Consistent neighbourhood selection for sparse high-dimensional graphs with the Lasso

Nicolai Meinshausen and Peter Bühlmann
Seminar für Statistik
ETH Zürich

May 14, 2004

Abstract

The pattern of zero entries in the covariance matrix of a multivariate normal distribution corresponds to conditional independence restrictions between variables. The structure is most conveniently summarized in a graphical model (Lauritzen 1996).

Covariance selection (Dempster 1972) aims at estimating those structural zeros from data. The complexity of standard covariance selection methods is, however, very high, making inference of all but low-dimensional graphs infeasible. Moreover, existence of the MLE estimate cannot be guaranteed and the performance of the method is poor if the number of observations is small compared to the number of variables.

We propose neighbourhood selection with the Lasso as a computationally attractive alternative to standard covariance selection for sparse high-dimensional graphs. Neighbourhood selection estimates the conditional independence restrictions separately for each node in the graph.

We show that the proposed neighbourhood selection scheme is consistent for sparse high-dimensional graphs. The consistency hinges on the choice of the penalty parameter. Maybe surprisingly, the oracle value for optimal prediction does not lead to a consistent neighbourhood estimate. It is proposed instead to control the probability of falsely joining some distinct connectivity components of the graph. This leads to consistent estimation for sparse graphs (with exponential rates), even when the number of variables grows like any power of the number of observations.

1 Introduction

Consider the p-dimensional multivariate normal distributed random variable $X = (X_1, \ldots, X_p) \sim \mathcal{N}(0, \Sigma)$. The conditional independence structure of the distribution can be conveniently represented by a graphical model (Γ, E), where $\Gamma = \{1, \ldots, p\}$ is the set of nodes and E the set of edges in $\Gamma \times \Gamma$. A pair (a, b) is contained in the edge set E if and only if X_a is conditionally dependent of X_b, given all remaining variables $X_{\Gamma\backslash\{a,b\}} = \{X_m, m \in \Gamma\backslash\{a, b\}\}$. Every pair of variables not contained in the edge set is conditionally independent, given all remaining variables and corresponds to a zero entry in the inverse covariance matrix (Lauritzen 1996).

Covariance selection was introduced by Dempster (1972) and aims at discovering the conditional independence restrictions (the graph) from a set of i.i.d. observations.

Covariance selection traditionally relies on the discrete optimization of an objective function (see e.g. Lauritzen 1996; Edwards 2000; or in the regression context e.g. Akaike 1970; Schwarz 1978; Shibata 1981; Rissanen 1986; Lienhart and Zuchini 1986; George 2000). Exhaustive search is computationally infeasible for all but very low-dimensional models. Usually, greedy forward or backward search is employed. In forward search, the initial estimate of the edge set is the empty set and edges are then added iteratively until a suitable stopping criterion is fulfilled. The selection (deletion) of a single edge in this search strategy requires an MLE fit (Speed and Kiiveri 1986) for $O(p^2)$ different models. The procedure is not well suited for high-dimensional graphs. The existence of the MLE is not guaranteed in general if the number of observations is smaller than the number of nodes (Buhl 1993). More disturbingly, the complexity of the procedure renders even greedy search strategies impractical for modestly sized graphs.

In contrast, neighbourhood selection with the Lasso, proposed in the following, relies on optimization of a convex function, applied consecutively to each node in the graph. The method is computationally very efficient and is consistent even for the high-dimensional setting, as will be shown.

Neighbourhood selection is a subproblem of covariance selection. The neighbourhood ne_a of a node $a \in \Gamma$ is the smallest subset of $\Gamma\backslash\{a\}$ so that X_a is conditionally independent of $X_{\Gamma\backslash(\mathrm{ne}_a \cup \{a\})}$, given X_{ne_a},

$$X_a \perp\!\!\!\perp X_{\Gamma\backslash(\mathrm{ne}_a \cup \{a\})} | X_{\mathrm{ne}_a}.$$

Hence, the neighbourhood of a node $a \in \Gamma$ consists of all nodes $b \in \Gamma\backslash\{a\}$ so that $(a, b) \in E$. Given n i.i.d. observations of X, neighbourhood selection aims at estimating (individually) the neighbourhood of any given variable (or node). The neighbourhood selection can be cast into a standard regression problem and can be solved efficiently with the Lasso (Tibshirani 1996), as will be shown in this paper.

The consistency of the proposed neighbourhood selection will be shown for sparse high-dimensional graphs, where the number of variables is potentially growing like any power of the number of observations (high-dimensionality) whereas the number of neighbours of any variable is growing at most slightly slower than the number of observations (sparsity). One area of application is the analysis of genetic regulatory (sub-) networks where graphical models are used to describe the interaction between dozens or hundreds of variables (genes) and sample size is in the dozens (see e.g. Toh and Horimoto 2002).

A number of studies (e.g. Huber 1973; Breiman and Freedman 1983; Portnoy 1984; Goldenshluger and Tsybakov 2001) have examined the case of regression with a growing number of parameters as sample size increases. The closest to our setting is the recent work of Greenshtein and Ritov (2003), who study consistent prediction in a triangular setup very similar to ours (see also Judistky and Nemirovski 2000). However, the problem of consistent estimation of the model structure, which is the relevant concept for graphical models, is very different and not treated in these studies.

We study in section 2 under which conditions, and at which rate, the neighbourhood estimate with the Lasso converges to the true neighbourhood. The choice of the penalty is crucial in the high-dimensional setting. Maybe surprisingly, the oracle penalty for optimal prediction turns out to be inconsistent for estimation of the true model. This solution might include an unbounded number of noise variables into the model. We motivate a different choice of the penalty such that the probability of falsely connecting two or more distinct connectivity components of the graph is controlled at very low levels. Asymptotically, the probability of estimating the correct neighbourhood converges exponentially to 1, even when the number of nodes in the graph is growing rapidly like any power of the number of observations. As a consequence, consistent estimation of the full edge set in a sparse high-dimensional graph is possible (section 3).

Encouraging numerical results are provided in section 4. The proposed estimate is shown to be both more accurate than the traditional forward selection MLE strategy and computationally much more efficient. The accuracy of the forward selection MLE fit is in particular poor if the number of nodes in the graph is comparable to the number of observations. In contrast, neighbourhood selection with the Lasso is shown to be able reasonably accurate for estimating graphs with several thousand nodes, using only a few hundred observations.

2 Neighbourhood Selection

Instead of assuming a fixed true underlying model, we adopt a more flexible approach similar to the triangular setup in Greenshtein and Ritov (2003). Both the number of nodes in the graphs (number of variables) and the distribution (the covariance matrix) depend in general on the number of observations, $\Gamma = \Gamma(n)$ and $\Sigma = \Sigma(n)$. The number of nodes in the graph is denoted by $p_n = |\Gamma(n)|$. The neighbourhood ne_a of a node $a \in \Gamma(n)$ is the smallest subset of $\Gamma(n)\backslash\{a\}$ so that X_a is conditionally independent of $X_{\Gamma(n)\backslash(\mathrm{ne}_a \cup \{a\})}$, given X_{ne_a},

$$X_a \perp\!\!\!\perp X_{\Gamma(n)\backslash(\mathrm{ne}_a \cup \{a\})} | X_{\mathrm{ne}_a},$$

see for example Lauritzen (1996). The neighbourhood depends in general on n as well. However, this dependence is often notationally suppressed in the following.
It is instructive to give a slightly different definition of a neighbourhood. For each node $a \in \Gamma(n)$ and an arbitrary subset $\Psi \subseteq \Gamma(n)\backslash\{a\}$, let $\theta^{a,\Psi} \in \mathbb{R}^{p_n}$ be the vector of coefficients for optimal prediction of the variable X_a, given $X_\Psi = \{X_k; k \in \Psi\}$,

$$\theta^{a,\Psi} = \arg \min_{\theta:\theta_k=0,\ \forall k \notin \Psi} E(X_a - \sum_{k \in \Gamma(n)} \theta_k X_k)^2. \tag{1}$$

For prediction of X_a, given all remaining variables $\{X_k, k \in \Gamma(n)\backslash\{a\}\}$, we use the shorthand notation $\theta^a = \theta^{a,\Gamma(n)\backslash\{a\}}$. The elements of θ^a are determined by the inverse covariance matrix (Lauritzen 1996). For $b \in \Gamma\backslash\{a\}$ and $K(n) = \Sigma(n)^{-1}$, it holds that $\theta^a_b = -K_{ab}(n)/K_{aa}(n)$. The set of non-zero coefficients of θ^a is identical to the set $\{b \in \Gamma\backslash\{a\} : K_{ab(n)} \neq 0\}$ of non-zero entries in the corresponding row vector of the inverse covariance matrix and defines precisely the set of neighbours of node a in the graph $\Gamma(n)$. The best predictor for X_a is thus a linear function of variables in the set of neighbours of the node a only. The set of neighbours of a node $a \in \Gamma(n)$ can hence be written as

$$\mathrm{ne}_a = \{b \in \Gamma(n) : \theta^a_b \neq 0\}.$$

Given n independent observations, $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}(0, \Sigma(n))$, neighbourhood selection tries to estimate the set of neighbours of a node $a \in \Gamma(n)$. As the optimal linear prediction of X_a has non-zero coefficients precisely for variables in the set of neighbours of the node a, it seems reasonable to try to exploit this relation.

Neighbourhood selection with the Lasso It is well known that the Lasso, introduced by Tibshirani (1996), and known as Basis Pursuit in the context of wavelet regression (Chen et al. 2001), has a parsimonious property (Knight and Fu 2000). When predicting a variable X_a with all remaining variables $\{X_k, k \in \Gamma(n)\backslash\{a\}\}$, the vanishing Lasso coefficient

estimates identify asymptotically the neighbourhood of node a in the graph, as shown in the following.

The Lasso estimate $\hat{\theta}^{a,\lambda}$ of θ^a is given by

$$\hat{\theta}^{a,\lambda} = \arg\min_{\theta:\theta_a=0} \Big(\frac{1}{n} \sum_{i=1}^{n} (X_a^{(i)} - \sum_{k \in \Gamma(n)} \theta_k X_k^{(i)})^2 + \lambda ||\theta||_1, \Big) \tag{2}$$

where $||\theta||_1 = \sum_{b \in \Gamma(n)} |\theta_b|$ is the l_1-norm of the coefficient vector. It follows from the proofs of the Theorems 1 and 2 that the solution of (2) is unique for most cases of interest. However, we do not necessarily require uniqueness for the following. If the solution of (2) is not unique, a convex set of solutions is obtained and the following results hold for any member of this convex set.

Normalization of all variables to common empirical variance is recommended for the estimator in (2). For notational simplicity, we present our theory with common population variance equal to one.

The neighbourhood estimate (parameterized by λ) is defined by the non-zero coefficient estimates of the l_1-penalized regression,

$$\hat{\text{ne}}_a^{\lambda} = \{b \in \Gamma(n) : \hat{\theta}_b^{a,\lambda} \neq 0\}.$$

Each choice of a penalty parameter λ specifies thus an estimate of the neighbourhood ne_a of node $a \in \Gamma(n)$ and one is left with the choice of a suitable penalty parameter. Larger values of the penalty tend to shrink the size of the estimated set, while more variables are in general included into $\hat{\text{ne}}_a^{\lambda}$ if the value of λ is diminished.

A first guess for a suitable penalty is to use the prediction-optimal penalty parameter. However, the prediction-oracle (and cross-validation) choice of the penalty parameter do not lead to consistent neighbourhood estimates, as will be shown in the following.

The prediction-oracle solution A seemingly useful choice of the penalty parameter is the (unavailable) prediction-oracle value,

$$\lambda_{\text{oracle}} = \arg\min_{\lambda} E(X_a - \sum_{k \in \Gamma(n) \setminus \{a\}} \hat{\theta}_k^{a,\lambda} X_k)^2.$$

The expectation is understood to be with respect to a new X, which is independent of the sample on which $\hat{\theta}^{a,\lambda}$ is estimated. The prediction-oracle penalty minimizes the predictive risk among all Lasso estimates. An estimate of λ_{oracle} is obtained by the cross-validated choice λ_{cv}.

For l_0-penalized regression it was shown by Shao (1993) that the cross-validated choice of the penalty parameter is consistent for model selection under certain conditions on the

size of the validation set. However, with the Lasso not even the prediction-oracle solution leads to consistent model selection, as shown in the following for a very simple example.

Example 1 *The number of variables is growing to infinity, that is* $p_n \to \infty$ *for* $n \to \infty$. *The covariance matrix is identical to the identity matrix except except for some pair* $(a, b) \in \Gamma(n) \times \Gamma(n)$, *for which* $\Sigma_{ab}(n) = \Sigma_{ba}(n) = s$, *for all* $n \in \mathbb{N}$ *and some* $0 < s < 1$.

Let $\hat{\text{ne}}_a^{\text{oracle}} = \{b \in \Gamma(n) : \hat{\theta}_b^{a,\lambda_{\text{oracle}}} \neq 0\}$ be the neighbourhood chosen by the oracle solution.

Proposition 1 *The probability of selecting the wrong neighbourhood with the prediction-oracle penalty converges to 1 for Example 1,*

$$P(\hat{\text{ne}}_a^{\text{oracle}} \neq \text{ne}_a) \to 1 \qquad \text{for } n \to \infty.$$

A proof is given in the appendix.
One might suspect that the reason for this result is that not all relevant predictor variables are included into the neighbourhood estimate, thereby reducing the variance of the prediction while accepting a slight increase in the bias. However, it follows from the proof of Proposition 1 that the opposite is true. Many noise variables are included into the neighbourhood estimate with the prediction-oracle solution. In fact, the probability of including noise variables with the prediction-oracle solution does not even vanish asymptotically for a fixed number of variables. This may be a disturbing result as it seems to suggest that consistent neighbourhood selection with the Lasso is hardly possible if not even the prediction-oracle solution is consistent.
However, consistent neighbourhood selection is possible with the Lasso for a different choice of the penalty parameter, as demonstrated in the following.

Consistent solutions The asymptotic properties of Lasso-type estimates in regression have been studied in detail by Knight and Fu (2000) for the more conventional setup with a fixed number of variables and increasing number of observations. Their results say that the penalty parameter should decay for an increasing number of observations at least as fast as $n^{-\frac{1}{2}}$ to obtain an optimal asymptotic distribution. It turns out that a more conservative approach is needed for consistent model selection, as motivated already by the previous example 1. A rate $\lambda \sim n^{-\frac{1}{2}+\varepsilon}$ with any $\varepsilon > 0$ is, however, sufficient for consistent neighbourhood selection, even when the number of variables is growing rapidly with the number of observations.

Assumptions We make a few assumptions to prove consistency of the neighbourhood selection with the Lasso. The main assumption is the sparsity of the graph. This entails that the size of the neighbourhood of any node in the graph is not growing faster than the number of observations.

A1 There exists some $\kappa < 1$ so that

$$\max_{a \in \Gamma(n)} |\mathrm{ne}_a| = O(n^\kappa) \qquad \text{for } n \to \infty.$$

A bounded variance of all variables would suffice for the proofs. However, in practice we recommend to scale the variables to common empirical variance. We mimic this scaling of all variables to common population variance.

A2 For all $a \in \Gamma(n)$ and $n \in \mathbb{N}$, $\Sigma_{aa}(n) = 1$.

Furthermore, we need positive definiteness of the covariance matrix to avoid collinearity. Let $K(n) = \Sigma(n)^{-1}$ be the inverse covariance matrix. Note that $K_{aa}(n)$ is the inverse of $\mathrm{Var}(X_a | X_{\Gamma(n)\setminus\{a\}})$, the variance of X_a conditional on all other variables (Lauritzen 1996).

A3 There exists $\omega^2 < \infty$ so that for all $a \in \Gamma(n)$ and $n \in \mathbb{N}$, $|K_{aa}(n)| \leq \omega^2$.

This assumption bounds from below the conditional variance of X_a for all $a \in \Gamma(n)$ and ensures positive definiteness of the covariance matrix.
Finally, consider the definition in (1) of the coefficients for optimal prediction of X_a, given a subset of variables $\{X_k;\ k \in \Psi\}$ and $\Psi \subseteq \Gamma(n)\setminus\{a\}$.

A4 There exists some $\vartheta_1 < 1$ so that for all $n \in \mathbb{N}$ and for all $a, b \in \Gamma(n)$ with $a \notin \mathrm{ne}_b \cup \{b\}$, $\sum_{k \in \mathrm{ne}_b \cup \{b\}} |\theta_k^{a, \mathrm{ne}_b \cup \{b\}}| < \vartheta_1$. There exists furthermore some $\vartheta_2 < \infty$ so that for all $n \in \mathbb{N}$ and for all $a, b \in \Gamma(n)$ with $a \in \mathrm{ne}_b \setminus \{b\}$, $\sum_{k \in \mathrm{ne}_b \cup \{b\}} |\theta_k^{a, \mathrm{ne}_b \cup \{b\}}| < \vartheta_2$.

This assumption is automatically fulfilled for trees, that is graphs without cycles. It is much more generally valid, though. Using Lemma 2 in the appendix, it can be seen that assumption A4 is fulfilled if the inverse covariance matrix $K(n) = \Sigma(n)^{-1}$ is diagonally dominant, meaning that there exists some $\vartheta_1 < 1$ so that

$$\sum_{b \in \Gamma(n)\setminus\{a\}} |K_{ab}(n)| \leq \vartheta_1 |K_{aa}(n)| \qquad \forall a \in \Gamma(n),\ n \in \mathbb{N}.$$

Using Gershgorins Theorem, a diagonally dominant inverse covariance matrix ensures that the minimal eigenvalue of the inverse covariance matrix is bounded away from zero for all $n \in \mathbb{N}$. Diagonal dominance is sufficient but not necessary for positive definiteness. However, there are examples where diagonal dominance is a necessary condition

for positive definiteness of the inverse covariance matrix. Consider for example a simple two-dimensional lattice with identical partial correlation between neighbouring nodes and periodic boundary conditions. Here, the boundary between positive definiteness and positive semi-definiteness is marked precisely by the value $\vartheta_1 = 1$. However, to keep results as general as possible, we use instead of diagonal dominance the less restrictive assumption A4.

Consistency Now we are ready to prove the first part of the consistency result.

Theorem 1 *Assume A1-A4. Let the penalty parameter satisfy* $\lambda \sim dn^{-\frac{1}{2}+\varepsilon}$ *with some* $0 < \varepsilon < 1/2$ *and* $d > 0$. *If* $p_n = O(n^\gamma)$ *with any* $\gamma > 0$, *there exists some constant* $c > 0$ *so that,*

$$P\Big(\hat{\text{ne}}_a^\lambda \not\subseteq \text{ne}_a\Big) = O(\exp(-cn^\varepsilon)) \qquad \textit{for } n \to \infty.$$

A proof is given in the appendix.

Theorem 1 states that the probability of (falsely) including any of the non-neighbouring variables of the node $a \in \Gamma(n)$ into the neighbourhood estimate is vanishing exponentially fast, even though the number of non-neighbouring variables may grow very rapidly with the number of observations.

The converse of Theorem 1 also holds true the magnitude of the partial correlations between neighbours are not vanishing to zero too fast, as shown in the following theorem. The partial correlation π_{ab} between X_a and X_b is the correlation conditional on all remaining variables $X_{\Gamma(n)\setminus\{a,b\}}$ and is identical to $\pi_{ab} = -K_{ab}(n)/(K_{aa}(n)K_{bb}(n))^{1/2}$, where $K(n) = \Sigma(n)^{-1}$; for details see Lauritzen (1996).

Theorem 2 *Assume the conditions of Theorem 1. Moreover, assume that there exists a constant* $\delta > 0$ *so that for every* $a \in \Gamma(n)$, $b \in \text{ne}_a$, *and* $n \in \mathbb{N}$, $|\pi_{ab}| \geq \delta n^{-\frac{1}{2}+\xi}$ *with some* $\max\{\varepsilon, \kappa/2\} < \xi < 1/2$ *(*ε *and* κ *as in Theorem 1). Then there exists some constant* $c > 0$ *so that*

$$P\Big(\text{ne}_a \neq \hat{\text{ne}}_a^\lambda\Big) = O(\exp(-cn^\varepsilon)) \qquad \textit{for } n \to \infty.$$

A proof is given in the appendix.

In summary, Theorems 1 and 2 show that the neighbourhood of any variable in a sparse high-dimensional graph can be estimated consistently with the Lasso.

3 Covariance Selection

It follows from section 2 that it is possible under certain conditions to estimate the neighbourhood of each node in the graph consistently, e.g.

$$P(\hat{\text{ne}}_a^\lambda = \text{ne}_a) \to 1 \qquad \text{for } n \to \infty.$$

The full graph is given by the set $\Gamma(n)$ of nodes and the edge set $E = E(n)$. The edge set contains those pairs $(a, b) \in \Gamma(n) \times \Gamma(n)$, for which the partial correlation between X_a and X_b is not zero. As the partial correlations are precisely non-zero for neighbours, the edge set $E \subseteq \Gamma(n) \times \Gamma(n)$ is given by

$$E = \{(a, b) : a \in \text{ne}_b \wedge b \in \text{ne}_a\}.$$

The first condition, $a \in \text{ne}_b$, implies in fact the second, $b \in \text{ne}_a$, and vice versa, so that the edge is as well identical to $\{(a, b) : a \in \text{ne}_b \vee b \in \text{ne}_a\}$. For an estimate of the edge set of a graph, we can apply neighbourhood selection to each node in the graph. A natural estimate of the edge set is then given by $\hat{E}^{\lambda,\wedge} \subseteq \Gamma(n) \times \Gamma(n)$, where

$$\hat{E}^{\lambda,\wedge} = \{(a, b) : a \in \hat{\text{ne}}_b^\lambda \wedge b \in \hat{\text{ne}}_a^\lambda\}.$$

Note that $a \in \hat{\text{ne}}_b^\lambda$ does not necessarily imply $b \in \hat{\text{ne}}_a^\lambda$ and vice versa. We can also define a second, less conservative, estimate of the edge set by $\hat{E}^{\lambda,\vee} = \{(a, b) : a \in \hat{\text{ne}}_b^\lambda \vee b \in \hat{\text{ne}}_a^\lambda\}$. The edge set estimates $\hat{E}^{\lambda,\wedge}$ and $\hat{E}^{\lambda,\vee}$ tend to give very similar results in our experience. The following theoretical result about consistency holds true for either of these edge set estimates and we refer to both collectively with the generic notation $\hat{E}^\lambda$.

Corrolary 1 *Under the conditions of Theorem 2, there exists some constant $c > 0$ so that*

$$P(\hat{E}^\lambda \neq E) = O(\exp(-cn^\varepsilon)) \qquad \textit{for } n \to \infty.$$

Proof: The result follows since $|\Gamma(n)|^2 = p_n^2 = O(n^{2\gamma})$ and neighbourhood selection has an exponentially fast convergence rate as described by Theorem 2.
Corrolary 1 says that the conditional independence structure of a multivariate normal distribution can be estimated consistently by combining the neighbourhood estimates for all variables. The procedure is moreover computationally efficient due to the convexity of the objective function.
Before providing some numerical results, we discuss in the following the choice of the penalty parameter.

Finite sample results and significance The previous results showed that consistent neighbourhood and covariance selection is possible with the Lasso in a high-dimensional setting. The asymptotic considerations give, however, little advice on how to choose a specific penalty parameter for a given problem.

Ideally, one would like to guarantee that pairs of variables which are not contained in the edge set enter the estimate of the edge set only with very low (pre-specified) probability. Unfortunately, it seems very difficult to obtain such a result as the probability of falsely including a pair of variables into the estimate of the edge set depends on the exact covariance matrix, which is in general unknown.

It is possible, however, to constrain the probability of (falsely) connecting two distinct connectivity components of the true graph. The connectivity component $C_a \subseteq \Gamma(n)$ of a node $a \in \Gamma(n)$ is the set of nodes which are connected to node a by a chain of edges. The neighbourhood ne_a is clearly part of the connectivity component C_a.

Let $\hat{C}_a^\lambda$ be the connectivity component of a in the estimated graph $(\Gamma, \hat{E}^\lambda)$. For any level $0 < \alpha < 1$, consider the choice of the penalty

$$\lambda(\alpha) = \frac{\hat{\sigma}_a}{\sqrt{n}} \tilde{\Phi}^{-1}\left(\frac{\alpha}{2p_n^2}\right), \tag{3}$$

where $\tilde{\Phi} = 1 - \Phi$ (Φ the c.d.f. of $\mathcal{N}(0,1)$) and $\hat{\sigma}_a^2 = \frac{1}{n}\sum_{i=1}^n (X_a^{(i)})^2$. The probability of falsely joining two distinct connectivity components with the estimate of the edge set is limited by the level α under the choice $\lambda = \lambda(\alpha)$ of the penalty parameter, as shown in the following theorem.

Theorem 3 *Under assumptions A2 and A3, using penalty parameter $\lambda(\alpha)$, it holds for all $n \in \mathbb{N}$ that*

$$P(\textit{there exists } \ a \in \Gamma(n) : \hat{C}_a^\lambda \not\subseteq C_a) \leq \alpha.$$

A proof is given in the appendix. This implies that if the edge set is empty, $E = \emptyset$, it is estimated by an empty set with high probability,

$$P(\hat{E}^\lambda = \emptyset) \geq 1 - \alpha.$$

Note that Theorem 3 is a finite sample result. The previous asymptotic results in Theorem 1 and 2 hold true if the level α is vanishing exponentially to zero for an increasing number of observations, leading to consistent edge set estimation. The above consideration offers a principled choice of the penalty parameter and allows a meaningful interpretation of the obtained results.

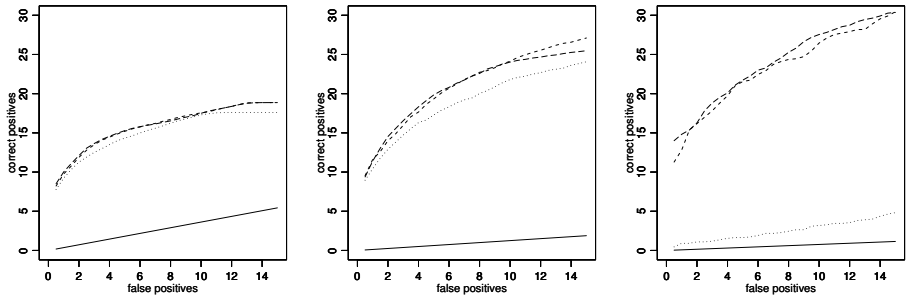

Figure 1: The average number of correct positives as a function of the number of false positives. The Lasso neighbourhood selection schemes (short dashed line for $\hat{E}^{\lambda,\wedge}$ and long dashed line for $\hat{E}^{\lambda,\vee}$), the forward selection MLE (dotted line) and finally, for comparison, the random guess solution (solid line) for $n = 40$ observations and $10, 20$ and 30 variables (from left to right).

4 Numerical examples

We use both the Lasso estimate from section 3 and forward selection MLE (Lauritzen 1996; Edwards 2000) to estimate sparse graphs.

We found it difficult to compare numerically neighbourhood selection with forward selection MLE for more than, say, thirty nodes in the graph. The high computational complexity of the forward selection MLE made the computations for such relatively low-dimensional problems very costly already. The Lasso scheme in contrast handled with ease graphs with more than 1000 nodes, using the recent algorithm developed in Efron et al. (2004).

Where comparison was feasible, the performance of the neighbourhood selection scheme was better. The difference was particularly pronounced if the ratio between observations to variables was low, as can be seen in Figure 1, which will be described in more detail below.

First we give an account of the generation of the underlying graphs, which we are trying to estimate. A realization of an underlying (random) graph is given in the left panel of Figure 2. The nodes of the graph are associated with a spatial location and the location of each node is distributed identically and uniformly on the two-dimensional square $[0, 1]^2$. Every pair of nodes is included initially in the edge set with a probability of $\varphi(d/\sqrt{p})$, where d is the Euclidean distance between the pair of variables and φ the density of the standard normal distribution. The maximum number of edges connecting to each node is limited to four to achieve the desired sparsity of the graph. Edges which connect to nodes which do not fulfill this constraint are removed randomly until the constraint is fulfilled

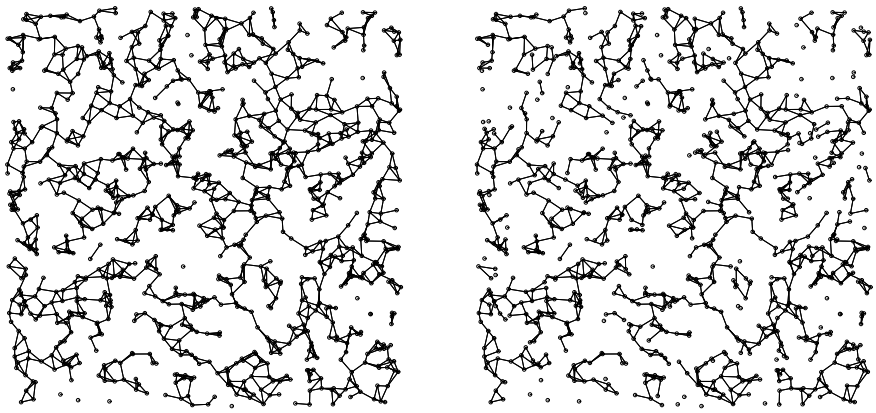

Figure 2: A realization of a graph with 1000 nodes, generated as described in the text, is shown in the right panel. The graph consists of 1000 nodes and 2256 edges out of a possible 449500 distinct pairs of variables. The estimated edge set $\hat{E}^{\lambda,\wedge}$, using (3) at level $\alpha = .05$, is shown in the left panel. The estimated edge set contains 1836 pairs of variables of which only 4 are falsely included. Not a single pair of disjoint connectivity components of the true graph has been (falsely) joined.

for all edges.

Initially all variables have identical conditional variance and the partial correlation between neighbours is set to 0.245 (absolute values less than 0.25 guarantee positive definiteness of the inverse covariance matrix), that is $\Sigma^{-1}_{aa} = 1$ for all nodes $a \in \Gamma$, $\Sigma^{-1}_{ab} = 0.245$ if there is an edge connecting a and b, $\Sigma^{-1}_{ab} = 0$ otherwise. Variables are then re-scaled so that the (unconditional) variance of each variable is identical to 1.

The average number of edges which are correctly included into the estimate of the edge set is shown in Figure 1 as a function of the number of edges which are falsely included. The accuracy of the forward selection MLE is comparable (or slightly worse) to the proposed Lasso neighbourhood selection if the number of nodes is much smaller than the number of observations. The accuracy of the forward selection MLE breaks down, however, if the number of nodes is comparable with the number of observations. This can be observed in the right panel of Figure 1, where forward selection MLE is only marginally better than random guessing. Computation of the forward selection MLE (using MIM, Edwards 2000) took on the same desktop up to several hundred times longer than the Lasso neighbourhood selection for the full graph. For more than 30 nodes, the differences are even more pronounced.

The Lasso neighbourhood selection can be applied to hundred- or thousand-dimensional

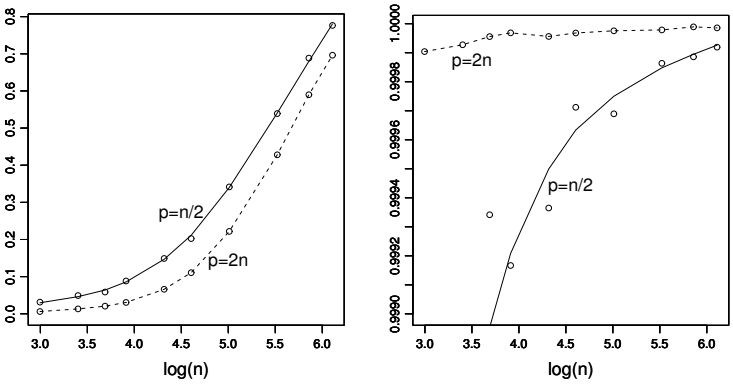

Figure 3: The percentage of all edges in E which are correctly included in $\hat{E}^{\lambda,\wedge}$ (left panel) for $p_n = 2n$ (broken line) and $p_n = n/2$ (solid line) variables. The percentage of edges not in E which are correctly not included in $\hat{E}^{\lambda,\wedge}$ (right panel). The absolute number of correct negatives is smaller for $p_n = 2n$ than for $p_n = n/2$ but the number of possible edges is vastly greater for the first case, leading to better relative proportion of correct negatives.

graphs, a realistic size for e.g. biological networks. A graph with 1000 nodes (following the same model as described above) and its estimate (with the Lasso neighbourhood selection using (3) at level $\alpha = .05$ and 500 observations) are shown in Figure 2. The average accuracy over 50 simulations is shown in the table below. Out of more than $4 \cdot 10^5$ pairs of variables, only about 5 are on average falsely included into the estimated edge set.

	$(a,b) \in \hat{E}^{\lambda,\wedge}$	$(a,b) \notin \hat{E}^{\lambda,\wedge}$	
$(a,b) \in E$	1459.5	509.5	1969
$(a,b) \notin E$	5.1	497525.9	497531
	1464.6	498035.4	499500

Next, the number of observations is varied together with the number of nodes in the graph. The proportion of all pairs of variables in the true edge set which are (correctly) included into the estimate of the edge set is shown in Figure 3 (averaged over 100 simulations). The proportion of all pairs of variables not included in the true edge set which are (correctly) not included into the estimated edge set is shown on the right side of the same figure. The estimates of the edge set are clearly increasing in accuracy with the number of observations, even though the number of observations per variable remains constant.
The result illustrates that the quality of neighbourhood selection with the Lasso is improving for an increasing number of observations even if the number of observations per variable is remaining constant, as expected from Theorems 1 and 2.

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22*, 203.

Breiman, L. and D. Freedman (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association 78*, 131–136.

Buhl, S. (1993). On the existence of maximum-likelihood estimators for graphical gaussian models. *Scandinavian Journal of Statistics 20*, 263–270.

Chen, S., S. Donoho, and M. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM Review 43*, 129–159.

Dempster, A. (1972). Covariance selection. *Biometrics 28*, 157–175.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *to appear in Annals of Statistics*.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association 95*, 1304–1308.

Goldenshluger, A. and A. Tsybakov (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Annals of Statistics 29*, 1601–1619.

Greenshtein, E. and Y. Ritov (2003). Persistency in high-dimensional predictor selection and the virtue of over-parametrization. Technical report, University of Haifa and Hebrew University.

Huber, P. (1973). Robust regression: asymptotics, conjectures, and monte carlo. *Annals of Statistics 1*, 799–821.

Judistky, A. and A. Nemirovski (2000). Functional aggregation for nonparametric regression. *Annals of Statistics 28*, 681–712.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Lienhart, H. and W. Zuchini (1986). *Model Selection*. Wiley, New York.

Osborne, M., B. Presnell, and B. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*, 319–337.

Portnoy, S. (1984). Asymptotic behavior of m-estimators of p regression paramateres if p^2/n is large. *Annals of Statistics 12*, 1298–1309.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics 14*, 1080–1100.

Schwarz, G. (1978). Estimating dimensions of a model. *Annals of Statistics 6*, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*, 486–494.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*, 45–54.

Speed, T. and H. Kiiveri (1986). Gaussian markov distributions over finite graphs. *Annals of Statistics 14*, 138–150.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Toh, H. and K. Horimoto (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics 18*, 287–297.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag.

5 Appendix

Definition 1 *As a generalization of (2), consider optimal prediction of X_a, given only a subset of variables $\{X_k; k \in \Psi\}$, where $\Psi \subseteq \Gamma(n)\backslash\{a\}$. The Lasso estimate $\hat{\theta}^{a,\Psi,\lambda}$ of $\theta^{a,\Psi}$ is given by*

$$\hat{\theta}^{a,\Psi,\lambda} = \arg \min_{\theta:\theta_k=0 \ \forall k \notin \Psi} \Big(\frac{1}{n} \sum_{i=1}^{n} (X_a^{(i)} - \sum_{k \in \Gamma(n)} \theta_k X_k^{(i)})^2 + \lambda ||\theta||_1 \Big). \tag{4}$$

The notation $\hat{\theta}^{a,\lambda}$ is thus just a shorthand for $\hat{\theta}^{a,\Gamma(n)\backslash\{a\},\lambda}$.

Lemma 1 *Let $G(\theta)$ be a p_n-dimensional vector with elements $G_b(\theta) = -\frac{2}{n} \sum_{i=1}^{n} (X_a^{(i)} - \sum_{k \in \Gamma(n)} \theta_k X_k^{(i)}) X_b^{(i)}$. A vector θ with $\theta_k = 0 \ \forall k \in \Gamma(n)\backslash\Psi$ is a solution to (4) iff for all $b \in \Psi$, $G_b(\theta) = \text{sign}(\theta_b)\lambda$ in case $\theta_b \neq 0$ and $|G_b(\theta)| \leq \lambda$ in case $\theta_b = 0$. Moreover, if the solution is not unique and $|G_b(\theta)| < \lambda$ for some solution θ, then $\theta_b = 0$ for all solutions of (4).*

Proof of Lemma 1 Denote the subdifferential of $\frac{1}{n} \sum_{i=1}^{n} (X_a^{(i)} - \sum_{k \in \Gamma(n)} \theta_k X_k^{(i)})^2 + \lambda ||\theta||_1$ with respect to θ by $D(\theta)$. The vector $\tilde{\theta}$ is a solution to (4) iff there exists an element

$d \in D(\tilde{\theta})$ so that $d_b = 0$, $\forall b \in \Psi$. $D(\theta)$ is given by $\{G(\theta) + \lambda e, e \in S\}$, where $S \subset \mathbb{R}^{p_n}$ is given by $S := \{e \in \mathbb{R}^{p_n} : e_b = \text{sign}(\theta_b)$ if $\theta_b \neq 0$ and $e_b \in [-1, 1]$ if $\theta_b = 0\}$. The first part of the claim follows. The second part follows from the proof of Theorem 3.1. in Osborne et al. (2000).

Proof of Theorem 1 The event $\hat{\text{ne}}_a^\lambda \not\subseteq \text{ne}_a$ implies that there exists some node $b \in \Gamma(n)\backslash\text{ne}_a$ in the set of non-neighbours of node a such that the estimated coefficient $\hat{\theta}_b^{a,\lambda}$ is not zero. Thus

$$P(\hat{\text{ne}}_a^\lambda \not\subseteq \text{ne}_a) \leq P(\exists b \in \Gamma(n)\backslash\text{ne}_a : \hat{\theta}_b^{a,\lambda} \neq 0). \tag{5}$$

Consider the Lasso estimate $\hat{\theta}^{a,\text{ne}_a,\lambda}$, which is by (4) constrained to have non-zero components only in the neighbourhood of node $a \in \Gamma(n)$. Using $|\text{ne}_a| = O(n^\kappa)$ with some $\kappa < 1$, we can assume w.l.o.g. that $|\text{ne}_a| \leq n$. This in turn implies, see e.g. Osborne et al. (2000), that $\hat{\theta}^{a,\text{ne}_a,\lambda}$ is a.s. a unique solution to (4) with $\Psi = \text{ne}_a$. Let $\mathcal{A}$ be the event

$$\max_{k \in \Gamma(n)\backslash\text{ne}_a} |G_k(\hat{\theta}^{a,\text{ne}_a,\lambda})| < \lambda.$$

Given the event $\mathcal{A}$, it follows from the first part of Lemma 1 that $\hat{\theta}^{a,\text{ne}_a,\lambda}$ is not only a solution of (4), with $\Psi = \text{ne}_a$, but as well a solution of (2). As $\hat{\theta}_b^{a,\text{ne}_a,\lambda} = 0$ for all $b \in \Gamma(n)\backslash\text{ne}_a$, it follows from the second part of Lemma 1, that $\hat{\theta}_b^{a,\lambda} = 0$, $\forall b \in \Gamma(n)\backslash\text{ne}_a$. Hence

$$\begin{aligned} P(\exists b \in \Gamma(n)\backslash\text{ne}_a : \hat{\theta}_b^{a,\lambda} \neq 0) &\leq 1 - P(\mathcal{A}) \\ &= P(\max_{k \in \Gamma(n)\backslash\text{ne}_a} |G_k(\hat{\theta}^{a,\text{ne}_a,\lambda})| \geq \lambda), \end{aligned}$$

where

$$G_b(\hat{\theta}^{a,\text{ne}_a,\lambda}) = -\frac{2}{n}\sum_{i=1}^{n}(X_a^{(i)} - \sum_{k \in \text{ne}_a} \hat{\theta}_k^{a,\text{ne}_a,\lambda} X_k^{(i)})X_b^{(i)}. \tag{6}$$

Using Bonferroni's inequality and $p_n = O(n^\gamma)$ for any $\gamma > 0$, it suffices to show that there exist constants $c, d > 0$ so that for all $b \in \Gamma(n)\backslash(\text{ne}_a \cup \{a\})$,

$$P(|G_b(\hat{\theta}^{a,\text{ne}_a,\lambda})| \geq \lambda) \leq d\exp(-cn^\varepsilon). \tag{7}$$

One can write for any $b \in \Gamma(n)\backslash(\text{ne}_a \cup \{a\})$,

$$X_b = \sum_{k \in \text{ne}_a} \theta_k^{b,\text{ne}_a} X_k + W_b, \tag{8}$$

where $W_b \sim \mathcal{N}(0, \sigma_b^2)$ for some $\omega^{-2} \leq \sigma_b^2 \leq 1$ and W_b is independent of $\{X_k;\ k \in \text{ne}_a \cup \{a\}\}$. By A4 it follows that for some $\vartheta_1 < 1$,

$$\sum_{k \in \text{ne}_a} |\theta_k^{b,\text{ne}_a}| \leq \vartheta_1. \tag{9}$$

Furthermore, by Lemma 1,

$$|\frac{2}{n}\sum_{i=1}^{n}(X_a^{(i)} - \sum_{m\in \mathrm{ne}_a} \hat{\theta}_m^{a,\mathrm{ne}_a,\lambda} X_m^{(i)})X_k^{(i)}| \leq \lambda \qquad \forall k \in \mathrm{ne}_a. \tag{10}$$

Using (8)-(10), the absolute value of the gradient G_b in equation (6) is hence bounded by

$$|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \leq \vartheta_1\lambda + |\frac{2}{n}\sum_{i=1}^{n}(X_a^{(i)} - \sum_{k\in \mathrm{ne}_a} \hat{\theta}_k^{a,\mathrm{ne}_a,\lambda} X_k^{(i)})W_b^{(i)}|. \tag{11}$$

Conditional on $X_{\mathrm{ne}_a\cup\{a\}}^{(1,\ldots,n)} = \{X_k^{(i)};\, k \in \mathrm{ne}_a \cup \{a\}, i = 1,\ldots,n\}$, the random variable $|\sum_{i=1}^{n}(X_a^{(i)} - \sum_{k\in\Gamma(n)} \hat{\theta}_k^{a,\mathrm{ne}_a,\lambda} X_k^{(i)})W_b^{(i)}|$ is normally distributed with mean zero and variance $\sigma_b^2 \sum_{i=1}^{n}(X_a^{(i)} - \sum_{k\in\mathrm{ne}_a} \hat{\theta}_k^{a,\mathrm{ne}_a,\lambda} X_k^{(i)})^2$. On the one hand, $\sigma_b^2 \leq 1$. On the other hand, by definition of $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$, $\sum_{i=1}^{n}(X_a^{(i)} - \sum_{k\in\mathrm{ne}_a} \hat{\theta}_k^{a,\mathrm{ne}_a,\lambda} X_k^{(i)})^2 \leq \sum_{i=1}^{n}(X_a^{(i)})^2$. Thus

$$|\frac{2}{n}\sum_{i=1}^{n}(X_a^{(i)} - \sum_{k\in \mathrm{ne}_a} \hat{\theta}_k^{a,\mathrm{ne}_a,\lambda} X_k^{(i)})W_b^{(i)}| \overset{st.}{\leq} |\frac{2}{n}\sum_{i=1}^{n} X_a^{(i)} W_b^{(i)}|,$$

where $\overset{st.}{\leq}$ denotes stochastically smaller or equal. Using (11), it follows that

$$P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \geq \lambda) \leq P(|\frac{2}{n}\sum_{i=1}^{n} X_a^{(i)} W_b^{(i)}| \geq (1-\vartheta_1)\lambda).$$

As W_b is independent of X_a, it follows that $E(X_a^{(i)} W_b^{(i)}) = 0$ for all $i \leq n$. Using the Gaussianity and bounded variance of both X_a and W_b, there exists some $g < \infty$ so that $E(\exp(|X_a^{(i)} W_b^{(i)}|)) \leq g$. Hence, using Bernstein's inequality, see e.g. Lemma 2.2.11 in van der Vaart and Wellner (1996), there exist indeed constants $c, d > 0$ so that for all $b \in \mathrm{ne}_a$, $P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \geq \lambda) \leq d\exp(-c(1-\vartheta_1)\sqrt{n}\lambda)$. The claim (7) follows, which completes the proof.

Proof of Theorem 2 First,

$$P(\mathrm{ne}_a = \hat{\mathrm{ne}}_a^{\lambda}) \geq 1 - P(\hat{\mathrm{ne}}_a^{\lambda} \not\subseteq \mathrm{ne}_a) - P(\mathrm{ne}_a \not\subseteq \hat{\mathrm{ne}}_a^{\lambda}).$$

It follows from Theorem 1 that the second term on the right hand side has the correct asymptotic behaviour and we can focus on the last term. The last term is identical to $P(\mathrm{ne}_a \not\subseteq \hat{\mathrm{ne}}_a^{\lambda}) = P(\exists b \in \mathrm{ne}_a :\ \hat{\theta}_b^{a,\lambda} = 0)$. Let $\mathcal{A}$ again be the event

$$\max_{k\in\Gamma(n)\setminus \mathrm{ne}_a} |G_k(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| < \lambda.$$

Given $\mathcal{A}$, we can conclude as in the proof of Theorem 1 that $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ and $\hat{\theta}^{a,\lambda}$ are unique solutions to (4) and (2) respectively, and $\hat{\theta}^{a,\mathrm{ne}_a,\lambda} = \hat{\theta}^{a,\lambda}$. Thus

$$P(\exists b \in \mathrm{ne}_a :\ \hat{\theta}_b^{a,\lambda} = 0) \quad \leq \quad P(\exists b \in \mathrm{ne}_a :\ \hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0)P(\mathcal{A}) + P(\mathcal{A}^c)$$

It follows from the proof of Theorem 1 that there exists some $c > 0$ so that $P(\mathcal{A}^c) = O(\exp(-cn^\varepsilon))$. Using Bonferroni's inequality, it hence remains to show that there exist $c, d > 0$ so that for all $b \in \mathrm{ne}_a$,

$$P(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0) \leq d\exp(-cn^\varepsilon), \tag{12}$$

which is equivalent to, using Lemma 1,

$$P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda})| \leq \lambda) \leq d\exp(-cn^\varepsilon) \qquad \forall b \in \mathrm{ne}_a. \tag{13}$$

We summarise briefly the calculations yielding the equivalence of (12) and (13). If the absolute value of the gradient $G_b(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda})$ is larger than λ, it follows by Lemma 1 that $\hat{\theta}^{a,\mathrm{ne}_a,\lambda} \neq \hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda}$. However, as $\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0$ would imply the equality $\hat{\theta}^{a,\mathrm{ne}_a,\lambda} = \hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda}$, it follows indeed that $\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} \neq 0$ as long as $|G_b(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda})| > \lambda$. Given the event $\mathcal{A}$, this implies that also $\hat{\theta}_b^{a,\lambda} \neq 0$, and it is hence sufficient to show (13).
We can write X_b as

$$X_b = \sum_{k\in\mathrm{ne}_a\setminus\{b\}} \theta_k^{b,\mathrm{ne}_a\setminus\{b\}} X_k + W_b, \tag{14}$$

where W_b is independent of $\{X_k;\, k \in \mathrm{ne}_a\setminus\{b\}\}$. Let in the following $R_a^{\lambda,(i)}$ be the residual

$$R_a^{\lambda,(i)} = X_a^{(i)} - \sum_{k\in\mathrm{ne}_a\setminus\{b\}} \hat{\theta}_k^{a,\mathrm{ne}_a\setminus\{b\},\lambda} X_k^{(i)} \qquad i = 1,\ldots,n.$$

Then, by straightforward calculation using (14)

$$G_b(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda}) = -\sum_{k\in\mathrm{ne}_a\setminus\{b\}} \theta_k^{b,\mathrm{ne}_a\setminus\{b\}}\Big(\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} X_k^{(i)}\Big) - \frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i)}. \tag{15}$$

By Lemma 1, for all $k \in \mathrm{ne}_a\setminus\{b\}$, $|G_k(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda})| = |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} X_k^{(i)}| \leq \lambda$. This together with (15) yields

$$|G_b(\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda})| \geq |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i)}| - \lambda \sum_{k\in\mathrm{ne}_a\setminus\{b\}} |\theta_k^{b,\mathrm{ne}_a\setminus\{b\}}|.$$

Using A4, there exists some $\vartheta_2 < \infty$ so that $\sum_{k\in\mathrm{ne}_a} |\theta_k^{a,\mathrm{ne}_a\setminus\{b\}}| \leq \vartheta_2$ and for proving (13) it is therefore sufficient to show that there exist $c, d > 0$ so that

$$P\Big(|\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i)}| \leq (\vartheta_2 + 1)\lambda\Big) \leq d\exp(-cn^\varepsilon) \qquad \forall b \in \mathrm{ne}_a. \tag{16}$$

Consider any $\overline{\kappa}$ with $\max\{\varepsilon, \kappa/2\} < \overline{\kappa}/2 < \xi$. To show (16), it is sufficient to prove that there exist $c, d > 0$ for any $g > 0$ so that

$$P\Big(|\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i)}| \leq g n^{-\frac{1-\overline{\kappa}}{2}}\Big) \leq d\exp(-cn^\varepsilon) \qquad \forall b \in \mathrm{ne}_a. \tag{17}$$

It holds for some random variable W_a, independent of X_{ne_a}, that

$$X_a = \sum_{k\in \mathrm{ne}_a} \theta_k^{a,\mathrm{ne}_a} X_k + W_a = \sum_{k\in \mathrm{ne}_a\setminus\{b\}} (\theta_k^{a,\mathrm{ne}_a} + \theta_b^{a,\mathrm{ne}_a}\theta_k^{b,\mathrm{ne}_a\setminus\{b\}})X_k + \theta_b^{a,\mathrm{ne}_a} W_b + W_a, \qquad (18)$$

having used (8). Note that W_a and W_b are independent normal distributed random variables with variances σ_b^2 and σ_a^2 respectively. By A3, $\omega^{-2} < \sigma_b^2, \sigma_a^2 \leq 1$.
Note that W_b is independent of $X_{\mathrm{ne}_a\setminus\{b\}}$ and W_a but not necessarily of X_a and $\hat{\theta}^{a,\mathrm{ne}_a\setminus\{b\},\lambda}$. With a little abuse of notation, split the n-dimensional vector of observations $(W_b^{(1)}, W_b^{(2)}, \ldots, W_b^{(n)})$, into the sum of two vectors $(W_b^{(1),\perp}, W_b^{(2),\perp}, \ldots, W_b^{(n),\perp})$ and $(W_b^{(1),\|}, W_b^{(2),\|}, \ldots, W_b^{(n),\|})$ where the latter vector is contained in the (at most $|\mathrm{ne}_a\setminus\{b\}|$-dimensional) space $V^{\|} \subseteq \mathbb{R}^n$, which is spanned by the vectors $(X_k^{(1)}, X_k^{(2)}, \ldots, X_k^{(n)})$ for all $k \in \mathrm{ne}_a\setminus\{b\}$. The remaining part $W_b^{\perp}$ is chosen orthogonal to this space (in the orthogonal complement of $V^{\|}$). The same notation is adopted for the residuals R_a^{λ}. Then, using the orthogonality property of $W_b^{(i),\perp}$,

$$\begin{aligned} |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i)}| &\geq |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i),\perp}| - |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i),\|}| \\ &\geq |\frac{2}{n}\sum_{i=1}^n \theta_b^{a,\mathrm{ne}_a}(W_b^{(i),\perp})^2| - |\frac{2}{n}\sum_{i=1}^n W_a^{(i)} W_b^{(i),\perp}| - |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i),\|}| \end{aligned} \qquad (19)$$

Consider the third term. By basic algebra,

$$|\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i)} W_b^{(i),\|}| = |\frac{2}{n}\sum_{i=1}^n R_a^{\lambda,(i),\|} W_b^{(i),\|}| \leq \frac{2}{n}\Big(\sum_{i=1}^n (R_a^{\lambda,(i),\|})^2\Big)^{\frac{1}{2}}\Big(\sum_{i=1}^n (W_b^{(i),\|})^2\Big)^{\frac{1}{2}}. \qquad (20)$$

The sum of squares of the residuals is increasing with increasing value of λ. Thus

$$\sum_{i=1}^n (R_a^{\lambda,(i),\|})^2 \leq \sum_{i=1}^n (X_a^{(i)})^2.$$

The residual sum of squares is hence stochastically smaller than a $\chi^2(n)$-distributed random variable. For the second term on the r.h.s. of (20), the expression $\sigma_b^{-2}\sum_{i=1}^n (W_b^{(i),\|})^2$ is $\chi^2(|\mathrm{ne}_a| - 1)$-distributed, which is stochastically smaller than a $\chi^2(|\mathrm{ne}_a|)$-distributed random variable. As $|\mathrm{ne}_a| = O(n^\kappa)$ and $\xi > \max\{\varepsilon, \kappa/2\}$, we can choose some $t^2 > 0$ and some $\overline{\kappa}$ with $\max\{\varepsilon, \kappa/2\} < \overline{\kappa}/2 < \xi$ so that $|\mathrm{ne}_a| \leq t^2 n^{\overline{\kappa}}$. The right hand side of (20) is then stochastically smaller than

$$2tn^{-\frac{1-\overline{\kappa}}{2}}(Z_1 Z_2)^{\frac{1}{2}},$$

where $Z_1 \sim \chi^2(n)/n$ and $Z_2 \sim \chi^2(t^2 n^{\overline{\kappa}})/(t^2 n^{\overline{\kappa}})$. Note that Z_1 and Z_2 are not necessarily independent. However, using Bonferroni's inequality, the properties of the χ^2-distribution, and using $\overline{\kappa}/2 > \varepsilon$, it follows for the third term on the right hand side of (19) that there

exist $c, d > 0$ so that

$$P(|\frac{2}{n}\sum_{i=1}^{n} R_a^{\lambda,(i)} W_b^{(i),\|}| \geq 4tn^{-\frac{1-\bar{\kappa}}{2}}) \leq d\exp(-cn^{\varepsilon}). \tag{21}$$

For the middle term on the right hand side of (19), due to Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner 1996), there exists $c, d > 0$ so that

$$P(|\frac{1}{n}\sum_{i=1}^{n} W_a^{(i)} W_b^{(i),\perp}| \geq n^{-\frac{1-\bar{\kappa}}{2}}) \leq d\exp(-cn^{\frac{\bar{\kappa}}{2}}). \tag{22}$$

To show (16), and hence complete the proof, it thus suffices by (21) and (22) to show that there exist $c, d > 0$ (possibly different from the one above) so that for the first term on the right hand side of (19), for any constant $g > 0$,

$$P(|\frac{2}{n}\sum_{i=1}^{n} \theta_b^{a,\mathrm{ne}_a}(W_b^{(i),\perp})^2| \leq gn^{-\frac{1-\bar{\kappa}}{2}}) \leq d\exp(-cn^{\varepsilon}). \tag{23}$$

By assumption, $|\pi_{ab}|$ is of order at least $n^{-\frac{1}{2}+\xi}$. Using additionally A2 and A3, this implies that there exist some $q > 0$ so that $|\theta_b^{a,\mathrm{ne}_a}| \geq qn^{-\frac{1}{2}+\xi}$ (using $\theta^a = \theta^{a,\mathrm{ne}_a}$). The term $\sum_{i=1}^{n}(W_b^{(i),\perp})^2$ follows a $\chi^2(n - |\mathrm{ne}_a|)$-distribution. As $|\mathrm{ne}_a| = O(n^{\kappa})$ with $\kappa < 1$, it follows for the term on the left hand side of (23), for $n \geq n_0$ with some $n_0 \in \mathbb{N}$, that

$$|\frac{2}{n}\sum_{i=1}^{n} \theta_b^{a,\mathrm{ne}_a}(W_b^{(i),\perp})^2| \overset{st.}{\geq} kn^{-\frac{3}{2}+\xi} Z_3,$$

where k is some positive constant, Z_3 is a $\chi^2(n/2)$-distributed random variable, and $\overset{st.}{\geq}$ denotes stochastically larger or equal. Thus, for some constant $c > 0$,

$$P(|\frac{2}{n}\sum_{i=1}^{n} \theta_b^{a,\mathrm{ne}_a}(W_b^{(i),\perp})^2| \leq gn^{-\frac{1-\bar{\kappa}}{2}}) \leq P(\frac{Z_3}{n/2} \leq cn^{\frac{\bar{\kappa}}{2}-\xi}).$$

from which the claim (23) follows by the properties of the χ^2-distribution as $\bar{\kappa}/2 < \xi$. This in turns shows that (17) holds and completes the proof.

Proof of Proposition 1 All diagonal elements of the covariance matrix $\Sigma(n)$ are equal to 1, while all off-diagonal elements vanish for all pairs except for $a, b \in \Gamma(n)$, where $\Sigma_{ab}(n) = s$ with $s > 0$. Assume w.l.o.g. that a corresponds to the first and b to the second variable. The best vector of coefficients θ^a for linear prediction of X_a is given by $\theta^a = (0, -K_{ab}(n)/K_{aa}(n), 0, 0, \ldots) = (0, s, 0, 0, \ldots)$, where $K(n) = \Sigma(n)^{-1}$. A necessary condition for $\hat{\mathrm{ne}}_a^{\lambda} = \mathrm{ne}_a$ is, that $\hat{\theta}^{a,\lambda} = (0, \tau, 0, 0, \ldots)$ is the oracle Lasso solution for some $\tau \neq 0$. In the following, we show first, that

$$P(\exists \lambda, \tau \geq s : \hat{\theta}^{a,\lambda} = (0, \tau, 0, 0, \ldots)) \to 0 \qquad n \to \infty. \tag{24}$$

The proof is then completed by showing in addition that $(0, \tau, 0, 0, \ldots)$ cannot be the *oracle* Lasso solution as long as $\tau < s$.

We begin by showing (24). If $\hat{\theta} = (0, \tau, 0, 0, \ldots)$ is a Lasso solution for some value of the penalty, it follows that, using Lemma 1 and positivity of τ,

$$\frac{1}{n}\sum_{i=1}^{n}(X_1^{(i)} - \tau X_2^{(i)})X_2^{(i)} \geq |\frac{1}{n}\sum_{i=1}^{n}(X_1^{(i)} - \tau X_2^{(i)})X_k^{(i)})| \qquad \forall k \in \Gamma(n), k > 2. \tag{25}$$

It is assumed for notational simplicity only that $\frac{1}{n}\sum_{i=1}^{n}(X_k^{(i)})^2 = 1$, for all $k \in \Gamma(n)$. This is w.l.o.g. due to the Bernstein-type exponential inequality. Under the made assumptions, $X_2, X_3, \ldots$ can be understood to be independently and identical distributed, while $X_1 = sX_2 + W_1$, with W_1 independent of $(X_2, X_3, \ldots)$. Substituting $X_1 = sX_2 + W_1$ in (25) yields

$$\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_2^{(i)} \geq |\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_k^{(i)} - (\tau - s)\frac{1}{n}\sum_{i=1}^{n} X_2^{(i)}X_k^{(i)}| + (\tau - s) \qquad \forall k \in \Gamma(n), k > 2.$$

The condition $\tau \geq s$ implies

$$\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_2^{(i)} = \max_{k\in\Gamma(n),k\geq 2}\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_k^{(i)}.$$

Let $U_2, U_3, \ldots, U_{p_n}$ be the random variables defined by $U_k = \sum_{i=1}^{n} W_1^{(i)}X_k^{(i)}$. As the random variables U_k, $k = 2, \ldots, p_n$ are exchangeable,

$$P(\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_2^{(i)} = \max_{k\in\Gamma(n),k\geq 2}\frac{1}{n}\sum_{i=1}^{n} W_1^{(i)}X_k^{(i)}) = (p_n - 1)^{-1}$$

and the claim (24) follows as $p_n \to \infty$ for $n \to \infty$. It hence suffices to show that $(0, \tau, 0, 0, \ldots)$ with $\tau < s$ cannot be the *oracle* Lasso solution. Let $\tau_{\max}$ be the maximal value of τ so that $(0, \tau, 0, \ldots)$ is a Lasso solution for some value $\lambda > 0$. By the previous assumption, $\tau_{\max} < s$. In this case, $(0, \tau, 0, \ldots)$ cannot be the oracle Lasso solution if $\tau < \tau_{\max}$. We show in the following that $(0, \tau_{\max}, 0, \ldots)$ can not be an oracle Lasso solution either.

Let $(0, \tau_{\max}, 0, 0, \ldots)$ be the Lasso solution $\hat{\theta}^{a,\lambda}$ for some $\lambda = \tilde{\lambda} > 0$. By appropriately reordering the variables $X_3, \ldots, X_{p_n}$, Lasso solutions for values $\lambda < \tilde{\lambda}$ are given by $\hat{\theta}^{a,\lambda} = (0, (\tau_{\max} + \delta), \pm\delta, 0, 0, \ldots)$ for any $0 < \delta < \delta_{\max}$ with some $\delta_{\max} > 0$. The sign of the third coefficient is equal to the sign of $\sum_{i=1}^{n}(X_1^{(i)} - \tau_{\max} X_2^{(i)})X_3^{(i)}$. Denote by L_δ the squared error loss for this solution. Then, for any $\delta \leq \delta_{\max}$.

$$\begin{aligned} L_\delta - L_0 &= E(X_1 - (\tau_{\max} + \delta)X_2 + \delta X_3)^2 - E(X_1 - \tau_{\max} X_2)^2, \\ &= (s - (\tau_{\max} + \delta))^2 + \delta^2 - (s - \tau_{\max})^2, \\ &= -2(s - \tau_{\max})\delta + 2\delta^2. \end{aligned}$$

It holds that $L_\delta - L_0 < 0$ for any $0 < \delta < \frac{1}{2}(s - \tau_{\max})$, which shows that $(0, \tau, 0, \ldots)$ cannot be the oracle solution for $\tau < s$. Together with (24), this completes the proof.

Proof of Theorem 3 A necessary condition for $\hat{C}_a^\lambda \not\subseteq C_a$) is that there exists an edge in $\hat{E}^\lambda$ joining two nodes in two different connectivity components. Hence

$$P(\exists a \in \Gamma(n) : \hat{C}_a^\lambda \not\subseteq C_a) \leq p_n \max_{a \in \Gamma(n)} P(\exists b \in \Gamma(n) \backslash C_a : b \in \hat{\text{ne}}_a^\lambda).$$

Using the same arguments as in the proof of Theorem 1,

$$P(\exists b \in \Gamma(n) \backslash C_a : b \in \hat{\text{ne}}_a^\lambda) \leq P(\max_{b \in \Gamma(n) \backslash C_a} |G_b(\hat{\theta}^{a, C_a, \lambda})| \geq \lambda),$$

where $\hat{\theta}^{a, C_a, \lambda}$, according to (4), has non-zero components only for variables in the connectivity component C_a of node a. Hence it is sufficient to show that

$$p_n^2 \max_{a \in \Gamma(n), b \in \Gamma(n) \backslash C_a} P(|G_b(\hat{\theta}^{a, C_a, \lambda})| \geq \lambda) \leq \alpha, \tag{26}$$

The gradient is given by $G_b(\hat{\theta}^{a, C_a, \lambda}) = -\frac{1}{n} \sum_{i=1}^n (X_a^{(i)} - \sum_{k \in C_a} \hat{\theta}_k^{a, C_a, \lambda} X_k^{(i)}) X_b^{(i)}$. It holds for all $k \in C_a$ that the variables X_b and X_k are independent as they are in different connectivity components. Hence, conditional on $X_{C_a}^{(1,\ldots,n)} = \{X_k^{(i)}; k \in C_a, i = 1, \ldots, n\}$, $G_b(\hat{\theta}^{a, C_a, \lambda}) \sim \mathcal{N}(0, R^2/n)$, where $R^2 = \frac{1}{n} \sum_{i=1}^n (X_a^{(i)} - \sum_{k \in C_a} \hat{\theta}_k^{a, C_a, \lambda} X_k^{(i)})^2$, which is smaller than or equal to $\hat{\sigma}_a^2 = \frac{1}{n} \sum_{i=1}^n (X_a^{(i)})^2$ by definition of $\hat{\theta}^{a, C_a, \lambda}$. Hence it holds for all $a \in \Gamma(n)$ and $b \in \Gamma(n) \backslash C_a$ that $P(|G_b(\hat{\theta}^{a, C_a, \lambda})| \geq \lambda | X_{C_a}^{(1,\ldots,n)}) \leq 2\tilde{\Phi}(\frac{\sqrt{n}}{\hat{\sigma}_a} \lambda)$, where $\tilde{\Phi} = 1 - \Phi$. Using the proposed $\lambda = \frac{\hat{\sigma}_a}{\sqrt{n}} \tilde{\Phi}^{-1}(\frac{\alpha}{2p_n^2})$, it follows that $P(|G_b(\hat{\theta}^{a, C_a, \lambda})| \geq \lambda | X_{C_a}^{(1,\ldots,n)}) \leq \alpha p_n^{-2}$, and therefore $P(|G_b(\hat{\theta}^{a, C_a, \lambda})| \geq \lambda) \leq \alpha p_n^{-2}$. Thus (26) follows, which completes the proof.

Lemma 2 *Let $K(n) = \Sigma(n)^{-1}$. If $\sum_{a \in \Gamma(n) \backslash \{a\}} |K_{ab}(n)| \leq \vartheta_1 |K_{aa}(n)|$, with $\vartheta_1 < 1$, $\forall a \in \Gamma(n)$, $\forall n \in \mathbb{N}$, then, for all $\Psi \subseteq \Gamma(n) \backslash \{a\}$, and all $n \in \mathbb{N}$,*

$$\sum_{k \in \Psi} |\theta_k^{a, \Psi}| \leq \vartheta_1. \tag{27}$$

Proof of Lemma 2 In the following, let for every subset $\Psi \subseteq \Gamma(n) \backslash \{a\}$, K^Ψ be the inverse of the submatrix $(\Sigma_{mk})_{m,k \in \Psi \cup \{a\}}$. Using $\theta_k^{a, \Psi} = -K_{ak}^\Psi(n) / K_{aa}^\Psi(n)$, the assumption $\sum_{a \in \Gamma(n) \backslash \{a\}} |K_{ab}(n)| \leq \vartheta_1 |K_{aa}(n)|$, $\forall a \in \Gamma(n)$ and $n \in \mathbb{N}$, is equivalent to $\sum_{k \in \Gamma(n) \backslash \{a\}} |\theta_k^{a, \Gamma(n) \backslash \{a\}}| \leq \vartheta_1$, so that (27) holds for $\Psi = \Gamma \backslash \{a\}$. By induction, it suffices to show that the claim holds then for $\Psi = \Gamma(n) \backslash \{a, b\}$ with any $b \in \Gamma(n) \backslash \{a\}$. Note that the optimal predictor of X_a (in the least-squares sense), given $X_{\Gamma(n) \backslash \{a\}}$, is

$$\sum_{k \in \Gamma(n) \backslash \{a\}} \theta_k^{a, \Gamma(n) \backslash \{a\}} X_k,$$

while the optimal predictor of X_b, given $X_{\Gamma(n)\setminus\{a,b\}}$, is

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} \theta_k^{b,\Gamma(n)\setminus\{a,b\}} X_k.$$

By the properties of the multivariate normal distribution we conclude that the optimal predictor of X_a, given $X_{\Gamma(n)\setminus\{a,b\}}$, is

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} \theta_k^{a,\Gamma(n)\setminus\{a,b\}} X_k = \sum_{k\in\Gamma(n)\setminus\{a,b\}} (\theta_k^{a,\Gamma(n)\setminus\{a\}} + \theta_b^{a,\Gamma(n)\setminus\{a\}} \theta_k^{b,\Gamma(n)\setminus\{a,b\}}) X_k,$$

Thus $\theta_k^{a,\Gamma\setminus\{a,b\}} = \theta_k^{a,\Gamma(n)\setminus\{a\}} + \theta_b^{a,\Gamma(n)\setminus\{a\}} \theta_k^{b,\Gamma(n)\setminus\{a,b\}}$ for all $k \in \Gamma(n)\setminus\{a,b\}$ and

$$\begin{aligned}\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| &\leq \sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a\}}| + |\theta_b^{a,\Gamma(n)\setminus\{a\}}| \sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{b,\Gamma(n)\setminus\{a,b\}}| \\ &= \sum_{k\in\Gamma(n)\setminus\{a\}} |\theta_k^{a,\Gamma(n)\setminus\{a\}}| + |\theta_b^{a,\Gamma(n)\setminus\{a\}}| \Big(\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{b,\Gamma(n)\setminus\{a,b\}}| - 1 \Big).\end{aligned}$$

Using $\sum_{k\in\Gamma(n)\setminus\{a\}} |\theta_k^{a,\Gamma(n)\setminus\{a\}}| \leq \vartheta_1$ and $\vartheta_1 < 1$, it follows that

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \leq |\theta_b^{a,\Gamma(n)\setminus\{a\}}| \Big(\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{b,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \Big). \tag{28}$$

By symmetry between a and b, it also holds that

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{b,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \leq |\theta_a^{b,\Gamma(n)\setminus\{b\}}| \Big(\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \Big). \tag{29}$$

Combining (28) and (29),

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \leq |\theta^{a,\Gamma(n)\setminus\{a\}}| |\theta_a^{b,\Gamma(n)\setminus\{b\}}| \Big(\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \Big).$$

As both $|\theta_b^{a,\Gamma(n)\setminus\{a\}}| \leq \sum_{k\in\Gamma(n)\setminus\{a\}} |\theta_k^{a,\Gamma(n)\setminus\{a\}}| \leq \vartheta_1 < 1$ and $|\theta_a^{b,\Gamma(n)\setminus\{b\}}| \leq \sum_{k\in\Gamma(n)\setminus\{b\}} |\theta_k^{b,\Gamma(n)\setminus\{b\}}| \leq \vartheta_1 < 1$, it follows that

$$\sum_{k\in\Gamma(n)\setminus\{a,b\}} |\theta_k^{a,\Gamma(n)\setminus\{a,b\}}| - \vartheta_1 \leq 0,$$

which completes the proof.