functions) and the invertibility of $\hat{\Theta}$ by definition yield the existence of a $\hat{Z} \in \mathbb{R}^{p \times p}$ with

$$\hat{\Sigma} - \hat{\Theta}^{-1} + \lambda \hat{Z} = 0,$$

where $\hat{Z}_{i,j} = \text{sign}\,\hat{\Theta}_{i,j}$ if $\hat{\theta}_{i.j} \neq 0$ and $\|\hat{Z}\|_\infty \leq 1$. Multiplying the upper display by $\hat{\Theta}$ yields $\hat{\Sigma}\hat{\Theta} - I = -\lambda\hat{Z}\hat{\Theta}$. Now recall $\text{rem}_{\text{reg}} = (\hat{\Theta} - \Theta_0)\eta(\hat{\Theta}) = -(\hat{\Theta} - \Theta_0)(\hat{\Sigma}\hat{\Theta} - I)$. Combining both with the decomposition in (2.5), Hölder's inequality, and recall the Oracle bounds:

Theorem 1 (Oracle Bounds). *Assume Conditions 3.1 and 3.2. Then, for* $\lambda \asymp \sqrt{n/\log p}$,

$$\|\hat{\Sigma} - \Sigma_0\|_\infty = O_{\mathbb{P}}(\lambda) = O_{\mathbb{P}}\left(\sqrt{n/\log p}\right), \quad |||\hat{\Theta} - \Theta_0|||_1 = O_{\mathbb{P}}\left((p+s)\lambda\right).$$

$$\begin{aligned}
&\| \text{rem} \|_\infty \leq \| \text{rem}_0 \|_\infty + \| \text{rem}_{\text{reg}} \|_\infty \\
&= \|\Theta_0(\hat{\Sigma} - \Sigma_0)(\hat{\Theta} - \Theta_0)\|_\infty + \|(\hat{\Theta} - \Theta_0)(\hat{\Sigma}\hat{\Theta} - I)\|_\infty \\
&= \|\Theta_0(\hat{\Sigma} - \Sigma_0)(\hat{\Theta} - \Theta_0)\|_\infty + \|(\hat{\Theta} - \Theta_0)\lambda\hat{Z}\hat{\Theta}\|_\infty \\
&\leq |||\hat{\Theta} - \Theta_0|||_1 \|\hat{\Sigma} - \Sigma_0\|_\infty |||\Theta_0|||_1 + \lambda\, |||\hat{\Theta} - \Theta_0|||_1\, \|\hat{Z}\|_\infty\, |||\hat{\Theta}|||_1 \\
&\leq O((p+s)\lambda^2)\, |||\Theta_0|||_1 + 2\lambda O((p+s)\lambda)\, |||\Theta_0|||_1.
\end{aligned} \tag{4.3}$$

The next step will be to prove

$$|||\Theta_0|||_1 \leq \sqrt{d+1}\Lambda_{\max}(\Theta_0),$$

meaning to relate the matrix norm to the sparsity and eigenvalues of the matrix.

$$\begin{aligned}
|||\Theta_0|||_1 &= \max_{1\leq j\leq p} \sum_{i=1}^{p} |\Theta^0_{i,j}| = \max_{1\leq j\leq p} \|\Theta^0_j\|_1 \\
&\leq \max_{1\leq j\leq p} \sqrt{d+1}\|\Theta^0_j|_2 \text{Hier Bild zu malen und ausführen} \\
&\leq \max_{1\leq j\leq p} \sqrt{d+1}\|\Theta^0\|_2 \\
&= \sqrt{d+1}\Lambda_{\max}(\Theta_0).
\end{aligned}$$

This in conjunction with (4.3) and Condition 3.1 gives

$$\|\text{rem}\|_\infty = O(\sqrt{d}(p+s)\lambda^2) = o(\sqrt{n}/\log p \cdot \log p/n) = o(n^{-1/2}),$$

this proves (4.1).

Normality in (4.2): By Slutsky, (4.1) and (2.5) we only need to show

$$\sqrt{n}\frac{S_{n,i,j}}{\text{Var}(S_{n,i,j})} \to \mathcal{N}(0,1),$$

where $S_n := \Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 = n^{-1}\Theta_0 X^T X \Theta_0 - \Theta_0$.
First note that the i-th row of $\Theta_0 X^T$ is given by $((\Theta^0_i)^T X^1, \ldots, (\Theta^0_i)^T X^n)$, where $X^k = (X^k_1, \ldots, X^k_p)$ denotes the k-th observation. (Ausführen auf Blatt). By symmetry the j-th column of $X\Theta_0$ is given by the j-th row of $\Theta_0 X^T$ transposed. Hence,

$$\begin{aligned}
S_{n,i,j} &= \frac{1}{n}\left((\Theta^0_i)^T X^1, \ldots, (\Theta^0_i)^T X^n\right) \cdot \left((\Theta^0_j)^T X^1, \ldots, (\Theta^0_j)^T X^n\right)^T - \Theta^0_{i,j} \\
&= \frac{1}{n}\sum_{k=1}^{n} (\Theta^0_i)^T X^k (\Theta^0_j)^T X^k - \Theta^0_{i,j} \\
&=: \frac{1}{n}\sum_{k=1}^{n} Z_{i,j,k}.
\end{aligned}$$

Now we are in the Lindeberg-Feller CLT setting! In order to verify the Lindeberg condition, we need an appropiate tail bound, which we will derive from Sub-Gaussianity and universal Eigenvalue bounds:

First, note

$$\begin{aligned}|(\Theta_i^0)^T X^k| &\leq \sum_{j=1}^{p} |\Theta_{i,j}^0 X_j^k| \\ &\leq \|\Theta_i^0\|_1 \max_{j\colon \Theta_{i,j}^0 \neq 0} |X_j^k| \\ &\leq \sqrt{d}\|\Theta_i^0\|_2 \max_{j\colon \Theta_{i,j}^0 \neq 0} |X_j^k|.\end{aligned}$$

This yields in conjunction with Condition 3.2 (Sub-Gaussianity)

$$\begin{aligned}\mathbb{P}(|(\Theta_i^0)^T X^k| > t) &\leq \mathbb{P}\Big(\max_{j\colon \Theta_{i,j}^0 \neq 0} |X_j^k| > \frac{t}{\sqrt{d}\|\Theta_i^0\|_2}\Big) \\ &\leq \sum_{j\colon \Theta_{i,j}^0 \neq 0} \mathbb{P}\Big(|X_j^k| > \frac{t}{\sqrt{d}\|\Theta_i^0\|_2}\Big) \\ &\leq d \max_{j\colon \Theta_{i,j}^0 \neq 0} \mathbb{P}\Big(|X_j^k| > \frac{t}{\sqrt{d}\|\Theta_i^0\|_2}\Big) \\ &= d \cdot O\bigg(\exp\Big(-\frac{t^2}{d\|\Theta_i^0\|_2^2}\Big)\bigg).\end{aligned}$$

This in turn gives

$$\mathbb{P}(|(\Theta_i^0)^T X^k (\Theta_j^0)^T X^k| \geq t) = d \cdot O\bigg(\exp\Big(-\frac{t}{\sqrt{d}\|\Theta_i^0\|_2}\Big)\bigg),$$

which after using $\|\Theta_i^0\|_2 \leq \|\Theta_0\|_2 = \Lambda_{\max}(\Theta^0) = O(1)$ by Condition 3.1 yields for $Z_{i,j,k}$ [recall $Z_{i,j,k} = (\Theta_i^0)^T X^k (\Theta_j^0)^T X^k - \Theta_{i,j}^0$]:

$$\mathbb{P}(|Z_{i,j,k}| > t) \leq c_1 d \exp\Big(-\frac{t - |\Theta_{i,j}^0|}{c_2 d}\Big) \leq c_1 d \exp\Big(-\frac{t}{c_3 d}\Big), \tag{4.4}$$

for c_1, c_2, c_3 depending on the bound of the maximal eigenvalue and not on n (since $|\Theta_{i,j}^0|$ is bounded and $d \geq 1$.)

We need to show

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{s_n^2} \mathbb{E}[Z_{i,j,k}^2 \mathbf{1}\{|Z_{i,j,k}| > \varepsilon s_n\}] := \lim_{n\to\infty} L_n = 0,$$

where $s_n^2 := n\sigma_{i,j}^2$. For fixed n, i, j the $Z_{i,j,k}$ are i.i.d, thus

$$L_n = \frac{1}{\sigma_{i,j}^2} \mathbb{E}[Z_{i,j,k}^2 \mathbf{1}\{|Z_{i,j,k}| > \varepsilon s_n\}]$$

and it is enough to show

$$\mathbb{E}[Z_{i,j,k}^2 \mathbf{1}\{|Z_{i,j,k}| > \varepsilon s_n\}] = o(1). \tag{4.5}$$

For $a > 0$ we may rewrite for a real random variable Y using Fubini

$$\begin{aligned} &\int (Y^2 - a^2)\mathbf{1}\{|Y| > a\} \, \mathrm{d}\mathbb{P} = \int \int 2u\mathbf{1}\{a < u < |Y|\} \, \mathbf{1}\{|Y| > a\} \, \mathrm{d}u \, \mathrm{d}\mathbb{P} \\ &= 2 \int_a^\infty \mathbb{P}(|Y| > u) \, \mathrm{d}u, \end{aligned} \tag{4.6}$$

now let $a := \varepsilon\sqrt{n}\sigma_{i,j}, Y := Z_{i,j,k}$ and rearrange (4.6) to obtain

$$\begin{aligned} &\mathbb{E}[Z_{i,j,k}^2 \mathbf{1}\{|Z_{i,j,k}| > \varepsilon s_n\}] \\ &= \varepsilon^2 n \sigma_{i,j}^2 \mathbb{P}(|Z_{i,j,k}| > \varepsilon\sqrt{n}\sigma_{i,j}) + 2 \int_{\varepsilon\sqrt{n}\sigma_{i,j}}^\infty \mathbb{P}(|Z_{i,j,k}| > u) \, \mathrm{d}u \\ &=: R_1 + R_2. \end{aligned}$$

Now for R_1 the tail bound (4.4) yields (leaving out the constants for the sake of readability)

$$R_1 \leq nd \cdot \exp\Big(- \frac{\sqrt{n}}{d}\Big),$$

now by the sparsity condition we have $\sqrt{d}(p+s) = o(\sqrt{n}/\log p)$, which gives $d^{3/2} = o(\sqrt{n}/\log p)$, as $s \geq d$ and thus $d = o(n^{1/3}/\log^{2/3} p)$. This results in

$$nd \exp\Big(- \frac{\sqrt{n}}{d}\Big) \lesssim n^{4/3} \exp\Big(- n^{1/2-1/3}\Big) = o(1),$$

hence $R_1 = o(1)$. For R_2 we obtain by using the tail bound (4.4) and substituting $u = \sqrt{n}\varepsilon\sigma_{i,j}t$ (thus $\mathrm{d}u = \sqrt{n}\varepsilon\sigma_{i,j} \, \mathrm{d}t$)

$$\begin{aligned} \frac{R_2}{2} &\leq \int_{\varepsilon\sqrt{n}\sigma_{i,j}}^\infty ud \exp\Big(- \frac{u}{d}\Big) \, \mathrm{d}u \\ &= \int_1^\infty dn\varepsilon^2\sigma_{i,j}^2 \exp\Big(- \frac{\sqrt{n}\varepsilon\sigma_{i,j}t}{d}\Big) \, \mathrm{d}t. \end{aligned}$$

Now by (recall from before) $d \leq n^{1/3}$ for $n \geq n_0$

$$\begin{aligned} &dn\varepsilon^2\sigma_{i,j}^2 \exp\Big(- \frac{\sqrt{n}\varepsilon\sigma_{i,j}t}{d}\Big) \\ &= n^{4/3}\varepsilon^2\sigma_{i,j}^2 \exp\Big(- n^{1/6}\varepsilon\sigma_{i,j}t\Big) \\ &= C(\varepsilon, L) n^{-(9/6-4/3)} t^{-9} \in \mathrm{L}^1([1,\infty)) \end{aligned}$$

uniformly in n, since the variances are bounded from below and above by Conditions 3.1 and 3.2.
Now by the upper two displays and Dominated Convergence we have $R_2 = o(1)$ and thus by (4.5) asymptotic normality holds.

□

5 Nodewise LASSO

We will briefly introduce the **Nodewise LASSO** estimator for the precision matrix and state an asymptotic normality result.

Idea: Estimate each column of the precision matrix by projecting every column of the design matrix on the rest. Meaning that these are p iterated independent minimizations.

The motivation for the estimators is as follows: For each $j = 1, \dots, p$ define the vector $\gamma_j^0 = \{\gamma_{j,k}^0, k \neq j\}$ as follows

$$\gamma_j^0 := \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} \mathbb{E} \|X_j - X_{-j} \cdot \gamma\|_2^2 / n,$$

where X_{-j} denotes the design matrix X without its j-th column. The error is called **noise level** and defined as $\tau_j^2 = \mathbb{E} \|X_j - X_{-j} \cdot \gamma_j^0\|_2^2 / n$. Now the following identity is central:

$$\Theta_j^0 = -(\gamma_{j,1}^0, \dots, \gamma_{j,j-1}^0, -1, \gamma_{j,j+1}^0, \dots, \gamma_{j,p}^0)^T / \tau_j^2. \tag{5.1}$$

This allows for estimating the precision matrix by estimating γ^0 and the corresponding noise levels. The authors use for estimating γ_j^0 the **square root LASSO** with weighted penalty:

$$\hat{\gamma}_j := \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} \|X_j - X_{-j}\gamma\|_2 / n + 2\lambda \|\hat{W}_{-j}\gamma\|_1,$$

where recall $\hat{W}^2 := \operatorname{diag} \hat{\Sigma}$. Now estimators for the noise level were defined as

$$\hat{\tau}_j^2 := \|X_j - X_{-j}\hat{\gamma}_j\|_2^2 / n, \quad \tilde{\tau}_j^2 := \hat{\tau}_j^2 + \lambda \hat{\tau}_j \|\hat{\gamma}_j\|_1.$$

Now plugging those into the identity (5.1) yields the nodewise estimator:

$$\hat{\Theta}_N := \begin{pmatrix} 1/\tilde{\tau}_1^2 & -\hat{\gamma}_{1,2}/\tilde{\tau}_1^2 & \dots & -\hat{\gamma}_{1,p}/\tilde{\tau}_1^2 \\ -\hat{\gamma}_{2,1}/\tilde{\tau}_2^2 & 1/\tilde{\tau}_2^2 & \dots & -\hat{\gamma}_{2,p}/\tilde{\tau}_2^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1}/\tilde{\tau}_p^2 & \dots & -\hat{\gamma}_{p,p-1}/\tilde{\tau}_p^2 & 1/\tilde{\tau}_p^2. \end{pmatrix}$$

Now the same as before, the de-sparsified estimator nodewise LASSO can be defined as

$$\hat{\Theta}_{D,N} := \hat{\Theta}_N + \hat{\Theta}_N^T - \hat{\Theta}_N^T \hat{\Sigma} \hat{\Theta}_N.$$

Theorem 5.1. *Assume Conditions 3.1, 3.2 and the sparsity Condition $d = o(\sqrt{n}/\log p)$. Then, for $\lambda \asymp \sqrt{\log p / n}$,*

$$\hat{\Theta}_{D,N} - \Theta_0 = -\Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 + \text{rem}, \quad \|\text{rem}\|_\infty = o_{\mathbb{P}}(n^{-1/2}).$$

Furthermore, for $i, j = 1, \dots, p$,

$$\sqrt{n} \frac{(\hat{\Theta}_D - \Theta_0)_{i,j}}{\sigma_{i,j}} \rightsquigarrow \mathcal{N}(0, 1).$$

The proof refers to another paper [Janková and Van de Geer, 2017], where again oracle bounds are derived and Berry-Esseen gets used.

Simulation Results and Discussion

- There were multiple papers comparing normal MLE to graphical LASSO, nodewise LASSO, nodewise square root LASSO and their de-sparsified counterparts. First, nodewise performed better than the whole graphical. Square root was comparable to normal LASSO. Both in the de-sparsified and not de-sparsified setting this trend was observable.
- Numerical results suggest to use nodewise LASSO estimation. Furthermore the iterated nodewise LASSO estimation is computational less expensive than the normal GLASSO.
- However the invertibility of $\hat{\Theta}_{D,N}$ has not been explored yet.
- The sparsity condition $d = o(\sqrt{n}/\log p)$ was shown to be the minimal sparsity condition to estimate Θ_0 at parametric rate $\sqrt{n}$ (in the Gaussian-model; not Sub-Gaussian) in an Annals paper by Ren et. al. [Ren et al., 2015].

Takeaway

- Estimation of the precision matrix with parametric rate is possible, if there are sparsity conditions
- Theoretically, the conditions are weaker if one uses nodewise LASSOs
- These sparsity conditions are essentially necessary to obtain parametric rates
- Central technique is the derivation of oracle bounds
- Shrinkage estimator might not be asymptotically normal but de-sparsifying may result in asymptotically normal estimator and thus confidence intervalls are at hand

References

[Janková and Van de Geer, 2017] Janková, J. and Van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162.

[Janková and van de Geer, 2019] Janková, J. and van de Geer, S. (2019). Inference in high-dimensional graphical models. In *Handbook of graphical models*, pages 325–349. Boca Raton, FL: CRC Press.

[Knight and Fu, 2000] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378.

[Pötscher and Leeb, 2009] Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082.

[Ren et al., 2015] Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Stat.*, 43(3):991–1026.

[Rothman, 2008] Rothman, Adam J., e. a. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515.

[van de Geer et al., 2014] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Camb. Ser. Stat. Probab. Math.* Cambridge: Cambridge Univ. Press.