Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs

BY ALI SHOJAIE AND GEORGE MICHAILIDIS

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

shojaie@umich.edu gmichail@umich.edu

SUMMARY

Directed acyclic graphs are commonly used to represent causal relationships among random variables in graphical models. Applications of these models arise in the study of physical and biological systems where directed edges between nodes represent the influence of components of the system on each other. Estimation of directed graphs from observational data is computationally NP-hard. In addition, directed graphs with the same structure may be indistinguishable based on observations alone. When the nodes exhibit a natural ordering, the problem of estimating directed graphs reduces to the problem of estimating the structure of the network. In this paper, we propose an efficient penalized likelihood method for estimation of the adjacency matrix of directed acyclic graphs, when variables inherit a natural ordering. We study variable selection consistency of lasso and adaptive lasso penalties in high-dimensional sparse settings, and propose an error-based choice for selecting the tuning parameter. We show that although the lasso is only variable selection consistent under stringent conditions, the adaptive lasso can consistently estimate the true graph under the usual regularity assumptions.

Some key words: Adaptive lasso; Directed acyclic graph; High-dimensional sparse graphs; Lasso; Penalized likelihood estimation; Small n large p asymptotics.

1. INTRODUCTION

Graphical models are efficient tools for the study of statistical models through a compact representation of the joint probability distribution of the underlying random variables. The nodes of the graph represent the random variables, while the edges capture the relationships among them. Both directed and undirected edges are used to represent interactions among random variables. However, there is a conceptual difference between these two types of graphs: while undirected graphs are used to represent conditional dependence, directed graphs often represent causal relationships (see [Pearl, 2000](#)). Directed acyclic graphs, also known as Bayesian networks, are a special class of directed graphs, where all the edges of the graph are directed and the graph has no cycles. Such graphs are the main focus of this paper and unless otherwise specified, any reference to directed graphs below refers to directed acyclic graphs.

Directed graphs are used in graphical models and belief networks and have been the focus of research in the computer science literature ([Pearl, 2000](#)). Important applications involving directed graphs also arise in the study of biological systems, including cell signalling pathways and gene regulatory networks ([Markowitz & Spang, 2007](#)).

Estimation of directed acyclic graphs is an NP-hard problem, and estimation of the direction of edges may not be possible due to observational equivalence. Most of the earlier estimation methods involve greedy algorithms that search through the space of possible graphs. A number

of methods are available for estimating directed graphs with a small to moderate number of nodes. The max-min hill-climbing algorithm (Tsamardinos et al., 2006), and the PC-algorithm (Spirtes et al., 2000) are two such examples. However, the space of directed graphs grows super-exponentially with the number of nodes (Robinson, 1977), and estimation using search-based methods, especially in high-dimensional settings, becomes impractical. Bayesian methods (e.g. Heckerman et al., 1995) are also computationally very intensive and therefore not particularly appropriate for high-dimensional settings. Recently, Kalisch & Bühlmann (2007) proposed an implementation of the PC-algorithm with polynomial complexity in high-dimensional sparse settings. When the variables inherit a natural ordering, estimation of directed graphs is reduced to estimating their structure or skeleton. Applications with a natural ordering of variables include estimation of causal relationships from temporal observations, estimation of transcriptional regulatory networks from gene expression data and settings where additional experimental data can determine the ordering of variables. Examples of such applications are given in § 6.

For Gaussian random variables, conditional independence relations among random variables are represented using an undirected graph, known as the conditional independence graph. The edges of this graph represent conditional dependencies among random variables, and correspond to nonzero elements of the inverse covariance matrix, also known as the precision matrix. Different penalization methods have been recently proposed to obtain sparse estimates of the precision matrix. Meinshausen & Bühlmann (2006) considered an approximation to the problem of sparse inverse covariance estimation using the lasso penalty. They showed, under appropriate assumptions, that their proposed method correctly determines the neighbourhood of each node. Banerjee et al. (2008) and Friedman et al. (2008) explored different aspects of the problem of estimating the precision matrix using the lasso penalty, while Yuan & Lin (2007) considered other choices for the penalty. Rothman et al. (2008) proved consistency in Frobenius norm, as well as in matrix norm, of the ℓ_1 -regularized estimate of the precision matrix when $p \gg n$, while Lam & Fan (2009) extended their result and considered estimation of matrices related to the precision matrix, including the Cholesky factor of the inverse covariance matrix, using general penalties. Penalization of the Cholesky factor of the inverse covariance matrix has also been considered by Huang et al. (2006) and Levina et al. (2008), who used the lasso penalty in order to obtain a sparse estimate of the inverse covariance matrix. This method is based on the regression interpretation of the Cholesky factorization model and therefore requires the variables to be ordered a priori.

In this paper, we consider the problem of estimating the skeleton of directed acyclic graphs, where the variables exhibit a natural ordering. The known ordering of variables is exploited to reformulate the likelihood as a function of the adjacency matrix of the graph, which results in an efficient algorithm for estimation of structure of directed graphs using penalized likelihood methods. Although the results of this paper are presented for the case of Gaussian observations, the proposed method can also be applied to non-Gaussian observations, provided the underlying causal mechanisms in the network are linear.

2. REPRESENTATION OF DIRECTED ACYCLIC GRAPHS

2.1. Background and notation

Consider a graph $\mathcal{G} = (V, E)$, where V corresponds to the set of nodes with p elements and $E \subset V \times V$ to the edge set. The nodes of the graph represent random variables $X_1, \dots, X_p$ and the edges capture the associations among them. An edge is called directed if $(j, i) \notin E$ whenever $(i, j) \in E$, and undirected when $(i, j) \in E$ if and only if $(j, i) \in E$. We denote by pa_i the set of parents of node i and for $j \in \text{pa}_i$, we denote $j \rightarrow i$. The skeleton of a directed graph is the

Fig. 1. Directed graphs. (a) A simple directed graph; and (b), an illustration of observational equivalence in directed graphs.

undirected graph that is obtained by replacing directed edges in E with undirected ones. We represent E using the adjacency matrix A of the graph; i.e. a $p \times p$ matrix whose (j, i) th entry is nonzero if there is an edge between nodes j and i .

The estimation of directed graphs is a challenging problem due to the so-called observational equivalence with respect to the same probability distribution. More specifically, regardless of the sample size, it may not be possible to infer the direction of causation among random variables from observational data. As an illustration, consider the simple graph in Fig. 1(a). Reversing the direction of all edges of the graph results in a new graph, which is isomorphic to the original graph, and hence not distinguishable from observations alone.

The second challenge in estimating directed graphs is that conditional independence among random variables may not reveal the skeleton. The notion of conditional independence in directed graphs is represented using the concept of d-separation (Pearl, 2000), or equivalently, the moral graph, obtained by removing the directions of the edges and joining the parents of each node (Lauritzen, 1996). Therefore, estimation of the conditional independence structure reveals the structure of the moral graph, which includes additional edges between parents of each node. For instance, X_2 and X_3 are connected in the moral graph of the graph in Fig. 1(b).

2.2. The latent variable model

The causal effect of random variables in a directed acyclic graph is often explained using structural equation models (Pearl, 2000). In particular,

$$X_i = f_i(\text{pa}_i, Z_i) \quad (i = 1, \dots, p), \quad (1)$$

where the random variables Z_i are the latent variables representing the unexplained variation in each node. To model the association among the nodes, we consider a simplification of (1) with linear f_i . More specifically, let ρ_{ij} represent the effect of node j on i for $j \in \text{pa}_i$. Then

$$X_i = \sum_{j \in \text{pa}_i} \rho_{ij} X_j + Z_i \quad (i = 1, \dots, p). \quad (2)$$

If the random variables are Gaussian, equations (1) and (2) are equivalent, in the sense that ρ_{ij} are the coefficients of the linear regression model of X_i on X_j , for $j \in \text{pa}_i$. It is known in the normal case that $\rho_{ij} = 0$, if $j \notin \text{pa}_i$.

Consider the graph in Fig. 1(a); denoting the influence matrix of the graph by Λ , (2) can be written in compact form as $X = \Lambda Z$, where for the example above, we have

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}.$$

Let the latent variables Z_i be independent with mean μ_i and variance σ_i^2 . Then, $E(X) = \Lambda\mu$ and $\Sigma = \text{var}(X) = \Lambda D \Lambda^\top$, where $D = \text{diag}(\sigma_i^2)$ and $\Lambda^\top$ denotes the transpose of the matrix Λ .

The following result from [Shojaie & Michailidis \(2009\)](#) establishes the relationship between the influence matrix Λ , and the adjacency matrix A of the graph. The second part of the lemma establishes a compact relationship between Λ and A in the case of directed graphs, which is explored in § 3 to directly formulate the problem of estimating the skeleton of the graph.

LEMMA 1. *For any graph $\mathcal{G}$, with the adjacency matrix A , and influence matrix Λ ,*

- (i) $\Lambda = A^0 + A^1 + A^2 + \cdots = \sum_{r=0}^{\infty} A^r$, where $A^0 \equiv I$.
- (ii) *If $\mathcal{G}$ is a directed acyclic graph, then Λ has full rank and $\Lambda = (I - A)^{-1}$.*

Remark 1. Part (ii) of Lemma 1 and the fact that $\Sigma = \Lambda D \Lambda^\top$ imply that for any directed acyclic graph, if $D_{ii} > 0$ for all i , then Σ is full rank. More specifically, let $\phi_j(\Sigma)$ denote the j th eigenvalue of matrix Σ . Then, $\phi_{\min}(\Sigma) > 0$, or $\phi_{\max}(\Sigma^{-1}) < \infty$. Similarly, since $\Sigma^{-1} = \Lambda^{-\top} D^{-1} \Lambda^{-1}$, then Λ having full rank implies that $\phi_{\min}(\Sigma^{-1}) > 0$, or equivalently $\phi_{\max}(\Sigma) < \infty$. This result also applies to all subgraphs of a directed graph.

The properties of the proposed latent variable model established in Lemma 1 are independent of the choice of probability distribution. In fact, since the latent variables Z_i in (2) are assumed independent, given the entries of the adjacency matrix, the distribution of each random variable X_i in the graph only depends on the values of pa_i . Hence, regardless of the choice of the probability distribution, the joint distribution of the random variables is compatible with $\mathcal{G}$ ([Pearl, 2000](#), p. 16). Therefore, based on the equivalence of conditional independence and d-separation, if the joint probability distribution of random variables on a directed graph is generated according to the latent variable model (2), zeros of the adjacency matrix, A , determine conditional independence relations among random variables. As mentioned before, under the normality assumption, the latent variable model is equivalent to the general structural equation model. Although we focus on Gaussian random variables in the remainder of this paper, the estimation procedure proposed in § 3 can be applied to a variety of other distributions, if one is willing to assume the linear structure in equation (2).

3. PENALIZED LIKELIHOOD ESTIMATION OF DIRECTED GRAPHS

3.1. Problem formulation

Consider the latent variable model of § 2.2 and denote by $\mathcal{X}$ the $n \times p$ data matrix. We assume, without loss of generality, that the X_i s are centred and scaled, so that $\mu_i = 0$ and $\sigma_i^2 = 1$ ($i = 1, \dots, p$).

Denote by $\Omega \equiv \Sigma^{-1}$ the precision matrix of a p -vector of Gaussian random variables and consider a general penalty function $J(\Omega)$. The penalized estimate of Ω is then given by

$$\hat{\Omega} = \arg \min_{\Omega > 0} \{-\log \det(\Omega) + \text{tr}(\Omega S) + \lambda J(\Omega)\}, \quad (3)$$

where $S = n^{-1} \mathcal{X}^\top \mathcal{X}$ denotes the empirical covariance matrix and λ is the tuning parameter controlling the size of the penalty. Applications in biological and social networks often involve sparse networks. It is therefore desirable to find a sparse solution for (3). This becomes more important in the small n , large p setting, where the unpenalized solution is unreliable. The lasso penalty and the adaptive lasso penalty ([Zou, 2006](#)) are singular at zero and therefore result in sparse solutions. We consider these two penalties in order to find a sparse estimate of the adjacency

matrix. However, the optimization algorithm proposed here can also be used with other choices of penalty function, if the penalty is applied to each individual component of the adjacency matrix.

Using the latent variable model of § 2.2, and the relationship between the covariance matrix and the adjacency matrix of directed graphs established in Lemma 1, the problem of estimating the adjacency matrix of the graph can be directly formulated as an optimization problem based on A . Specifically, if the underlying graph is directed, and the ordering of the variables is known, then, A is a lower triangular matrix with zeros on the diagonal. Let $\mathcal{A} = \{A : A_{ij} = 0, j \geq i\}$. Then, using the facts that $\det(A) = 1$ and $\sigma_i^2 = 1$, A can be estimated by:

$$\hat{A} = \arg \min_{A \in \mathcal{A}} [\text{tr} \{(I - A)^T(I - A)S\} + \lambda J(A)]. \quad (4)$$

In this paper, we consider the general weighted lasso problem, where

$$J(A) = \sum_{i,j=1:p, j < i} w_{ij} |A_{ij}|.$$

The lasso and adaptive lasso problems are special cases of this general penalty. In the case of the lasso, $w_{ij} = 1$. In the original proposal of Zou (2006), the weights for the adaptive lasso are obtained by setting $w_{ij} = |\tilde{A}_{ij}|^{-\gamma}$, for some initial estimate of the adjacency matrix $\tilde{A}$ and some power γ . We consider the following modification of the original weights:

$$w_{ij} = 1 \vee |\tilde{A}_{ij}|^{-\gamma}, \quad (5)$$

where the initial estimates $\tilde{A}$ are obtained from the regular lasso estimates, and $x \vee y$ represents the maximum of x and y . Aside from the truncation of weights from below, which is implemented to facilitate the study of asymptotic properties, the main difference between the adaptive lasso penalty using (5) and the proposal of Zou (2006) is the use of the lasso estimates to construct the weights. In §§ 4 and 5, we show that this modification, which could also be considered a two-stage or iterated lasso penalty, results in improvements over the regular lasso penalty in terms of both asymptotic properties and numerical performance.

The objective functions for the lasso and adaptive lasso problems are convex. However, since the ℓ_1 penalty is nondifferentiable, these problems can be reformulated using matrices $A_+ = \max(A, 0)$ and $A_- = -\min(A, 0)$. To that end, let W be the $p \times p$ matrix of weights for the adaptive lasso, or the matrix of ones for the lasso estimation problem. Problem (4) can then be formulated as

$$\min_{A_+, A_- \geq 0} \text{tr} \{S(I - A_+ + A_-)^T(I - A_+ + A_-) + \lambda(A_+ + A_-)W + \Delta(A_+ + A_-)1_{u^+}\}, \quad (6)$$

where ≥ 0 is interpreted componentwise, Δ is a large positive number and 1_{u^+} is the indicator matrix for upper triangular elements of a $p \times p$ matrix, including the diagonal elements. The last term of the objective function, $\text{tr} \{\Delta(A_+ + A_-)1_{u^+}\}$, prevents the upper triangular elements of the matrices A_+ and A_- being nonzero.

Problem (6) corresponds to a quadratic optimization problem with nonnegativity constraints and can be solved using standard interior point algorithms. However, such algorithms do not scale well with dimension and are only applicable if p ranges in the hundreds. In § 3.2, we present an alternative formulation of the problem, which leads to more efficient algorithms.

3.2. Optimization algorithm

Consider again the problem of estimating the adjacency matrix of directed graphs with the general weighted lasso penalty. Let $\mathbf{a}_i$ be the i th row of matrix A , and denote by I' the set of

indices up to $l - 1$, i.e. $l^- = \{j : j = 1, \dots, l - 1\}$. Then (4) can be written as

$$\hat{A} = \arg \min_{A \in \mathcal{A}} \left\{ \sum_{i=1}^p (\mathbf{a}_i S \mathbf{a}_i^\top - 2 \mathbf{a}_i S_i + \lambda |\mathbf{a}_i| W_i) \right\}. \quad (7)$$

It can be seen that the objective function in (7) is separable and therefore it suffices to solve the optimization problem over each row of matrix A . Then, taking advantage of the lower triangular structure of A and noting that $A_{11} = 0$, solving (7) is equivalent to solving the $p - 1$ optimization problems:

$$\hat{A}_{i,i^-} = \arg \min_{\theta \in \mathbb{R}^{i-1}} \left\{ \theta^\top S_{i^-,i^-} \theta - 2 S_{i,i^-} \theta + \lambda \sum_{j=1}^{i-1} |\theta_j| w_{ij} \right\} \quad (i = 2, \dots, p). \quad (8)$$

In addition, $S_{i^-,i^-} = n^{-1} (\mathcal{X}_{i^-})^\top \mathcal{X}_{i^-}$ and $S_{i,i^-} = n^{-1} (\mathcal{X}_i)^\top \mathcal{X}_{i^-}$, and hence the problem in (8) can be reformulated as the following ℓ_1 -regularized least-squares problems:

$$\hat{A}_{i,i^-} = \arg \min_{\theta \in \mathbb{R}^{i-1}} \left\{ n^{-1} \|\mathcal{X}_i - \mathcal{X}_{i^-} \theta\|_2^2 + \lambda \sum_{j=1}^{i-1} |\theta_j| w_{ij} \right\} \quad (i = 2, \dots, p). \quad (9)$$

The formulation in (9) indicates that the i th row of matrix A includes the coefficient of projection of X_i on X_j ($j = 1, \dots, i - 1$), in agreement with the discussion in § 2.2. It also reveals a connection between estimation of the underlying graphs and the neighbourhood selection approach of Meinshausen & Bühlmann (2006): when the underlying graph is directed, the approximate solution of the neighbourhood selection problem is exact, if the regression model is fitted on the set of parents of each node instead of all other nodes in the graph.

Using (9), the problem of estimating directed graphs can be solved very efficiently. In fact, it suffices to solve $p - 1$ lasso problems for estimation of least-squares coefficients, with dimensions ranging from 1 to $p - 1$. To solve these problems, we use the efficient pathwise coordinate optimization algorithm of Friedman et al. (2007), implemented in the R-package glmnet (R Development Core Team, 2010). The proposed procedure is summarized in Algorithm 1.

Algorithm 1. Penalized likelihood estimation of directed graphs.

- (1) Given the ordering $\mathcal{O}$, order the columns of observation matrix $\mathcal{X}$ in increasing order.
- (2) For $i = 2, 3, \dots, p$,
 - 2.1. Denote $y = \mathcal{X}_i$, $X = \mathcal{X}_{i^-}$. Given the weight matrix W , let $w = W_{i,i^-}$, and solve

$$\hat{A}_{i,i^-} = \arg \min \left\{ n^{-1} \|y - X\theta\|_2^2 + \lambda_i \sum_{j=1}^{i-1} |\theta_j| w_j \right\}$$

3.3. Analysis of computational complexity

As mentioned in § 1, the space of all possible directed graphs is super-exponential in the number of nodes and hence it is not surprising that the PC-algorithm, without any restriction on the space, has exponential complexity. Kalisch & Bühlmann (2007) recently proposed an efficient implementation of the PC-algorithm for sparse graphs; its complexity when the maximal neighbourhood size q is small, is bounded with high probability by $O(p^q)$. Although this is a considerable improvement over the original algorithm, in many applications it can become fairly expensive. For instance, gene regulatory networks and signaling pathways include many hub genes, which lead to large values for q .

The reformulation of the directed graph estimation problem in (9) requires solving $p - 1$ lasso regression problems. The cost of solving a lasso problem comprised of k covariates and n observations using the pathwise coordinate optimization algorithm is $O(nk)$; hence, the total cost of estimating the adjacency matrix of the graph is $O(np^2)$, which is the same as the cost of calculating the empirical covariance matrix S . Moreover, the formulation in (9) includes a set of non-overlapping subproblems. Therefore, for problems with very large number of nodes and/or observations, the performance of the algorithm can be further improved by parallelizing the estimation of these subproblems. The adaptive lasso version of the problem is similarly solved using the modification of the regular lasso problem proposed in Zou (2006), which results in the same computational cost as the regular lasso problem.

To evaluate the performance of these algorithms, we compared the average CPU time, over 10 simulation runs, for estimation of directed graphs with different numbers of nodes, $p = 100, 1000$, and different sample sizes, $n = 100, 1000$. The computational time for the PC-algorithm increases with larger values of the average neighbourhood size and the significance level α . Therefore, to control the computational complexity of the PC-algorithm, these parameters are set to 5 and 0.01, respectively. In addition, in order to compare equivalent quantities, we only consider the CPU time that the PC-algorithm requires for estimation of the graph skeleton. Based on this simulation study, while for $p = 100$, the computation time for the PC-algorithm is comparable to the time for the penalized likelihood algorithm, in a graph with $p = 1000$ and $n = 1000$, the average CPU time for the PC-algorithm could be up to two orders of magnitude larger than the equivalent time for Algorithm 1.

4. ASYMPTOTIC PROPERTIES

4.1. Preliminaries

Next, we establish theoretical properties of the lasso and adaptive lasso estimates of the adjacency matrix of directed graphs. Asymptotic properties of the lasso-type estimates with fixed design matrices have been studied by a number of researchers (Knight & Fu, 2000; Zou, 2006; Huang et al., 2008), while random design matrices have been considered by Meinshausen & Bühlmann (2006). Rothman et al. (2008) and Lam & Fan (2009), among others, have studied asymptotic properties of estimates of covariance and precision matrices.

As discussed in §3.2, the problem of estimating the adjacency matrix of a directed graph is equivalent to solving the $p - 1$ non-overlapping penalized least-squares problems described in (9). In order to study the asymptotic properties of the proposed estimators, we focus on the asymptotic consistency of network estimation, i.e. the probability of correctly estimating the network structure, in terms of Type I and Type II errors. We allow the total number of nodes in the graph to grow as an arbitrary polynomial function of the sample size, while assuming that the true underlying network is sparse.

4.2. Assumptions

Let $X = (X_1, \dots, X_p)$ be a collection of p zero-mean Gaussian random variables with covariance matrix Σ , and let $\mathcal{X}$ and S be defined as in §3.1. To simplify the notation, denote by $\theta^i = A_{i,-}$ the entries of the i th row of A to the left of the diagonal. Further, let $\theta^{i,\mathcal{I}}$ be the estimate for the i th row, with values outside the set of indices $\mathcal{I}$ set to zero; i.e. $\theta^{i,\mathcal{I}} \equiv A_{i,-}$ and $A_{i,j} = 0$, $j \notin \mathcal{I}$.

The following assumptions are used in establishing the consistency of network estimation.

Assumption 1. For some $a > 0$, $p = p(n) = O(n^a)$ as $n \rightarrow \infty$, and there exists a $0 \leq b < 1$ such that $\max_{i \in V} \text{card}(\text{pa}_i) = O(n^b)$ as $n \rightarrow \infty$.

Assumption 2. There exists $\nu > 0$ such that for all $n \in \mathbb{N}$ and all $i \in V$, $\text{var}(X_i | X_{i^-}) \geq \nu$.

Assumption 3. There exists $\delta > 0$ and some $\xi > b$, with b defined above, such that for all $i \in V$ and for every $j \in \text{pa}_i$, $|\pi_{ij}| \geq \delta n^{-(1-\xi)/2}$, where π_{ij} is the partial correlation between X_i and X_j after removing the effect of the remaining variables.

Assumption 4. There exists $\Psi < \infty$ such that for all $n \in \mathbb{N}$ and every $i \in V$ and $j \in \text{pa}_i$, $\|\theta^{j, \text{pa}_i}\|_2 \leq \Psi$.

Assumption 5. There exists $\kappa < 1$ such that for all $i \in V$ and $j \notin \text{pa}_i$, $|\sum_{k \in \text{pa}_i} \text{sign}(\theta_k^{i, \text{pa}_i}) \theta_k^{j, \text{pa}_i}| < \kappa$.

Assumption 1 determines the permissible rates of increase in the number of variables and the neighbourhood size, as a function of n , Assumption 2 prevents singular or near-singular covariance matrices, and Assumption 3 guarantees that true partial correlations are bounded away from zero.

Assumption 4 limits the magnitude of the shared ancestral effect between each node in the network and any of its parents. This is less restrictive than the equivalent assumption for the neighbourhood selection problem, where the effects over all neighbouring nodes are assumed to be bounded. In fact, in the case of gene regulatory networks, empirical data indicate that the average number of upstream-regulators per gene is less than two (Leclerc, 2008). Thus, the number of parents of each node is small, while each hub node can affect many downstream nodes.

Assumption 5 is referred to as neighbourhood stability and is equivalent to the irrerepresentability assumption of Huang et al. (2008). It has been shown that the lasso estimates are not in general variable selection consistent if this assumption is violated. Huang et al. (2008) considered the adaptive lasso estimates with general initial weights and proved their variable selection consistency under a weaker form of irrerepresentability assumption, referred to as adaptive irrerepresentability. We will show that when the initial weights for the adaptive lasso are derived from the regular lasso estimates as in (5), the assumption of neighbourhood stability, and the less stringent Assumption 4 are not required for establishing variable selection consistency of the adaptive lasso. This relaxation in assumptions required for variable selection consistency, is a result of the consistency of the regular lasso estimates, and the special structure of directed graphs. Similar results can be obtained for the adaptive lasso estimates of the precision matrix, as well as regression models with fixed and random design matrices, under additional mild assumptions.

4.3. Asymptotic consistency of directed graph estimation

Our first result studies the variable selection consistency of the lasso estimate. In this theorem, (i) corresponds to sign consistency, (ii) and (iii) establish control of Type I and Type II errors and (iv) addresses the consistency of network estimation. We denote by $\hat{E}$ the estimate of the edge set of the graph, and by $x \sim y$ the asymptotic equivalence between x and y .

THEOREM 1. Suppose that Assumption 1–5 hold and $\lambda \sim dn^{-(1-\zeta)/2}$ for some $b < \zeta < \xi$ and $d > 0$. Then for the lasso estimate there exist constants $c_{(i)}, \dots, c_{(iv)} > 0$ such that for all $i \in V$, as $n \rightarrow \infty$,

$$(i) \text{ pr}\{\text{sign}(\hat{\theta}_j^{i, \text{pa}_i}) = \text{sign}(\theta_j^{i, \text{pa}_i}) \text{ for all } j \in \text{pa}_i\} = 1 - O\{\exp(-c_{(i)}n^\zeta)\},$$

- (ii) $\text{pr}(\hat{\text{pa}}_i \subseteq \text{pa}_i) = 1 - O\{\exp(-c_{(ii)}n^\zeta)\},$
- (iii) $\text{pr}(\text{pa}_i \subseteq \hat{\text{pa}}_i) = 1 - O\{\exp(-c_{(iii)}n^\zeta)\}$ and
- (iv) $\text{pr}(\hat{E} = E) = 1 - O\{\exp(-c_{(iv)}n^\zeta)\}.$

Proof. The proof of this theorem follows from arguments similar to those presented in Meinshausen & Bühlmann (2006) with minor modifications and replacing conditional independence for undirected graphs with d-separation for directed graphs, and hence is omitted. $\square$

The next result establishes similar properties for the adaptive lasso estimates, without the assumptions of neighbourhood stability. The proof of Theorem 3 makes use of the consistency of sparse estimates of the Cholesky factor of covariance matrices, established in Theorem 9 of Lam & Fan (2009). For completeness, we restate a simplified version of the theorem for our lasso problem, for which $\sigma_i = 1$ ($i = 1, \dots, p$), and the eigenvalues of the covariance matrix are bounded; see Remark 1. Throughout this section, we denote by s the total number of nonzero elements of the true adjacency matrix A .

THEOREM 2 (Lam & Fan 2009). *If $n^{-1}(s+1)\log p = o(1)$ and $\lambda = O\{(\log p/n)^{1/2}\}$, then $\|\hat{A} - A\|_F = O_p\{(n^{-1}s \log p)^{1/2}\}.$*

It can be seen from Theorem 2 that the lasso estimates are consistent as long as $n^{-1}(s+1)\log p = o(1)$. To take advantage of this result, we replace Assumption 1 with the following assumption.

Assumption 1'. For some $a > 0$, $p = p(n) = O(n^a)$ as $n \rightarrow \infty$. Also, $\max_{i \in V} \text{card}(\text{pa}_i) = O(n^b)$ as $n \rightarrow \infty$, where $sn^{2b-1} \log n = o(1)$ as $n \rightarrow \infty$.

Assumption 1' further restricts the number of parents of each node and also enforces a restriction on the total number of nonzero elements of the adjacency matrix. Condition $sn^{2b-1} \log n = o(1)$ implies that $b < 1/2$. Therefore, although the consistency of the adaptive lasso in Theorem 3 is established without making any further assumptions on the structure of the network, it is achieved at the price of requiring a higher degree of sparsity in the network.

THEOREM 3. *Consider the adaptive lasso estimation problem, where the initial weights are calculated using the regular lasso estimates of the adjacency matrix of the graph in (9). Suppose Assumptions 1', 2 and 3 hold and $\lambda \sim dn^{-(1-\zeta)/2}$ for some $b < \zeta < \xi$ and $d > 0$. Also suppose that the initial lasso estimates are found using a penalty parameter λ^0 that satisfies $\lambda^0 = O\{(\log p/n)^{1/2}\}$. Then there exist constants $c_{(i)}, \dots, c_{(iv)} > 0$ such that for all $i \in V$, as $n \rightarrow \infty$, (i)–(iv) in Theorem 1 hold.*

Proof. See the Appendix. $\square$

4.4. Choice of the tuning parameter

Both lasso and adaptive lasso estimates of the adjacency matrix depend on the choice of the tuning parameter λ . Different methods have been proposed for selecting the value of the tuning parameter, including crossvalidation (Rothman et al., 2008) and the Bayesian information criterion (Yuan & Lin, 2007). However, choices of λ that result in the optimal classification error do not guarantee a small error for network reconstruction. We propose next a choice of λ for the general weighted lasso problem with weights w_{ij} . Let Z_q^* denote the $(1-q)$ th quantile of standard normal distribution, and define

$$\lambda_i(\alpha) = 2n^{-1/2} Z_{\alpha/\{2p(i-1)\}}^* \quad (10)$$

The following result establishes that such a choice controls the probability of falsely joining two distinct ancestral sets, defined next.

DEFINITION 1. *For every node $i \in V$, the ancestral set of node i , AN_i , consists of all nodes j , such that j is an ancestor of i or i is an ancestor of j or i and j have a common ancestor k .*

THEOREM 4. *Under the assumptions of Theorems 1 and 3 above, for the lasso and adaptive lasso, respectively, for all $n \in \mathbb{N}$ the solution of the general weighted lasso estimation problem with tuning parameter determined in (10) satisfies*

$$\text{pr}(\text{there exists } i \in V : \hat{AN}_i \not\subseteq AN_i) \leq \alpha.$$

Proof. See the Appendix. □

Theorem 4 is true for all values of p and n , but it does not provide any guarantee for the probability of false positive error for individual edges in the graph. We also need to determine the optimal choice of penalty parameter λ^0 for the first phase of the adaptive lasso, where the weights are estimated using the lasso. Since the goal of the first phase is to achieve prediction consistency, crossvalidation can be used to determine the optimal choice of λ^0 . On the other hand, it is easy to see that the error-based proposal in (10) satisfies the requirement of Theorem 2 and can therefore be used to define λ^0 . It is, however, recommended to use a higher value of significance level in estimating the initial weights, in order to prevent an oversparse solution.

5. PERFORMANCE ANALYSIS

5.1. Preliminaries

In this section, we consider examples of estimating directed graphs of a varying number of edges from randomly generated data. To randomly generate data, one needs to generate lower-triangular adjacency matrices with sparse nonzero elements, ρ_{ij} . In order to control the computational complexity of the PC-algorithm, we use the random directed graph generator in the R-package pcalg (Kalisch & Bühlmann, 2007), which generates graphs with given values of the average neighbourhood size. The sparsity levels of graphs with different sizes are set according to the theoretical bounds in § 4, as well as the recommendations of Kalisch & Bühlmann (2007) for the neighbourhood size. We use an average neighbourhood size of five, while limiting the total number of true edges to be equal to the sample size n .

Different measures of structural difference can be used to evaluate the performance of estimators. The structural Hamming distance represents the number of edges that are not in common between the estimated and true graphs, i.e. $\text{SHD} = \text{card}(\hat{E} \setminus E) + \text{card}(E \setminus \hat{E})$, where $\hat{E}$ and E are defined as in Theorem 1. The main drawback of this measure is its dependency on the number of nodes, as well as the sparsity of the network. The second measure of goodness of estimation considered here is the Matthew's correlation coefficient, which is commonly used to assess the performance of binary classification methods, and is defined as

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})\}^{1/2}}, \quad (11)$$

where TP, TN, FP and FN denote the total number of true positive, true negative, false positive and false negative edges, respectively. The value of Matthew's correlation coefficient ranges from -1 to 1 with larger values corresponding to better fits, and -1 and 1 representing worst and best fits, respectively. Finally, in order to compare the performance of different estimation methods with theoretical bounds established in § 4.3, we also report the values of true and false positive rates.

Fig. 2. True directed graph along with estimates from Gaussian observations. The greyscale represents the percentage of inclusion of edges.

The performances of the PC-algorithm and our proposed estimators based on the choice of tuning parameter in (10), vary with different values of significance level α . In the following experiments, we first investigate the appropriate choice of α for each estimator. We then compare the performance of the estimators with an optimal choice of α . The results reported in this section are based on estimates obtained from 100 replications; to offset the effect of numerical instability, we consider an edge present only if $|\hat{A}_{ij}| > 10^{-4}$.

5.2. Estimation of directed graphs from normally distributed observations

We begin with an example that illustrates the differences between estimation of directed graphs and conditional independence graphs. The first two images in Fig. 2 represent a randomly generated directed graph of size $p = 50$ and the greyscale image of the average precision matrix estimated based on a sample of size $n = 100$ using the graphical lasso algorithm (Friedman et al., 2008). The image is obtained by calculating the proportion of times that a specific edge is present in 100 replications. To control the probability of falsely connecting two components of the graph, the value of the tuning parameter for the graphical lasso is defined based on the error-based proposal in Banerjee et al. (2008, Theorem 2). It can be seen that the conditional independence graph has many more edges, 8% false positive rate compared to 1% for the lasso and adaptive lasso, and does not reveal the true structure of the underlying directed graph. Therefore, although methods of estimating conditional independence graphs are computationally efficient, they should not be used in applications, such as estimation of gene regulatory networks, where the underlying graph is directed.

In simulations throughout this section, the sample size is fixed at $n = 100$, and estimators are evaluated for an increasing number of nodes, $p = 50, 100, 200$. Figure 3 shows the mean and standard deviation of the Hamming distances, expressed in base 10 logarithmic scale, for estimates based on the PC-algorithm, as well as the proposed lasso, and adaptive lasso methods for different values of the tuning parameter α and different network sizes. It can be seen that for all values of p and α , the adaptive lasso estimate produces the best results, and the proposed penalized likelihood methods outperform the PC-algorithm. This difference becomes more significant as the size of the network increases.

As mentioned in § 2, it is not always possible to estimate the direction of the edges of a directed graph and therefore, the estimate from the PC-algorithm may include undirected edges. Since our penalized likelihood methods assume knowledge of the ordering of variables and estimate the structure of the network, in the simulations considered here, we only estimate the skeleton of the network using the PC-algorithm. We then use the ordering of the variables to determine the direction of the edges. The performance of the PC-algorithm for estimation of partially completed directed graphs may therefore be worse than the results reported here.

Fig. 3. Logarithm, in base 10, of the Hamming distances for estimation of directed graphs using the PC-algorithm (black solid), lasso (black dashes) and adaptive lasso (grey dot-dashes) from normal observations.

Fig. 4. Matthew's correlation coefficient, true and false positives for estimation of directed graphs using the PC-algorithm (black solid), lasso (black dashes) and adaptive lasso (grey dot-dashes) from normal observations.

In our simulation results, observations are generated according to the linear structural equation model (2) with standard normal latent variables and $\rho_{ij} = \rho = 0.8$. Additional simulation studies with different values of σ and ρ indicate that changes in σ do not have a considerable effect on the performance of the proposed models. As the magnitude of ρ decreases, the performance of the proposed methods, as well as the PC-algorithm deteriorates, but the findings remain unchanged.

The above simulation results suggest that the optimal performance of the PC-algorithm is achieved when $\alpha = 0.01$. The performance of the lasso and adaptive lasso methods is less sensitive to the choice of α ; however, a value of $\alpha = 0.10$ seems to deliver more reliable estimates. Our extended simulations indicate that the performance of the adaptive lasso does not vary significantly with the value of power γ and therefore we present the results for $\gamma = 1$.

Figure 2 represents images of estimated and true directed graphs created based on the above considerations for tuning parameters for $p = 50$. Similar results were also observed for larger values of p , $p = 100, 200$, and are excluded due to space considerations. Plots in Fig. 4 compare the performance of the three methods with the optimal settings of tuning parameters, over a range of values of p . It can be seen that the values of Matthew's correlation coefficient confirm the above findings based on the Hamming distance. On the other hand, false positive and true positive rates only focus on one aspect of estimation at a time and do not provide a clear distinction between the methods.

The representation of conditional independence in directed graphs adapted in our algorithm is not restricted to normally distributed random variables; if the underlying structural equations are

Fig. 5. Matthew's correlation coefficient, true and false positives for estimation of directed graphs using the PC-algorithm (black solid), lasso (black dashes) and adaptive lasso (grey dot-dashes) with random ordering.

linear, the method proposed in this paper can correctly estimate the underlying graph. In order to assess the sensitivity of the estimates to the underlying distribution, we performed two simulation studies with nonnormal observations. In both simulations, observations were generated according to a linear structural model. In the first simulation, the latent variables were generated from a mixture of a standard normal and a t -distribution with three degrees of freedom, while in the second simulation, a t -distribution with four degrees of freedom was used. The performance of the proposed algorithm for nonnormal observations was similar to the case of Gaussian observations, with the adaptive lasso providing the best estimates, and the performance of penalized methods improving in sparse settings.

5.3. Sensitivity to perturbations in the ordering of the variables

Algorithm 1 assumes a known ordering of the variables. The superior performance of the proposed penalized likelihood methods in comparison to the PC-algorithm may be explained by the fact that additional information about the order of the variables significantly simplifies the problem of estimating directed graphs. Therefore, when such additional information is available, estimates using the PC-algorithm suffer from a disadvantage. However, as the underlying network becomes more sparse, the network includes fewer complex structures and the ordering of variables should play a less significant role.

Next, we study the performances of the proposed methods, and the PC-algorithm in problems where the ordering of variables is unknown. To this end, we generate normally distributed observations from the latent variable model of § 2.2. We then randomly permute the order of variables in the observation matrix and use the permuted matrix to estimate the original directed graph. Figure 5 illustrates the performance of the three methods for choices of α described in § 5.2. It can be seen that for small, dense networks, the PC-algorithm outperforms the proposed methods. This is expected since the change in the order of variables causes the proposed algorithm to include unnecessary moral edges, while failing to recognize some of the existing associations. However, as the size of the network and correspondingly the degree of sparsity increase, the local structures become simpler and therefore the ordering of the variables becomes less crucial. Thus, the performance of penalized likelihood algorithms is improved compared to that of the PC-algorithm. For the high-dimensional sparse case, where the computational cost of the PC-algorithm becomes more significant, the penalized likelihood methods provide better estimates.

Fig. 6. Known and estimated networks for human cell signalling data. True and false edges are marked with solid and dashed arrows, respectively. MCC, Matthew's correlation coefficient; FN, False negative; FP, False positive; SHD, structural Hamming distance.

6. REAL DATA APPLICATION

6.1. Analysis of cell signalling pathway data

[Sachs et al. \(2003\)](#) carried out a set of flow cytometry experiments on signaling networks of human immune system cells. The ordering of the connections between pathway components was established based on perturbations in cells using molecular interventions and we consider it to be known a priori. The dataset includes $p = 11$ proteins and $n = 7466$ samples.

[Friedman et al. \(2008\)](#) analyzed this dataset using the graphical lasso algorithm. They estimated the graph for a range of values of the ℓ_1 penalty and reported moderate agreement, around 50% false positive and false negative rates, between one of their estimates and the findings of [Sachs et al. \(2003\)](#). True and estimated signaling networks using the PC-algorithm, with $\alpha = 0.01$, and the lasso and adaptive lasso algorithms, with $\alpha = 0.1$, along with the corresponding performance measures, are given in Fig. 6. The estimated network using the PC-algorithm includes a number of undirected edges. As in the simulation studies, we only estimate the structure of the network using the PC-algorithm and determine the direction of edges by enforcing the ordering of nodes. It can be seen that the adaptive lasso and lasso provide estimates that are closer to the true structure.

6.2. Transcription regulatory network of *E. coli*

Transcriptional regulatory networks play an important role in controlling the gene expression in cells and incorporating the underlying regulatory network results in more efficient estimation and inference ([Shojaie & Michailidis, 2009, 2010](#)). [Kao et al. \(2004\)](#) proposed the network component analysis method to infer the transcriptional regulatory network of *Escherichia coli*, *E. coli*. They also provided whole genome expression data over time, with $n = 24$, as well as information about the known regulatory network of *E. coli*.

In this application, the set of transcription factors are known a priori and the goal is to find connections among transcription factors and regulated genes through analysis of whole genome transcriptomic data. The algorithm proposed in this paper can be used by exploiting the natural hierarchy of transcription factors and regulated genes. [Kao et al. \(2004\)](#) provide gene expression data for seven transcription factors and 40 regulated genes, i.e. $p = 47$. Figure 7 presents the known regulatory network of *E. coli* along with the estimated networks and the corresponding performance measures using the PC-algorithm, lasso and adaptive lasso. The values of α are set as in § 6.1. The relatively poor performance of the algorithms in this example can be partially attributed to the small sample size. However, it is also known that no single source of

ACKNOWLEDGEMENT

Thanks go to the authors of [Kalisch & Bühlmann \(2007\)](#) and [Friedman et al. \(2007\)](#) for making the R-packages `pcalg` and `glmnet` available. We are especially thankful to Markus Kalisch and

Trevor Hastie for help with technical difficulties with these packages. We would like to thank the editor, an associate editor, and a referee, whose comments and suggestions significantly improved the clarity of the manuscript. This research was partially funded by grants from the National Institutes of Health, U.S.A.

APPENDIX

Technical proofs

The following lemma is a consequence of the Karush–Kuhn–Tucker conditions for the general weighted lasso problem and is used in the proof of Theorems 3 and 4.

LEMMA A1. Let $\hat{\theta}^{i,\mathcal{I}}$ be the general weighted lasso estimate of $\theta^{i,\mathcal{I}}$, i.e.

$$\hat{\theta}^{i,\mathcal{I}} = \arg \min_{\theta: \theta_k=0, k \notin \mathcal{I}} \left\{ n^{-1} \|\mathcal{X}_i - \mathcal{X}\theta\|_2^2 + \lambda \sum_{k=1}^p |\theta_k| w_{ik} \right\}. \quad (\text{A1})$$

Define

$$G_j(\theta) = -2n^{-1} \mathcal{X}_j^\top (\mathcal{X}_i - \mathcal{X}\theta),$$

and let w_i be the vector of initial weights in the adaptive lasso estimation problem. Then a vector $\hat{\theta}$ with $\hat{\theta}_k = 0$, for all $k \notin \mathcal{I}$ is a solution of (A1) if and only if for all $j \in \mathcal{I}$, $G_j(\theta) = -\text{sign}(\hat{\theta}_j) w_{ij} \lambda$ if $\hat{\theta}_j \neq 0$ and $|G_j(\theta)| \leq w_{ij} \lambda$ if $\hat{\theta}_j = 0$. Moreover, if the solution is not unique and $|G_j(\theta)| < w_{ij} \lambda$ for some solution $\hat{\theta}$, then $\hat{\theta}_j = 0$ for all solutions of (A1).

Proof. The proof of the lemma is identical to the proof of Lemma (A1) in Meinshausen & Bühlmann (2006), except for inclusion of general weights w_{ij} , and is therefore omitted. $\square$

Proof of Theorem 3. To prove (i), note that by Bonferroni's inequality, and the fact that $\text{card}(\text{pa}_i) = o(n)$ as $n \rightarrow \infty$, it suffices to show that there exists some $c_{(i)} > 0$ such that for all $i \in V$ and for every $j \in \text{pa}_i$, $\text{pr}\{\text{sign}(\hat{\theta}_j^{i,\text{pa}_i}) = \text{sign}(\theta_j^{i,\text{pa}_i})\} = 1 - O\{\exp(-c_{(i)} n^\zeta)\}$ as $n \rightarrow \infty$.

Let $\hat{\theta}^{i,\text{pa}_i}(\beta)$ be the estimate of θ^{i,pa_i} in (A1), with the j th component fixed at a constant value β ,

$$\hat{\theta}^{i,\text{pa}_i}(\beta) = \arg \min_{\theta \in \Theta_\beta} \left\{ n^{-1} \|\mathcal{X}_i - \mathcal{X}\theta\|_2^2 + \lambda \sum_{k=1}^p |\theta_k| w_k \right\}, \quad (\text{A2})$$

where $\Theta_\beta \equiv \{\theta \in \mathbb{R}^p : \theta_j = \beta, \theta_k = 0, k \notin \text{pa}_i\}$. Note that for $\beta = \hat{\theta}_j^{i,\text{pa}_i}$, $\hat{\theta}^{i,\text{pa}_i}(\beta)$ is identical to $\hat{\theta}^{i,\text{pa}_i}$. Thus, if $\text{sign}(\hat{\theta}_j^{i,\text{pa}_i}) \neq \text{sign}(\theta_j^i)$, there would exist some β with $\text{sign}(\beta) \text{sign}(\theta_j^i) \leq 0$ such that $\hat{\theta}^{i,\text{pa}_i}(\beta)$ is a solution to (A2). Since $\theta_j^i \neq 0$, for all $j \in \text{pa}_i$, it suffices to show that for all β with $\text{sign}(\beta) \text{sign}(\theta_j^i) < 0$, with high probability, $\hat{\theta}^{i,\text{pa}_i}(\beta)$ cannot be a solution to (A2).

Without loss of generality, we consider the case where $\theta_j^i > 0$; $\theta_j^i < 0$ can be shown similarly. Then if $\beta \leq 0$, from Lemma A1, $\hat{\theta}^{i,\text{pa}_i}(\beta)$ can be a solution to (A2) only if $G_j\{\hat{\theta}^i(\beta)\} \geq -\lambda w_{ij}$. Hence, it suffices to show that for some $c_{(i)} > 0$ and all $j \in \text{pa}_i$ with $\theta_j^i > 0$,

$$\text{pr} \left(\sup_{\beta \leq 0} [G_j\{\hat{\theta}^i(\beta)\} < -\lambda w_{ij}] \right) = 1 - O\{\exp(-c_{(i)} n^\zeta)\}, \quad n \rightarrow \infty.$$

Define, $\mathcal{R}_i(\beta) = \mathcal{X}_i - \mathcal{X}\hat{\theta}^i(\beta)$. Then for every $j \in \text{pa}_i$ we can write

$$X_j = \sum_{k \in \text{pa}_i \setminus \{j\}} \theta_k^{j,\text{pa}_i \setminus \{j\}} X_k + Z_j, \quad (\text{A3})$$

where Z_j is independent of $\{X_k; k \in \text{pa}_i \setminus \{j\}\}$. Then by (A3),

$$G_j\{\hat{\theta}^i(\beta)\} = -2n^{-1}Z_j^\top \mathcal{R}_i(\beta) - \sum_{k \in \text{pa}_i \setminus \{j\}} \theta^{j, \text{pa}_i \setminus \{j\}} 2n^{-1} \mathcal{X}_k^\top \mathcal{R}_i(\beta).$$

By Lemma A1, it follows that for all $k \in \text{pa}_i \setminus \{j\}$, $|G_k\{\hat{\theta}^i(\beta)\}| = |2n^{-1} \mathcal{X}_k^\top \mathcal{R}_i(\beta)| \leq \lambda w_{ik}$, and hence,

$$G_j\{\hat{\theta}^i(\beta)\} \leq -2n^{-1}Z_j^\top \mathcal{R}_i(\beta) + \lambda \sum_{k \in \text{pa}_i \setminus \{j\}} |\theta^{j, \text{pa}_i \setminus \{j\}}| w_{ik}.$$

Using the fact that $|\theta^{j, \text{pa}_i \setminus \{j\}}| \leq 1$, it suffices to show that

$$\text{pr} \left[\sup_{\beta \leq 0} \{-2n^{-1}Z_j^\top \mathcal{R}_i(\beta)\} < -\lambda \sum_{k \in \text{pa}_i} w_{ik} \right] = 1 - O\{\exp(-c_{(i)} n^\zeta)\}, \quad n \rightarrow \infty.$$

It is shown in Lemma A.2. of Meinshausen & Bühlmann (2006) that for any $q > 0$, there exists $c_{(i)} > 0$ such that for all $j \in \text{pa}_i$ with $\theta_j^i > 0$,

$$\text{pr} \left[\inf_{\beta \leq 0} \{2n^{-1}Z_j^\top \mathcal{R}_i(\beta)\} > q\lambda \right] = 1 - O\{\exp(-c_{(i)} n^\zeta)\}, \quad n \rightarrow \infty. \quad (\text{A4})$$

However, by definition $w_{ik} \geq 1$ and therefore, $\sum_{k \in \text{pa}_i} w_{ik} \geq \text{card}(\text{pa}_i) \geq 1$, and (i) follows from (A4).

To prove (ii), note that the event $\hat{\text{pa}}_i \not\subseteq \text{pa}_i$ is equivalent to the event that there exists a node $j \in i^- \setminus \text{pa}_i$ such that $\hat{\theta}_j^i \neq 0$. In other words, denoting the latter event by $\mathcal{D}$, $\text{pr}(\hat{\text{pa}}_i \subseteq \text{pa}_i) = 1 - \text{pr}(\mathcal{D})$.

However, by Lemma A1, and since $w_{ij} \geq 1$,

$$\begin{aligned} \text{pr}(\mathcal{D}) &= \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } |G_j(\hat{\theta}^{i, \text{pa}_i})| \geq w_{ij}\lambda) \\ &\leq \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } |G_j(\hat{\theta}^{i, \text{pa}_i})| \geq q\lambda \text{ and } w_{ij}\lambda \leq q\lambda \text{ for some } q \geq 1) \\ &\leq \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } w_{ij} \leq q \text{ for some } q \geq 1). \end{aligned}$$

But $w_{ij} = 1 \vee |\tilde{\theta}_j^i|^{-\gamma}$, with $\tilde{\theta}_j^i$ the lasso estimate of the adjacency matrix from (9). Hence, letting $\mathcal{F}$ be the event that there exists $j \in i^- \setminus \text{pa}_i$ such that $w_{ij} \leq q$ for some $q > 0$, and using Lemma A1 we can write

$$\begin{aligned} \text{pr}(\mathcal{F}) &= \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } |\tilde{\theta}_j^i| \geq q^{-1/\gamma} \text{ for some } q \geq 1) \\ &\leq \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } |\tilde{\theta}_j^i| \geq q' \text{ for some } q' > 0) \\ &\leq \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } \tilde{\theta}_j^i \neq 0) \\ &= \text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } |G_j(\tilde{\theta}^{i, \text{pa}_i})| \geq \lambda^0). \end{aligned}$$

Since $\text{card}(\text{pa}_i) = o(n)$, we can assume, without loss of generality, that $\text{card}(\text{pa}_i) < n$, which implies that $\tilde{\theta}^{i, \text{pa}_i}$ is an almost sure unique solution to (A1) with $\mathcal{I} = \text{pa}_i$. Let $\mathcal{E} = \{\max_{j \in i^- \setminus \text{pa}_i} |G_j(\tilde{\theta}^{i, \text{pa}_i})| < \lambda^0\}$. Then conditional on the event $\mathcal{E}$, it follows from the first part of Lemma A1 that $\tilde{\theta}^{i, \text{pa}_i}$ is also a solution of the unrestricted weighted lasso problem (A1) with $\mathcal{I} = i^-$. Since $\tilde{\theta}_j^{i, \text{pa}_i} = 0$, for all $j \in i^- \setminus \text{pa}_i$, it follows from the second part of Lemma A1 that $\tilde{\theta}_j^i = 0$, for all $j \in i^- \setminus \text{pa}_i$. Hence,

$$\text{pr}(\text{there exists } j \in i^- \setminus \text{pa}_i \text{ such that } \tilde{\theta}_j^i \neq 0) \leq 1 - \text{pr}(\mathcal{E}) = \text{pr} \left\{ \max_{j \in i^- \setminus \text{pa}_i} |G_j(\tilde{\theta}^{i, \text{pa}_i})| \geq \lambda^0 \right\},$$

where $G_j(\tilde{\theta}^{i, \text{pa}_i}) = -2n^{-1} \mathcal{X}_j^\top (\mathcal{X}_i - \mathcal{X} \tilde{\theta}^{i, \text{pa}_i})$.

Since $\text{card}(V) = O(n^a)$ for some $a > 0$, Bonferroni's inequality implies that to verify (ii) it suffices to show that there exists a constant $c_{(ii)} > 0$ such that for all $j \in i^- \setminus \text{pa}_i$,

$$\text{pr}(|G_j(\tilde{\theta}^{i, \text{pa}_i})| \geq \lambda^0) = O\{\exp(-c_{(ii)} n^\zeta)\}, \quad n \rightarrow \infty,$$

where $R_j \sim N(0, \sigma_j^2)$, $\sigma_j^2 \leq 1$ and R_j is independent of X_l , $l \in \text{pa}_i$. Similarly, with R_i satisfying the same requirements as R_j , we get $X_i = \sum_{k \in \text{pa}_i} \theta_k^{i, \text{pa}_i} X_k + R_i$. Denote by $\mathcal{X}_{\text{pa}_i}$ the columns of $\mathcal{X}$ corresponding to pa_i and let θ_{pa_i} be the column vector of coefficients with dimension $\text{card}(\text{pa}_i)$ corresponding to pa_i . Then,

$$\begin{aligned} \text{pr}\{|G_j(\tilde{\theta}^{i, \text{pa}_i})| \geq \lambda^0\} &= \text{pr}\{|-2n^{-1} \mathcal{X}_j^T (\mathcal{X}_i - \mathcal{X} \tilde{\theta}^{i, \text{pa}_i})| \geq \lambda^0\} \\ &= \text{pr}\{|-2n^{-1} \{\mathcal{X}_{\text{pa}_i} \theta_{\text{pa}_i}^{j, \text{pa}_i} + \mathcal{R}_j\}^T \{\mathcal{X}_{\text{pa}_i} (\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}) + \mathcal{R}_i\}| \geq \lambda^0\} \\ &\leq \text{pr}\{|-2n^{-1} (\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i})^T \mathcal{X}_{\text{pa}_i}^T \mathcal{X}_{\text{pa}_i} \theta_{\text{pa}_i}^{j, \text{pa}_i}| \geq \lambda^0/3\} \\ &\quad + \text{pr}\{|-2n^{-1} (\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i})^T \mathcal{X}_{\text{pa}_i}^T \mathcal{R}_j| \geq \lambda^0/3\} \\ &\quad + \text{pr}\{|-2n^{-1} (\mathcal{X}_{\text{pa}_i} \theta_{\text{pa}_i}^{j, \text{pa}_i} + \mathcal{R}_j)^T \mathcal{R}_i| \geq \lambda^0/3\} \\ &\equiv \text{I} + \text{II} + \text{III}, \end{aligned}$$

say. Let $\mathbf{1}_{\text{pa}_i}$ denote a vector of 1s of dimension $\text{card}(\text{pa}_i)$. Then using the fact that $|\theta_l^{j, \text{pa}_i}| \leq 1$, for all $l \in \text{pa}_i$, we can write $\text{I} \leq \text{pr}\{2\|\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}\|_\infty n^{-1} (\mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i})^T \mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i} \geq \lambda^0/3\}$. Then $\mathcal{X}_{\text{pa}_i}^T \mathcal{X}_{\text{pa}_i} \sim \mathbb{W}_{\text{card}(\text{pa}_i)}(\Sigma_{\text{pa}_i}, n)$ where $\mathbb{W}_m(\Sigma, n)$ denotes a Wishart distribution with mean $n\Sigma$. Hence, from properties of the Wishart distribution, we get $(\mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i})^T \mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i} \sim \mathbb{W}_1(\mathbf{1}_{\text{pa}_i}^T \Sigma_{\text{pa}_i} \mathbf{1}_{\text{pa}_i}, n)$.

Since pa_i also forms a directed acyclic graph, the eigenvalues Σ_{pa_i} are bounded, see Remark 1, and hence

$$\mathbf{1}_{\text{pa}_i}^T \Sigma_{\text{pa}_i} \mathbf{1}_{\text{pa}_i} \leq \text{card}(\text{pa}_i) \phi_{\max}(\Sigma_{\text{pa}_i}). \quad (\text{A5})$$

Therefore, if $Z \sim \chi_1^2$, then $n^{-1} (\mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i})^T \mathcal{X}_{\text{pa}_i} \mathbf{1}_{\text{pa}_i}$ is stochastically smaller than $\text{card}(\text{pa}_i) \phi_{\max}(\Sigma_{\text{pa}_i}) Z$. On the other hand, by Theorem 2, $\|A - \tilde{A}\|_F = O_p\{(n^{-1} s \log p)^{1/2}\}$, and hence,

$$\|\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}\|_\infty = O_p\{(n^{-1} s \log p)^{1/2}\}. \quad (\text{A6})$$

Noting that $\text{card}(\text{pa}_i) = O(n^b)$, $b < 1/2$ and $p = O(n^a)$, $a > 0$, (A5) and (A6) imply that

$$\|\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}\|_\infty \text{card}(\text{pa}_i) \phi_{\max}(\Sigma_{\text{pa}_i}) = O_p\{(sn^{2b-1} a \log n)^{1/2}\}.$$

By Assumption 1', $sn^{2b-1} \log n = o(1)$ and hence by Slutsky's Theorem and properties of the χ^2 -distribution, there exists $c_{(\text{I})} > 0$ such that for all $j \in i^- \setminus \text{pa}_i$, $\text{I} = O\{\exp(-c_{(\text{I})} n^\xi)\}$ as $n \rightarrow \infty$.

Using a similar argument, $\text{II} \leq \text{pr}(2n^{-1} \|\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}\|_\infty |\mathbf{1}_{\text{pa}_i}^T \mathcal{X}_{\text{pa}_i}^T \mathcal{R}_j| \geq \lambda^0/3)$. But columns of $\mathcal{X}_{\text{pa}_i}$ have mean zero and are all independent of $\mathcal{R}_j$, so it suffices to show that there exists $c_{(\text{II})} > 0$ such that for all $j \in i^- \setminus \text{pa}_i$ and for all $k \in \text{pa}_i$,

$$\text{pr}\{2n^{-1} \|\theta_{\text{pa}_i}^{i, \text{pa}_i} - \tilde{\theta}_{\text{pa}_i}^{i, \text{pa}_i}\|_\infty \text{card}(\text{pa}_i) |\mathcal{X}_k^T \mathcal{R}_j| \geq \lambda^0/3\} = O\{\exp(-c_{(\text{II})} n^\xi)\} \quad \text{as } n \rightarrow \infty. \quad (\text{A7})$$

By (A6) and Assumption 1', the random variable on the left-hand side of (A7) is stochastically smaller than $2n^{-1} |\mathcal{X}_k^T \mathcal{R}_j|$. By independence of X_k and R_j , $E(X_k R_j) = 0$. Also, using Gaussianity of both X_k and R_j , there exists $g < \infty$ such that $E\{\exp(|X_k R_j|)\} \leq g$. Since $\lambda^0 = O\{(\log p/n)^{1/2}\}$, by Bernstein's inequality (Van der Vaart & Wellner, 1996), $\text{pr}(2n^{-1} |\mathcal{X}_k^T \mathcal{R}_j| > \lambda^0/3) \leq \exp(-c_{(\text{II})} n^\xi)$ for some $c_{(\text{II})} > 0$ and hence (A7) is satisfied.

Finally, $\text{pr}\{|-2n^{-1} (\mathcal{X}_{\text{pa}_i} \theta_{\text{pa}_i}^{j, \text{pa}_i} + \mathcal{R}_j)^T \mathcal{R}_i| \geq \lambda^0/3\} = \text{pr}\{|-2n^{-1} \mathcal{X}_i^T \mathcal{R}_j| \geq \lambda^0/3\}$, and using the Bernstein's inequality we conclude that there exists $c_{(\text{III})} > 0$ such that for all $j \in i^- \setminus \text{pa}_i$ and for all $k \in \text{pa}_i$, $\text{III} = O\{\exp(-c_{(\text{III})} n^\xi)\}$ as $n \rightarrow \infty$. The proof of (ii) is then completed by taking $c_{(ii)}$ to be the minimum of $c_{(\text{I})}, \dots, c_{(\text{III})}$.

To prove (iii), note that $\text{pr}(\text{pa}_i \subseteq \hat{\text{pa}}_i) = 1 - \text{pr}(\text{there exists } j \in \text{pa}_i \text{ such that } \hat{\theta}_j^i = 0)$, and let $\mathcal{E} = \{\max_{k \in i^- \setminus \text{pa}_i} |G_j(\hat{\theta}^{i, \text{pa}_i})| < \lambda w_{ij}\}$. It then follows from an argument similar to the proof of (ii) that conditional on $\mathcal{E}$, $\hat{\theta}^{i, \text{pa}_i}$ is an almost sure unique solution of the unrestricted adaptive lasso problem (A1) with

$\mathcal{I} = i^-$. Therefore,

$$\text{pr}(\text{there exists } j \in \text{pa}_i \text{ such that } \hat{\theta}_j^i = 0) \leq \text{pr}(\text{there exists } j \in \text{pa}_i \text{ such that } \hat{\theta}_j^i = 0) + \text{pr}(\mathcal{E}^c).$$

From (i), there exists a $c_1 > 0$ such that $\text{pr}(\text{there exists } j \in \text{pa}_i \text{ such that } \hat{\theta}_j^i = 0) = O\{\exp(-c_1 n^\zeta)\}$ and it was shown in (ii) that $\text{pr}(\mathcal{E}^c) = O\{\exp(-c_2 n^\zeta)\}$ for some $c_2 > 0$. Thus (iii) follows from Bonferroni's inequality.

The claim in (iv) follows from (ii) and (iii), and Bonferroni's inequality as $p = O(n^a)$. $\square$

Proof of Theorem 4. We first show that if $AN_i \cap AN_j = \emptyset$, then i and j are independent. Since $\Sigma = \Lambda \Lambda^\top$ and Λ is lower triangular,

$$\Sigma_{ij} = \sum_{k=1}^{\min(i,j)} \Lambda_{ik} \Lambda_{jk}. \quad (\text{A8})$$

We assume without loss of generality that $i < j$. The argument for $j > i$ is similar. Suppose for all $k = 1, \dots, i$, that $\Lambda_{ik} = 0$ or $\Lambda_{jk} = 0$. Then by (A8) i and j are independent. However, by Lemma 1, Λ_{jk} is the influence of k th node on j , and this is zero only if there is no path from k to j . If i is an ancestor of j , we have $\Sigma_{ij} \neq 0$. On the other hand, if there is no node $k \in i^-$ such that k influences both i and j , i.e. k is a common ancestor of i and j , then for all $k = 1, \dots, i$ we have $\Lambda_{ik} \Lambda_{jk} = 0$ and the claim follows.

Using Bonferroni's inequality twice and Lemma A1, we get

$$\begin{aligned} \text{pr}(\text{there exists } i \in V \text{ such that } \hat{AN}_i \not\subseteq AN_i) &\leq p \max_{i \in V} \text{pr}(\text{there exists } j \in i^- \setminus AN_i \text{ such that } j \in \hat{\text{pa}}_i) \\ &\leq p(i-1) \max_{i \in V, j \in i^- \setminus AN_i} \text{pr}(j \in \hat{\text{pa}}_i) \\ &\leq p(i-1) \max_{i \in V, j \in i^- \setminus AN_i} \text{pr}\{|G_j(\hat{\theta}^{i, AN_i})| \geq \lambda w_{ij}\}. \end{aligned}$$

However, by definition $w_{ij} \geq 1$, and hence it suffices to show that

$$(i-1)p \max_{i \in V, j \in i^- \setminus AN_i} \text{pr}\{|G_j(\hat{\theta}^{i, AN_i})| \geq \lambda\} \leq \alpha.$$

Note that $G_j(\hat{\theta}^{i, AN_i}) = -2n^{-1} \mathcal{X}_j^\top (\mathcal{X}_i - \mathcal{X} \hat{\theta}^{i, AN_i})$ and X_j is independent of X_k for all $k \in AN_i$. Therefore, conditional on $\mathcal{X}_{AN_i}$, $G_j(\hat{\theta}^{i, AN_i}) \sim (0, 4R^2/n)$, where $R^2 = n^{-1} \|\mathcal{X}_i - \mathcal{X} \hat{\theta}^{i, AN_i}\|_2^2 \leq n^{-1} \|\mathcal{X}_i\|_2^2 = 1$, by definition of $\hat{\theta}^{i, AN_i}$ and the fact that columns of the data matrix are scaled.

It follows that for all $j \in i^- \setminus AN_i$, $\text{pr}\{|G_j(\hat{\theta}^{i, AN_i})| \geq \lambda \mid \mathcal{X}_{AN_i}\} \leq 2\{1 - \Phi(n^{1/2}\lambda/2)\}$, where Φ is the cumulative distribution function for standard normal random variable. Using the choice of λ proposed in (10), we get $\text{pr}\{|G_j(\hat{\theta}^{i, AN_i})| \geq \lambda \mid \mathcal{X}_{AN_i}\} \leq \alpha\{(i-1)p\}^{-1}$, and the result follows. $\square$

REFERENCES

- BANERJEE, O., EL GHAOU, L. & D'ASPREMONT ALEXANDRE (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.
- FRIEDMAN, J., HASTIE, T., HÖFLING, H. & TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–32.
- FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.
- HECKERMAN, D., GEIGER, D. & CHICKERING, D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* **20**, 197–243.
- HUANG, J., LIU, N., POURAHMADI, M. & LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.
- HUANG, J., MA, S. & ZHANG, C. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–18.
- KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8**, 613–36.

- KAO, K., YANG, Y., BOSCOLO, R., SABATTI, C., ROYCHOWDHURY, V. & LIAO, J. (2004). Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc. Nat. Acad. Sci.* **101**, 641–6.
- KNIGHT, K. & FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.
- LAM, C. & FAN, J. (2009). Sparsity and rate of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–78.
- LAURITZEN, S. (1996). *Graphical Models*. Oxford: Oxford University Press.
- LECLERC, R. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molec. Syst. Biol.* **4**, 1–6.
- LEVINA, E., ROTHMAN, A. & ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Statist.* **2**, 245–63.
- MARKOWETZ, F. & SPANG, R. (2007). Inferring cellular networks—a review. *BMC Bioinformatics* **8**, S5.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–62.
- PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
- R DEVELOPMENT CORE TEAM (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. URL: <http://www.R-project.org>.
- ROBINSON, R. (1977). Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V: Proc. Fifth Australian Conf., R. Melbourne Inst. Technol.*, Ed. C. H. C. Little, pp. 28–43. Berlin: Springer.
- ROTHMAN, A., BICKEL, P., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.
- SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. & NOLAN, G. (2003). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 504–6.
- SHOJAIE, A. & MICHAILIDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *J. Comp. Biol.* **16**, 407–26.
- SHOJAIE, A. & MICHAILIDIS, G. (2010). Network enrichment analysis in complex experiments. *Statist. Appl. Genet. Molec. Biol.* **9**, Article 22.
- SPIRITES, P., GLYMOUR, C. & SCHEINES, R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
- TSAMARDINOS, I., BROWN, L. & ALIFERIS, C. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65**, 31–78.
- VAN DER VAART, A. & WELLNER, J. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.
- YUAN, M. & LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–36.
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[Received July 2009. Revised March 2010]