Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Consistency results and confidence intervals for adaptive ℓ_1-penalized estimators of the high-dimensional sparse precision matrix

Valeriy Avanesov , Jörg Polzehl , Karsten Tabelow

submitted: February 11, 2016

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: valeriy.avanesov@wias-berlin.de
joerg.polzehl@wias-berlin.de
karsten.tabelow@wias-berlin.de

No. 2229
Berlin 2016

2010 *Mathematics Subject Classification.* 62J07, 62P10.

Key words and phrases. adaptive ℓ_1 penalty, precision matrix, high-dimensional statistics, sparsity, confidence intervals, functional connectivity.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

In this paper we consider the adaptive ℓ_1-penalized estimators for the precision matrix in a finite-sample setting. We show consistency results and construct confidence intervals for the elements of the true precision matrix. Additionally, we analyze the bias of these confidence intervals. We apply the estimator to the estimation of functional connectivity networks in functional Magnetic Resonance data and elaborate the theoretical results in extensive simulation experiments.

1. Introduction

Throughout the (long) history of neuroscience the question whether brain function is localized or distributed over the human brain was subject to intense discussion (Finger 1994). The two concepts are nowadays summarized as functional segregation on the one hand and functional integration on the other (Friston 1994). With the advance of imaging techniques like functional Magnetic Resonance Imaging (fMRI) most studies focused on the localization of function (Friston 2011; Poldrack, Mumford, and Nichols 2011). However, soon a growing number of studies revealed a plethora of new findings with respect to functional integration of the working brain (Sporns 2011; Sporns 2013). These are typically described in terms of functional or effective connectivity. While the former relates to statistical correlation between neurophysiological events, the latter refers to explicit influence among neural systems (Friston 2011). The statistical problems considered in this paper belong to the category of functional connectivity (FC). Yet, the interpretation of FC networks from fMRI data poses problems as the data is only a relative indirect measure of neural activity (Buxton, Wong, and Frank 1998; Huettel, Song, and McCarthy 2014).

Several methods are used to describe FC networks, see, e.g., Smith et al. 2011 or Poldrack, Mumford, and Nichols 2011 for reviews: Among them are matrix factorization methods like Principal Component Analysis (Friston et al. 1993) or Independent Component Analysis (ICA) (Kiviniemi et al. 2003; Mantini et al. 2007), but also techniques like beta-series correlation (Rissman, Gazzaley, and D'Esposito 2004), psychophysiological interaction (Friston et al. 1997) and others. However, very popular and simple is the approach to reflect the functional connectivity by the correlation of the fMRI time series assigned to some suitable nodes defined based on the results of activation-based fMRI, ICA, or brain atlas. We denote the number of nodes by p. In order to eliminate the correlations that are only mediated via some common nodes partial correlations can be considered. Typically, before the estimation of the FC network correlations due to global artifacts (motion, field drift) or due to activation should be removed (Poldrack, Mumford, and Nichols 2011). In this paper we will solely concentrate on the estimation of correlations by means of the inverse covariance or precision matrix (Allen et al. 2012). The estimation generally makes use of the observation that FC networks are of small-world type and sparse (Sporns 2011), which can be incorporated by regularization.

Currently most studies focus on static FC network estimation (Sporns 2013), but interest in network dynamics is growing fast. This is particularly important for research on neural diseases but also in the context of learning research with first attempts to characterize the re-configuration of the brain (Bassett et al. 2011). This requires the estimation of confidence sets for FC networks.

Specifically, in this paper we consider an i.i.d. sample $X_1, \ldots X_n \in \mathbb{R}^p$ with zero mean and n is the length of the fMRI time series. Let X be the $n \times p$ matrix of samples. The FC network is then characterized by the covariance matrix Σ or the precision matrix $\Theta = \Sigma^{-1}$. An estimate is obtained by minimizing over the cone S^p_{++} of positive-definite $p \times p$ matrices:

$$\arg \min_{\Theta \in S^p_{++}} \left[\mathrm{tr}(\Theta \hat{\Sigma}) - \log \det \Theta + p_\lambda(\Theta) \right]$$

with some suitable penalization $p_\lambda(\Theta)$ and the empirical covariance $\hat{\Sigma} = \frac{1}{n} X^T X$.

In order to address the problem ℓ_1-penalization approaches which were initially suggested by Tibshirani 1994 may be used imposing the required sparsity on the estimate:

$$\hat{\Theta} = \arg \min_{\Theta \in S^p_{++}} \left[\mathrm{tr}(\Theta \hat{\Sigma}) - \log \det \Theta + \|\Lambda * \Theta\|_1 \right] \tag{1.1}$$

where Λ is a $p \times p$ matrix of non-negative off-diagonal elements and zero diagonal ones and $\cdot * \cdot$ denotes matrix element-wise product.

There are consistency results of such estimators for samples of finite size (Ravikumar et al. 2011) along with asymptotic confidence intervals for the elements of the true precision matrix for the case of equal amount of penalization applied to each element of precision matrix (Janková and Geer 2015). On the other hand, there are some experimental evidences that adaptive penalization approaches may perform better (Fan, Feng, and Wu 2009). In this paper, we provide the consistency results for adaptive ℓ_1-penalized estimators of precision matrix and we also construct the confidence intervals based on these estimators for the elements of the true precision matrix. We show that the bias introduced by the penalization and the non-normality of the constructed confidence intervals depends only on the largest amount of penalization applied to non-zero elements of the true precision matrix. All the results are obtained in a finite-sample-size setting.

The paper is organized as follows. Section 2 introduces the notation used throughout the paper. Section 3 lists main theoretical results of the paper. Namely, sub-Section 3.1 contains non-trivial assumptions, sub-Sections 3.2 and 3.3 give the definition and consistency results for adaptive approaches such as classical adaptive graphical lasso (Zou 2006) (Fan, Feng, and Wu 2009) and SCAD lasso (Zou and Li 2008) (Fan, Feng, and Wu 2009) (Fan and Li 2001) respectively and the sub-Section 3.4 comes up with the definition of a de-sparsified estimator and provides the results estimating its distribution which gives rise to confidence intervals construction along with hypotheses testing. In Section 4 we provide the proofs of the claimed results. Finally, Section 5 describes our experimental study.

2. Notation

We denote the empirical covariance matrix as $\hat{\Sigma} = \frac{1}{n} X^T X$, the true covariance matrix as Σ^* and their difference as $W = \hat{\Sigma} - \Sigma^*$. Throughout the paper we assume that the true precision matrix ${\Sigma^*}^{-1}$ exists and we denote it as Θ^*.

Also define the set of non-zero entries of Θ^* as $S = \{(i,j) : \Theta^*_{ij} \neq 0\}$ and its complement as $S^c = \{1..p\}^2 \setminus S$.

We also use the following notations for matrix norms: $\|A\|_1 = \sum_{i,j} |A_{ij}|$, $\|A\|_\infty = \max_{i,j} |A_{ij}|$ and $|||A|||_\infty = |||A^T|||_1 = \max_j \|A_{\cdot j}\|_1$.

For a matrix A its vectorization is denoted as $\bar{A}$ or, equivalently as $\operatorname{vec} A$.

Let $\Gamma^* = \Sigma^* \otimes \Sigma^*$ where $\cdot \otimes \cdot$ stands for Kronecker product, $\kappa_{\Gamma^*} = |||(\Gamma^*_{SS})^{-1}|||_\infty$, $\kappa_{\Sigma^*} = |||\Sigma^*|||_\infty$, $\kappa_{\Theta^*} = |||\Theta^*|||_1$.

Our main results assume lower bounds on the smallest absolute values of non-zero elements of the true precision matrix which is denoted as $\theta_{min} = \min_{i,j:\Theta_{ij}\neq 0} |\Theta^*_{ij}|$.

Other values we keep track on are the maximum number of non-zero elements in a row of the true precision matrix $d = \max_i |\{j \colon \Theta^*_{ij} \neq 0\}|$ and the minimal penalization parameter corresponding to zero elements of the true precision matrix $\rho = \min\limits_{(i,j)\in S^c} \Lambda_{ij}$.

3. Main results

3.1. Irrepresentability assumption

Assumption 1.

$$\exists \alpha \in (0,1] \quad s.t. \quad \max_{e \in S^c} \left\| \Gamma^*_{eS} (\Gamma^*_{SS})^{-1} \right\|_1 \leq (1-\alpha)$$

The irrepresentability assumption is usually interpreted as follows (Janková and Geer 2015) (Ravikumar et al. 2011).

Define a centered random variable for each edge $(i,j) \in \{1..p\}^2$

$$Y_{(i,j)} = X_{1i} X_{ij} - \mathsf{E}[X_{1i} X_{1j}]$$

then covariances of these variables may be expressed in terms of matrix Γ^* as

$$\mathbf{cov}(Y_{(i,j)}, Y_{(k,l)}) = \Gamma^*_{(i,j),(k,l)} + \Gamma^*_{(j,i),(k,l)}.$$

So the Assumption 1 requires low correlation between edges from active set S and its complement S_c. The higher the constant α is, the stricter upper bound is assumed.

3.2. Adaptive graphical lasso

3.2.1. Definition

Let $\hat{\Theta}^{init}$ be a solution of optimization problem (1.1) with penalization parameters $\Lambda_{ij} = \lambda_{init}$ for $i \neq j$.

Then, the adaptive graphical lasso estimator $\hat{\Theta}^{ada}$ is defined as the solution of the optimization problem (1.1) with tuning parameters $\Lambda_{ij}^{ada} = \lambda_{init} \frac{1}{|\hat{\Theta}_{ij}^{init}|^{\gamma}}$ for $i \neq j$ where $\gamma \in (0, 1]$ ($\gamma = 0.5$ is usually used). If $\hat{\Theta}_{ij}^{init} = 0$, we define $\Lambda_{ij} = +\infty$, thereby excluding the corresponding variable from optimization and forcing it to equal zero.

3.2.2. Consistency result

Theorem 1. *Assume the conditions of the Lemma 8 hold. Furthermore, suppose*

$$d \leq \frac{\delta_n}{6\left(\delta_n + \frac{\lambda_n}{(\theta_{min} - r)^{\gamma}}\right)^2 \max\{\kappa_{\Gamma^*}\kappa_{\Sigma^*}, \kappa_{\Gamma^*}^2\kappa_{\Sigma^*}^3\}} \tag{3.1}$$

Then on the set $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_{\infty} < \delta_n\right\}$ *the following holds:*
$\left\|\hat{\Theta}^{ada} - \Theta^*\right\|_{\infty} \leq 2\kappa_{\Gamma^*}\left(\delta_n + \frac{\lambda_n}{(\theta_{min} - r)^{\gamma}}\right)$ *and* $\Theta_{ij}^* = 0 \Leftrightarrow \hat{\Theta}_{ij}^{ada} = 0$.

Remark 1. *The main results in the paper are conditioned on the set* $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_{\infty} < \delta_n\right\}$. *The lower bound for the probability of the set* $\mathcal{T}$ *under sub-Gaussianity assumption is provided by Lemma 11.*

3.3. SCAD graphical lasso

3.3.1. Definition

SCAD was suggested in Fan and Li 2001 and was applied for sparse precision matrix estimation in Lam and Fan 2009 as an alternative adaptive penalization approach.

Consider the following optimization problem:

$$\hat{\Theta} = \arg\min_{\Theta \in S_{++}^p} \left[\operatorname{tr}(\Theta\hat{\Sigma}) - \log\det\Theta + \sum_{i \neq j} SCAD_{\lambda,a}(|\theta_{ij}|)\right]$$

for some positive λ and a (usually $a = 3.7$ is used) with the first derivative of $SCAD_{\lambda,a}(\cdot)$ defined as

$$SCAD'_{\lambda,a}(x) = \lambda \left\{ I(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x \geq \lambda) \right\}$$

where $(\cdot)_+$ denotes a positive cut: $(x)_+ = \max\{0, x\}$.

In order to solve this non-convex optimization problem, the following approximate recurrent algorithm was suggested in Fan, Feng, and Wu 2009

$$\hat{\Theta}^{(k)} = \arg \max_{\Theta \in S^p_{++}} \operatorname{tr}(\Theta\hat{\Sigma}) - \log\det\Theta + \sum_{i,j} SCAD'_{\lambda,a}(|\theta_{ij}^{(k-1)}|)|\theta_{ij}| \tag{3.2}$$

where $\hat{\Theta}^{(0)}$ is obtained as a solution of (1.1) with $\Lambda_{ij} = \lambda \; \forall i \neq j$.

As one can see, $SCAD'_{\lambda,a}(x) = 0$ for x large enough, so the problem (3.2) may have no optimum in case if $\hat{\Sigma}$ is singular. In this section we therefore assume that $\hat{\Sigma}$ is non-singular. However, this assumption may be dropped if we replace $SCAD'_{\lambda,a}(\cdot)$ with $I(SCAD'_{\lambda,a}(\cdot) = 0)\epsilon + I(SCAD'_{\lambda,a}(\cdot) > 0)SCAD'_{\lambda,a}(\cdot)$ for some $\epsilon > 0$.

Also, denote the limiting point of the algorithm as $\hat{\Theta}^{SCAD} = \lim_{k\to\infty} \hat{\Theta}^{(k)}$.

We denote the penalization matrix used at k-th iteration as $\Lambda_{ij}^{(k)} = SCAD'_{\lambda,a}(|\theta_{ij}^{(k-1)}|)$ and its minimal value corresponding to zero elements of true precision matrix as $\rho^{(k)} = \min_{(i,j)\in S^c} \Lambda_{ij}^{(k)}$.

On the other hand the paper Zou and Li 2008 provides asymptotic properties of one-step estimate $\hat{\Theta}^{OSSCAD} = \hat{\Theta}^{(1)}$.

3.3.2. SCAD graphical lasso consistency results

Theorem 2. *Assume the conditions of the Lemma 8. Also suppose that the matrix $\hat{\Sigma}$ is non-singular.*

Then on the set $\mathcal{T} = \left\{ \left\| \hat{\Sigma} - \Sigma^* \right\|_\infty < \delta_n \right\}$ the following holds:

$$\left\| \hat{\Theta}^{OSSCAD} - \Theta^* \right\|_\infty \leq 2\kappa_{\Gamma^*} \left(\delta_n + SCAD'_{\lambda,a}(\theta_{min} - r) \right)$$

and $\Theta^*_{ij} = 0 \Leftrightarrow \hat{\Theta}^{OSSCAD}ij = 0$.

Theorem 3. *Assume the conditions of Theorem 2.*

Then on the set $\mathcal{T} = \left\{ \left\| \hat{\Sigma} - \Sigma^* \right\|_\infty < \delta_n \right\}$ the following holds:

$$\left\| \hat{\Theta}^{SCAD} - \Theta^* \right\|_\infty \leq 2\kappa_{\Gamma^*} \left(\delta_n + \left(\frac{a\lambda_n - \theta_{min} + 2\kappa_{\Gamma^*}\delta_n}{2\kappa_{\Gamma^*} + a - 1} \right)_+ \right)$$

and $\Theta^*_{ij} = 0 \Rightarrow \hat{\Theta}^{SCAD}_{ij} = 0$.

3.4. Inference result

In this section we aim to construct confidence intervals for true values of precision matrix Θ^*_{ij}. In order to do so we mostly follow the approach suggested in Geer et al. 2014 and applied to the problem of estimation of high-dimensional precision matrix in Janková and Geer 2015.

Consider the stationarity condition corresponding to the problem (1.1):

$$-\hat{\Theta}^{-1} + \hat{\Sigma} + \Lambda * Z = 0$$

where $Z \in \partial \left\|\Theta\right\|_1$.

Multiply from both sides by $\hat{\Theta}$:

$$\hat{\Theta}\hat{\Sigma}\hat{\Theta} - \hat{\Theta} + \hat{\Theta}(\Lambda * Z)\hat{\Theta} = 0$$

By rearranging obtain

$$\hat{\Theta} + \hat{\Theta}(\Lambda * Z)\hat{\Theta} = \Theta^* - \Theta^* W \Theta^* + r \tag{3.3}$$

where

$$r = -(\hat{\Theta} - \Theta^*)W\Theta^* - (\hat{\Theta}\hat{\Sigma} - I_p)(\hat{\Theta} - \Theta^*)$$

Finally, we define a de-sparsified estimator as

$$\begin{aligned} \hat{T} &:= 2\hat{\Theta} - \hat{\Theta}\hat{\Sigma}\hat{\Theta} \\ &= \hat{\Theta} + \hat{\Theta}(\Lambda * Z)\hat{\Theta} \\ &= \Theta^* - \Theta^* W \Theta^* + r \end{aligned} \tag{3.4}$$

Theorem 4. *Suppose, assumptions of Lemma 5 hold. Moreover, suppose,* $\mathrm{p}_{\mathcal{T}} := \mathbb{P}\left\{\mathcal{T}\right\} > 0$

Then, for all (i, j) *the following upper and lower bounds hold:*

$$\begin{aligned} \Phi\left(\frac{\sigma_{ij}}{\sqrt{n}}c - \frac{R}{\sigma_{ij}\sqrt{n}}\right) - \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}} - 2(1 - \mathrm{p}_{\mathcal{T}}) \leq \mathbb{P}\left\{\hat{T}_{ij} - \Theta^*_{ij} < c \mid \mathcal{T}\right\} \\ \leq \Phi\left(\frac{\sigma_{ij}}{\sqrt{n}}c + \frac{R}{\sigma_{ij}\sqrt{n}}\right) + \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}} + 2(1 - \mathrm{p}_{\mathcal{T}}) \end{aligned}$$

where $A < 0.4748$, $\sigma_{ij}^2 = Var[\Theta^*_i X_k \Theta^*_j X_k - \Theta^*_{ij}]$ *and* μ_{ij3} *is the third moment of* $|Z_{ijk}|$ *(see* (4.17)*) and* R *is defined by* (4.15)*.*

4. Proofs

In order to prove the claimed consistency results we employ the primal-dual witness technique which suggest to consider the following optimization problem:

$$\tilde{\Theta} = \arg \min_{\substack{\Theta \in S^p_{++} \\ \Theta_{S^c}=0}} \Big[\mathrm{tr}(\Theta \hat{\Sigma}) - \log\det\Theta + \|\Lambda * \Theta\|_1 \Big] \tag{4.1}$$

The only difference between the problems (1.1) and (4.1) is that the latter one forces all zero elements to be estimated as zero, e.g. $\tilde{\Theta}_{S^c} = 0$. The main idea of the technique is to show that $\tilde{\Theta} = \hat{\Theta}$ on some set of high probability.

We use $\Delta = \tilde{\Theta} - \Theta^*$ to denote the mis-tie between the true precision matrix and the solution of the problem (4.1).

In our derivations we also make use of properties of the residuals of the first-order Taylor expansion of the gradient of the log-det functional which take form:

$$R(\Delta) = \tilde{\Theta}^{-1} - {\Theta^*}^{-1} + {\Theta^*}^{-1}\Delta{\Theta^*}^{-1}$$

4.1. Existence and uniqueness of solutions of problems (1.1) and (4.1)

Since we are about to investigate the properties of solutions of the problems (1.1) and (4.1), we first need to give sufficient conditions for their existence and uniqueness.

The lemma below is a slightly generalized version of Lemma 3 given in Ravikumar et al. 2011 and can be proven by exactly the same argument.

Lemma 1. *Let $\forall i \neq j\ \Lambda_{ij} > 0$, $\Lambda_{ii} = 0$ and $\Sigma_{ii} > 0\ \forall i$, then the problems (1.1) and (4.1) have unique solutions.*

We also give sufficient conditions which do not include positiveness of all non-diagonal elements of Λ but in turn rely on non-singularity of the sample covariance matrix $\hat{\Sigma}$.

Lemma 2. *Suppose, $\hat{\Sigma}$ is non-singular. Then the problems (1.1) and (4.1) have unique solutions.*

Proof. We give the proof for the problem (1.1). The uniqueness of the solution for the problem (4.1), as well as for a problem with any set of non-diagonal values of Θ restricted to zero (in case it does not violate symmetry) can be established by the same argument.

By Lagrange duality we can rewrite the problem (1.1) in form

$$\hat{\Theta} = \min_{\substack{\Theta \in S^p_{++} \\ \|C(\Lambda)*\Theta\|_1 \leq 1}} \Big[\mathrm{tr}(\Theta \hat{\Sigma}) - \log\det\Theta \Big]$$

for some $C_{ij}(\Lambda) < +\infty$ for $\Lambda_{ij} > 0$ and $C_{ij}(\Lambda) = 0$ for $\Lambda_{ij} = 0$.

Now, since $\hat{\Sigma}$ is non-singular and it is a covariance matrix, it is positive-definite. Thus, there exists and orthogonal transform S such that $S^T\hat{\Sigma}S = D = diag(d_1...d_p)$ and $\forall i\ d_i > 0$.

Then, by using the fact that $\operatorname{tr}\Theta\hat{\Sigma} = \operatorname{tr} S^T\Theta\hat{\Sigma}S$ and by noting that $\det S = 1$, we further rewrite the problem as

$$\hat{\Theta} = \min_{\substack{\Theta' \in S^p_{++} \\ \left\|C(\Lambda)*(S\Theta'S^T)\right\|_1 \leq 1}} [\operatorname{tr}(\Theta' D) - \log\det\Theta']$$

where $\Theta' = S^T\Theta S$. Here we have also used the fact that $\Theta' \in S_{++}$ iff. $\Theta \in S_{++}$.

Now we just substitute the definition of the trace:

$$\hat{\Theta} = \min_{\substack{\Theta' \in S^p_{++} \\ \left\|C(\Lambda)*(S\Theta'S^T)\right\|_1 \leq 1}} \left[\sum_i d_i\Theta'_{ii} - \log\det\Theta'\right]$$

But, due to the fact that $d_i > 0$ by Lagrange duality we finally obtain

$$\hat{\Theta} = \min_{\substack{\Theta' \in S^p_{++} \\ \left\|C(\Lambda)*(S\Theta'S^T)\right\|_1 \leq 1 \\ \forall i\ |\Theta'_{ii}| \leq C_i(d_i)}} - \log\det\Theta' \tag{4.2}$$

for some $C_i(d_i) < +\infty$.

So, the diagonal elements of Θ' are bounded. Therefore, its trace is bounded, thus the sum of its eigenvalues is bounded, so the feasible set is compact. Thus (recalling the convexity of the log-det functional) the optimum exists and is unique.

Using the fact of equivalence of the problems (4.2) and (1.1) we obtain the claimed statement.

□

4.2. Proof of adaptive lasso consistency result

Lemma 3 (generalization of Lemma 6, Ravikumar et al. 2011)**.** *Suppose that*

$$r := 2\kappa_{\Gamma^*}(\|W\|_\infty + \|\Lambda_S\|_\infty) \leq \min\left\{\frac{1}{3\kappa_{\Sigma^*}d}, \frac{1}{3\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}d}\right\} \tag{4.3}$$

then

$$\left\|\Theta^* - \tilde{\Theta}\right\|_\infty \leq r$$

Proof (adaptation of the one given in Ravikumar et al. 2011). The problem (4.1) has a unique solution, thus the gradient condition holds:

$$G(\Theta_S) \coloneqq -[\Theta^{-1}]_S + \hat{\Sigma}_S + \Lambda_S * Z_S = 0$$

where Z_S denotes an element of the sub-gradient: $Z_S \in \partial_S \left\|\tilde{\Theta}\right\|_1$.

Now we define a continuous function $F\colon B(r) \to \mathbb{R}^{|S|}$ (where $B(r)$ stands for a zero-centered $|S|$-dimensional l_∞ ball of radius r)

$$F(\Delta_S) \coloneqq -(\Gamma^*_{SS})^{-1}\overline{G}(\Theta^* + \Delta_S) + \overline{\Delta}_S$$

We now claim that $F(B(r)) \subseteq B(r)$.

First, rewrite the expression for $G(\tilde{\Theta}_S)$ as

$$G(\Theta^*_S + \Delta_S) = [-[(\Theta^* + \Delta)^{-1}]_S + [\Theta^{*-1}]_S] + W_S + \Lambda_S * Z_S \tag{4.4}$$

By Lemma 9 (which applies due to assumption (4.3) and the choice of Δ) we have

$$\overline{R}(\Delta_S)_S = \operatorname{vec}((\Theta^* + \Delta)^{-1} - \Theta^{*-1})_S + \Gamma^*_{SS}\overline{\Delta}_S = \operatorname{vec}(\Theta^{*-1}\Delta\Theta^{*-1}\Delta J\Theta^{*-1})_S \tag{4.5}$$

Using (4.4) and (4.5) obtain

$$F(\overline{\Delta}_S) = \underbrace{(\Gamma^*_{SS})^{-1}\operatorname{vec}(\Theta^{*-1}\Delta\Theta^{*-1}\Delta J\Theta^{*-1})_S}_{T_1} - \underbrace{(\Gamma^*_{SS})^{-1}(\overline{W}_S + \overline{\Lambda}_S * \overline{Z}_S)}_{T_2}$$

Clearly, $\|T_2\|_\infty \leq \kappa_{\Gamma^*}(\|W\|_\infty + \|\Lambda\|_\infty) = r/2$

As for T_1, by Lemma 9, we have,

$$\|T_1\|_\infty \leq \frac{3}{2}d\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}\|\Delta\|^2_\infty \leq \frac{3}{2}d\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}r^2$$

and again, by assumption (4.3), we obtain $\|T_1\|_\infty \leq r/2$.

Now, we have shown that the continuous function $F(\cdot)$ maps a ball $B(r)$ onto itself. Thus, we can apply the fixed-point theorem. But obviously, this function has a fixed point iff. $\exists\Delta_S \in B(r)\colon G(\Theta^*_S + \Delta_S) = 0$ which is a sufficient and necessary condition for $\Theta^* + \Delta$ to be a solution of optimization problem (4.1) and thus $\left\|\Theta^* - \tilde{\Theta}\right\|_\infty \leq r$.

□

Proof of Theorem 1. First, we note that $\hat{\Theta}^{init}_{ij} = 0$ iff. $\theta^*_{ij} = 0$ (by Lemma 8).

Thus, by the choice of Λ^{ada}, $\hat{\Theta}^{ada} = \tilde{\Theta}^{ada}$, so we can analyze the problem (4.1). Note, that this also implies the fact that $\Theta^*_{ij} = 0 \Leftrightarrow \hat{\Theta}^{ada}_{ij} = 0$

Also, by Lemma 8 $\left\|\hat{\Theta}^{ada} - \Theta^*\right\|_\infty \leq r$. Thus,

$$\left\|\Lambda^{ada}_S\right\|_\infty \leq \frac{\lambda_n}{\left(\min\limits_{i,j\colon \hat{\Theta}^{init}_{ij} \neq 0} \hat{\Theta}^{init}_{ij}\right)^\gamma} \leq \frac{\lambda_n}{(\theta_{min} - r)^\gamma} \tag{4.6}$$

Lemma 3 applies to the problem (4.1) with tuning parameters Λ^{ada} due to the sparsity bound (3.1) and the bound we have just obtained. Thus, $\left\|\Theta^* - \tilde{\Theta}^{ada}\right\|_\infty \leq 2\kappa_{\Gamma^*}(\|W\|_\infty + \left\|\Lambda^{ada}_S\right\|_\infty)$.

Substituting the bound (4.6), recalling that we are considering the set $\mathcal{T}$ and that $\hat{\Theta}^{ada} = \tilde{\Theta}^{ada}$ we obtain the claimed bound.

□

4.3. Proof of SCAD graphical lasso consistency result

Lemma 4 (generalization of Lemma 4, Ravikumar et al. 2011)**.** *Let*

$$\max\{\|W\|_\infty, \|R(\Delta)\|_\infty\} \leq \frac{\alpha}{8}\rho$$

and

$$\frac{\|\Lambda_S\|_\infty}{\rho} \leq 1 \tag{4.7}$$

Also, suppose Assumption 1 holds for some $\alpha \in (0, 1]$.
Then $\hat{\Theta} = \tilde{\Theta}$.

Proof (adaptation of the one given in Ravikumar et al. 2011). First, rewrite the stationarity condition for the problem (1.1) as

$$\Theta^{*-1}\Delta\Theta^{*-1} + W - R(\Delta) + \Lambda * Z = 0$$

By vectorizing obtain:

$$\Gamma^*\overline{\Delta} + \overline{W} - \overline{R} + \overline{\Lambda * Z} = 0$$

Now, using the fact that $\Delta_{S^c} = 0$ rewrite it in terms of disjoint decomposition:

$$\Gamma^*_{SS}\overline{\Delta_S} + \overline{W_S} - \overline{R}_S + \overline{\Lambda_S * Z_S} = 0 \tag{4.8}$$

$$\Gamma^*_{\mathrm{S^c}\,S}\overline{\Delta_S} + \overline{W_{S^c}} - \overline{R}_{\mathrm{S^c}} + \overline{\Lambda_{\mathrm{S^c}} * Z_{S^c}} = 0 \tag{4.9}$$

Solving (4.8) we obtain

$$\overline{\Delta_S} = -(\Gamma^*_{SS})^{-1}[\overline{W_S} - \overline{R}_S + \overline{\Lambda_S * Z_S}]$$

Now, by solving (4.9) for $\overline{Z}_{\mathrm{S^c}}$ and by substituting $\overline{\Delta_S}$:

$$\begin{aligned}\overline{Z}_{\mathrm{S^c}} &= -[\Gamma^*_{\mathrm{S^c}\,S}\overline{\Delta_S} + \overline{W_{S^c}} - \overline{R}_{\mathrm{S^c}}] \oslash \overline{\Lambda}_{\mathrm{S^c}} \\ &= \left[(I - \Gamma^*_{\mathrm{S^c}\,S}{\Gamma^*_{SS}}^{-1})(\overline{W}_S + \overline{R}_S) - \Gamma^*_{\mathrm{S^c}\,S}{\Gamma^*_{SS}}^{-1}\overline{\Lambda}_S * \overline{Z}_S\right]\end{aligned}$$

Where $\cdot \oslash \cdot$ denotes matrix element-wise division.

Now we take the ℓ_∞ norm of both sides and recall Assumption 1

$$\begin{aligned}\left\|\overline{Z}_{\mathrm{S^c}}\right\|_\infty &\leq \frac{2-\alpha}{\rho}(\|W\|_\infty + \|R\|_\infty) + (1-\alpha)\frac{\|\Lambda_S\|_\infty}{\rho} \\ &\leq \frac{2}{\rho}(\|W\|_\infty + \|R\|_\infty) + (1-\alpha) \\ &\leq \frac{2}{\rho}\left(\frac{2\alpha}{8}\rho\right) + (1-\alpha) \\ &= 1 - \frac{\alpha}{2} \\ &< 1\end{aligned}$$

The strict dual feasibility condition holds. Therefore, we have $\hat{\Theta} = \tilde{\Theta}$.

□

Lemma 5 (generalization of Theorem 1, Ravikumar et al. 2011)**.** *Consider a distribution satisfying Assumption 1 with some $\alpha \in (0, 1]$, let $\hat{\Theta}$ be the solution of optimization problem* (1.1)*.*

Suppose also the following restrictions on the penalization parameters Λ hold

$$\|\Lambda_S\|_\infty \leq \frac{8}{\alpha}\delta_n$$

and

$$\rho \geq \frac{8}{\alpha}\delta_n$$

Furthermore, suppose the following sparsity assumption holds:

$$d \leq \frac{\delta_n}{6(\delta_n + \|\Lambda_S\|_\infty)^2 \max\{\kappa_{\Gamma^*}\kappa_{\Sigma^*}, \kappa_{\Gamma^*}^2\kappa_{\Sigma^*}^3\}} \tag{4.10}$$

Then on the set $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_\infty < \delta_n\right\}$ *the following hold:*

$$\left\|\hat{\Theta} - \Theta^*\right\|_\infty \leq r_\Lambda := 2\kappa_{\Gamma^*}(\delta_n + \|\Lambda_S\|_\infty)$$

and

$$\Theta^*_{ij} = 0 \Rightarrow \hat{\Theta}_{ij} = 0 \tag{4.11}$$

Proof. First, we show that Lemma 3 applies.

The inequality

$$2\kappa_{\Gamma^*}(\|W\|_\infty + \|\Lambda_S\|_\infty) \leq \min\left\{\frac{1}{3\kappa_{\Sigma^*}d}, \frac{1}{3\kappa_{\Sigma^*}^3\kappa_{\Gamma^*}d}\right\} \tag{4.12}$$

holds due to assumption (4.10). Therefore, we have a bound

$$\left\|\tilde{\Theta} - \Theta^*\right\|_\infty \leq 2\kappa_{\Gamma^*}(\|W\|_\infty + \|\Lambda_S\|_\infty) \tag{4.13}$$

Now, we show the applicability of Lemma 4.

First, observe that

$$\|W\|_\infty \leq \delta_n \leq \frac{\alpha}{8}\rho$$

In order to bound $R(\Delta)$ we use Lemma 9 which applies due to bounds (4.13) and (4.12) and make use of the sparsity bound (4.10):

$$\|R(\Delta)\|_\infty \leq \frac{3}{2}d\,\|\Delta\|_\infty^2\,\kappa_{\Sigma^*}^3 \leq \delta_n \leq \frac{\alpha}{8}\rho$$

The assumption (4.7) of Lemma 4 clearly holds as well.

Thus, $\tilde{\Theta} = \hat{\Theta}$ which combined with (4.13) gives the bound claimed along the with the sparsistency property (4.11).

□

Proof of Theorem 2. Since, the conditions of the Lemma 8 hold, $\left\|\hat{\Theta}^{(0)} - \Theta^*\right\|_\infty \leq r$ and $\Theta^*_{ij} = 0 \Leftrightarrow \Theta^{(0)}_{ij} = 0$.

Therefore, $\left\|\Lambda_S^{(1)}\right\|_\infty \leq \lambda_n \leq \frac{8}{\alpha}\delta_n$ and $\rho^{(1)} = \lambda_n \geq \frac{8}{\alpha}\delta_n$ and, due to $\hat{\Sigma}$ being non-singular, the problem (3.2) has a unique solution. Thus, Lemma 5 applies here giving the bound for

$\hat{\Theta}^{OSSCAD}$. Moreover, due to the bound (A.1) we have $\Theta^*_{ij} = 0 \Leftrightarrow \hat{\Theta}^{OSSCAD}_{ij} = 0$ (since the bound for $\hat{\Theta}^{(1)}_{ij}$ is not less strict that the one for $\hat{\Theta}^{(0)}_{ij}$). □

Proof of Theorem 3. Theorem 2 provides the bound for $\hat{\Theta}^{(1)}$ along with the sparsistency property: $\Theta^*_{ij} = 0 \Leftrightarrow \hat{\Theta}^{(1)}_{ij} = 0$.

Following the same argument we prove the following bound for every $\hat{\Theta}^{(k)}$:

$$\left\|\hat{\Theta}^{(k)} - \Theta^*\right\|_\infty \leq 2\kappa_{\Gamma^*}\left(\delta_n + \left\|\Lambda^{(k)}_S\right\|_\infty\right) \tag{4.14}$$

and we have the following recurrent expression for $\Lambda^{(k)}_S$

$$\left\|\Lambda^{(k)}_S\right\|_\infty \leq \left(\frac{a\lambda_n - |\theta_{min} - 2\kappa_{\Gamma^*}(\delta_n + \Lambda^{(k-1)}_S)|}{a-1}\right)_+$$

Some algebra yields

$$\left\|\Lambda^{(k)}_S\right\|_\infty \xrightarrow{k\to\infty} \left\|\Lambda^{(\infty)}_S\right\|_\infty \leq \left(\frac{a\lambda_n - \theta_{min} + 2\kappa_{\Gamma^*}\delta_n}{2\kappa_{\Gamma^*} + a - 1}\right)_+$$

And the passage to the limit in inequality (4.14) yields the claimed bound. The second statement of the theorem follows from (A.1).

□

4.4. Proof of the inference result

The next lemma bounds the remainder r on the set $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_\infty < \delta_n\right\}$.

Lemma 6. *Suppose, assumptions of Lemma 5 hold. Then, on the set* $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_\infty < \delta_n\right\}$ *it holds that*

$$\|r\|_\infty \leq R := 2dr_\Lambda(\kappa_{\Theta^*}\delta_n + (d\delta_n + \kappa_{\Sigma^*})r_\Lambda) \tag{4.15}$$

Proof.

$$\begin{aligned}
\|r\|_\infty &\leq \left\|(\hat{\Theta} - \Theta^*)W\Theta^*\right\|_\infty + \left\|(\hat{\Theta}\hat{\Sigma} - I_p)(\hat{\Theta} - \Theta^*)\right\|_\infty \\
&\leq |||(\hat{\Theta} - \Theta^*)|||_1(\|W\Theta^*\|_\infty + \left\|(\hat{\Theta}\hat{\Sigma} - I_p)\right\|_\infty) \\
&\leq dr_\Lambda(\|W\Theta^*\|_\infty + \left\|\hat{\Theta}\hat{\Sigma} - I_p\right\|_\infty)
\end{aligned}$$

$$\begin{aligned}\left\|(\hat{\Theta}\hat{\Sigma} - I_p)\right\|_\infty &= \left\|(\hat{\Sigma} - \Sigma^*)(\hat{\Theta} - \Theta^*) + \Sigma^*(\hat{\Theta} - \Theta^*) + (\hat{\Sigma} - \Sigma^*)\Theta^*\right\|_\infty \\ &\leq (d\delta_n + \kappa_{\Sigma^*}) \left\|\hat{\Theta} - \Theta^*\right\|_\infty + \|W\Theta^*\|_\infty\end{aligned}$$

$$\begin{aligned}\|r\|_\infty &\leq dr_\Lambda(2\|W\Theta^*\|_\infty + (d\delta_n + \kappa_{\Sigma^*}) \left\|\hat{\Theta} - \Theta^*\right\|_\infty) \\ &\leq 2dr_\Lambda(\kappa_{\Theta^*}\delta_n + (d\delta_n + \kappa_{\Sigma^*})r_\Lambda)\end{aligned}$$

□

The next lemma shows that conditioning on a set of high probability does not significantly change the cumulative distribution function of a random variable

Lemma 7. *Let x be a random variable and let A be some set of probability $p_A > 0$. Then*

$$|\,\mathbb{P}\{x < c\} - \mathbb{P}\{x < c \,|\, A\}| \leq 2(1 - p_A)$$

Proof.

$$\begin{aligned}|\,\mathbb{P}\{x < c\} - \mathbb{P}\{x < c \,|\, A\}| &= |\,\mathbb{P}\{x < c \mid A\}\,\mathbb{P}(A) + \mathbb{P}\{x < c \mid \overline{A}\}\,\mathbb{P}(\overline{A}) - \mathbb{P}\{x < c \,|\, A\}| \\ &\leq (1 - p_A)\,\mathbb{P}\{x < c \,|\, A\} + \mathbb{P}\{x < c \mid \overline{A}\}(1 - p_A) \\ &\leq (1 - p_A) + (1 - p_A) = 2(1 - p_A)\end{aligned}$$

□

Proof of Theorem 4. Using equation (3.3), and the definition of $\hat{T}$ (3.4) we obtain for all (i, j)

$$\sqrt{n}(\hat{T}_{ij} - \Theta^*_{ij}) = \frac{1}{\sqrt{n}} \sum_k Z_{ijk} + \frac{r}{\sqrt{n}} \tag{4.16}$$

where

$$Z_{ijk} := \Theta^*_i X_k \Theta^*_j X_k - \Theta^*_{ij} \tag{4.17}$$

Observe that Z_{ijk} are i.i.d. (for (i, j) fixed) and $\mathsf{E}[Z_{ijk}] = 0$.

Now we divide both sides of (4.16) by $\sigma_{ij} := \sqrt{Var[Z_{ijk}]}$

$$\sqrt{n}(\hat{T}_{ij} - \Theta^*_{ij})/\sigma_{ij} = \underbrace{\frac{1}{\sigma_{ij}\sqrt{n}} \sum_k Z_{ijk}}_{S} + \frac{r\sqrt{n}}{\sigma_{ij}}$$

The cumulative distribution function of S can be estimated by Berry-Esseen inequality (Korolev and Shevtsova 2010)

$$|\,\mathbb{P}\{S < c\} - \Phi(c)| \leq \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}}$$

with $A < 0.4748$.

Now from Lemma 7 we have

$$|\,\mathbb{P}\{S < c\} - \mathbb{P}\{S < c|\,\mathcal{T}\}| \leq 2(1 - \mathrm{p}_{\mathcal{T}})$$

Combining the latter two inequalities yields

$$|\,\mathbb{P}\{S < c|\,\mathcal{T}\} - \Phi(c)| \leq \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}} + 2(1 - \mathrm{p}_{\mathcal{T}})$$

Next we make use of the bound for the residual r provided by Lemma 6.

$$\mathbb{P}\left\{S + \frac{r\sqrt{n}}{\sigma_{ij}} < c \mid \mathcal{T}\right\} \leq \mathbb{P}\left\{S - \frac{R\sqrt{n}}{\sigma_{ij}} < c \mid \mathcal{T}\right\}$$

$$\mathbb{P}\left\{\hat{T}_{ij} - \Theta_{ij}^* < c \mid \mathcal{T}\right\} \leq \Phi\left(\frac{\sigma_{ij}}{\sqrt{n}}c + \frac{R\sqrt{n}}{\sigma_{ij}}\right) + \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}} + 2(1 - \mathrm{p}_{\mathcal{T}})$$

And in the same way the lower bound can be obtained as:

$$\mathbb{P}\left\{\hat{T}_{ij} - \Theta_{ij}^* < c \mid \mathcal{T}\right\} \geq \Phi\left(\frac{\sigma_{ij}}{\sqrt{n}}c - \frac{R\sqrt{n}}{\sigma_{ij}}\right) - \frac{A\mu_{ij3}}{\sigma_{ij}^3\sqrt{n}} - 2(1 - \mathrm{p}_{\mathcal{T}})$$

□

5. Simulation experiments

5.1. Functional connectivity network from experimental data

For our examples we rely on a functional network that we determined from an experimental fMRI dataset in a recent study (Puschmann, Brechmann, and Thiel 2013) that examined learning-dependent plasticity in the human auditory cortex. There, fMRI data with a total of 1680 EPI volumes were acquired with a 3 T Siemens MAGNETOM Trio MRI scanner (Siemens AG, Erlangen, Germany) with an eight-channel head array. We randomly selected a dataset from a single subject. The subject performed a learning experiment with auditory stimuli, number comparison task and reward. The details of the experiment and data acquisition can be found in

Puschmann, Brechmann, and Thiel 2013. We do not repeat them here, because we used the fMRI data only to obtain a realistic network with a natural sparsity pattern for the simulation experiments.

In order to define suitable nodes for the functional connectivity network we used the parcellation atlas defined in a recent study (Finn et al. 2015) which is available online at the BioImage Suite NITRC page [1]. We normalized the atlas to the motion-corrected functional dataset using SPM12 [2] with standard parameters. Mean time courses of the $p = 256$ regions-of-interest were determined to estimate a functional connectivity network. In order to exclude changes in the network due to the learning effect in the experiment, only the last $n = 300$ time points were used. The network analysis was conducted on the residuals of linear modeling common in fMRI experiments (Poldrack, Mumford, and Nichols 2011). This way a matrix X^* of size 256×300 of real data was acquired.

5.2. Software

All simulations were performed with the R language and environment for statistical computing and graphics (R Core Team 2016). The following R packages were used: `oro.nifti` (Whitcher, Schmid, and Thornton 2011) was used in order to work with the format the data were stored in, as an implementation of graphical lasso the package `glasso` (Friedman, Hastie, and Tibshirani 2014) was used, `igraph` package (Csardi and Nepusz 2006) was used in order to manipulate and visualize graphs, sampling from multivariate normal distribution was conducted by `MASS` package (Venables and Ripley 2002), and an implementation of partial correlation matrix estimator by Pearson method was borrowed from `ppcor` package (Kim 2012).

5.3. Simulation study

5.3.1. The way the data are simulated

First, we extracted a precision matrix which was then considered to be a true matrix from the real data X^*. In order to do so we first used thresholded graphical lasso with some penalization parameter λ_1 and a threshold parameter of 0.1 (e. g. we set all the parameters estimated by graphical lasso smaller than 0.1 by absolute value to zero) applied to all the non-diagonal elements of the precision matrix on the whole sample X^*. This way we obtained a sparse precision matrix Θ_1 which may be seen as an adjacency matrix of some graph with 256 vertices: $|V| = 256$. Next we chose the largest component of the graph $C \subset V$. Finally, we used thresholded graphical lasso with the same threshold parameter and a possibly different penalization parameter λ_2 taking into account only the nodes included in the main component, e.g. we applied graphical lasso to the sample $X^*_{C.}$, which produces a precision matrix Θ^* of size $|C| \times |C|$ which is treated as a true matrix in our simulation. It is easy to see that, λ_1 controls the size of the true network and λ_2 controls its sparsity.

[1] `https://www.nitrc.org/frs/?group_id=51`
[2] `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`

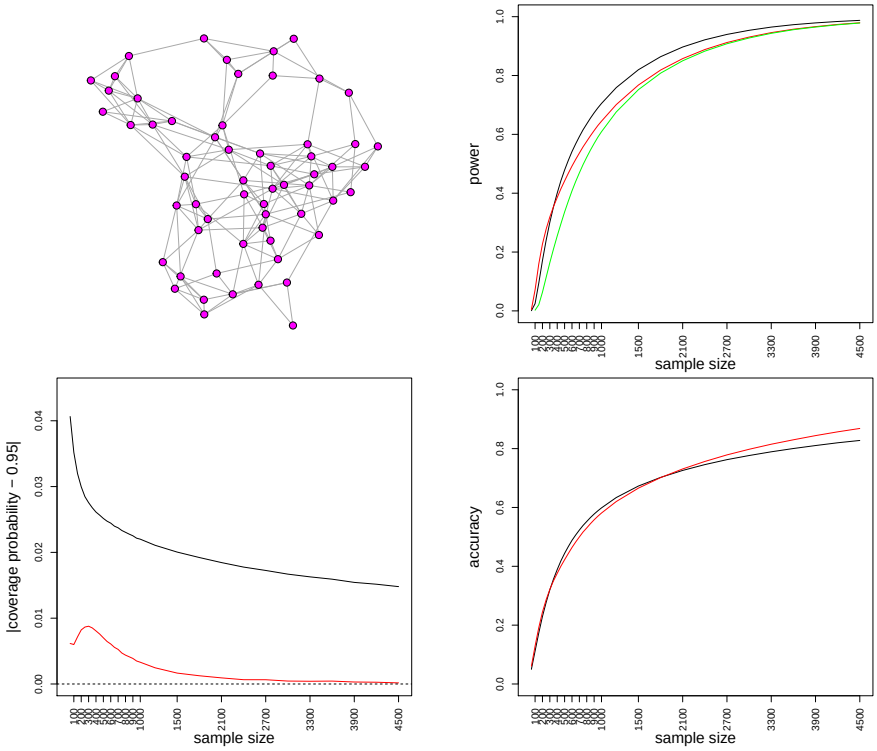

Figure 1 – *Upper-left: Graph obtained with* $\lambda_1 = 0.6$, $\lambda_2 = 0.3$. *Then* $p = 60$ *and* $sparsity = 0.100$. *Upper-right: power of hypothesis testing for adaptive (red), non-adaptive (black) and the classical approach (green). Lower-left: coverage probability for adaptive (red) and non-adaptive (black) approach. Lower-right: accuracy of classification between zero and non-zero parameters using adaptive (red) and non-adaptive approach (black).*

Simulated data were drown independently from a Gaussian distribution $\mathcal{N}(0, {\Theta^*}^{-1})$ varying n from 50 to 4500.

In all the experiments involving either adaptive or non-adaptive graphical lasso the penalization parameter was chosen as $\lambda = \sqrt{\frac{\log p}{n}}$ which is an asymptotically optimal choice (Janková and Geer 2015). In all the experiments one-step SCAD graphical lasso was used as an adaptive approach.

5.3.2. Hypothesis testing

For each non-diagonal element of the precision matrix the null-hypothesis $\mathbb{H}_0^{ij} = \{\Theta_{ij}^* = 0\}$ can be tested against an alternative $\mathbb{H}_1^{ij} = \{\Theta_{ij}^* \neq 0\}$. In order to do so a de-sparsified estimator $\hat{T}_{ij} \rightsquigarrow \mathcal{N}(\Theta_{ij}^*, \sigma_{ij}^2)$ was used with σ_{ij}^2 replaced by the suitable estimator

$$\hat{\sigma}_{ij}^2 := \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}_{ij}^2$$

(see also Lemma 10). Finally, Bonferroni-Hochberg multiplicity correction was applied and the power of the test was computed.

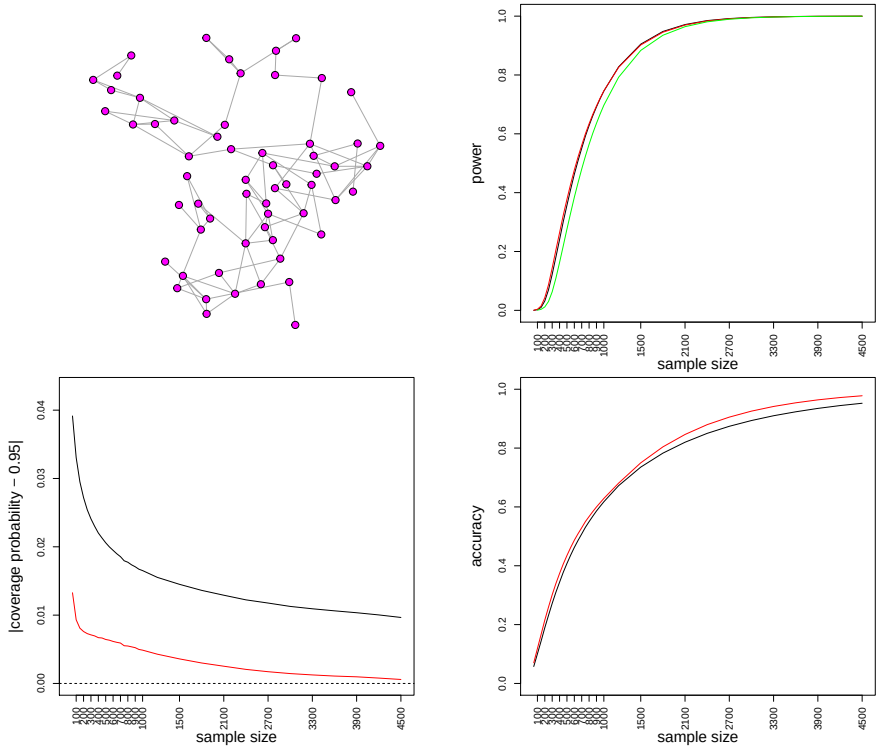

Figure 2 – *As Fig. 1 but with* $\lambda_1 = 0.6$, $\lambda_2 = 0.6$, $p = 60$, $sparsity = 0.050$

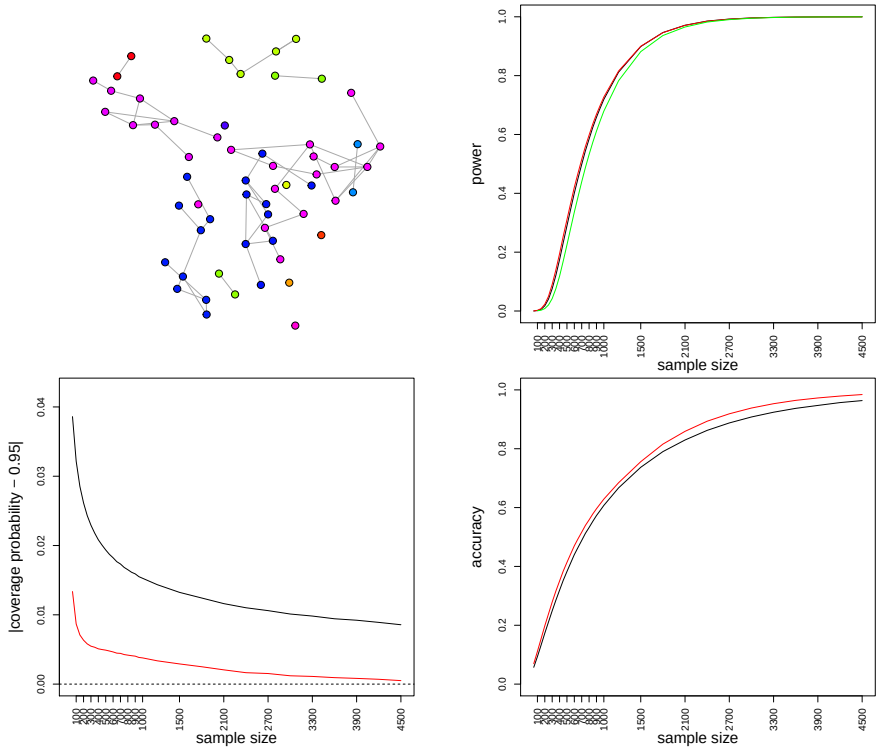

Figure 3 – *As Fig. 1 but* $\lambda_1 = 0.6$, $\lambda_2 = 0.65$, $p = 60$, $sparsity = 0.034$

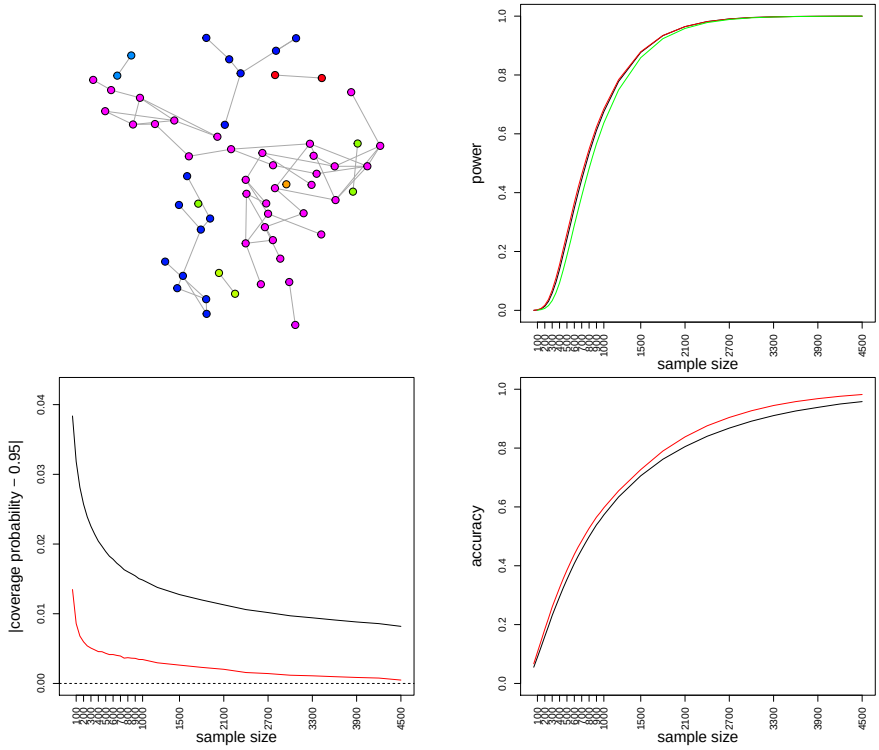

Figure 4 – *As Fig. 1 but* $\lambda_1 = 0.6$, $\lambda_2 = 0.67$, $p = 60$, $sparsity = 0.030$

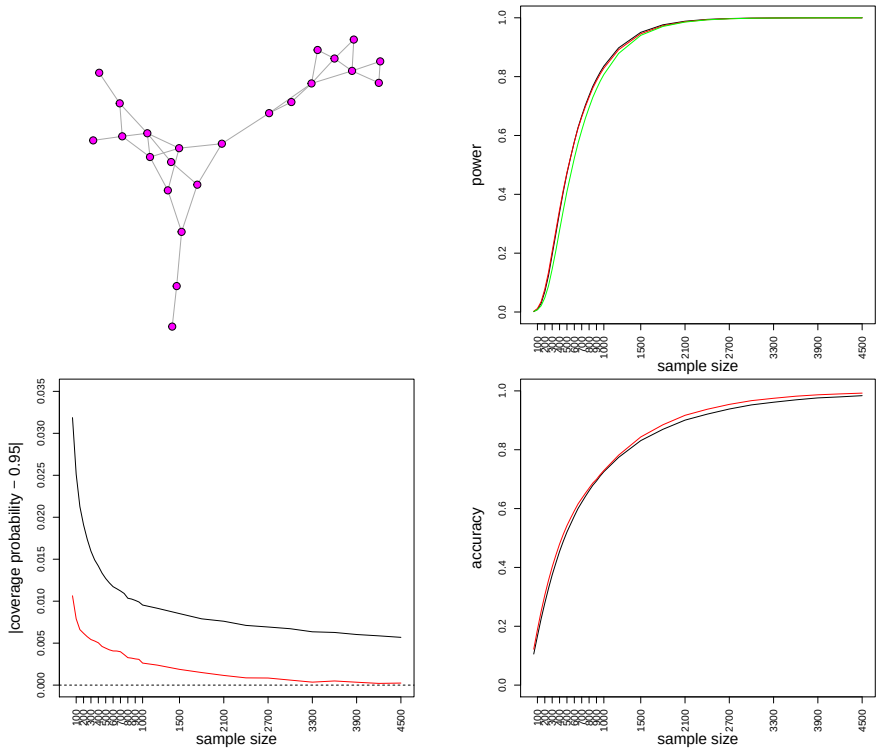

Figure 5 – *As Fig. 1 but* $\lambda_1 = 0.65$, $\lambda_2 = 0.65$, $p = 23$, $sparsity = 0.130$

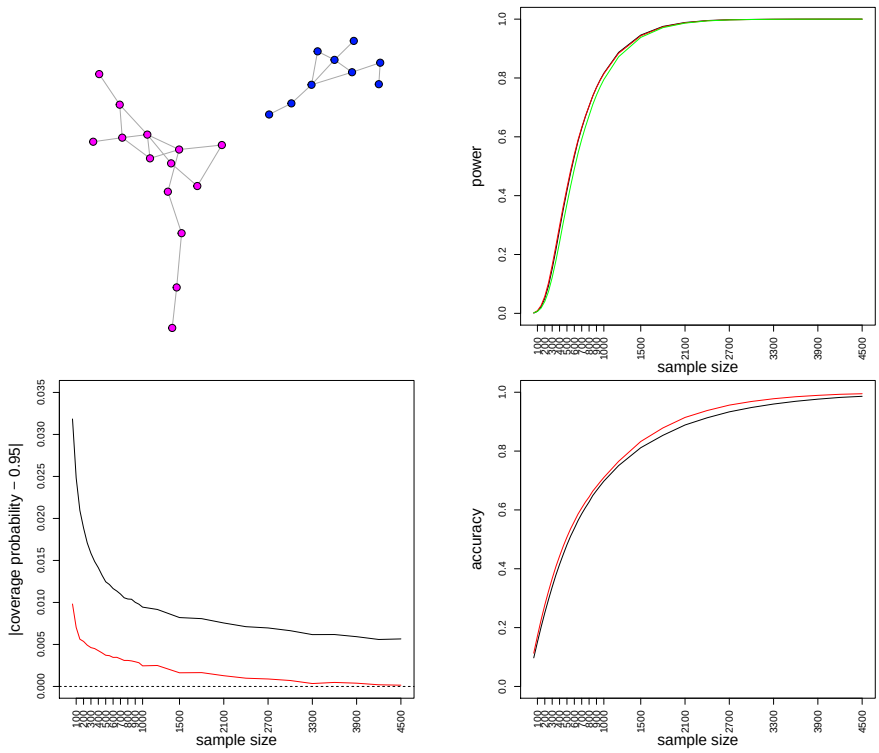

Figure 6 – *As Fig. 1 but* $\lambda_1 = 0.65$, $\lambda_2 = 0.67$, $p = 23$, $sparsity = 0.107$

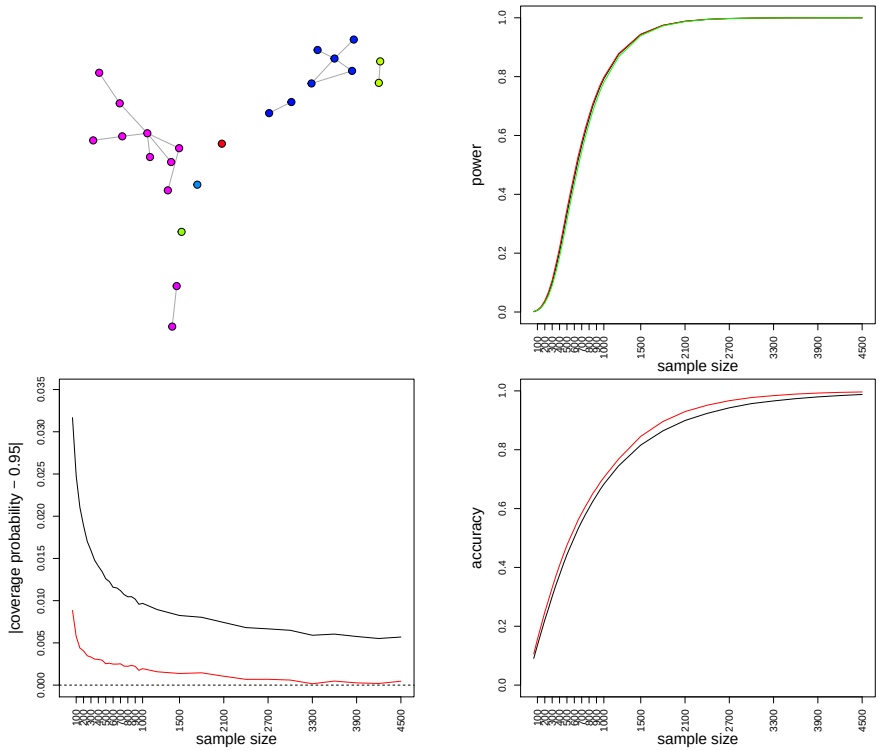

Figure 7 – *As Fig. 1 but* $\lambda_1 = 0.65$, $\lambda_2 = 0.7$, $p = 23$, $sparsity = 0.063$

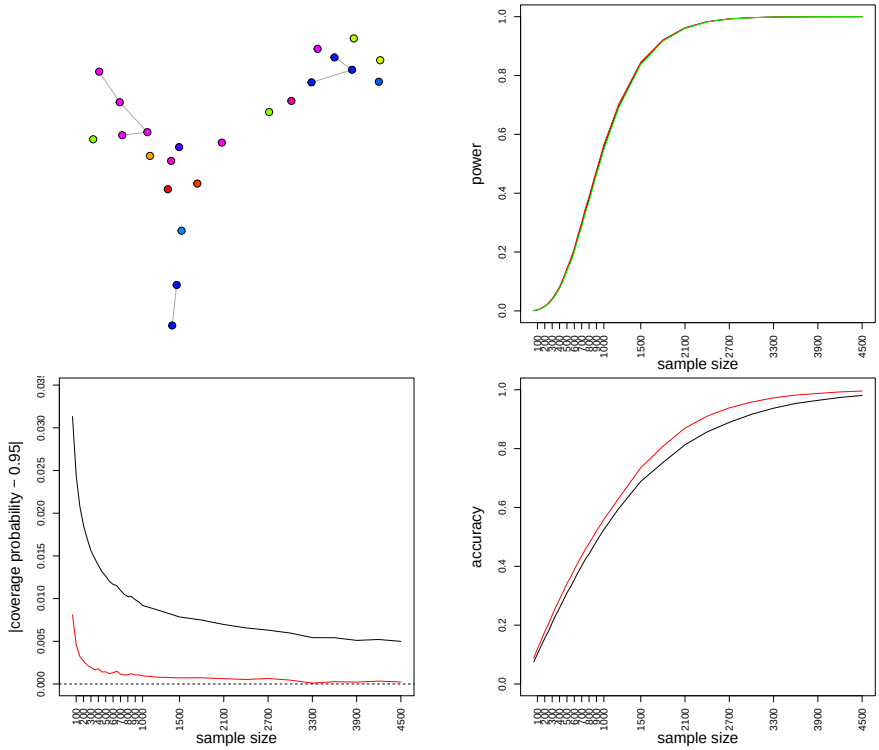

Figure 8 – *As Fig. 1 but* $\lambda_1 = 0.65$, $\lambda_2 = 0.75$, $p = 23$, $sparsity = 0.024$

In our experiments we compared tests based on the de-sparsified estimator produced by ℓ_1-penalized estimator, the adaptive estimator and on the classical approach employing Fisher z-transform $z(\cdot)$ on the elements of the partial correlation matrix. The classical approach can be summarized as follows: the partial correlation matrix was estimated with the Pearson method. Fisher z-transform was applied afterward producing approximately normally distributed values $z_{ij} \rightsquigarrow \mathcal{N}(z(\rho^*_{ij}), n - p - 1)$ where ρ^* is a true partial correlation matrix. Clearly, $\rho^*_{ij} = 0$ iff. $\Theta^*_{ij} = 0$, so one can use values z_{ij} as a test statistic.

The power (the fraction of null-hypotheses $\mathbb{H}_0^{ij}$ rejected for non-zero elements of Θ^*) of these tests were compared.

5.3.3. Confidence interval

Using the de-sparsified estimator $\hat{T}_{ij}$ approximate $(1 - \beta)100\%$ confidence intervals for the individual values of precision matrix were constructed as

$$I_{ij}^{\beta,n}(\hat{\Theta}) = [\hat{T}_{ij} - \Phi^{-1}(1 - \beta/2)\sigma_{ij}/\sqrt{n}, \hat{T}_{ij} + \Phi^{-1}(1 - \beta/2)\sigma_{ij}/\sqrt{n}]$$

where $\Phi(\cdot)$ stands for cumulative distribution function of standard normal distribution and $\Phi^{-1}(\cdot)$ denotes its inverse.

In order to compare the approaches we estimated the mean probability for the interval to cover the true value $\frac{1}{p(p-1)} \sum_{i \neq j} \mathbb{P}\{\Theta^*_{ij} \in I_{ij}^{\beta,n}(\hat{\Theta})\}$ and compared its absolute deviation from $1 - \beta$

for the approaches under comparison.

In our experiments the confidence level was chosen as $\beta = 0.05$.

5.3.4. Classification between zero and non-zero elements

We compared the ability of the adaptive and non-adaptive approaches to classify between zero and non-zero parameters. In order to improve accuracy we applied Fisher z-transform to the partial correlation matrix corresponding to the precision matrix produced by the methods being compared. The distribution of the estimated partial correlation coefficients is highly skewed (for $|\rho| \gg 0$) with variances depending on the correlation coefficients. Fisher z-transform makes variance parameters independent and improves the normal approximation. The values produced by Fisher z-transform were compared with 0.05 (half of the threshold parameter, see sub-Section 5.3.1) in order to classify the correlation coefficient equals as zero or not.

The accuracies of such classifiers were compared.

5.3.5. Description of the figures

The graphs obtained in the manner described in Section 5.3.1 along with the results obtained (see Sections 5.3.2–5.3.4 for details) are given in Figures 1 – 8. The values of the penalization parameters λ_1 and λ_2 used to produce these graphs along the number of vertexes p and its sparsity (the fraction of non-zero off-diagonal element of Θ^*) are given in the captions. The upper-left plots represent the extracted graph. In all these plots each vertex occupies the same spot and disconnected components are shown in different colors. The upper-right plots report the powers of hypothesis testing based on the adaptive, non-adaptive graphical lasso and on a classical approach (see sub-Section 5.3.2). The lower-left plots compare the coverage probabilities of the constructed confidence intervals using the estimator based on the adaptive and non-adaptive estimator (see sub-Section 5.3.3). The lower-right plots represent the comparison of accuracies of classification between zero and non-zero parameters based on the adaptive and non-adaptive approach (see sub-Section 5.3.4). The performance of non-adaptive graphical lasso is show in black, of the adaptive approach in red and the performance of the classical approach (on the plots reporting the powers of statistical tests) is show in green.

6. Discussion

The experiments showed that the tests based on the classical approach are always outperformed by those based on graphical lasso approaches (apart from the cases where $n \gg p$ where all the approaches perform nearly perfect). At the same time adaptive graphical lasso tends to notably outperform non-adaptive graphical lasso in case of short samples, though sometimes (in case of a denser true precision matrix, see Figure 1) non-adaptive approach performs better for sufficiently large samples.

The confidence intervals constructed using the adaptive graphical lasso estimate exhibit cov-

erage probabilities significantly closer to the desired confidence level in comparison to those obtained using non-adaptive graphical lasso estimates.

In experiments on the accuracy of classification between zero and non-zero parameters adaptive approach performs notably better for all sizes of sample n apart for the case of a denser precision matrix (see Figure 1).

We believe, that superiority of adaptive graphical lasso over a non-adaptive graphical lasso is related to the fact that non-adaptive lasso penalizes all the values with the same penalty parameter λ whereas adaptive graphical lasso might reduce penalization of non-zero parameters which leads to the reduction of bias brought in by penalization (compare Theorem 2 and Lemma 8). At the same time, non-normality of the de-sparsified estimator depends on the largest penalization parameter corresponding to a non-zero element $\|\Lambda_S\|_\infty$ (see Theorem 4), which in case of non-adaptive graphical lasso equals λ while in case of an adaptive approach might be smaller.

Acknowledgments

We thank André Brechmann for providing the experimental fMRI data used as basis for the simulation and for extensive discussions on the background.

Appendix A Consistency result for the ℓ_1-penalized estimator by Ravikumar et al. 2011

Lemma 8 (Theorem 1, Ravikumar et al. 2011)**.** *Consider a distribution satisfying Assumption 1 with some $\alpha \in (0, 1]$, let $\hat{\Theta}$ be a solution of the optimization problem* (1.1) *with tuning parameters $\Lambda_{ij} = \lambda_n = \frac{8}{\alpha}\delta_n$ for $i \neq j$.*

Furthermore, suppose the following sparsity assumption:

$$d \leq \frac{\delta_n}{6(\delta_n + \lambda_n)^2 \max\{\kappa_{\Gamma^*}\kappa_{\Sigma^*}, \kappa_{\Gamma^*}^2\kappa_{\Sigma^*}^3\}}$$

Also assume that

$$\theta_{min} > r := 2\kappa_{\Gamma^*}(\delta_n + \lambda_n) \tag{A.1}$$

Then on the set $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_\infty < \delta_n\right\}$ the following holds: $\left\|\hat{\Theta} - \Theta^*\right\|_\infty \leq r$ and $\Theta^*_{ij} = 0 \Leftrightarrow \hat{\Theta}_{ij} = 0$.

Appendix B The bound for $R(\triangle)$ by Ravikumar et al. 2011

Lemma 9 (Lemma 5, Ravikumar et al. 2011)**.** *Suppose,* $\|\Delta\|_\infty \leq \frac{1}{3\kappa_{\Sigma^*}d}$*. Then the matrix* $J := \sum_{k=0}^{\infty}(-1)^k(\Theta^{*^{-1}}\Delta)^k$ *satisfies the bound* $\|J\|_\infty \leq 3/2$ *and the matrix*

$$R(\Delta) = \Theta^{*^{-1}}\Delta\Theta^{*^{-1}}\Delta J\Theta^{*^{-1}}$$

is bounded as

$$\|R(\Delta)\|_\infty \leq \frac{3}{2}d\,\|\Delta\|_\infty^2\,\kappa_{\Sigma^*}^3$$

Appendix C The estimation $\hat{\sigma}_{ij}^2$ for σ_{ij}^2

Lemma 10 (generalization of Lemma 2 by Janková and Geer 2015)**.** *Assume conditions of Lemma 5. Moreover, let* $X_i \sim \mathcal{N}(0, \Sigma^*)$*. Define the estimator* $\hat{\sigma}_{ij}^2$ *as*

$$\hat{\sigma}_{ij}^2 := \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}_{ij}^2$$

Then on set $\mathcal{T} = \left\{\left\|\hat{\Sigma} - \Sigma^*\right\|_\infty < \delta_n\right\}$

$$|\hat{\sigma}_{ij}^2 - \sigma_{ij}^2| \leq 2r_\Lambda(2\nu_{\Theta^*} + r_\Lambda)$$

where $\nu_{\Theta^*} = \|\Theta^*\|_\infty$

Proof. $X_i \sim \mathcal{N}(0, \Sigma^*)$, then $\Theta^* X \sim N(0, \Theta^*)$. Some algebra yields

$$\sigma_{ij}^2 = \Theta_{ii}^*\Theta_{jj}^* + \Theta_{ij}^{*2}$$

Therefore

$$|\hat{\sigma}_{ij}^2 - \sigma_{ij}^2| \leq |\hat{\Theta}_{ii}\hat{\Theta}_{jj} - \Theta_{ii}^*\Theta_{jj}^*| + |\hat{\Theta}_{ij}^2 - \Theta_{ij}^{*2}|$$

Now using the bond provided by Lemma 5 we can bound the terms on the right hand

$$|\hat{\Theta}_{ii}\hat{\Theta}_{jj} - \Theta_{ii}^*\Theta_{jj}^*| \leq (\Theta_{ii}^* + \Theta_{jj}^*)r_\Lambda + r_\Lambda^2$$

$$\begin{aligned}|\hat{\Theta}_{ij}^2 - \Theta_{ij}^{*2}| &= |(\hat{\Theta}_{ij} - \Theta_{ij}^*)(\hat{\Theta}_{ij} + \Theta_{ij}^*)| \\ &\leq r_\Lambda(2\Theta_{ij}^* + r_\Lambda)\end{aligned}$$

And finally

$$
\begin{aligned}
|\hat{\sigma}_{ij}^2 - \sigma_{ij}^2| &\leq r_\Lambda(2\nu_{\Theta^*} + r_\Lambda) + 2\nu_{\Theta^*} r_\Lambda + r_\Lambda^2 \\
&= 2r_\Lambda(2\nu_{\Theta^*} + r_\Lambda)
\end{aligned}
$$

□

Appendix D Probability of the set $\mathcal{T}$

Assumption 2 (Sub-Gaussianity condition)**.** *Denote the normalized components of the vector X_1 as $\xi_i = \frac{X_{1i}}{\sqrt{\Sigma^*_{ii}}}$. Then, we say that the Sub-Gaussianity condition holds for vector X_1 if*

$$
\exists K > 0 : \forall i \; \mathsf{E} \exp\left(\frac{\xi_i^2}{K^2}\right) \leq 2
$$

Lemma 11 (by Ravikumar et al. 2011 in form by Janková and Geer 2015)**.** *Let Assumption 2 hold for some $K > 0$. Then for*

$$
\delta(n, r) = 8(1 + 12K^2) \max_i \Sigma^*_{ii} \sqrt{2\frac{\log(4r)}{n}}
$$

and for any $\gamma > 2$ and for n such that $\delta(n, p^\gamma) < 8(1 + 12K^2) \max_i \Sigma^*_{ii}$ we have

$$
\mathbb{P}\left\{ \left\| \hat{\Sigma} - \Sigma^* \right\|_\infty \leq \delta(n, p^\gamma) \right\} \geq 1 - \frac{1}{p^{\gamma-2}}
$$

References

Allen, E., E. Damaraju, S. Plis, E. Erhardt, T. Eichele, and V. Calhoun (2012). „Tracking Whole-Brain Connectivity Dynamics in the Resting State." *Cereb. Cortex* 24.3, pp. 663–676.

Bassett, D., N. Wymbs, M. Porter, P. Mucha, J. Carlson, and S. Grafton (2011). „Dynamic reconfiguration of human brain networks during learning". *Proc. Natl. Acad. Sci. USA* 108, pp. 7641–7646.

Buxton, R., E. Wong, and L. Frank (1998). „Dynamics of blood flow and oxygenation changes during brain activation: The balloon model". *Magn. Res. Med.* 39.6, pp. 855–864.

Csardi, G. and T. Nepusz (2006). „The igraph software package for complex network research". *InterJournal* Complex Systems, p. 1695. URL: `http://igraph.org`.

Fan, J., Y. Feng, and Y. Wu (2009). „Network exploration via the adaptive LASSO and SCAD penalties". *Ann. Appl. Stat.* 3.2, pp. 521–541.

Fan, J. and R. Li (2001). *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*.

Finger, S. (1994). *Origins of Neuroscience: A History of Explorations into Brain Function*. Oxford University Press.

Finn, E. S., X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable (2015). „Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity". *Nat. Neurosci.* 18, pp. 1664–1671.

Friedman, J., T. Hastie, and R. Tibshirani (2014). *glasso: Graphical lasso- estimation of Gaussian graphical models*. R package version 1.8. URL: `https://CRAN.R-project.org/package=glasso`.

Friston, K. J. (2011). „Functional and effective connectivity: a review." *Brain Connect.* 1, pp. 13–36.

Friston, K. J., C. D. Frith, P. F. Liddle, and R. S. J. Frackowiak (1993). „Functional Connectivity: The Principal-Component Analysis of Large (PET) Data Sets". *J. Cereb. Blood Flow Metab.* 13.1, pp. 5–14.

Friston, K. J., C. Buechel, G. R. Fink, J. Morris, E. Rolls, and R. J. Dolan (1997). „Psychophysiological and Modulatory Interactions in Neuroimaging". *NeuroImage* 6, pp. 218–229.

Friston, K. J. (1994). „Functional and Effective Connectivity in Neuroimaging: A Synthesis". *Hum. Brain Mapp.* 2, pp. 56–78.

Geer, S. van de, P. Bühlmann, Y. Ritov, and R. Dezeure (2014). „On asymptotically optimal confidence regions and tests for high-dimensional models". *Ann. Statist.* 42.3, pp. 1166–1202.

Huettel, S., A. Song, and G. McCarthy (2014). *Functional Magnetic Resonance Imaging*. 3rd. Sinauer Associates, Inc.

Janková, J. and S. van de Geer (2015). „Confidence intervals for high-dimensional inverse covariance estimation". *Electron. J. Statist.* 9.1, pp. 1205–1229.

Kim, S. (2012). *ppcor: Partial and Semi-partial (Part) correlation*. R package version 1.0. URL: `https://CRAN.R-project.org/package=ppcor`.

Kiviniemi, V., J.-H. Kantola, J. Jauhiainen, A. Hyvärinen, and O. Tervonena (2003). „Independent component analysis of nondeterministic fMRI signal sources". *NeuroImage* 19, pp. 253–260.

Korolev and I. G. Shevtsova (2010). „On the Upper Bound for the Absolute Constant in the Berry-Esseen Inequality“. *Theory Probab. Appl.* 54.4, pp. 638–658.

Lam, C. and J. Fan (2009). „Sparsistency and rates of convergence in large covariance matrix estimation“. *Ann. Statist.* 37.6B, pp. 4254–4278. URL: http://dx.doi.org/10.1214/09-AOS720.

Mantini, D., M. G. Perrucci, C. D. Gratta, G. L. Romani, and M. Corbetta (2007). „Electrophysiological signatures of resting state networks in the human brain“. *Proc. Natl. Acad. Sci. USA* 104.32, pp. 13170–13175.

Poldrack, R. A., J. A. Mumford, and T. E. Nichols (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Puschmann, S., A. Brechmann, and C. M. Thiel (2013). „Learning-dependent plasticity in human auditory cortex during appetitive operant conditioning“. *Hum. Brain Mapp.* 34.11, pp. 2841–2851.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). „High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence“. *Electron. J. Statist.* 5, pp. 935–980.

Rissman, J., A. Gazzaley, and M. D'Esposito (2004). „Measuring functional connectivity during distinct stages of a cognitive task.“ *NeuroImage* 23.2, pp. 752–763.

Smith, S. M., K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich (2011). „Network modelling methods for FMRI“. *NeuroImage* 54, pp. 875–91.

Sporns, O. (2013). „Network attributes for segregation and integration in the human brain“. *Curr. Opin. Neurobio* 23, pp. 162–171.

Sporns, O. (2011). *Networks of the brain*. The MIT Press.

Tibshirani, R. (1994). „Regression Shrinkage and Selection Via the Lasso“. *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. 4th. New York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Whitcher, B., V. J. Schmid, and A. Thornton (2011). „Working with the DICOM and NIfTI Data Standards in R“. *J. Stat. Softw.* 44.6, pp. 1–28.

Zou, H. (2006). „The adaptive lasso and its oracle properties“. *J. Amer. Statist. Assoc.* 101, pp. 1418–1429.

Zou, H. and R. Li (2008). „One-step sparse estimates in nonconcave penalized likelihood models“. *Ann. Statist.* 36.4, pp. 1509–1533.