The Annals of Statistics

CONSISTENT NEIGHBORHOOD SELECTION FOR SPARSE HIGH-DIMENSIONAL GRAPHS WITH THE LASSO

By Nicolai Meinshausen and Peter Bühlmann

ETH Zürich

The pattern of zero entries in the inverse 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 neighborhood selection with the Lasso as a computationally attractive alternative to standard covariance selection for sparse high-dimensional graphs. Neighborhood selection estimates the conditional independence restrictions separately for each node in the graph. We show that the proposed neighborhood selection scheme is consistent for sparse high-dimensional graphs. The consistency hinges on the choice of the penalty parameter. The oracle value for optimal prediction does not lead to a consistent neighborhood 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 the number of observations raised to an arbitrary power.

1. Introduction.

Consider the p-dimensional multivariate normal distributed random variable

$$X = (X_1, \dots, X_p) \sim \mathcal{N}(\mu, \Sigma).$$

AMS 2000 subject classifications: Primary 62J07; secondary 62H20, 62F12

Keywords and phrases: Linear regression, Covariance selection, Gaussian graphical models, Penalized regression

Assuming that the covariance matrix Σ is non-singular, the conditional independence structure of the distribution can be conveniently represented by a graphical model $\mathcal{G} = (\Gamma, E)$, 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 $X_{\Gamma \backslash \{a,b\}} = \{X_k, k \in \Gamma \backslash \{a, b\}\}$. 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; Linhart and Zucchini, 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, neighborhood 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.

Neighborhood selection is a subproblem of covariance selection. The neighborhood ne_a of a node $a \in \Gamma$ is the smallest subset of $\Gamma \backslash \{a\}$ so that, given all variables X_{ne_a} in the neighbourhood, X_a is conditionally independent of all remaining variables. The neighborhood 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, neighborhood selection aims at estimating (individually) the neighborhood of any given variable (or node). The neighborhood 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 neighborhood 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 neighbors 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 (2004), who study consistent prediction in a triangular setup very similar to ours (see also Juditsky 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 neighborhood estimate with the Lasso converges to the true neighborhood. The choice of the penalty is crucial in the high-dimensional setting. 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 neighborhood 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, neighborhood 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. Neighborhood Selection.

Instead of assuming a fixed true underlying model, we adopt a more flexible approach similar to the triangular setup in Greenshtein and Ritov (2004). Both the number of nodes in the graphs (number of variables), denoted by $p(n) = |\Gamma(n)|$, and the distribution (the covariance matrix) depend in general on the number of observations, so that $\Gamma = \Gamma(n)$ and $\Sigma = \Sigma(n)$. The neighborhood 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 all remaining variables. Denote the closure of node $a \in \Gamma(n)$ by $\mathrm{cl}_a := \mathrm{ne}_a \cup \{a\}$. Then

$$X_a \perp \{X_k, k \in \Gamma(n)\backslash\mathrm{cl}_a\} \,|\, X_{\mathrm{ne}_a}.$$

For details see Lauritzen (1996). The neighborhood depends in general on n as well. However, this dependence is notationally suppressed in the following.

It is instructive to give a slightly different definition of a neighborhood. For each node $a \in \Gamma(n)$, consider optimal prediction of X_a, given all remaining variables. Let $\theta^a \in \mathbb{R}^{p(n)}$ be the vector of coefficients for optimal prediction,

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

As a generalization of (1), which will be of use later, consider optimal prediction of X_a, given only a subset of variables $\{X_k; k \in \mathcal{A}\}$, where $\mathcal{A} \subseteq \Gamma(n)\backslash\{a\}$. The optimal prediction is characterized by the vector $\theta^{a,\mathcal{A}}$,

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

The elements of θ^a are determined by the inverse covariance matrix (Lauritzen, 1996). For $b \in \Gamma\backslash\{a\}$ and $K(n) = \Sigma^{-1}(n)$, 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 $\{b \in \Gamma(n)\backslash\{a\} : K_{ab}(n) \neq 0\}$ of non-zero entries in the corresponding row vector of the inverse covariance matrix and defines precisely the set of neighbors of node a. The best predictor for X_a is thus a linear function of variables in the set of neighbors of the node a only. The set of neighbors 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 of $X \sim \mathcal{N}(0, \Sigma(n))$, neighborhood selection tries to estimate the set of neighbors 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 neighbors of the node a, it seems reasonable to try to exploit this relation.

2.1. *Neighborhood 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 neighborhood of node a in the graph, as shown in the following. Let the $n \times p(n)$-dimensional matrix $\mathbf{X}$ contain n independent observations of X, so that the columns $\mathbf{X}_a$ correspond for all $a \in \Gamma(n)$ to the vector of n independent observations of X_a. Let $\langle \cdot, \cdot \rangle$ be the usual inner product on $\mathbb{R}^n$ and $\| \cdot \|_2$ the corresponding norm.

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

$$\hat{\theta}^{a,\lambda} = \arg \min_{\theta : \theta_a = 0} (n^{-1} \|\mathbf{X}_a - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_1), \tag{3}$$

where $\|\theta\|_1 = \sum_{b \in \Gamma(n)} |\theta_b|$ is the l_1-norm of the coefficient vector. Normalization of all variables to common empirical variance is recommended for the estimator in (3).

Other regression estimates have been proposed, which are based on the l_p-norm, where p is typically in the range $[0, 2]$ (Frank and Friedman, 1993). A value of $p = 2$ leads to the ridge estimate, while $p = 0$ corresponds to traditional model selection. It is well known that the estimates have a parsimonious property (with some components being exactly zero) for $p \leq 1$ only, while the optimization problem in (3) is only convex for $p \geq 1$. Hence l_1-constrained empirical risk minimization occupies a unique position, as

$p = 1$ is the only value of p for which variable selection takes place while the optimization problem is still convex and hence feasible for high-dimensional problems.

The solution to (3) is not necessarily unique. However, if uniqueness fails, the set of solutions is still convex and all our results about neighborhoods (in particular Theorems 1 and 2) hold for any solution of (3).

The neighborhood estimate (parameterized by λ) is defined by the nonzero coefficient estimates of the l_1-penalized regression,

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

Each choice of a penalty parameter λ specifies thus an estimate of the neighborhood 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 $\hat{\mathrm{ne}}_a^\lambda$ if the value of λ is diminished.

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

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

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 $\lambda_{\mathrm{oracle}}$ 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. The prediction-oracle solution does not lead to consistent model selection for the Lasso, as shown in the following for a very simple example.

PROPOSITION 1. *Let the number of variables grow to infinity, $p(n) \to \infty$, for $n \to \infty$. Assume that the covariance matrices $\Sigma(n)$ are identical to the identity matrix except for some pair $(a, b) \in \Gamma(n) \times \Gamma(n)$, for which $\Sigma_{ab}(n) = \Sigma_{ba}(n) = s$, for some $0 < s < 1$ and all $n \in \mathbb{N}$. The probability* *of selecting the wrong neighborhood for node a converges to 1 under the prediction-oracle penalty,*

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

A proof is given in the appendix. It follows from the proof of Proposition 1 that many noise variables are included into the neighborhood 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. If the penalty is chosen larger than the prediction-optimal value, consistent neighborhood selection is possible with the Lasso, as demonstrated in the following.

2.3. *Assumptions.* We make a few assumptions to prove consistency of neighborhood selection with the Lasso. We always assume availability of n independent observations from $X \sim \mathcal{N}(0, \Sigma)$.

High-dimensionality. The number of variables is allowed to grow like the number of observations n raised to an arbitrarily high power.

Assumption 1. *There exists $\gamma > 0$, so that*

$$p(n) = O(n^{\gamma}) \qquad \textit{for } \ n \to \infty.$$

It is in particular allowed for the following analysis that the number of variables is very much larger than the number of observations, $p(n) \gg n$.

Non-singularity. We make two regularity assumptions for the covariance matrices.

Assumption 2. *For all $a \in \Gamma(n)$ and $n \in \mathbb{N}$, $Var(X_a) = 1$. There exists $v^2 > 0$, so that for all $n \in \mathbb{N}$ and $a \in \Gamma(n)$,*

$$Var(X_a | X_{\Gamma(n) \setminus \{a\}}) \geq v^2.$$

Common variance can always be achieved by appropriate scaling of the variables. A scaling to common (empirical) variance of all variables is desirable, as the solutions would otherwise depend on the chosen units or dimensions in which they are represented. The second part of the assumption explicitly excludes singular or nearly singular covariance matrices. For singular covariance matrices, edges are not uniquely defined by the distribution and it is hence not surprising that nearly singular covariance matrices are not suitable for consistent variable selection. Note, however, that the empirical covariance matrix is a.s. singular if $p(n) > n$, which is allowed in our analysis.

Sparsity. The main assumption is the sparsity of the graph. This entails a restriction on the size of the neighborhoods of variables.

ASSUMPTION 3. *There exists some* $0 \le \kappa < 1$ *so that*

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

This assumption limits the maximal possible rate of growth for the size of neighborhoods.

For the next sparsity condition, consider again the definition in (2) of the optimal coefficient $\theta^{b,\mathcal{A}}$ for prediction of X_b, given variables in the set $\mathcal{A} \subset \Gamma(n)$.

ASSUMPTION 4. *There exists some* $\vartheta < \infty$ *so that for all neighboring nodes* $a, b \in \Gamma(n)$ *and all* $n \in \mathbb{N}$,

$$\|\theta^{a,\mathrm{ne}_b\setminus\{a\}}\|_1 \le \vartheta.$$

This assumption is e.g. fulfilled if Assumption 2 holds and the size of the overlap of neighborhoods is bounded by an arbitrarily large number from above for neighboring nodes. That is, if there exists some $m < \infty$ so that for all $n \in \mathbb{N}$,

$$\max_{a,b\in\Gamma(n), b\in\mathrm{ne}_a} |\mathrm{ne}_a \cap \mathrm{ne}_b| \le m \qquad \text{for } n \to \infty, \tag{4}$$

then Assumption 4 is fulfilled. To see this, note that Assumption 2 gives a finite bound for the l_2-norm of $\theta^{a,\mathrm{ne}_b\setminus\{a\}}$, while (4) gives a finite bound for the l_0-norm. Taken together, Assumption 4 is implied.

Magnitude of partial correlations. The next assumption bounds the magnitude of partial correlations from below. The partial correlation π_{ab} between variables X_a and X_b is the correlation after having eliminated the linear effects from all remaining variables $\{X_k, k \in \Gamma(n)\setminus\{a,b\}\}$; for details see Lauritzen (1996).

Assumption 5. *There exists a constant $\delta > 0$ and some $\xi > \varepsilon$, with ε as in Theorem 1, so that for every $(a, b) \in E$,*

$$|\pi_{ab}| \geq \delta n^{-(1-\xi)/2}.$$

It will be shown below that Assumption 5 cannot be relaxed in general. Note that neighborhood selection for node $a \in \Gamma(n)$ is equivalent to simultaneously testing the null hypothesis of zero partial correlation between variable X_a and all remaining variables X_b, $b \in \Gamma(n)\backslash\{a\}$. The null hypothesis of zero partial correlation between two variables can be tested by using the corresponding entry in the normalised inverse empirical covariance matrix. A graph estimate based on such tests has been proposed by Drton and Perlman (2004). Such a test can only be applied, however, if the number of variables is smaller than the number of observations, $p(n) \leq n$, as the empirical covariance matrix is singular otherwise. Even if $p(n) = n - c$ for some constant $c > 0$, Assumption 5 would have to hold with $\xi = 1$ to have a positive power of rejecting false null hypotheses for such an estimate, that is partial correlations would have to be bounded by a positive value from below. In the current high-dimensional setting, we are able to achieve power converging to 1 with the proposed Lasso neighborhood selection even if the number of predictor variables is growing like the number of observations n raised to an arbitrarily high power, that is $p(n) = O(n^\gamma)$ with some $\gamma > 0$, and in particular $p(n) \gg n$.

Neighborhood stability. The last assumption is referred to as neighborhood stability. Using the definition of $\theta^{a,\mathcal{A}}$ in equation (2), define for all $a, b \in \Gamma(n)$,

$$S_a(b) := \sum_{k \in \mathrm{ne}_a} \mathrm{sign}(\theta_k^{a,\mathrm{ne}_a})\theta_k^{b,\mathrm{ne}_a}. \tag{5}$$

The assumption of neighborhood stability restricts the magnitude of the quantities $S_a(b)$ for non-neighboring nodes $a, b \in \Gamma(n)$.

Assumption 6. *There exists some $\delta < 1$ so that for all $a, b \in \Gamma(n)$ with $b \notin \mathrm{ne}_a$,*

$$|S_a(b)| < \delta.$$

It is shown in Proposition 3 that this assumption cannot be relaxed.

We give in the following a more intuitive condition which essentially implies Assumption 6. This will justify the term neighborhood stability. Consider the definition in equation (1) of the optimal coefficients θ^a for prediction of X_a, For $\eta > 0$, define $\theta^a(\eta)$ as the optimal set of coefficients under an additional l_1-penalty,

$$\theta^a(\eta) := \arg \min_{\theta:\theta_a=0} \; E(X_a - \sum_{k\in\Gamma(n)} \theta_k X_k)^2 + \eta\|\theta\|_1. \tag{6}$$

For increasing values of the penalty parameter η one expects that the magnitudes of the coefficients of $\theta^a(\eta)$ are diminished and are eventually, for large values of η, equal to zero. The neighborhood ne_a of node a was defined as the set of non-zero coefficients of θ^a, $\mathrm{ne}_a = \{k \in \Gamma(n) : \theta^a_k = 0\}$. Define the disturbed neighborhood $\mathrm{ne}_a(\eta)$ as

$$\mathrm{ne}_a(\eta) := \{k \in \Gamma(n) : \theta^a_k(\eta) = 0\}.$$

It clearly holds that $\mathrm{ne}_a = \mathrm{ne}_a(0)$. The assumption of neighborhood stability is fulfilled if there exists some infinitesimal small perturbation η, which may depend on n, so that the disturbed neighborhood $\mathrm{ne}_a(\eta)$ is identical to the undisturbed neighborhood $\mathrm{ne}_a(0)$.

Proposition 2. *If there exists some $\eta > 0$ so that $\mathrm{ne}_a(\eta) = \mathrm{ne}_a(0)$, then $|S_a(b)| \le 1$ for all $b \in \Gamma(n)\backslash\mathrm{ne}_a$.*

A proof is given in the appendix.

In light of Proposition 2 it seems that Assumption 6 is a very weak condition. To give one example, Assumption 6 is automatically fulfilled under the much stronger assumption that the graph does not contain cycles. We give a brief brief reasoning for this. Consider two non-neighboring nodes nodes a and b. If the nodes are in different connectivity components there is nothing left to show as $S_a(b) = 0$. If they are in the same connectivity components, then there exists one node $k \in \mathrm{ne}_a$ that separates b from $\mathrm{ne}_a\backslash\{k\}$. This follows from the fact that there is just one unique path between any two variables in the same connectivity component if the graph does not contain cycles. Using the global Markov property, the random variable X_b is independent of $X_{\mathrm{ne}_a\backslash\{k\}}$, given X_k. The random variable $E(X_b|X_{\mathrm{ne}_a})$ is thus a function of X_k only. As the distribution is Gaussian,

$E(X_b|X_{\mathrm{ne}_a}) = \theta_k^{b,\mathrm{ne}_a} X_k$. By non-singularity of the covariance matrix, Assumption 2, $Var(X_b|X_{\mathrm{ne}_a}) = v^2$ for some $v^2 > 0$. It follows by Assumption 2 that $Var(X_b) = v^2 + (\theta_k^{b,\mathrm{ne}_a})^2 = 1$ and hence $\theta_k^{b,\mathrm{ne}_a} = \sqrt{1-v^2} < 1$, which implies that Assumption 6 is indeed fulfilled if the graph does not contain cycles.

We mention that Assumption 6 is likewise fulfilled if the inverse covariance matrices $\Sigma^{-1}(n)$ are for each $n \in \mathbb{N}$ diagonally dominant. A matrix is said to be diagonally dominant iff, for each row, the sum of the absolute values of the non-diagonal elements is smaller than the absolute value of the diagonal element. The proof of this is straightforward but tedious and hence omitted.

2.4. *Controlling type I errors.* The asymptotic properties of Lasso-type estimates in regression have been studied in detail by Knight and Fu (2000) for a fixed number of variables. Their results say that the penalty parameter λ should decay for an increasing number of observations at least as fast as $n^{-1/2}$ to obtain a $n^{1/2}$-consistent estimate. It turns out that a slower rate is needed for consistent model selection in the high-dimensional case where $p(n) \gg n$. However, a rate $n^{-(1-\varepsilon)/2}$ with any $\kappa < \varepsilon < 1$, with κ defined as in Assumption 3, is sufficient for consistent neighborhood selection, even when the number of variables is growing rapidly with the number of observations.

THEOREM 1. *Let the Assumptions of section 2.3 be fulfilled. Let the penalty parameter satisfy $\lambda_n \sim dn^{-(1-\varepsilon)/2}$ with some $\kappa < \varepsilon < 1$ and $d > 0$. There exists some constant $c > 0$ so that for all $a \in \Gamma(n)$,*

$$P(\hat{\mathrm{ne}}_a^\lambda \subseteq \mathrm{ne}_a) = 1 - O(\exp(-cn^\varepsilon)) \qquad \textit{for } n \to \infty.$$

A proof is given in the appendix.

Theorem 1 states that the probability of (falsely) including any of the non-neighboring variables of the node $a \in \Gamma(n)$ into the neighborhood estimate is vanishing exponentially fast, even though the number of non-neighboring variables may grow very rapidly with the number of observations. It is shown in the following that Assumption 6 cannot be relaxed.

PROPOSITION 3. *If there exists some $a, b \in \Gamma(n)$ with $b \notin \mathrm{ne}_a$ and $|S_a(b)| > 1$, then*

$$P(\hat{\mathrm{ne}}_a^\lambda \subseteq \mathrm{ne}_a) \to 0 \qquad \textit{for } n \to \infty.$$

A proof is given in the appendix. Assumption 6 of neighborhood stability is hence critical for the success of Lasso neighborhood selection.

2.5. *Controlling type II errors.* So far it was shown that the probability of falsely including variables into the neighborhood can be controlled by the Lasso. The question arises if the probability of including all neighboring variables into the neighborhood estimate is converging to 1 for $n \to \infty$.

THEOREM 2. *Let the Assumptions of Theorem 1 be fulfilled. Then there exists some constant $c > 0$ so that*

$$P(\mathrm{ne}_a \subseteq \hat{\mathrm{ne}}_a^\lambda) = 1 - O(\exp(-cn^\varepsilon)) \qquad \textit{for } n \to \infty.$$

A proof is given in the appendix.

It may be of interest whether Assumption 5 could be relaxed, so that edges are detected even if the partial correlation is vanishing at a rate $n^{-(1-\xi)/2}$ for some $\xi < \varepsilon$. The following proposition says that $\xi > \varepsilon$ is a necessary condition if a stronger version of Assumption 4 holds, which is e.g. still fulfilled for forests and trees.

PROPOSITION 4. *Let the Assumption of Theorem 1 be fulfilled with $\vartheta < 1$ in Assumption 4. For $a \in \Gamma(n)$, let there be some $b \in \Gamma(n)\backslash\{a\}$ with $\pi_{ab} \neq 0$ and $|\pi_{ab}| = O(n^{-(1-\xi)/2})$ for $n \to \infty$ for some $\xi < \varepsilon$. Then*

$$P(b \in \hat{\mathrm{ne}}_a^\lambda) \to 0 \qquad \textit{for } n \to \infty.$$

Theorem 2 and Proposition 4 say that edges between nodes which partial correlation vanishes at a rate $n^{-(1-\xi)/2}$ are, with probability converging to 1 for $n \to \infty$, detected if $\xi > \varepsilon$ and are undetected if $\xi < \varepsilon$. The results do not cover the case $\xi = \varepsilon$, which remains a challenging question for further research.

All results so far have treated the distinction between zero and non-zero partial correlations only. We remark that it follows from the proofs of the previous results that the signs of partial correlations of neighboring nodes can be estimated consistently under the same assumptions and and with the same rates.

3. Covariance Selection.

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

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

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 neighbors, the edge set $E \subseteq \Gamma(n) \times \Gamma(n)$ is given by

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

The first condition, $a \in \mathrm{ne}_b$, implies in fact the second, $b \in \mathrm{ne}_a$, and vice versa, so that the edge is as well identical to $\{(a, b) : a \in \mathrm{ne}_b \vee b \in \mathrm{ne}_a\}$. For an estimate of the edge set of a graph, we can apply neighborhood 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

$$\hat{E}^{\lambda,\wedge} = \{(a, b) : a \in \hat{\mathrm{ne}}_b^\lambda \wedge b \in \hat{\mathrm{ne}}_a^\lambda\}. \tag{7}$$

Note that $a \in \hat{\mathrm{ne}}_b^\lambda$ does not necessarily imply $b \in \hat{\mathrm{ne}}_a^\lambda$ and vice versa. We can hence also define a second, less conservative, estimate of the edge set by

$$\hat{E}^{\lambda,\vee} = \{(a, b) : a \in \hat{\mathrm{ne}}_b^\lambda \vee b \in \hat{\mathrm{ne}}_a^\lambda\}. \tag{8}$$

Note that there are in total $2^{(p^2-p)/2}$ distinct graphs for a p-dimensional variable. However, for each of the p nodes there are only 2^{p-1} distinct potential neighborhoods. By braking the graph selection problem into a consecutive series of neighborhood selection problems, the complexity of the search is thus reduced substantially at the price of potential inconsistencies between neighborhood estimates. Graph estimates that apply this strategy for complexity reduction are sometimes called dependency networks (Heckerman et al., 2000). The complexity of the proposed neighborhood selection for one node with the Lasso is reduced further to $O(np \min\{n, p\})$, as the Lars procedure of (Efron et al., 2004) requires $O(\min\{n, p\})$ steps, each of complexity $O(np)$. For high-dimensional problems as in Theorems 1 and 2, where the number of variables grows like $p(n) \sim cn^\gamma$ for some $c > 0$ and $\gamma > 1$, this is equivalent to $O(p^{2+2/\gamma})$ computations for the whole graph. The complexity of the proposed method scales thus approximately quadratic with the number of nodes for large values of γ.

The discrepancies between the estimates (7) and (8) are quite small in our experience. Asymptotically the difference between both estimates vanishes, as seen in the following corollary. We refer to both edge set estimates collectively with the generic notation $\hat{E}^{\lambda}$, as the following result holds true for both of them.

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

$$P(\hat{E}^{\lambda} = E) = 1 - O(\exp(-cn^{\varepsilon})) \qquad \textit{for } n \to \infty.$$

The claim follows since $|\Gamma(n)|^2 = p(n)^2 = O(n^{2\gamma})$ by Assumption 1 and neighborhood selection has an exponentially fast convergence rate as described by Theorem 2. Corollary 1 says that the conditional independence structure of a multivariate normal distribution can be estimated consistently by combining the neighborhood estimates for all variables. 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 neighborhood and covariance selection is possible with the Lasso in a high-dimensional setting. However, the asymptotic considerations give 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 neighborhood 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

$$\lambda(\alpha) = \frac{2\hat{\sigma}_a}{\sqrt{n}} \tilde{\Phi}^{-1}(\frac{\alpha}{2p(n)^2}), \tag{9}$$

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

Theorem 3. *Let the Assumptions of section 2.3 be fulfilled. Using penalty parameter $\lambda(\alpha)$, it holds for all $n \in \mathbb{N}$ that*

$$P(\exists\, a \in \Gamma(n) : \hat{C}_a^\lambda \nsubseteq C_a) \leq \alpha.$$

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.

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 neighborhood selection with forward selection MLE for more than 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 neighborhood 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.

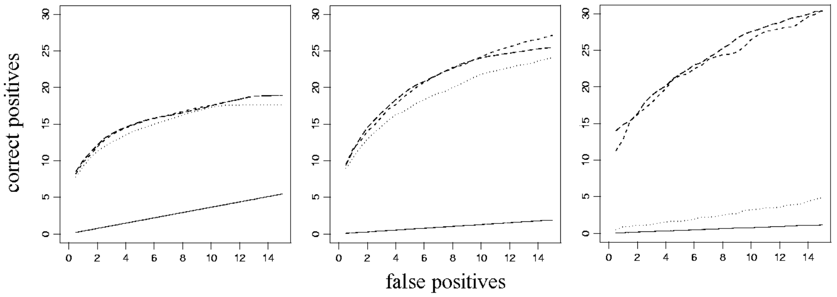

Figure 1: The average number of correct positives as a function of the number of false positives. The Lasso neighborhood selection schemes (short dashed line for the estimate $\hat{E}^{\lambda,\wedge}$, long dashed line for the estimate $\hat{E}^{\lambda,\vee}$), the forward selection MLE (dotted line)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

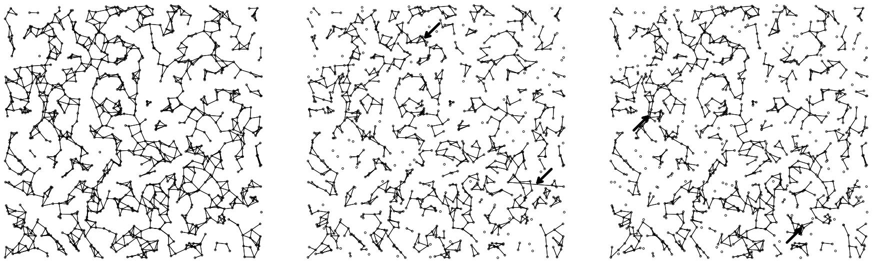

Figure 2: A realization of a graph is shown on the left, generated as described in the text. The graph consists of 1000 nodes and 1747 edges out of 449500 distinct pairs of variables. The estimated edge set, using estimate (7) at level $\alpha = .05$, see equation (9), is shown in the middle. There are 2 erroneously included edges, marked by an arrow, while 1109 edges are correctly detected. For estimate (8) and an adjusted level as described in the text, the result is shown on the right. Again two edges are erroneously included. Not a single pair of disjoint connectivity components of the true graph has been (falsely) joined by either estimate.

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 for all edges. Initially all variables have identical conditional variance and the partial correlation between neighbors 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. The diagonal elements of the corresponding covariance matrix are in general larger than 1. To achieve constant variance, all variables are finally rescaled so that the diagonal elements of Σ are all unity. Using the Cholesky transformation of the covariance matrix, n independent samples are drawn from the corresponding Gaussian distribution.

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 to the proposed Lasso neighborhood 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 approximately equal to 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 neighborhood selection for the full graph. For more than 30 nodes, the differences are even more pronounced.

The Lasso neighborhood selection can be applied to hundred- or thousand-dimensional graphs, a realistic size for e.g. biological networks. A graph with 1000 nodes (following the same model as described above) and its estimates (7) and (8), using 600 observations, are shown in Figure 2. A level $\alpha = .05$ is used for estimate $\hat{E}^{\lambda,\vee}$. For better comparison, the level α was adjusted to $\alpha = .064$ for the estimate $\hat{E}^{\lambda,\wedge}$, so that both estimates lead to the same number of included edges. There are two erroneous edge inclusions, while 1109 out of all 1747 edges have been correctly identified by either estimate. Of these 1109 edges, 907 are common to both estimates while 202 are just present in either (7) or (8).

Table 1: *The average number of correctly and falsely included and excluded edges.*

	$\hat{E} = \hat{E}^{\lambda,\vee}$		$\hat{E} = \hat{E}^{\lambda,\wedge}$		
	pure Gaussian				
	$(a,b) \in \hat{E}$	$(a,b) \notin \hat{E}$	$(a,b) \in \hat{E}$	$(a,b) \notin \hat{E}$	total
$(a,b) \in E$	1451	310	1393	394	1761
$(a,b) \notin E$	13	497726	71	497642	497739
total	1464	498036	1464	498036	
	additional t_2-distributed noise				
	$(a,b) \in \hat{E}$	$(a,b) \notin \hat{E}$	$(a,b) \in \hat{E}$	$(a,b) \notin \hat{E}$	total
$(a,b) \in E$	1285	471	1235	521	1756
$(a,b) \notin E$	19	497725	69	497675	497744
total	1304	498196	1304	498196	

The average accuracy over 10 simulations, using 500 independent observations for each simulation, is shown in Table 1. Out of more than 10^5 pairs of variables, only about 13 are on average falsely included into the estimated edge set with the estimate $\hat{E}^{\lambda,\vee}$ at level $\alpha = .05$. For better comparison, the level for estimate $\hat{E}^{\lambda,\wedge}$ was chosen again so that both estimates contain the same amount of edges, leading to an average level of $\alpha = .065$. If adjusted in this way, the difference between both estimates are in general small, even if $\hat{E}^{\lambda,\vee}$ is performing slightly better.

Next, long-tailed noise is added to the observations to examine if results are critically dependent on the assumption of Gaussianity. Instead of n i.i.d.

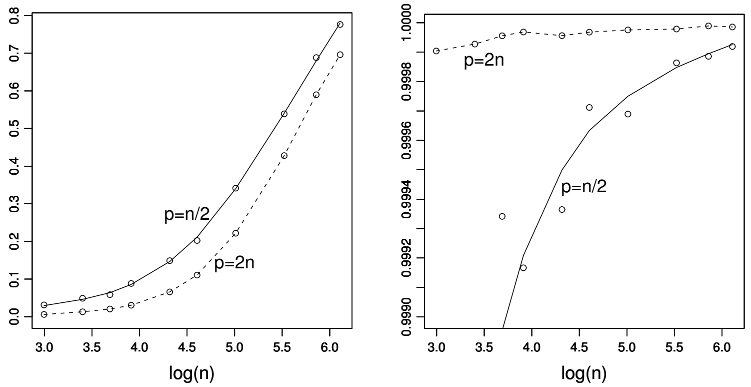

Figure 3: Left panel: the percentage of all edges in E which are correctly included in $\hat{E}^{\lambda,\wedge}$ as a function of $\log(n)$; for $p(n) = 2n$ variables, broken line, and $p(n) = n/2$, solid line. Right panel: the percentage of edges not in E which are correctly not included in $\hat{E}^{\lambda,\wedge}$ as a function of $\log(n)$. 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.

observations of $X \sim \mathcal{N}(0, \Sigma)$ there are n i.i.d observations of $X + 0.1Z$, where the components of Z are independent and follow a t_2-distribution. The results about the correctness of the estimation of conditional independence are shown in the lower part of Table 1. The dominant effect is a reduction in the number of included edges, while the proportion of false rejections is increasing only slightly from 0.8% to 1.4% for $\hat{E}^{\lambda,\vee}$ and from 4.8% to 5.2% for $\hat{E}^{\lambda,\wedge}$.

Finally, 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 variables is increasing at the same rate as the number of observations. The result illustrates that the quality of neighborhood 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.

5. Proofs.

5.1. *Notation and useful lemmata.* As a generalization of (3), the Lasso estimate $\hat{\theta}^{a,\mathcal{A},\lambda}$ of $\theta^{a,\mathcal{A}}$, defined in (2), is given by

$$\hat{\theta}^{a,\mathcal{A},\lambda} = \arg \min_{\theta:\theta_k=0\ \forall k\notin\mathcal{A}} (n^{-1}\|\mathbf{X}_a - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1). \tag{10}$$

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

Lemma 1. *Given $\theta \in \mathbb{R}^{p(n)}$, let $G(\theta)$ be a $p(n)$-dimensional vector with elements*

$$G_b(\theta) = -2n^{-1}\langle \mathbf{X}_a - \mathbf{X}\theta, \mathbf{X}_b\rangle.$$

A vector $\hat{\theta}$ with $\hat{\theta}_k = 0, \forall k \in \Gamma(n)\setminus\mathcal{A}$ is a solution to (10) iff for all $b \in \mathcal{A}$, $G_b(\hat{\theta}) = -\text{sign}(\hat{\theta}_b)\lambda$ in case $\hat{\theta}_b \neq 0$ and $|G_b(\hat{\theta})| \leq \lambda$ in case $\hat{\theta}_b = 0$. Moreover, if the solution is not unique and $|G_b(\hat{\theta})| < \lambda$ for some solution $\hat{\theta}$, then $\hat{\theta}_b = 0$ for all solutions of (10).

Proof. Denote the subdifferential of

$$n^{-1}\|\mathbf{X}_a - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1$$

with respect to θ by $D(\theta)$. The vector $\hat{\theta}$ is a solution to (10) iff there exists an element $d \in D(\hat{\theta})$ so that $d_b = 0, \forall b \in \mathcal{A}$. $D(\theta)$ 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). □

Lemma 2. *Let $\hat{\theta}^{a,\text{ne}_a,\lambda}$ be defined for every $a \in \Gamma(n)$ as in (10). Under the assumptions of Theorem 1, it holds for all $a \in \Gamma(n)$,*

$$P(\text{sign}(\hat{\theta}_b^{a,\text{ne}_a,\lambda}) = \text{sign}(\theta_b^a),\ \forall b \in \text{ne}_a) = 1 - O(\exp(-cn^{\varepsilon})) \qquad \textit{for } n \to \infty$$

For the sign-function, it is is understood that sign(0) = 0. The Lemma says, in other words, that if one could restrict the Lasso estimate to have non-zero coefficients in the neighborhood of node a only, then the signs of the partial correlations in this neighborhood are estimated consistently under the made assumptions.

Proof. Using Bonferroni's inequality, and $|ne_a| = o(n)$ for $n \to \infty$, it suffices to show that for every $a, b \in \Gamma(n)$ with $b \in \mathrm{ne}_a$,

$$P(\mathrm{sign}(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda}) = \mathrm{sign}(\theta_b^a)) = 1 - O(\exp(-cn^{\varepsilon})) \qquad \text{for } n \to \infty$$

Consider the definition of $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ in (10),

$$\hat{\theta}^{a,\mathrm{ne}_a,\lambda} = \arg\min_{\theta:\theta_k=0\ \forall k\notin \mathrm{ne}_a} (n^{-1}\|\mathbf{X}_a - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1). \tag{11}$$

Consider now that component b of this estimate is fixed at a constant value β. Denote this new estimate by $\tilde{\theta}^{a,b,\lambda}(\beta)$,

$$\tilde{\theta}^{a,b,\lambda}(\beta) = \arg\min_{\theta\in\Theta_b(\beta)} (n^{-1}\|\mathbf{X}_a - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1), \tag{12}$$

where

$$\Theta_b(\beta) := \{\theta \in \mathbb{R}^{p(n)} : \theta_b = \beta; \theta_k = 0, \forall k \notin \mathrm{ne}_a\}.$$

There exists always a value β (namely $\beta = \hat{\theta}_b^{a,\mathrm{ne}_a,\lambda}$) so that $\tilde{\theta}^{a,b,\lambda}(\beta)$ is identical to $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$. Thus, if $\mathrm{sign}(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda}) \neq \mathrm{sign}(\theta_b^a)$, there exists some β with $\mathrm{sign}(\beta)\mathrm{sign}(\theta_b^a) \leq 0$ so that $\tilde{\theta}^{a,b,\lambda}(\beta)$ is a solution to (11). Using $\mathrm{sign}(\theta_b^a) \neq 0$ for all $b \in \mathrm{ne}_a$, it is thus sufficient to show that for every β with $\mathrm{sign}(\beta)\mathrm{sign}(\theta_b^a) < 0$, $\tilde{\theta}^{a,b,\lambda}(\beta)$ cannot be a solution to (11) with high probability.

To show this, we focus in the following on the case where $\theta_b^a > 0$ for notational simplicity. The case $\theta_b^a < 0$ follows analogous. If $\theta_b^a > 0$, it follows by Lemma 1 that $\tilde{\theta}^{a,b,\lambda}(\beta)$ with $\tilde{\theta}_b^{a,b,\lambda}(\beta) = \beta \leq 0$ can only be a solution to (11) if $G_b(\tilde{\theta}^{a,b,\lambda}(\beta)) \geq -\lambda$. Hence it suffices to show that for all $b \in \mathrm{ne}_a$ with $\theta_b^a > 0$, for $n \to \infty$,

$$P(\sup_{\beta\leq 0}\{G_b(\tilde{\theta}^{a,b,\lambda}(\beta))\} < -\lambda) = 1 - O(\exp(-cn^{\varepsilon})). \tag{13}$$

Let in the following $\mathbf{R}_a^{\lambda}(\beta)$ be the n-dimensional vector of residuals,

$$\mathbf{R}_a^{\lambda}(\beta) := \mathbf{X}_a - \mathbf{X}\tilde{\theta}^{a,b,\lambda}(\beta). \tag{14}$$

We can write X_b as

$$X_b = W_b + \sum_{k\in \mathrm{ne}_a\setminus\{b\}} \theta_k^{b,\mathrm{ne}_a\setminus\{b\}} X_k, \tag{15}$$

where W_b is independent of $\{X_k;\, k \in \mathrm{ne}_a \backslash \{b\}\}$. By straightforward calculation, using (15),

$$G_b(\tilde{\theta}^{a,b,\lambda}(\beta)) = -2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b\rangle - \sum_{k \in \mathrm{ne}_a\backslash\{b\}} \theta_k^{b,\mathrm{ne}_a\backslash\{b\}}(2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{X}_k\rangle).$$

By Lemma 1, for all $k \in \mathrm{ne}_a\backslash\{b\}$, $|G_k(\tilde{\theta}^{a,b,\lambda}(\beta))| = |2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{X}_k\rangle| \le \lambda$. This together with the equation above yields

$$G_b(\tilde{\theta}^{a,b,\lambda}(\beta)) \le -2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b\rangle + \lambda\|\theta^{b,\mathrm{ne}_a\backslash\{b\}}\|_1. \tag{16}$$

Using Assumption 4, there exists some $\vartheta < \infty$, so that $\|\theta^{b,\mathrm{ne}_a\backslash\{b\}}\|_1 \le \vartheta$. For proving (13) it is therefore sufficient to show that there exists $c > 0$ so that for all $b \in \mathrm{ne}_a$ with $\theta_b^a > 0$ and any $g > 0$, for $n \to \infty$,

$$P(\inf_{\beta \le 0}\{2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b\rangle\} > g\lambda) = 1 - O(\exp(-cn^\varepsilon)). \tag{17}$$

With a little abuse of notation, let $V^\| \subseteq \mathbb{R}^n$ be the, at most $(|\mathrm{ne}_a| - 1)$-dimensional, space which is spanned by the vectors $\{\mathbf{X}_k,\, k \in \mathrm{ne}_a\backslash\{b\}\}$ and let $V^\perp$ be the orthogonal complement of $V^\|$ in $\mathbb{R}^n$. Split the n-dimensional vector $\mathbf{W}_b$ of observations of W_b into the sum of two vectors

$$\mathbf{W}_b = \mathbf{W}_b^\perp + \mathbf{W}_b^\|, \tag{18}$$

where $\mathbf{W}_b^\|$ is contained in the space $V^\| \subseteq \mathbb{R}^n$, while the remaining part $\mathbf{W}_b^\perp$ is chosen orthogonal to this space (in the orthogonal complement $V^\perp$ of $V^\|$). The inner product in (17) can be written as

$$2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b\rangle = 2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^\perp\rangle + 2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^\|\rangle. \tag{19}$$

By Lemma 3, there exists for every $g > 0$ some $c > 0$ so that for $n \to \infty$,

$$P(\inf_{\beta \le 0}\{2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^\|\rangle/(1 + |\beta|)\} > -g\lambda) = 1 - O(\exp(-cn^\varepsilon))$$

To show (17), it hence suffices to proof that there exists for every $g > 0$ some $c > 0$ so that for $n \to \infty$,

$$P(\inf_{\beta \le 0}\{2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^\perp\rangle - g(1 + |\beta|)\lambda\} > 0) = 1 - O(\exp(-cn)). \tag{20}$$

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

$$X_a = \sum_{k \in \mathrm{ne}_a} \theta_k^a X_k + W_a.$$

Note that W_a and W_b are independent normal distributed random variables with variances σ_a^2 and σ_b^2 respectively. By Assumption 2, $0 < v^2 \le \sigma_b^2, \sigma_a^2 \le 1$. Note furthermore that W_b and $X_{\mathrm{ne}_a \setminus \{b\}}$ are independent. However, W_a and X_a are not necessarily independent and neither are W_a and $\hat{\theta}^{a,\mathrm{ne}_a \setminus \{b\},\lambda}$. Using $\theta^a = \theta^{a,\mathrm{ne}_a}$ and (15),

$$X_a = \sum_{k \in \mathrm{ne}_a \setminus \{b\}} (\theta_k^a + \theta_b^a \theta_k^{b,\mathrm{ne}_a \setminus \{b\}}) X_k + \theta_b^a W_b + W_a. \tag{21}$$

Using (21), the definition of the residuals in (14), and the orthogonality property of $\mathbf{W}_b^\perp$,

$$\begin{aligned} 2n^{-1}\langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^\perp \rangle &= 2n^{-1}(\theta_b^a - \beta)\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp \rangle + 2n^{-1}\langle \mathbf{W}_a, \mathbf{W}_b^\perp \rangle, \\ &\ge 2n^{-1}(\theta_b^a - \beta)\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp \rangle - |2n^{-1}\langle \mathbf{W}_a, \mathbf{W}_b^\perp \rangle|. \end{aligned} \tag{22}$$

The second term, $|2n^{-1}\langle \mathbf{W}_a, \mathbf{W}_b^\perp \rangle|$, is stochastically smaller than $|2n^{-1}\langle \mathbf{W}_a, \mathbf{W}_b \rangle|$. Due to independence of W_a and W_b, $E(W_a W_b) = 0$. Using Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner, 1996), and $\lambda \sim dn^{-(1-\varepsilon)/2}$ with $\varepsilon > 0$, there exists a $c > 0$ for every $g > 0$ so that

$$P(|2n^{-1}\langle \mathbf{W}_a, \mathbf{W}_b^\perp \rangle| \ge g\lambda) = O(\exp(-cn)). \tag{23}$$

Instead of (20), it is thus sufficient to show that there exists $c > 0$ for every $g > 0$ so that for $n \to \infty$,

$$P(\inf_{\beta \le 0} \{2n^{-1}(\theta_b^a - \beta)\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp \rangle - g(1 + |\beta|)\lambda\} > 0) = 1 - O(\exp(-cn)). \tag{24}$$

Note that $\sigma_b^{-2}\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp \rangle$ follows a $\chi^2_{n-|\mathrm{ne}_a|}$-distribution. As $|\mathrm{ne}_a| = o(n)$ and $\sigma_b^2 \ge v^2$ (by Assumption 2), it follows for $n \ge n_0$ with some $n_0 \in \mathbb{N}$, that

$$2n^{-1}\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp \rangle \overset{st.}{\ge} k(Z/n),$$

where $\overset{st.}{\ge}$ denotes stochastically larger or equal, k is some positive constant, and Z is a $\chi^2_{n/2}$-distributed random variable. By the properties of the χ^2-distribution, there exist some $t, c > 0$ so that $P(Z/n < t) = O(\exp(-cn))$.

To show (24) it hence suffices to proof that for every $d > 0$ there exists some $n_0(d) \in \mathbb{N}$ so that for all $n \geq n_0$,

$$\inf_{\beta \leq 0} \{(\theta_b^a - \beta)d - g(1 + |\beta|)\lambda\} > 0. \tag{25}$$

By Assumption 5, $|\pi_{ab}|$ is of order at least $n^{-(1-\xi)/2}$. Using additionally Assumption 2, this implies that there exist some $q > 0$ so that $\theta_b^a \geq q n^{-(1-\xi)/2}$. As $\lambda \sim d n^{-(1-\varepsilon)/2}$ and, by Assumption 5, $\xi > \varepsilon$, it follows that for every $d > 0$ and large enough values of n,

$$\theta_b^a d - g\lambda > 0$$

It remains to show that for any $d > 0$ there exists some $n_0(d)$ so that for all $n \geq n_0$,

$$\inf_{\beta \leq 0} \{-\beta d - g|\beta|\lambda\} > 0.$$

This follows as $\lambda \to 0$ for $n \to \infty$, which completes the proof. □

Lemma 3. *Assume the conditions of section 2.3 hold true. Let $\mathbf{R}_a^\lambda(\beta)$ be defined as in (14) and $\mathbf{W}_b^{\|}$ as in (18). Then there exists $c > 0$ so that for all $a, b \in \Gamma(n)$ and any $g > 0$ and $n \to \infty$,*

$$P(\sup_{\beta \in \mathbb{R}} |2n^{-1} \langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^{\|} \rangle| / (1 + |\beta|) < g\lambda) = 1 - O(\exp(-cn^\varepsilon)).$$

Proof. By Schwartz's inequality,

$$|2n^{-1} \langle \mathbf{R}_a^\lambda(\beta), \mathbf{W}_b^{\|} \rangle| / (1 + |\beta|) \leq 2n^{-1/2} \|\mathbf{W}_b^{\|}\|_2 \frac{n^{-1/2} \|\mathbf{R}_a^\lambda(\beta)\|_2}{1 + |\beta|}. \tag{26}$$

The sum of squares of the residuals is increasing with increasing value of λ. Thus, $\langle \mathbf{R}_a^\lambda(\beta), \mathbf{R}_a^\lambda(\beta) \rangle \leq \langle \mathbf{R}_a^0(\beta), \mathbf{R}_a^0(\beta) \rangle$. By definition of $\mathbf{R}_a^\lambda$ in (14), and using (12),

$$\langle \mathbf{R}_a^0(\beta), \mathbf{R}_a^0(\beta) \rangle = \langle \mathbf{X}_a - \beta \mathbf{X}_b, \mathbf{X}_a - \beta \mathbf{X}_b \rangle,$$

and hence

$$\langle \mathbf{R}_a^0(\beta), \mathbf{R}_a^0(\beta) \rangle \leq (1 + |\beta|)^2 \max\{\|\mathbf{X}_a\|_2^2, \|\mathbf{X}_b\|_2^2\}.$$

Therefore

$$P(\sup_{\beta \in \mathbb{R}} \frac{n^{-1/2} \|\mathbf{R}_a^\lambda(\beta)\|_2}{1 + |\beta|} > d) = P(n^{-1/2} \max\{\|\mathbf{X}_a\|_2, \|\mathbf{X}_b\|_2\} > d).$$

Note that both $\|\mathbf{X}_a\|_2^2$ and $\|\mathbf{X}_b\|_2^2$ are χ_n^2-distributed. Thus there exists $d > 1$ and $c > 0$ so that

$$P(\sup_{\beta\in\mathbb{R}} \frac{n^{-1/2}\|\mathbf{R}_a^\lambda(\beta)\|_2}{1+|\beta|} > d) = O(\exp(-cn)) \qquad \text{for } n \to \infty. \tag{27}$$

It remains to be shown that there exists some $c > 0$ for every $g > 0$ so that

$$P(n^{-1/2}\|\mathbf{W}_b^{\|}\|_2 > g\lambda) = O(\exp(-cn^\varepsilon)) \qquad \text{for } n \to \infty. \tag{28}$$

The expression $\sigma_b^{-2}\langle \mathbf{W}_b^{\|}, \mathbf{W}_b^{\|}\rangle$ is $\chi^2_{|\mathrm{ne}_a|-1}$-distributed. As $\sigma_b \le 1$ and $|\mathrm{ne}_a| = O(n^\kappa)$, it follows that $n^{-1/2}\|\mathbf{W}_b^{\|}\|_2$ is stochastically smaller than

$$tn^{-(1-\kappa)/2}(Z/n^\kappa)^{1/2},$$

where Z is $\chi^2_{n^\kappa}$-distributed. Thus for every $g > 0$,

$$P(n^{-1/2}\|\mathbf{W}_b^{\|}\|_2 > g\lambda) \le P((Z/n^\kappa) > (g/t)^2 n^{(1-\kappa)}\lambda^2).$$

As $\lambda \sim dn^{-(1-\varepsilon)/2}$ for some $d > 0$ and, by Assumption 5, $\varepsilon > \kappa$, it follows that $n^{1-\kappa}\lambda^2 \ge hn^{\varepsilon-\kappa}$ for some $h > 0$ and $n \to \infty$. By the properties of the χ^2-distribution, claim (28) follows. This completes the proof. □

5.2. *Proof of Proposition 2.* The subdifferential of the argument in (6),

$$E(X_a - \sum_{m\in\Gamma(n)} \theta_m^a X_m)^2 + \eta\|\theta^a\|_1,$$

with respect to θ_k^a, $k \in \Gamma(n)\backslash\{a\}$, is given by

$$-2E((X_a - \sum_{m\in\Gamma(n)} \theta_m^a X_m)X_k) + \eta e_k,$$

where $e_k \in [-1,1]$ if $\theta_k^a = 0$, and $e_k = \operatorname{sign}(\theta_k^a)$ if $\theta_k^a \ne 0$. Using that $\mathrm{ne}_a(\eta) = \mathrm{ne}_a$, it follows as in Lemma 1 for all $k \in \mathrm{ne}_a$,

$$2E((X_a - \sum_{m\in\Gamma(n)} \theta_m^a(\eta)X_m)X_k) = \eta \operatorname{sign}(\theta_k^a), \tag{29}$$

and, for $b \notin \mathrm{ne}_a$,

$$|2E((X_a - \sum_{m\in\Gamma(n)} \theta_m^a(\eta)X_m)X_b)| \le \eta. \tag{30}$$

A variable X_b with $b \notin \mathrm{ne}_a$ can be written as

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

where W_b is independent of $\{X_k, k \in \mathrm{cl}_a\}$. Using this in (30) yields

$$|2 \sum_{k \in \mathrm{ne}_a} \theta_k^{b,\mathrm{ne}_a} E((X_a - \sum_{m \in \Gamma(n)} \theta_m^a(\eta) X_m) X_k)| \leq \eta.$$

Using (29) and $\theta^a = \theta^{a,\mathrm{ne}_a}$, it follows that

$$|\sum_{k \in \mathrm{ne}_a} \theta_k^{b,\mathrm{ne}_a} \mathrm{sign}(\theta_k^{a,\mathrm{ne}_a})| \leq 1,$$

which completes the proof.

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

$$P(\hat{\mathrm{ne}}_a^\lambda \subseteq \mathrm{ne}_a) = 1 - P(\exists b \in \Gamma(n) \backslash \mathrm{cl}_a : \hat{\theta}_b^{a,\lambda} \neq 0). \tag{31}$$

Consider the Lasso estimate $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$, which is by (10) constrained to have non-zero components only in the neighborhood of node $a \in \Gamma(n)$. Using $|\mathrm{ne}_a| = O(n^\kappa)$ with some $\kappa < 1$, we can assume w.l.o.g. that $|\mathrm{ne}_a| \leq n$. This in turn implies, see e.g. Osborne et al. (2000), that $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ is a.s. a unique solution to (10) with $\mathcal{A} = \mathrm{ne}_a$. Let $\mathcal{E}$ be the event

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

Given the event $\mathcal{E}$, it follows from the first part of Lemma 1 that $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ is not only a solution of (10), with $\mathcal{A} = \mathrm{ne}_a$, but as well a solution of (3), where $\mathcal{A} = \Gamma(n) \backslash \{a\}$. As $\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0$ for all $b \in \Gamma(n) \backslash \mathrm{cl}_a$, it follows from the second part of Lemma 1, that $\hat{\theta}_b^{a,\lambda} = 0$, $\forall b \in \Gamma(n) \backslash \mathrm{cl}_a$. Hence

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

where

$$G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda}) = -2n^{-1} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{X}_b \rangle. \tag{32}$$

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

$$P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \geq \lambda) = O(\exp(-cn)). \tag{33}$$

One can write for any $b \in \Gamma(n) \backslash \mathrm{cl}_a$,

$$X_b = \sum_{m \in \mathrm{ne}_a} \theta_m^{b,\mathrm{ne}_a} X_m + W_b, \tag{34}$$

where $W_b \sim \mathcal{N}(0, \sigma_b^2)$ for some $\sigma_b^2 \leq 1$ and W_b is independent of $\{X_m;\, m \in \mathrm{cl}_a\}$. Hence

$$\begin{aligned} G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda}) &= -2n^{-1} \sum_{m \in \mathrm{ne}_a} \theta_m^{b,\mathrm{ne}_a} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{X}_m \rangle \\ &\quad -2n^{-1} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle. \end{aligned}$$

By Lemma 2, with probability $1 - O(\exp(-cn^\varepsilon))$,

$$\mathrm{sign}(\hat{\theta}_k^{a,\mathrm{ne}_a,\lambda}) = \mathrm{sign}(\theta^{a,\mathrm{ne}_a}), \ \forall k \in \mathrm{ne}_a. \tag{35}$$

and in this case it holds by Lemma 1 that

$$2n^{-1} \sum_{m \in \mathrm{ne}_a} \theta_m^{b,\mathrm{ne}_a} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{X}_m \rangle = (\sum_{m \in \mathrm{ne}_a} \mathrm{sign}(\theta_m^{a,\mathrm{ne}_a}) \theta_m^{b,\mathrm{ne}_a}) \lambda.$$

If (35) holds true, the gradient is given by

$$\begin{aligned} G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda}) &= -(\sum_{m \in \mathrm{ne}_a} \mathrm{sign}(\theta_m^{a,\mathrm{ne}_a}) \theta_m^{b,\mathrm{ne}_a}) \lambda \\ &\quad -2n^{-1} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle. \end{aligned} \tag{36}$$

Using Assumption 6 and Proposition 2, there exists some $\delta < 1$ so that

$$|\sum_{m \in \mathrm{ne}_a} \mathrm{sign}(\theta_m^{a,\mathrm{ne}_a}) \theta_m^{b,\mathrm{ne}_a}| \leq \delta.$$

The absolute value of the coefficient G_b of the gradient in equation (32) is hence bounded with probability $1 - O(\exp(-cn^\varepsilon))$ by

$$|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \leq \delta\lambda + |2n^{-1} \langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle|. \tag{37}$$

Conditional on $\mathbf{X}_{\mathrm{cl}_a} = \{\mathbf{X}_k;\, k \in \mathrm{cl}_a\}$, the random variable

$$\langle \mathbf{X}_a - \mathbf{X}\hat{\theta}_k^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle$$

is normally distributed with mean zero and variance $\sigma_b^2 \|\mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}\|_2^2$. On the one hand, $\sigma_b^2 \le 1$. On the other hand, by definition of $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$,

$$\|\mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}\|_2 \le \|\mathbf{X}_a\|_2.$$

Thus

$$|2n^{-1}\langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle| \overset{st.}{\le} |2n^{-1}\langle \mathbf{X}_a, \mathbf{W}_b \rangle|,$$

where $\overset{st.}{\le}$ denotes stochastically smaller or equal. Using (37), it remains to be shown that there exists $c > 0$ so that for any $\delta < 1$,

$$P(|2n^{-1}\langle \mathbf{X}_a, \mathbf{W}_b \rangle| \ge (1-\delta)\lambda) = O(\exp(-cn^{\varepsilon})).$$

As W_b and X_a are independent, $E(X_a W_b) = 0$. Using the Gaussianity and bounded variance of both X_a and W_b, there exists some $g < \infty$ so that $E(\exp(|X_a W_b|)) \le g$. Hence, using Bernstein's inequality, see e.g. Lemma 2.2.11 in van der Vaart and Wellner (1996), and using the boundedness of λ, there exists indeed $c > 0$ so that for all $b \in \mathrm{ne}_a$, $P(|2n^{-1}\langle \mathbf{X}_a, \mathbf{W}_b \rangle| \ge (1-\delta)\lambda) = O(\exp(-cn\lambda^2))$. The claim (33) follows, which completes the proof. □

5.4. *Proof of Proposition 3.* Following a similar argument as in Theorem 1 up to equation (33), it is sufficient to show that for every a, b with $b \in \Gamma(n) \backslash \mathrm{cl}_a$ and $|S_a(b)| > 1$ it holds that

$$P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| > \lambda) \to 1 \qquad \text{for } n \to \infty. \tag{38}$$

Using (36) in the proof of Theorem 1, one can conclude that for some $\delta > 1$,

$$|G_b(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| \ge \delta\lambda - |2n^{-1}\langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle|. \tag{39}$$

Using the identical argument as in the proof of Theorem 1 below (37), it holds for the second term that for any $g > 0$,

$$P(|2n^{-1}\langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,\mathrm{ne}_a,\lambda}, \mathbf{W}_b \rangle| > g\lambda) \to 0 \qquad \text{for } n \to \infty,$$

which together with $\delta > 1$ in (39) shows that (38) holds. This completes the proof.

5.5. *Proof of Theorem 2.* First, $P(\mathrm{ne}_a \subseteq \hat{\mathrm{ne}}_a^\lambda) = 1 - P(\exists b \in \mathrm{ne}_a : \hat{\theta}_b^{a,\lambda} = 0)$. Let $\mathcal{E}$ be again the event

$$\max_{k \in \Gamma(n) \setminus \mathrm{cl}_a} |G_k(\hat{\theta}^{a,\mathrm{ne}_a,\lambda})| < \lambda. \tag{40}$$

Given $\mathcal{E}$, 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 (10) and (3) respectively, and $\hat{\theta}^{a,\mathrm{ne}_a,\lambda} = \hat{\theta}^{a,\lambda}$. Thus

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

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

$$P(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0) = O(\exp(-cn^\varepsilon)). \tag{41}$$

This follows from Lemma 2, which completes the proof.

5.6. *Proof of Proposition 4.* The proof of Proposition 4 is to a large extent analogous to the proof of Theorems 1 and 2. Let $\mathcal{E}$ be again the event (40). Given event $\mathcal{E}$, as in (40), we can conclude as before that $\hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ and $\hat{\theta}^{a,\lambda}$ are unique solutions to (10) and (3) respectively, and $\hat{\theta}^{a,\mathrm{ne}_a,\lambda} = \hat{\theta}^{a,\lambda}$. Thus, for any $b \in \mathrm{ne}_a$,

$$P(b \notin \hat{\mathrm{ne}}_a^\lambda) = P(\hat{\theta}_b^{a,\lambda} = 0) \ge P(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0) P(\mathcal{E}).$$

As $P(\mathcal{E}) \to 1$ for $n \to \infty$ by Theorem 1, it suffices to show that for all $b \in \mathrm{ne}_a$ with $|\pi_{ab}| = O(n^{-(1-\xi)/2})$ and $\xi < \varepsilon$,

$$P(\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0) \to 1 \qquad \text{for } n \to \infty$$

This is fulfilled if

$$P(|G_b(\hat{\theta}^{a,\mathrm{ne}_a \setminus \{b\},\lambda})| < \lambda) \to 1 \qquad \text{for } n \to \infty, \tag{42}$$

as $|G_b(\hat{\theta}^{a,\mathrm{ne}_a \setminus \{b\},\lambda})| < \lambda$ implies that $\hat{\theta}^{a,\mathrm{ne}_a \setminus \{b\},\lambda} = \hat{\theta}^{a,\mathrm{ne}_a,\lambda}$ and hence $\hat{\theta}_b^{a,\mathrm{ne}_a,\lambda} = 0$. Using (16), it holds that

$$|G_b(\hat{\theta}^{a,\mathrm{ne}_a \setminus \{b\},\lambda})| \le |2n^{-1} \langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b \rangle| + \lambda \|\theta^{b,\mathrm{ne}_a \setminus \{b\}}\|.$$

It holds by assumption that $\|\theta^{b,\mathrm{ne}_a\setminus\{b\}}\|_1 < 1$. It is thus sufficient to show that for any $g > 0$,

$$P(|2n^{-1}\langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b\rangle| < g\lambda) \to 1 \qquad \text{for } n \to \infty. \tag{43}$$

Analogous to (19), we can write

$$|2n^{-1}\langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b\rangle| \le |2n^{-1}\langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b^\perp\rangle| + |2n^{-1}\langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b^\|\rangle| \tag{44}$$

Using Lemma 3 it follows for the last term on the right hand side that for every $g > 0$,

$$P(|2n^{-1}\langle \mathbf{R}_a^\lambda(0), \mathbf{W}_b^\|\rangle| < g\lambda) \to 1 \qquad \text{for } n \to \infty.$$

Using (22) and (23), it hence remains to be shown that for every $g > 0$,

$$P(|2n^{-1}\theta_b^{a,\mathrm{ne}_a}\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp\rangle| < g\lambda) \to 1 \qquad \text{for } n \to \infty. \tag{45}$$

We have already seen above that the term $\sigma_b^{-2}\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp\rangle$ follows a $\chi^2_{n-|\mathrm{ne}_a|}$-distribution and is hence stochastically smaller than a χ^2_n-distributed random variable. By Assumption 2, $\sigma_b^2 \le 1$. Furthermore, using Assumption 2 and $|\pi_{ab}| = O(n^{-(1-\xi)/2})$, it holds that $|\theta_b^{a,\mathrm{ne}_a}| = O(n^{-(1-\xi)/2})$. Hence, with $\lambda \sim dn^{-(1-\varepsilon)/2}$, it follows that for some constant $k > 0$, $\lambda/|\theta_b^{a,\mathrm{ne}_a}| \ge kn^{(\varepsilon-\kappa)/2}$. Thus, for some constant $c > 0$,

$$P(|2n^{-1}\theta_b^{a,\mathrm{ne}_a}\langle \mathbf{W}_b^\perp, \mathbf{W}_b^\perp\rangle| < g\lambda) \ge P(Z/n > cn^{(\varepsilon-\xi)/2}), \tag{46}$$

where Z follows a χ^2_n distribution. By the properties of the χ^2-distribution and the assumption $\xi < \varepsilon$, the r.h.s. in (46) converges to 1 for $n \to \infty$, from which equation (45) and hence the claim follows.

5.7. *Proof of Proposition 1.* All diagonal elements of the covariance matrices $\Sigma(n)$ 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 $0 < s < 1$. 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 $\theta^a = (0, -K_{ab}/K_{aa}, 0, 0, \ldots) = (0, s, 0, 0, \ldots)$, where $K = \Sigma^{-1}(n)$. A necessary condition for $\widehat{\mathrm{ne}}_a^\lambda = \mathrm{ne}_a$ is that $\hat\theta^{a,\lambda} = (0, \tau, 0, 0, \ldots)$ is the oracle Lasso solution for some $\tau \ne 0$. In the following, we show first, that

$$P(\exists\lambda, \tau \ge s : \hat\theta^{a,\lambda} = (0, \tau, 0, 0, \ldots)) \to 0 \qquad n \to \infty. \tag{47}$$

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 (47). 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,

$$\langle \mathbf{X}_1 - \tau \mathbf{X}_2, \mathbf{X}_2 \rangle \geq |\langle \mathbf{X}_1 - \tau \mathbf{X}_2, \mathbf{X}_k \rangle| \qquad \forall k \in \Gamma(n), k > 2. \tag{48}$$

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 (48) yields for all $k \in \Gamma(n)$ with $k > 2$,

$$\langle \mathbf{W}_1, \mathbf{X}_2 \rangle - (\tau - s)\langle \mathbf{X}_2, \mathbf{X}_2 \rangle \geq |\langle \mathbf{W}_1, \mathbf{X}_k \rangle - (\tau - s)\langle \mathbf{X}_2, \mathbf{X}_k \rangle|.$$

Let $U_2, U_3, \ldots, U_{p(n)}$ be the random variables defined by $U_k = \langle \mathbf{W}_1, \mathbf{X}_k \rangle$. Note that the random variables U_k, $k = 2, \ldots, p(n)$, are exchangeable. Let furthermore

$$D = \langle \mathbf{X}_2, \mathbf{X}_2 \rangle - |\langle \mathbf{X}_2, \mathbf{X}_k \rangle|.$$

The inequality above implies then

$$U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D.$$

To show the claim, it thus suffices to show that

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \to 0 \qquad \text{for } n \to \infty. \tag{49}$$

Using $\tau - s > 0$,

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \leq P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) + P(D < 0).$$

Using the assumption that $s < 1$, there exists some $\epsilon > 0$ so that $E(D) > \epsilon$, and hence, using a Bernstein-type inequality,

$$P(D < 0) \to 0 \qquad \text{for } n \to \infty.$$

Furthermore, as $U_2, \ldots, U_{p(n)}$ are exchangeable,

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) = (p(n) - 1)^{-1} \to 0 \qquad \text{for } n \to \infty,$$

which shows that (49) holds. The claim (47) follows.

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, $\tau_{\max} < s$. For $\tau < \tau_{\max}$, the vector $(0, \tau, 0, \ldots)$ cannot be the oracle Lasso solution. We show in the following that $(0, \tau_{\max}, 0, \ldots)$ can not be an oracle Lasso solution either. Suppose that $(0, \tau_{\max}, 0, 0, \ldots)$ is the Lasso solution $\hat{\theta}^{a,\lambda}$ for some $\lambda = \tilde{\lambda} > 0$. As $\tau_{\max}$ is the maximal value such that $(0, \tau, 0, \ldots)$ is a Lasso solution, there exists some $k \in \Gamma(n) > 2$, such that

$$|n^{-1}\langle \mathbf{X}_1 - \tau_{\max}\mathbf{X}_2, \mathbf{X}_2\rangle| = |n^{-1}\langle \mathbf{X}_1 - \tau_{\max}\mathbf{X}_2, \mathbf{X}_k\rangle|,$$

and the value of both components G_2 and G_3 of the gradient is equal to $\tilde{\lambda}$. By appropriately reordering the variables we can assume that $k = 3$. Furthermore, it holds a.s. that

$$\max_{k\in\Gamma(n), k>3} |\langle \mathbf{X}_1 - \tau_{\max}\mathbf{X}_2, \mathbf{X}_k\rangle| < \tilde{\lambda}.$$

Hence, for sufficiently small $\delta\lambda \geq 0$, a Lasso solution for the penalty $\tilde{\lambda} - \delta\lambda$ is given by

$$(0, \tau_{\max} + \delta\theta_2, \delta\theta_3, 0, \ldots).$$

Let H_n be the empirical covariance matrix of (X_2, X_3). Assume w.l.og. that $n^{-1}\langle \mathbf{X}_1 - \tau_{\max}\mathbf{X}_2, \mathbf{X}_k\rangle > 0$ and $n^{-1}\langle \mathbf{X}_2, \mathbf{X}_2\rangle = n^{-1}\langle \mathbf{X}_3, \mathbf{X}_3\rangle = 1$. Following e.g. Efron et al. (2004), p. 417, the components $(\delta\theta_2, \delta\theta_3)$ are then given by $H_n^{-1}(1,1)^T$, from which it follows that $\delta\theta_2 = \delta\theta_3$, which we abbreviate by $\delta\theta$ in the following (one can accommodate a negative sign of $n^{-1}\langle \mathbf{X}_1 - \tau_{\max}\mathbf{X}_2, \mathbf{X}_k\rangle$ by reversing the sign of $\delta\theta_3$). Denote by L_δ the squared error loss for this solution. Then, for sufficiently small $\delta\theta$,

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

It holds that $L_{\delta\theta} - L_0 < 0$ for any $0 < \delta\theta < 1/2(s - \tau_{\max})$, which shows that $(0, \tau, 0, \ldots)$ cannot be the oracle solution for $\tau < s$. Together with (47), this completes the proof. □

5.8. *Proof of Theorem 3.* A necessary condition for $\hat{C}_a^\lambda \nsubseteq C_a$ is that there exists an edge in $\hat{E}^\lambda$ joining two nodes in two different connectivity components. Hence

$$P(\exists a \in \Gamma(n) : \hat{C}_a^\lambda \nsubseteq C_a) \leq p(n) \max_{a \in \Gamma(n)} P(\exists b \in \Gamma(n) \backslash C_a : b \in \widehat{\mathrm{ne}}_a^\lambda).$$

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

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

where $\hat{\theta}^{a,C_a,\lambda}$, according to (10), has non-zero components only for variables in the connectivity component C_a of node a. Hence it is sufficient to show that

$$p(n)^2 \max_{a \in \Gamma(n), b \in \Gamma(n) \backslash C_a} P(|G_b(\hat{\theta}^{a,C_a,\lambda})| \geq \lambda) \leq \alpha, \tag{50}$$

The gradient is given by $G_b(\hat{\theta}^{a,C_a,\lambda}) = -2n^{-1}\langle \mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,C_a,\lambda}, \mathbf{X}_b \rangle$. 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 $\mathbf{X}_{C_a} = \{\mathbf{X}_k; k \in C_a\}$,

$$G_b(\hat{\theta}^{a,C_a,\lambda}) \sim \mathcal{N}(0, R^2/n),$$

where $R^2 = 4n^{-1}\|\mathbf{X}_a - \mathbf{X}\hat{\theta}^{a,c_a,\lambda}\|_2^2$, which is smaller than or equal to $\hat{\sigma}_a^2 = 4n^{-1}\|\mathbf{X}_a\|_2^2$ 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

$$P(|G_b(\hat{\theta}^{a,C_a,\lambda})| \geq \lambda | \mathbf{X}_{C_a}) \leq 2\tilde{\Phi}(\sqrt{n}\lambda/(2\hat{\sigma}_a)),$$

where $\tilde{\Phi} = 1 - \Phi$. It follows for the λ proposed in (9) that $P(|G_b(\hat{\theta}^{a,C_a,\lambda})| \geq \lambda | \mathbf{X}_{C_a}) \leq \alpha p(n)^{-2}$, and therefore $P(|G_b(\hat{\theta}^{a,C_a,\lambda})| \geq \lambda) \leq \alpha p(n)^{-2}$. Thus (50) follows, which completes the proof. □

References.

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

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

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

Drton, M. and Perlman, M. (2004). Model selection for gaussian concentration graphs. *Biometrika* **91** 591–602.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–451.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

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

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

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. and Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research* **1** 49–75.

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

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

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

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

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, New York.

Osborne, M., Presnell, B. and Turlach, B. (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 the dimension 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 Kiiveri, H. (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 Horimoto, K. (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 Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag.

Seminar für Statistik, ETH Zürich,
CH-8092 Zürich
e-mail: nicolai@stat.math.ethz.ch