Biometrika (2014), **101**, 2, *pp.* 253–268
Printed in Great Britain
doi: 10.1093/biomet/asu009
Advance Access publication 12 May 2014

Direct estimation of differential networks

By SIHAI DAVE ZHAO

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.
sihai@mail.med.upenn.edu

T. TONY CAI

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
tcai@wharton.upenn.edu

and HONGZHE LI

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.
hongzhe@upenn.edu

Summary

It is often of interest to understand how the structure of a genetic network differs between two conditions. In this paper, each condition-specific network is modelled using the precision matrix of a multivariate normal random vector, and a method is proposed to directly estimate the difference of the precision matrices. In contrast to other approaches, such as separate or joint estimation of the individual matrices, direct estimation does not require those matrices to be sparse, and thus can allow the individual networks to contain hub nodes. Under the assumption that the true differential network is sparse, the direct estimator is shown to be consistent in support recovery and estimation. It is also shown to outperform existing methods in simulations, and its properties are illustrated on gene expression data from late-stage ovarian cancer patients.

Some key words: Differential network; Graphical model; High dimensionality; Precision matrix.

1. Introduction

A complete understanding of the molecular basis of disease will require characterization of the network of interdependencies between genetic components. There are many types of network that may be considered, such as protein-protein interaction networks or metabolic networks (Emmert-Streib & Dehmer, 2011), but the focus of this paper is on transcriptional regulatory networks. In many cases, interest centres not on a particular network but rather on whether and how the network changes between disease states. Indeed, differential networking analysis has recently emerged as an important complement to differential expression analysis (de la Fuente, 2010; Ideker & Krogan, 2012). For example, Hudson et al. (2009) studied a mutant breed of cattle known to differ from wild-type cattle by a mutation in the myostatin gene. The myostatin gene was not differentially expressed between the two breeds, but Hudson et al. (2009) showed that a

differential network analysis could correctly identify it as the gene containing the causal mutation. In another example, using an experimental technique called differential epistasis mapping, Bandyopadhyay et al. (2010) demonstrated large-scale changes in the genetic networks of yeast cells after perturbation by a DNA-damaging agent.

Transcriptional networks are frequently modelled as Gaussian graphical models (Markowetz & Spang, 2007). Gene expression levels are assumed to be jointly Gaussian, so that two expression levels are conditionally independent given the other genes if and only if the corresponding entry of the precision matrix, or inverse covariance matrix, is zero. Representing gene expression levels as nodes and conditional dependency relationships as edges in a graph results in a Gaussian graphical model (Lauritzen, 1996). A differential network can be modelled as a set of changes in this graph structure between two conditions.

However, there may be cases where the conditional dependency relationships between pairs of genes change in magnitude but not in structure. For example, two genes may be positively conditionally dependent in one group but negatively conditionally dependent in the other; the supports of the precision matrices of the two groups would be identical and would not reflect these potentially biologically significant differences in magnitude. For this reason, in the present paper two genes are instead defined to be connected in the differential network if the magnitude of their conditional dependency relationship changes between two groups. More precisely, consider independent observations of the expression levels of p genes from two groups of subjects: $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n_X$ from one group and $Y_i = (Y_{i1}, \ldots, Y_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n_Y$ from the other, where $X_i \sim N(\mu_X, \Sigma_X)$ and $Y_i \sim N(\mu_Y, \Sigma_Y)$. The differential network is defined to be the difference between the two precision matrices, denoted by $\Delta_0 = \Sigma_Y^{-1} - \Sigma_X^{-1}$. The entries of Δ_0 can also be interpreted as the differences in the partial covariances of each pair of genes between the two groups. This type of model for a differential network has been adopted by others as well, for example Li et al. (2007) and Danaher et al. (2014), and in a 2013 unpublished technical report by N. Städler and S. Mukherjee (arXiv:1308.2771).

2. Previous approaches

There are currently two main types of approach to estimating Δ_0. The most straightforward one is to separately estimate Σ_X^{-1} and Σ_Y^{-1} and then subtract the estimates. A naive estimate of a single precision matrix can be obtained by inverting the sample covariance matrix. However, in most experiments the number of gene expression probes exceeds the number of subjects. In this high-dimensional data setting, the sample covariance matrix is singular and alternative methods are needed to estimate the precision matrix. Theoretical and computational work has shown that estimation is possible under the key assumption that the precision matrix is sparse, meaning that each row and each column has relatively few nonzero entries (Friedman et al., 2008; Ravikumar et al., 2008; Yuan, 2010; Cai et al., 2011).

The second type of approach is to jointly estimate Σ_X^{-1} and Σ_Y^{-1}, assuming that they share common features. For example, Chiquet et al. (2011), Guo et al. (2011) and Danaher et al. (2014) penalized the joint loglikelihood of the X_i and Y_i using penalties such as the group lasso (Yuan & Lin, 2006) and group bridge (Huang et al., 2009; Wang et al., 2009), which encourage the estimated precision matrices to have similar supports. Danaher et al. (2014) also introduced the fused graphical lasso, which uses a fused lasso penalty (Tibshirani et al., 2005) to encourage the entries of the estimated precision matrices to have similar magnitudes.

Most of these approaches assume that both Σ_X^{-1} and Σ_Y^{-1} are sparse, but real transcriptional networks often contain hub nodes (Barabási & Oltvai, 2004; Barabási et al., 2011), or genes that interact with many other genes. The rows and columns of Σ_X^{-1} and Σ_Y^{-1} corresponding to hub

nodes have many nonzero entries and violate the sparsity condition. The method of Danaher et al. (2014) is one exception that does not require individual sparsity. Its estimates $\hat{\Sigma}_X^{-1}$ and $\hat{\Sigma}_Y^{-1}$ minimize

$$\sum_{g\in\{X,Y\}} n_g\{\log\det\Sigma_g^{-1} - \mathrm{tr}(\hat{\Sigma}_g\Sigma_g^{-1})\} - \lambda_1 \sum_{g\in\{X,Y\}}\sum_{j\neq k}|\omega_{jk}^g| + \lambda_2\sum_{jk}|\omega_{jk}^X - \omega_{jk}^Y|, \quad (1)$$

where $\hat{\Sigma}_X$ and $\hat{\Sigma}_Y$ are sample covariance matrices of the X_i and Y_i, ω_{jk}^X and ω_{jk}^Y are the (j,k)th entries of Σ_X^{-1} and Σ_Y^{-1}, and $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix, respectively. The first term of (1) is the joint likelihood of the X_i and Y_i, and the second and third terms constitute a fused lasso-type penalty. The parameters λ_1 and λ_2 control the sparsity of the individual precision matrix estimates and the similarities of their entries, respectively, and when λ_1 is set to zero (1) does not require $\hat{\Sigma}_X^{-1}$ or $\hat{\Sigma}_Y^{-1}$ to be sparse. A referee pointed out that methods recently introduced by Mohan et al. (2012) were also designed for estimating networks containing hubs; however, theoretical performance guarantees for these methods have not yet been derived.

The direct estimation method proposed in this paper does not require Σ_X^{-1} and Σ_Y^{-1} to be sparse and does not require separate estimation of these precision matrices. Theoretical performance guarantees are provided for differential network recovery and estimation, and simulations show that when the separate networks include hub nodes, direct estimation is more accurate than fused graphical lasso or separate estimation.

3. Direct estimation of the difference of two precision matrices

3·1. *Constrained optimization approach*

We use $|\cdot|$ to denote elementwise norms and $\|\cdot\|$ to denote matrix norms. For a $p\times 1$ vector $a=(a_1,\ldots,a_p)^{\mathrm{T}}$, define $|a|_0$ to be the number of nonzero elements of a, and let $|a|_1=\sum_j|a_j|$, $|a|_2=(\sum_j a_j^2)^{1/2}$ and $|a|_\infty=\max_j|a_j|$. For a $p\times p$ matrix A with entries a_{jk}, let $|A|_0$ be the number of nonzero entries of A, $|A|_1=\sum_{j,k}|a_{jk}|$, $|A|_\infty=\max_{j,k}|a_{jk}|$, $\|A\|_1=\max_k\sum_j|a_{jk}|$, $\|A\|_\infty=\max_j\sum_k|a_{jk}|$, $\|A\|_2=\sup_{|a|_2\leqslant 1}|Aa|_2$ and $\|A\|_{\mathrm{F}}=(\sum_{j,k}a_{jk}^2)^{1/2}$.

Let $\hat{\Sigma}_X=n_X^{-1}\sum_i(X_i-\bar{X})(X_i-\bar{X})^{\mathrm{T}}$, where $\bar{X}=n_X^{-1}\sum_i X_i$, and define $\hat{\Sigma}_Y$ similarly. Since the true Δ_0 satisfies $\Sigma_X\Delta_0\Sigma_Y-(\Sigma_X-\Sigma_Y)=0$, a sensible estimation procedure would solve $\hat{\Sigma}_X\Delta\hat{\Sigma}_Y-(\hat{\Sigma}_X-\hat{\Sigma}_Y)=0$ for Δ. When $\min(n_X,n_Y)<p$ there are an infinite number of solutions, but accurate estimation is still possible when Δ_0 is sparse. Motivated by the constrained ℓ_1 minimization approach to precision matrix estimation of Cai et al. (2011), one estimator can be obtained by solving

$$\arg\min|\Delta|_1 \quad \text{subject to} \quad |\hat{\Sigma}_X\Delta\hat{\Sigma}_Y-\hat{\Sigma}_X+\hat{\Sigma}_Y|_\infty\leqslant\lambda_n$$

and then symmetrizing the solution. This is equivalent to a linear program, as for any three $p\times p$ matrices A, B and C, $\mathrm{vec}(ABC)=(C^{\mathrm{T}}\otimes A)\,\mathrm{vec}(B)$ where $\otimes$ denotes the Kronecker product and $\mathrm{vec}(B)$ denotes the $p^2\times 1$ vector obtained by stacking the columns of B. Therefore Δ_0 could be estimated by solving and then symmetrizing

$$\arg\min|\Delta|_1 \quad \text{subject to} \quad |(\hat{\Sigma}_Y\otimes\hat{\Sigma}_X)\,\mathrm{vec}(\Delta)-\mathrm{vec}(\hat{\Sigma}_X-\hat{\Sigma}_Y)|_\infty\leqslant\lambda_n. \quad (2)$$

This approach directly estimates the difference matrix without even implicitly estimating the individual precision matrices. The key is that sparsity is assumed for Δ_0 and not for Σ_X^{-1} or Σ_Y^{-1}.

Direct estimation thus allows the presence of hub nodes in the individual networks and can still achieve accurate support recovery and estimation in high dimensions, as will be discussed in § 4. A similar direct estimation approach was proposed by Cai & Liu (2011) for high-dimensional linear discriminant analysis. Linear discriminant analysis depends on the product of a precision matrix and the difference between two mean vectors, and Cai & Liu (2011) showed that direct estimation of this product is possible even in cases where the precision matrix or the mean difference are not individually estimable.

3·2. *A modified problem*

The linear program (2) has a $p^2 \times p^2$ constraint matrix $\hat{\Sigma}_Y \otimes \hat{\Sigma}_X$ and can become computationally demanding for large p. A modified procedure can alleviate this burden by requiring the estimate to be symmetric. Denote the (j, k)th entry of a matrix Δ by δ_{jk}, and define β to be the $p(p+1)/2 \times 1$ vector with $\beta = (\delta_{jk})_{1 \leqslant j \leqslant k \leqslant p}$. Estimating a symmetric Δ is thus equivalent to estimating β, which has only $p(p+1)/2$ parameters. Define the $p^2 \times p(p+1)/2$ matrix S with columns indexed by $1 \leqslant j \leqslant k \leqslant p$ and with rows indexed by $l = 1, \ldots, p$ and $m = 1, \ldots, p$, so that each entry is labelled by $S_{lm,jk}$. For $j \leqslant k$, let $S_{jk,jk} = S_{kj,jk} = 1$ and set all other entries of S equal to zero. For example, when $p = 3$,

$$S = \begin{pmatrix}
S_{11,11} & S_{11,12} & S_{11,13} & S_{11,22} & S_{11,23} & S_{11,33} \\
S_{21,11} & S_{21,12} & S_{21,13} & S_{21,22} & S_{21,23} & S_{21,33} \\
S_{31,11} & S_{31,12} & S_{31,13} & S_{31,22} & S_{31,23} & S_{31,33} \\
S_{12,11} & S_{12,12} & S_{12,13} & S_{12,22} & S_{12,23} & S_{21,33} \\
S_{22,11} & S_{22,12} & S_{22,13} & S_{22,22} & S_{22,23} & S_{11,33} \\
S_{32,11} & S_{32,12} & S_{32,13} & S_{32,22} & S_{32,23} & S_{32,33} \\
S_{13,11} & S_{13,12} & S_{13,13} & S_{13,22} & S_{13,23} & S_{23,33} \\
S_{23,11} & S_{23,12} & S_{23,13} & S_{23,22} & S_{23,23} & S_{13,33} \\
S_{33,11} & S_{33,12} & S_{33,13} & S_{33,22} & S_{33,23} & S_{33,33}
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

When Δ is symmetric, some calculation shows that $(\hat{\Sigma}_Y \otimes \hat{\Sigma}_X)\,\mathrm{vec}(\Delta) = (\hat{\Sigma}_Y \otimes \hat{\Sigma}_X)S\beta$. Furthermore, if $\beta_0 = (\delta^0_{jk})_{1 \leqslant j \leqslant k \leqslant p}$, where δ^0_{jk} is the (j, k)th entry of Δ_0, then by Lemma A1 β_0 is the unique solution to $S^{\mathrm{T}}(\Sigma_Y \otimes \Sigma_X)S\beta - S^{\mathrm{T}}\,\mathrm{vec}(\Sigma_X - \Sigma_Y) = 0$. Therefore one reasonable way to estimate a sparse β_0 in high dimensions is to solve

$$\hat{\beta} = \arg\min |\beta|_1 \quad \text{subject to} \quad |S^{\mathrm{T}}(\hat{\Sigma}_Y \otimes \hat{\Sigma}_X)S\beta - S^{\mathrm{T}}\,\mathrm{vec}(\hat{\Sigma}_X - \hat{\Sigma}_Y)|_\infty \leqslant \lambda_n.$$

However, the inequality constraints can be improved. Let E be the $p \times p$ matrix such that $\mathrm{vec}(E) = (\hat{\Sigma}_Y \otimes \hat{\Sigma}_X)S\beta$. These constraints treat the diagonals and off-diagonals of $E - (\hat{\Sigma}_X - \hat{\Sigma}_Y)$ differently, with the diagonals constrained roughly half as much as the off-diagonals.

Therefore the remainder of this paper considers the estimate of Δ_0 obtained by solving

$$\hat{\beta} = \arg\min |\beta|_1 \quad \text{subject to} \quad \begin{cases} |S^{\mathrm{T}}\hat{\Sigma}S\beta - S^{\mathrm{T}}\hat{b}|_{\mathrm{O}\infty} \leqslant \lambda_n, \\ |S^{\mathrm{T}}\hat{\Sigma}S\beta - S^{\mathrm{T}}\hat{b}|_{\mathrm{D}\infty} \leqslant \lambda_n/2, \end{cases} \tag{3}$$

where $\hat{\Sigma} = \hat{\Sigma}_Y \otimes \hat{\Sigma}_X$, $\hat{b} = \mathrm{vec}(\hat{\Sigma}_X - \hat{\Sigma}_Y)$ and, for a $p(p+1)/2 \times 1$ vector c, $|c|_{\mathrm{O}\infty}$ denotes the sup-norm of the entries of c corresponding to the off-diagonal elements of its matrix form, while $|c|_{\mathrm{D}\infty}$ denotes the sup-norm of the entries corresponding to the diagonal elements. The matrix form of $\hat{\beta}$ will be denoted by $\hat{\Delta}$. Compared with (2), the estimator (3) requires only a

$p(p+1)/2 \times p(p+1)/2$ constraint matrix, but needs a stronger theoretical condition to guarantee support recovery and estimation consistency, which is discussed in § 4.

3·3. *Implementation*

The estimator (3) can be computed by slightly modifying code from the R (R Development Core Team, 2014) package flare, recently developed by Li et al. (2013) to implement a variety of high-dimensional linear regression and precision matrix estimation methods. Their algorithm uses the alternating direction method of multipliers; for a thorough discussion see Boyd et al. (2011). To apply their algorithm, rewrite (3) as

$$\hat{\beta} = \arg\min_{r,\beta} f(r) + |\beta|_1 \quad \text{subject to} \quad r + S^{\mathrm{T}}\hat{\Sigma}S\beta = S^{\mathrm{T}}\hat{b},$$

where $f(r)$ equals infinity if $|r|_{\mathrm{O}\infty} > \lambda_n$ or $|r|_{\mathrm{D}\infty} > \lambda_n/2$ and zero otherwise. The augmented Lagrangian is then

$$L_\rho(r, \beta, y) = f(r) + |\beta|_1 + u^{\mathrm{T}}(r + S^{\mathrm{T}}\hat{\Sigma}S\beta - S^{\mathrm{T}}\hat{b}) + (\rho/2)\|r + S^{\mathrm{T}}\hat{\Sigma}S\beta - S^{\mathrm{T}}\hat{b}\|_2^2,$$

where u is the Lagrange multiplier and $\rho > 0$ is a penalty parameter specified by the user. The alternating direction method of multipliers obtains the solution by using the updates

$$\begin{aligned}
r^{(t+1)} &= \arg\min_{r} \|u^{(t)}/\rho + S^{\mathrm{T}}\hat{b} - S^{\mathrm{T}}\hat{\Sigma}S\beta^{(t)} - r\|_2^2/2 + f(r)/\rho,\\
\beta^{(t+1)} &= \arg\min_{\beta} \|u^{(t)}/\rho - r^{(t+1)} + S^{\mathrm{T}}\hat{b} - S^{\mathrm{T}}\hat{\Sigma}S\beta\|_2^2/2 + |\beta|_1/\rho,\\
u^{(t+1)} &= u^{(t)} + \rho(S^{\mathrm{T}}\hat{b} - r^{(t+1)} - S^{\mathrm{T}}\hat{\Sigma}S\beta^{(t+1)}),
\end{aligned}$$

for each iteration t. The flare package incorporates several strategies to speed convergence, such as using a closed-form expression for $r^{(t+1)}$, employing a hybrid coordinate descent and linearization procedure to obtain $\beta^{(t+1)}$, and dynamically adjusting ρ at each iteration.

The direct estimation approach can be tuned using an approximate Akaike information criterion. For the loss functions

$$L_\infty(\lambda_n) = |\hat{\Sigma}_X\hat{\Delta}(\lambda_n)\hat{\Sigma}_Y - \hat{\Sigma}_X + \hat{\Sigma}_Y|_\infty, \quad L_{\mathrm{F}}(\lambda_n) = \|\hat{\Sigma}_X\hat{\Delta}(\lambda_n)\hat{\Sigma}_Y - \hat{\Sigma}_X + \hat{\Sigma}_Y\|_{\mathrm{F}}, \tag{4}$$

where $\hat{\Delta}(\lambda_n)$ makes explicit the dependence of the estimator on the tuning parameter, λ_n is chosen to minimize

$$(n_X + n_Y)L(\lambda_n) + 2k, \tag{5}$$

where $L(\lambda_n)$ represents either L_∞ or L_{F} and k is the effective degrees of freedom, which can be approximated by $k = |\hat{\beta}|_0$, or the number of nonzero elements in the upper triangular part of $\hat{\Delta}$. The loss functions (4) focus on the supremum and Frobenius norms in light of Theorems 2 and 3 in § 4, but other norms could be used as well.

4. Theoretical properties

Let σ_{jk}^X and σ_{jk}^Y be the (j,k)th entries of Σ_X and Σ_Y, respectively. Define $\sigma_{\max}^X = \max_j \sigma_{jj}^X$ and $\sigma_{\max}^Y = \max_j \sigma_{jj}^Y$. Good performance of direct estimation requires the following conditions.

Condition 1. The true difference matrix Δ_0 has $s < p$ nonzero entries in its upper triangular part, and $|\Delta_0|_1 \leqslant M$, where M does not depend on p.

Condition 2. With s defined as in Condition 1, the constants $\mu_X = \max_{j \neq k} |\sigma^X_{jk}|$ and $\mu_Y = \max_{j \neq k} |\sigma^Y_{jk}|$ must satisfy $\mu = 4\max(\mu_X \sigma^Y_{\max}, \mu_Y \sigma^X_{\max}) \leqslant \sigma^S_{\min}(2s)^{-1}$, where $\sigma^S_{\min} = \min_{j,k}(\sigma^Y_{jj}\sigma^X_{jj}, \sigma^Y_{kk}\sigma^X_{jj} + 2\sigma^Y_{kj}\sigma^X_{jk} + \sigma^Y_{jj}\sigma^X_{kk})$.

Condition 1 requires the difference matrix to have essentially constant sparsity, which is reasonable because genetic networks are not expected to differ much between two conditions. Condition 2 requires that the true covariances between the covariates are not too high, and can hold even when Σ_X^{-1} and Σ_Y^{-1} are not sparse. Actually it is sufficient to require only that the magnitude of the largest off-diagonal entry of $S^{\mathrm{T}}(\Sigma_Y \otimes \Sigma_X)S$ be less than $\sigma^S_{\min}/2s$, but Condition 2 is more interpretable.

Condition 2 is closely related to the mutual incoherence property introduced by Donoho & Huo (2001), but is more complicated in the current setting because it involves a linear function of the Kronecker product of two covariance matrices. Solving (2) instead of (3) would require only $\max(\mu_X \sigma^Y_{\max}, \mu_Y \sigma^X_{\max}) \leqslant \min_{j,k}(\sigma^X_{jj}\sigma^Y_{kk})(2\tilde{s})^{-1}$, with $\tilde{s}$ equal to the total number of nonzero entries of Δ_0. If in addition $\sigma^X_{jj} = \sigma^Y_{jj} = 1$ for all j, $\max(\mu_X, \mu_Y) \leqslant (2\tilde{s})^{-1}$ would be required, which is similar to imposing the usual mutual incoherence condition on Σ_X and Σ_Y. Condition 2 is more restrictive, but (3) is easy to compute and still gives good finite-sample results.

Under these conditions, a thresholded version of the direct estimator $\hat{\Delta}$ can successfully recover the support of Δ_0. Let the (j,k)th entries of Δ_0 and $\hat{\Delta}$ be δ^0_{jk} and $\hat{\delta}_{jk}$, respectively. For a threshold $\tau_n > 0$, define the estimator

$$\hat{\Delta}_{\tau_n} = \{\hat{\delta}_{jk} I(|\hat{\delta}_{jk}| > \tau_n)\}.$$

Let the (j,k)th entry of $\hat{\Delta}_{\tau_n}$ be $\hat{\delta}^{\tau_n}_{jk}$, and define the function

$$\mathrm{sgn}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases}$$

Then, if $\mathcal{M}(\hat{\Delta}_{\tau_n}) = \{\mathrm{sgn}(\hat{\delta}^{\tau_n}_{jk}) : j = 1, \ldots, p;\ k = 1, \ldots, p\}$ and $\mathcal{M}(\Delta_0) = \{\mathrm{sgn}(\delta^0_{jk}) : j = 1, \ldots, p;\ k = 1, \ldots, p\}$ are vectors of the signs of the entries of the estimated and true difference matrices, respectively, the following theorem holds.

THEOREM 1. *Suppose that Conditions 1 and 2 hold, and let* $\sigma^S_{\min}$ *and* μ *be as defined in Condition 2. If* $\min(n_X, n_Y) > \log p$,

$$\tau_n \geqslant \frac{1}{\sigma^S_{\min}} \left\{ 1 + \frac{\sigma^S_{\min}}{\sigma^S_{\min} - (2s-1)\mu} \right\} \{M(|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C) + 1\} 8C \left\{ \frac{\log p}{\min(n_X, n_Y)} \right\}^{1/2}$$

and $\min_{j,k:\delta^0_{jk} \neq 0} |\delta^0_{jk}| > 2\tau_n$, *then* $\mathcal{M}(\hat{\Delta}_{\tau_n}) = \mathcal{M}(\Delta_0)$ *with probability at least* $1 - 8p^{-\tau}$, *where* C *is defined in Lemma A2 in the Appendix.*

Theorem 1 states that with high probability, $\hat{\Delta}_{\tau_n}$ can recover not only the support of Δ_0 but also the signs of its nonzero entries, as long as those entries are sufficiently large. In other words, in the context of genetic networks, $\hat{\Delta}_{\tau_n}$ can correctly identify genes whose conditional dependencies change in magnitude between two conditions, as well as the directions of those changes, as long

as $\min(n_X, n_Y)$ is large relative to $\log p$. In practice, the threshold τ_n can be treated as a tuning parameter. In simulations and data analysis, τ_n was set to 0·0001.

A thresholding step is natural in practice because small entries of $\hat{\Delta}$ are probably noisy estimates of zero. This step could be avoided by imposing an irrepresentability condition on $\Sigma_X \otimes \Sigma_Y$, similar to the conditions assumed in the proofs of the selection consistencies of the lasso (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006) and the graphical lasso (Ravikumar et al., 2008). However, these types of condition are stronger than the mutual incoherence-type property assumed in Condition 2, as discussed in Lounici (2008). The thresholded estimators are pursued in this paper because of their milder theoretical requirements.

In addition to identifying the entries of Σ_X^{-1} and Σ_Y^{-1} that change, $\hat{\Delta}$ can correctly quantify these changes, in the sense of being consistent for Δ_0 in the Frobenius norm.

THEOREM 2. *Suppose that Conditions 1 and 2 hold, and define* $\sigma^S_{\min}$ *and* μ *as in Condition 2. If* $\min(n_X, n_Y) > \log p$ *and*

$$\lambda_n = \{M(|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C) + 1\}4C \left\{ \frac{\log p}{\min(n_X, n_Y)} \right\}^{1/2},$$

then

$$\|\hat{\Delta} - \Delta_0\|_{\mathrm{F}} \leqslant \frac{(5s)^{1/2}}{\sigma^S_{\min}} \left\{ 1 + \frac{\sigma^S_{\min}}{\sigma^S_{\min} - (2s-1)\mu} \right\} 2\lambda_n$$

with probability at least $1 - 8p^{-\tau}$, *where C is defined in Lemma A2 in the Appendix.*

The proofs of Theorems 1 and 2 rely on the following bound on the elementwise ℓ_∞ norm of the estimation error.

THEOREM 3. *Suppose that Conditions 1 and 2 hold, and define* $\sigma^S_{\min}$ *and* μ *as in Condition 2. If* $\min(n_X, n_Y) > \log p$ *and*

$$\lambda_n = \{M(|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C) + 1\}4C \left\{ \frac{\log p}{\min(n_X, n_Y)} \right\}^{1/2},$$

then

$$|\hat{\Delta} - \Delta_0|_\infty \leqslant \frac{1}{\sigma^S_{\min}} \left\{ 1 + \frac{\sigma^S_{\min}}{\sigma^S_{\min} - (2s-1)\mu} \right\} 2\lambda_n$$

with probability at least $1 - 8p^{-\tau}$, *where C is defined in Lemma A2 in the Appendix.*

Similar theoretical properties have been derived for separate and joint approaches to estimating differential networks (Cai et al., 2011; Guo et al., 2011). However, these require sparsity conditions on each Σ^{-1}, such as $\|\Sigma^{-1}\|_\infty \leqslant M' < \infty$, which can be violated if the individual networks contain hub nodes. In contrast, Theorems 1–3 can still hold in the presence of hubs.

5. SIMULATIONS

5·1. *Settings*

Simulations were conducted to compare the direct estimation given by (3), the fused graphical lasso given in (1), and separate estimation using the procedure of Cai et al. (2011). Data were

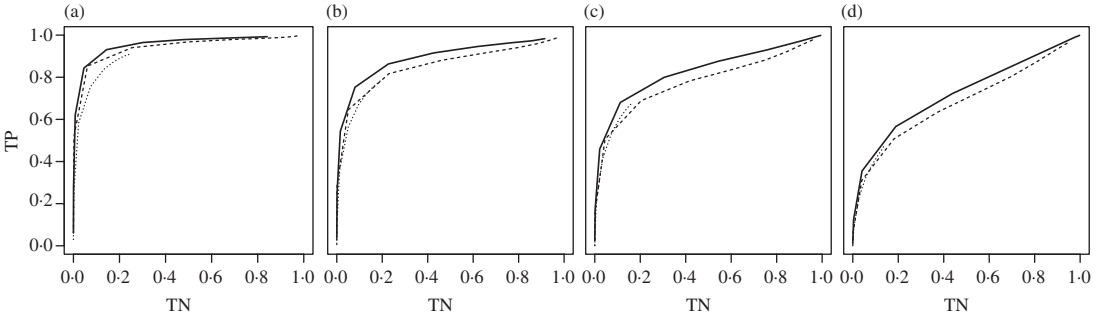

Fig. 1. Receiver operating characteristic curves for support recovery of $\Delta_0 = \Sigma_Y^{-1} - \Sigma_X^{-1}$, with (a) $p = 40$, (b) $p = 60$, (c) $p = 90$, and (d) $p = 120$. In each panel, TP and TN are the true positive and true negative rates, respectively, defined in § 5·2; the solid line represents the thresholded direct estimator, the dashed line represents the thresholded fused graphical lasso estimator with $\lambda_1 = 0$, and the dotted line represents the fused graphical lasso estimator with $\lambda_1 = 0{\cdot}1$.

generated with $p = 40$, 60, 90 and 120, and $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ were generated from $N(0, \Sigma_X)$ and $N(0, \Sigma_Y)$, respectively, with $n_X = n_Y = 100$.

For each p, the support of Σ_X^{-1} was first generated according to a network with $p(p-1)/10$ edges and a power-law degree distribution with an expected power parameter of 2, which should mimic real-world networks (Newman, 2003). This still gives a relatively sparse network, since only 20% of all possible edges are present, but the power-law structure creates hub nodes, which make certain rows and columns nonsparse.

Next, the value of each nonzero entry of Σ_X^{-1} was generated from a uniform distribution with support $[-0{\cdot}5, -0{\cdot}2] \cup [0{\cdot}2, 0{\cdot}5]$. To ensure positive definiteness, each row was divided by two when $p = 40$, three when $p = 60$, four when $p = 90$, and five when $p = 120$. The diagonals were then set equal to one and the matrix was symmetrized by averaging it with its transpose. The differential network Δ_0 was generated such that the largest 20%, by magnitude, of the connections of the top two hub nodes of Σ_X^{-1} changed sign between Σ_X^{-1} and Σ_Y^{-1}. In other words, Δ_0 was a sparse matrix, with zero entries everywhere except for 20% of the entries in two rows and columns.

Each method was tuned using an approximate Akaike information criterion. Direct estimation was tuned using (5) and one of the loss functions in (4). For a fair comparison, the fused graphical lasso was tuned in the same way, after searching across all combinations of three values of λ_1 and ten values of λ_2. Small values of λ_1 were used because the true precision matrices were nonsparse. Separate estimation was tuned by searching across ten different values of the tuning parameter to minimize $\text{AIC}_X = n_X \operatorname{tr}(\hat{\Sigma}_X \hat{\Omega}_X) - n_X \log\det(\hat{\Omega}_X) + 2|\hat{\Omega}_X|_0$, where $\hat{\Sigma}_X$ was the sample covariance matrix of the X_i and $\hat{\Omega}_X$ was the estimated precision matrix. The same was done for the Y_i, with AIC_Y defined similarly. Results were averaged over 250 replications.

5·2. *Results*

Figure 1 shows the receiver operating characteristic curves of the three estimation methods. Let $\hat{\delta}_{jk}$ be the (j, k)th entry of a given estimator $\hat{\Delta}$, and let δ^0_{jk} be the (j, k)th entry of the true Δ_0. The true positive and true negative rates of $\hat{\Delta}$ are defined as

$$\text{TP} = \frac{\sum_{jk} I(\hat{\delta}_{jk} \neq 0,\ \delta^0_{jk} \neq 0)}{\sum_{jk} I(\delta^0_{jk} \neq 0)}, \qquad \text{TN} = \frac{\sum_{jk} I(\hat{\delta}_{jk} = 0,\ \delta^0_{jk} = 0)}{\sum_{jk} I(\delta^0_{jk} = 0)},$$

Table 1. *Average true discovery rates over* 250 *simulations, with standard errors given in parentheses*

	$\hat{\Delta}_{\tau_n}$		$\hat{\Delta}_{\mathrm{FGL}\tau_n}$		$\hat{\Delta}_{\mathrm{S}\tau_n}$
p	L_∞	L_F	L_∞	L_F	
40	77 (16)	29 (16)	83 (12)	27 (17)	2 (0)
60	76 (19)	66 (21)	74 (19)	65 (30)	1 (0)
90	66 (25)	80 (31)	55 (26)	12 (28)	1 (0)
120	48 (39)	61 (45)	33 (25)	1 (7)	1 (0)

$\hat{\Delta}_{\tau_n}$, thresholded direct estimator; $\hat{\Delta}_{\mathrm{FGL}\tau_n}$, thresholded fused graphical lasso estimator; $\hat{\Delta}_{\mathrm{S}\tau_n}$, thresholded separate estimator; the direct and fused graphical lasso estimators were tuned using (5) and either L_∞ or L_F from (4).

respectively. Different points on the curves correspond to different tuning parameter values. The curves for the fused graphical lasso estimator (1) were plotted by varying λ_2. The λ_1 parameter, which controls the sparsity of the individual precision matrix estimates, was fixed at a small value because the individual matrices were not sparse. For a fair comparison with the thresholded direct estimator, the fused graphical lasso was thresholded at 0·0001. The separate estimator performed poorly and its curves are not plotted. Figure 1 shows that direct estimation compares favourably with the fused graphical lasso.

The true discovery and nondiscovery rates of the three estimators were studied as well, which are defined as

$$\mathrm{TD} = \frac{\sum_{jk} I(\hat{\delta}_{jk} \neq 0,\ \delta^0_{jk} \neq 0)}{\sum_{jk} I(\hat{\delta}_{jk} \neq 0)}, \qquad \mathrm{TND} = \frac{\sum_{jk} I(\hat{\delta}_{jk} = 0,\ \delta^0_{jk} = 0)}{\sum_{jk} I(\hat{\delta}_{jk} = 0)},$$

respectively. These rates are taken to be zero when their denominators are equal to zero. In the analysis of genomic data, minimizing the number of false discoveries is a major concern. The direct and fused graphical lasso estimators were thresholded at 0·0001, and the separate estimator was thresholded at 0·0002. In all settings, the true nondiscovery rates of the direct and fused graphical lasso estimators were close to 100%; separate estimation frequently did not identify any zero entries in the differential network. The true discovery rates are reported in Table 1, which also compares the effects of tuning using different loss functions (4). For direct estimation, tuning using L_∞ gave the best true discovery rates for smaller p, while L_F was preferable for larger p. For the fused graphical lasso, L_∞ was always the better choice. Using either L_∞ or L_F, direct estimation performed well compared to the fused graphical lasso and separate estimation, especially for larger p.

The Frobenius norm estimation accuracies of the unthresholded estimators tuned using the loss functions L_∞ and L_F are reported in Table 2. The different loss functions gave comparable results. Direct estimation was much more accurate than separate estimation and slightly more accurate than the fused graphical lasso. It is possible for direct estimation to simultaneously give markedly better support recovery but similar estimation compared to the fused graphical lasso, because estimation error depends on the magnitudes of the estimated entries, while support recovery depends only on whether the entries are nonzero. For example, suppose that $\hat{\Delta}$ had the same support as the true Δ_0 but that each nonzero entry had magnitude 0·01. Then, under the $p = 120$ simulation setting, $\|\hat{\Delta} - \Delta_0\|_\mathrm{F} = 1{\cdot}56$. The estimation error of this $\hat{\Delta}$ is even higher than those in Table 2, but it exactly recovers the true support.

The good performance of direct estimation comes at the price of some computational convenience. The memory required by the large constraint matrix behaves like $O(p^4)$, although the simulations generally needed no more than one gigabyte of memory when $p = 120$. In a 2012

Table 2. *Average estimation errors in the Frobenius norm over* 250 *simulations, with standard errors given in parentheses*

	$\hat{\Delta}$		$\hat{\Delta}_{\mathrm{FGL}}$		$\hat{\Delta}_{\mathrm{S}}$
p	L_∞	L_{F}	L_∞	L_{F}	
40	1·67 (0·13)	1·46 (0·22)	1·72 (0·06)	1·50 (0·15)	12·96 (0·78)
60	1·68 (0·07)	1·62 (0·11)	1·68 (0·05)	1·69 (0·06)	28·45 (2·01)
90	1·68 (0·04)	1·71 (0·03)	1·68 (0·03)	1·72 (0·03)	57·30 (2·75)
120	1·55 (0·02)	1·55 (0·01)	1·54 (0·02)	1·55 (0·00)	102·84 (4·06)

$\hat{\Delta}$, direct estimator; $\hat{\Delta}_{\mathrm{FGL}}$, fused graphical lasso estimator; $\hat{\Delta}_{\mathrm{S}}$, separate estimator; the direct and fused graphical lasso estimates were tuned using (5) with either L_∞ or L_{F} from (4).

unpublished technical report (arXiv:1208.3922), M. Hong and Z.-Q. Luo proved the global linear convergence of the alternating direction method of multipliers applied to problems like (3). However, each iteration of the proposed algorithm requires roughly $O(sp^4)$ computations, where s is the number of nonzero entries in the upper triangular part of Δ_0, as defined in Condition 1. The simulations required on average 51, 853, 6231 and 51 589 seconds when $p = 40$, 60, 90 and 120, respectively. On the other hand, these memory and time requirements are still reasonable in practice. Recently, Pang et al. (2013) have developed an even faster algorithm for constrained ℓ_1 minimization problems such as (3), available in the R package fastclime (R Development Core Team, 2014; Pang et al., 2013), which should reduce the computational burden of direct estimation.

6. Gene expression study of ovarian cancer

The proposed approach was applied to gene expression data collected from patients with stage III or stage IV ovarian cancer. Using these data, Tothill et al. (2008) identified six molecular subtypes of ovarian cancer, which they labelled C1 to C6. They found that the C1 subtype, characterized by differential expression of genes associated with stromal and immune cell types, was associated with much shorter survival times.

The proposed direct estimation procedure was applied to investigate whether this poor prognosis subtype was also associated with differential wiring of genetic networks. The subjects were divided into a C1 group, with 78 patients, and a C2–C6 group, with 113 patients. Several pathways from the KEGG pathway database (Ogata et al., 1999; Kanehisa et al., 2012) were studied to determine whether any differences existed in the conditional dependency relationships of the gene expression levels between the subtypes. All probe sets corresponding to the same gene symbol were first averaged to obtain gene-level expression measurements.

Direct estimation and fused graphical lasso were tuned using (5) with the loss function L_∞, because in simulations this gave the best results for the fused graphical lasso and good results for direct estimation. Separate estimation was tuned as described in § 3·3. The direct and fused graphical lasso estimators were thresholded at 0·0001 to recover the differential network. The separate estimator was not simply thresholded at 0·0002, as simulations showed that this method gave poor true discovery rates; instead, two genes were defined as being linked in the differential network if they were connected in one group but not the other, or if they were connected in both groups but their conditional dependency relationship changed sign. The procedure of Cai et al. (2011) thresholded at 0·0001 was used to recover the individual networks.

Two illustrative examples are reported in Fig. 2. Only genes included in at least one edge, or which saw a change in partial variance between the two subtypes, were included in the figure. In the results of separate estimation, only genes in the differential network estimated by direct estimation were labelled. To interpret the results, the most highly connected genes in the differential networks were considered to be important.

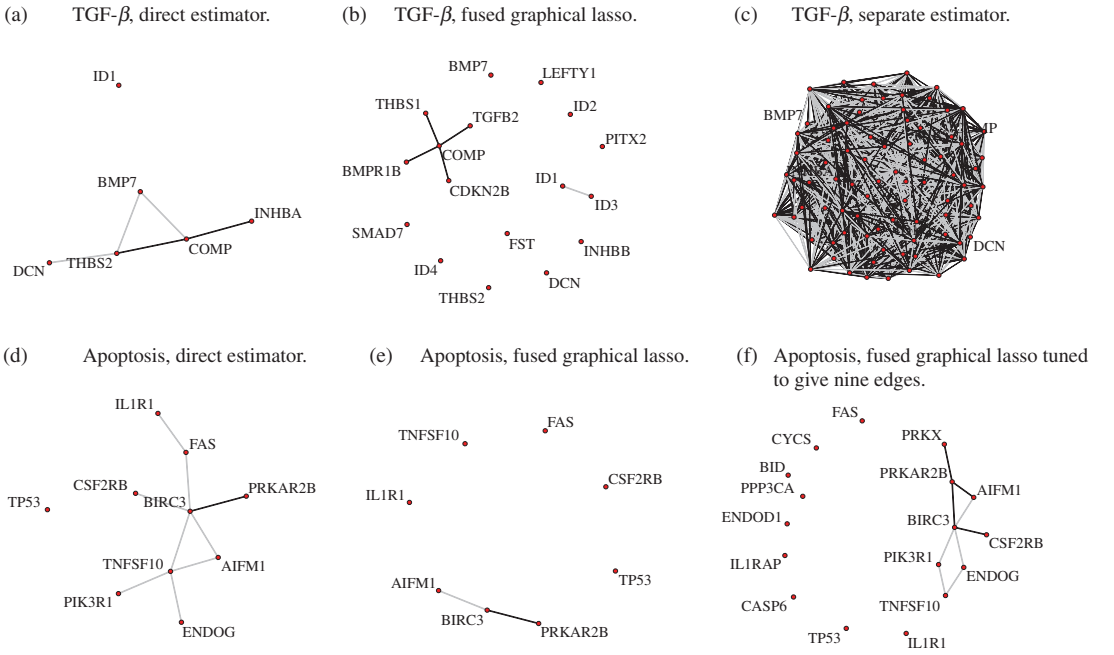

Fig. 2. Estimates of the differential networks between ovarian cancer subtypes: (a)–(c) KEGG 04350, TGF-β pathway; (d)–(f) KEGG 04210, apoptosis pathway. The direct and fused graphical lasso estimators were thresholded, and the separate estimator was further sparsified; see § 6. Black edges show an increase in conditional dependency from ovarian cancer subtype C1 to subtypes C2–C6; grey edges show a decrease. The estimators in (a)–(e) were tuned using L_∞ with (5), and the estimator in (f) was tuned with $\lambda_1 = 0$ and λ_2 to give the same number of edges as (d); see § 6. The separate estimator is not shown for the apoptosis pathway.

Figures 2(a)–(c) illustrate estimates of the differential network of the TGF-β signalling pathway, which at 82 genes is larger than the sample size of the C1 group. Direct estimation suggested the presence of two hub genes, COMP and THBS2, which have both been found to be related to resistance to platinum-based chemotherapy in epithelial ovarian cancer (Marchini et al., 2013). The fused graphical lasso gave the same number of edges in the differential network as direct estimation and only suggested the importance of COMP. It was hard to draw meaningful conclusions from the results of separate estimation, because the denseness of the estimated network made it difficult to identify a small number of important genes.

Figures 2(d)–(f) give the estimates for the apoptosis pathway, which at 87 genes was also larger than the sample size of the C1 group. The separate estimator again resulted in a dense network and is not included in Fig. 2. Direct estimation pointed to BIRC3 and TNFSF10 as being important genes. Indeed, TNFSF10 encodes the TRAIL protein, which has been studied a great deal because of its potential as an anticancer drug (Yagita et al., 2004; Bellail et al., 2009) and, in particular, as a therapy for ovarian cancer (Petrucci et al., 2012; Kipps et al., 2013). BIRC3 can inhibit TRAIL-induced apoptosis (Johnstone et al., 2008) and has also been considered for use as a therapeutic target in cancer (Vucic & Fairbrother, 2007). Figure 2(e) shows the fused graphical lasso estimator tuned in the same way as the direct estimator, and suggests only BIRC3 as being important. For a fairer comparison with direct estimation, Fig. 2(f) depicts the fused graphical lasso estimator after fixing $\lambda_1 = 0$ and adjusting λ_2 to achieve the same level of sparsity as Fig. 2(d). The result is similar to Fig. 2(d), although it suggests that BIRC3 and PRKAR2B, a protein kinase, are important, rather than BIRC3 and TRAIL.

7. Discussion

Instead of modelling a differential network as the difference of two precision matrices, as proposed above, another possibility is to use the difference between two directed acyclic graphs. These graphs are natural models for single transcriptional regulatory networks, with nodes representing gene expression levels and edges indicating how the nodes are causally related to each other. Biological changes to a network can be thought of as interventions on some of its nodes, which result in changes to the graphical structure (Hauser & Bühlmann, 2014). Frequently, however, only observational data are available for gene expression, so it is difficult to estimate the underlying causal structures (Kalisch & Bühlmann, 2007; Maathuis et al., 2009). It would be interesting to develop a direct estimation method for differential networks with interventional data.

For method (3) to have good properties in high dimensions, Δ_0 must be sparse. While reasonable, this assumption will be violated if the biological differences between two groups manifest as global changes that affect a large number of gene-gene dependencies. If some proportion of these global changes are of sufficient magnitude, the method should still be able to detect their presence, though it may not recover all of the changes or accurately estimate their magnitudes. The most challenging case for the proposed method occurs when the network changes are numerous but small. A new statistic could be defined to quantify the degree of global change between two precision matrices, but so far there is little consensus as to what statistic would be most biologically meaningful.

Finally, while the focus has been on directly estimating the difference between two precision matrices, there are situations where interest may centre on how a transcriptional regulatory network differs between K conditions, where $K > 2$. The proposed method could of course be used to estimate all pairwise differential networks, but this could be time-consuming. Another possibility would be to estimate the difference between each precision matrix and some common precision matrix, which could be taken to be the inverse of the pooled covariance matrix of all K groups. In other words, if $\hat{\Sigma}_k$ were the sample covariance matrix of the kth group, let $\hat{\Sigma}_{\mathrm{P}} = \sum_k w_k \hat{\Sigma}_k$ be some weighted average of the $\hat{\Sigma}_k$ and consider solving

$$\hat{\Delta}_k = \arg\min |\Delta|_1 \quad \text{subject to} \quad |\hat{\Sigma}_k \Delta \hat{\Sigma}_{\mathrm{P}} - \hat{\Sigma}_k + \hat{\Sigma}_{\mathrm{P}}|_\infty \leqslant \lambda_n.$$

If the difference matrix $\Delta_k = \Sigma_{\mathrm{P}}^{-1} - \Sigma_k^{-1}$ were sparse, where $\Sigma_{\mathrm{P}} = E(\hat{\Sigma}_{\mathrm{P}})$ and Σ_k^{-1} is the precision matrix of the kth group, $\hat{\Delta}_k$ would be a direct estimate of Δ_k. The differential network between the jth and kth group could then be estimated as $\hat{\Delta}_j - \hat{\Delta}_k$.

Acknowledgement

This research was supported by the U.S. National Institutes of Health and the National Science Foundation. The authors thank Xingguo Li for sharing his expertise regarding the computational complexity of the proposed algorithm.

Appendix

Proofs of theorems

Lemma A1. *The matrix $S^{\mathrm{T}}(\Sigma_Y \otimes \Sigma_X)S$ is invertible, where Σ_Y and Σ_X are $p \times p$ covariance matrices and S is as defined in* § 3·2.

Proof. Since Σ_X and Σ_Y are positive definite, there exists a full-rank matrix $\Sigma^{1/2}$ such that $\Sigma_Y \otimes \Sigma_X = (\Sigma^{1/2})^{\mathrm{T}}\Sigma^{1/2}$. Furthermore, from its construction S has full column rank, so $\mathrm{rank}(S) = p(p+1)/2$. Therefore

$$\mathrm{rank}\{S^{\mathrm{T}}(\Sigma_Y \otimes \Sigma_X)S\} = \mathrm{rank}\{S^{\mathrm{T}}(\Sigma^{1/2})^{\mathrm{T}}\Sigma^{1/2}S\} = \mathrm{rank}(\Sigma^{1/2}S) = \mathrm{rank}(S).$$

Since $S^{\mathrm{T}}(\Sigma_Y \otimes \Sigma_X)S$ is $p(p+1)/2 \times p(p+1)/2$, it is of full rank and therefore invertible. □

The next lemma comes from the proofs of Theorems 1(a) and 4(a) in Cai et al. (2011).

LEMMA A2. *Let $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n$ be independent and identically distributed random vectors with $E(X_i) = (\mu_1, \ldots, \mu_j)^{\mathrm{T}}$, and let $\bar{X} = n^{-1}X_i$ and $\hat{\Sigma} = n^{-1}\sum_i (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$. If there exists some $0 < \eta < 1/4$ such that $\log p/n \leqslant \eta$ and $E[\exp\{t(X_{ij} - \mu_j)^2\}] \leqslant K < \infty$ for all $|t| \leqslant \eta$ and $j = 1, \ldots, p$, then*

$$|\hat{\Sigma} - \Sigma|_\infty \leqslant C(\log p/n)^{1/2}$$

with probability at least $1 - 4p^{-\tau}$, where $C = 2\eta^{-2}(2 + \tau + \eta^{-1}\mathrm{e}^2K^2)^2$ and $\tau > 0$.

LEMMA A3. *Let $\Sigma = \Sigma_Y \otimes \Sigma_X$. Label the entries of $S^{\mathrm{T}}\Sigma S$ as $\sigma^S_{j'k',jk}$ $(1 \leqslant j' \leqslant k' \leqslant p;\ 1 \leqslant j \leqslant k \leqslant p)$. Then*

$$\begin{aligned}
\sigma^S_{j'k',jk} &= \sigma^Y_{k'k}\sigma^X_{j'j} + \sigma^Y_{k'j}\sigma^X_{j'k} + \sigma^Y_{j'k}\sigma^X_{k'j} + \sigma^Y_{j'j}\sigma^X_{k'k}, && j' \neq k',\ j \neq k;\\
\sigma^S_{j'k',jj} &= \sigma^Y_{k'j}\sigma^X_{j'j} + \sigma^Y_{j'j}\sigma^X_{k'j}, && j' \neq k',\ j = k;\\
\sigma^S_{j'j',jk} &= \sigma^Y_{j'k}\sigma^X_{j'j} + \sigma^Y_{j'j}\sigma^X_{j'k}, && j' = k',\ j \neq k;\\
\sigma^S_{j'j',jj} &= \sigma^Y_{j'j}\sigma^X_{j'j}, && j' = k',\ j = k.
\end{aligned}$$

Proof. Label the entries of Σ as $\sigma_{l'm',lm}$ $(l' = 1, \ldots, p;\ m' = 1, \ldots, p;\ l = 1, \ldots, p;\ m = 1, \ldots, p)$ and the entries of S as $S_{lm,jk}$ $(l = 1, \ldots, p;\ m = 1, \ldots, p;\ 1 \leqslant j \leqslant k \leqslant p)$, as in § 3·3. By the definition of the Kronecker product, $\sigma_{l'm',lm} = \sigma^Y_{m'm}\sigma^X_{l'l}$ and $\sigma^S_{j'k',jk} = \sum_{l',m',l,m} S_{l'm',j'k'}\sigma_{l'm',lm}S_{lm,jk}$, so the lemma follows from the definition of the entries of S. □

Proof of Theorem 3. Let the entries of Δ_0 be denoted by δ^0_{jk}, and define the $p(p+1)/2 \times 1$ vector $\beta_0 = (\delta^0_{jk})_{1 \leqslant j \leqslant k \leqslant p}$. Define Σ as in Lemma A3, and let $\hat{b} = \mathrm{vec}(\hat{\Sigma}_X - \hat{\Sigma}_Y)$, $b = \mathrm{vec}(\Sigma_X - \Sigma_Y)$ and $h = \hat{\beta} - \beta_0$. The bound on $|\hat{\Delta} - \Delta_0|_\infty = |h|_\infty$ is obtained by following Lounici (2008).

Denote the ath component of $S^{\mathrm{T}}\Sigma Sh$ by $(S^{\mathrm{T}}\Sigma Sh)_a$, the (a, b)th entry of $S^{\mathrm{T}}\Sigma S$ by σ^S_{ab}, and the bth component of h by h_b. Also let $\mu = \max_{a \neq b}|\sigma^S_{ab}|$. Then $(S^{\mathrm{T}}\Sigma Sh)_a = \sum_{b=1}\sigma^S_{ab}h_b = \sigma^S_{aa}h_a + \sum_{b \neq a}\sigma^S_{ab}h_b$, which implies that

$$|\sigma^S_{aa}h_a| \leqslant |S^{\mathrm{T}}\Sigma Sh|_\infty + \mu\,|\sigma^S_{ab}|\sum_{b \neq a}|h_b|. \tag{A1}$$

The diagonal terms σ^S_{aa} can be relabelled as $\sigma^S_{jk,jk}$, where j may equal k, and from Lemma A3 they must satisfy $\sigma^S_{jk,jk} \geqslant \sigma^S_{\min}$, with $\sigma^S_{\min}$ as defined in Condition 2. The off-diagonal terms σ^S_{ab}, $a \neq b$, can be relabelled as $\sigma^S_{j'k',jk}$ with $j' \neq j$ or $k' \neq k$, and by Lemma A3 they must satisfy $\sigma^S_{j'k',jk} \leqslant 4\max(\mu_X\sigma^X_{\max}, \mu_Y\sigma^Y_{\max}) = \mu$, with μ_X and μ_Y defined as in Condition 2. Using these facts and Condition 2, (A1) becomes

$$|h|_\infty \leqslant \frac{1}{\sigma^S_{\min}}\left(|S^{\mathrm{T}}\Sigma Sh|_\infty + \frac{\sigma^S_{\min}}{2s}|h|_1\right). \tag{A2}$$

The method of Cai et al. (2010b) is used to bound $|h|_1$. Let T_0 be the set of indices corresponding to the support of β_0, and for any $p \times 1$ vector $a = (a_1, \ldots, a_p)^{\mathrm{T}}$ let a_{T_0} be the vector with components $a_{T_0 j} = 0$ for $j \notin T_0$ and $a_{T_0 j} = a_j$ for $j \in T_0$.

First it must be shown that β_0 is in the feasible set with high probability. Since X_i and Y_i are both Gaussian, they satisfy the conditions of Lemma A2, and thus $|\hat{\Sigma}_X - \Sigma_X|_\infty$ and $|\hat{\Sigma}_Y - \Sigma_Y|_\infty$ are both less

than $C\{\log p/\min(n_X, n_Y)\}^{1/2}$ with probability at least $1 - 8p^{-\tau}$. Then

$$\begin{aligned}|S^{\mathrm{T}}\hat{\Sigma}S\beta_0 - S^{\mathrm{T}}\hat{b}|_\infty &\leqslant |S^{\mathrm{T}}(\hat{\Sigma} - \Sigma)S\beta_0|_\infty + |S^{\mathrm{T}}(\hat{b} - b)|_\infty \\ &\leqslant \|S^{\mathrm{T}}\|_\infty|\hat{\Sigma} - \Sigma|_\infty\|S\|_1|\beta_0|_1 + \|S^{\mathrm{T}}\|_\infty(|\hat{\Sigma}_X - \Sigma_X|_\infty + |\hat{\Sigma}_Y - \Sigma_Y|_\infty) \\ &\leqslant 4M|\hat{\Sigma} - \Sigma|_\infty + 4C\{\log p/\min(n_X, n_Y)\}^{1/2},\end{aligned}$$

where $\|S\|_1 = 2$ by the definition of S and $|\beta_0|_1 \leqslant M$ by Condition 1. Next, from the proof of Lemma A3, each entry of Σ can be written as $\sigma^X_{l'l}\sigma^Y_{m'm}$, and so

$$\begin{aligned}|\hat{\sigma}^X_{l'l}\hat{\sigma}^Y_{m'm} - \sigma^X_{l'l}\sigma^Y_{m'm}| &= |\sigma^X_{l'l}(\hat{\sigma}^Y_{m'm} - \sigma^Y_{m'm}) + (\hat{\sigma}^X_{l'l} - \sigma^X_{l'l})\sigma^Y_{m'm} + (\hat{\sigma}^X_{l'l} - \sigma^X_{l'l})(\hat{\sigma}^Y_{m'm} - \sigma^Y_{m'm})| \\ &\leqslant [|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C\{\log p/\min(n_X, n_Y)\}^{1/2}]C\{\log p/\min(n_X, n_Y)\}^{1/2} \\ &\leqslant (|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C)\, C\{\log p/\min(n_X, n_Y)\}^{1/2},\end{aligned}$$

since $\min(n_X, n_Y) > \log p$. Then β_0 is feasible with probability at least $1 - 8p^{-\tau}$ if $\lambda_n = \{M(|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C) + 1\}4C\{\log p/\min(n_X, n_Y)\}^{1/2}$.

Now $|h|_1$ can be bounded. By the definition of (3), $|\beta_0|_1 - |\hat{\beta}|_1 \geqslant 0$. This implies that $|\beta_{0T_0}|_1 - (|\hat{\beta}_{T_0}|_1 + |\hat{\beta}_{T_0^{\mathrm{c}}}|_1) \geqslant 0$. Using the triangle inequality, $|\beta_{0T_0} - \hat{\beta}_{T_0}|_1 \geqslant |\hat{\beta}_{T_0^{\mathrm{c}}}|_1$ or, in other words, $|h_{T_0^{\mathrm{c}}}|_1 \leqslant |h_{T_0}|_1$. Therefore $|h|_1 \leqslant 2|h_{T_0}|_1 \leqslant 2s^{1/2}|h_{T_0}|_2$. To bound $|h_{T_0}|_2$, observe, following Cai et al. (2009), that for any s-sparse vector c,

$$|c^{\mathrm{T}}S^{\mathrm{T}}\Sigma Sc| \geqslant \sum_a \sigma^S_{aa}c_a^2 - \left|\sum_{a\neq b}\sigma^S_{ab}c_ac_b\right| \geqslant \sigma^S_{\min}|c|_2^2 - \mu\sum_{a\neq b}|c_ac_b| \geqslant \sigma^S_{\min}|c|_2^2 - \mu(s-1)|c|_2^2.$$

This implies that

$$\begin{aligned}|h^{\mathrm{T}}_{T_0}S^{\mathrm{T}}\Sigma Sh| &\geqslant |h^{\mathrm{T}}_{T_0}S^{\mathrm{T}}\Sigma Sh_{T_0}| - \left|h^{\mathrm{T}}_{T_0}S^{\mathrm{T}}\Sigma Sh_{T_0^{\mathrm{c}}}\right| \geqslant \{\sigma^S_{\min} - (s-1)\mu\}|h_{T_0}|_2^2 - \left|\sum_{a,b}\sigma^S_{ab}h_{T_0a}h_{T_0^{\mathrm{c}}b}\right| \\ &\geqslant \{\sigma^S_{\min} - (s-1)\mu\}|h_{T_0}|_2^2 - \mu|h_{T_0}|_1|h_{T_0^{\mathrm{c}}}|_1 \geqslant \{\sigma^S_{\min} - (s-1)\mu\}|h_{T_0}|_2^2 - \mu|h_{T_0}|_1^2 \\ &\geqslant \{\sigma^S_{\min} - (2s-1)\mu\}|h_{T_0}|_2^2.\end{aligned}$$

Together with $|h^{\mathrm{T}}_{T_0}S^{\mathrm{T}}\Sigma Sh| \leqslant |h_{T_0}|_1|S^{\mathrm{T}}\Sigma Sh|_\infty \leqslant s^{1/2}|h_{T_0}|_2|S^{\mathrm{T}}\Sigma Sh|_\infty$, this implies that

$$|h|_1 \leqslant 2s^{1/2}|h_{T_0}|_2 \leqslant \frac{2s|S^{\mathrm{T}}\Sigma Sh|_\infty}{\sigma^S_{\min} - (2s-1)\mu},$$

so (A2) becomes

$$|h|_\infty \leqslant \frac{1}{\sigma^S_{\min}}\left\{1 + \frac{\sigma^S_{\min}}{\sigma^S_{\min} - (2s-1)\mu}\right\}|S^{\mathrm{T}}\Sigma Sh|_\infty.$$

Bounding $|S^{\mathrm{T}}\Sigma Sh|_\infty$ uses the proof that β_0 is feasible, because

$$\begin{aligned}|S^{\mathrm{T}}\Sigma Sh|_\infty &= |S^{\mathrm{T}}\Sigma S\hat{\beta} - S^{\mathrm{T}}b|_\infty \\ &\leqslant |S^{\mathrm{T}}\hat{\Sigma}S\hat{\beta} - S^{\mathrm{T}}\hat{b}|_\infty + |S^{\mathrm{T}}(\hat{\Sigma} - \Sigma)S\hat{\beta}_1|_\infty + |S^{\mathrm{T}}(\hat{b} - b)|_\infty \\ &\leqslant \lambda_n + \|S^{\mathrm{T}}\|_\infty|\hat{\Sigma} - \Sigma|_\infty\|S\|_1|\beta_0|_1 + \|S^{\mathrm{T}}\|_\infty(|\hat{\Sigma}_X - \Sigma_X|_\infty + |\hat{\Sigma}_Y - \Sigma_Y|_\infty) \\ &\leqslant \lambda_n + \{4M(|\sigma^X_{l'l}| + |\sigma^Y_{m'm}| + C) + 4\}\{\log p/\min(n_X, n_Y)\}^{1/2} = 2\lambda_n\end{aligned}$$

when $|\hat{\Sigma}_X - \Sigma_X|_\infty$ and $|\hat{\Sigma}_Y - \Sigma_Y|_\infty$ are both less than $C\{\log p/\min(n_X, n_Y)\}^{1/2}$. □

Proof of Theorem 1. Let $\hat{\delta}_{jk}^{\tau_n}$ be the (j,k)th entry of $\hat{\Delta}_{\tau_n}$. Then

$$\mathrm{pr}\{\mathcal{M}(\hat{\Delta}_{\tau_n}) = \mathcal{M}(\Delta_0)\} = \mathrm{pr}\left[\left\{\max_{j,k:\delta_{jk}^0=0} |\hat{\delta}_{jk}^{\tau_n}| = 0\right\} \cap \left\{\min_{j,k:\delta_{jk}^0>0} \hat{\delta}_{jk}^{\tau_n} > 0\right\} \cap \left\{\max_{j,k:\delta_{jk}^0<0} \hat{\delta}_{jk}^{\tau_n} < 0\right\}\right].$$

Suppose $\delta_{jk}^0 > 0$. Then $\hat{\delta}_{jk} = \delta_{jk}^0 - (\hat{\delta}_{jk} - \delta_{jk}^0) > 2\tau_n - \tau_n$ with probability going to 1, by Theorem 3, so $\hat{\delta}_{jk}^{\tau_n} = \hat{\delta}_{jk} > 0$. Next, suppose $\delta_{jk}^0 < 0$. Then $\hat{\delta}_{jk} = \delta_{jk}^0 - (\hat{\delta}_{jk} - \delta_{jk}^0) < -2\tau_n + \tau_n$ with probability going to 1, so $\hat{\delta}_{jk}^{\tau_n} = \hat{\delta}_{jk} < 0$. Finally, for $\delta_{jk}^0 = 0$, $|\hat{\delta}_{jk}| = |\hat{\delta}_{jk} - \delta_{jk}^0| \leqslant \tau_n$ with probability going to 1, so $\hat{\delta}_{jk}^{\tau_n} = 0$. □

Proof of Theorem 2. Solutions to this type of ℓ_1 constrained optimization problem have $|h_{T_0}|_1 \geqslant |h_{T_0^c}|_1$. Cai et al. (2010a) used this property of h, along with their Lemma 3, to show that $|h|_2 \leqslant 2|h_{T_0 \cup T^*}|_2$, where T^* is the set of indices corresponding to the $s/4$ largest components of $h_{T_0^c}$. Then $|h|_2 \leqslant 2(1{\cdot}25s)^{1/2}|h|_\infty$, and combining this with Theorem 3 completes the proof. □

References

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W.-K., Aebersold, R., Keogh, M.-C., Krogan, N. J. & Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Sci. Signal.* **330**, 1385–9.

Barabási, A.-L., Gulbahce, N. & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Rev. Genet.* **12**, 56–68.

Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–13.

Bellail, A. C., Qi, L., Mulligan, P., Chhabra, V. & Hao, C. (2009). TRAIL agonists on clinical trials for cancer therapy: The promises and the challenges. *Rev. Recent Clin. Trials* **4**, 34–41.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundat. Trends Mach. Learn.* **3**, 1–122.

Cai, T. & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.* **106**, 1566–77.

Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

Cai, T., Wang, L. & Xu, G. (2010a). Shifting inequality and recovery of sparse signals. *IEEE Trans. Sig. Proces.* **58**, 1300–8.

Cai, T., Wang, L. & Xu, G. (2010b). Stable recovery of sparse signals and an oracle inequality. *IEEE Trans. Info. Theory* **56**, 3516–22.

Cai, T., Xu, G. & Zhang, J. (2009). On recovery of sparse signals via ℓ_1 minimization. *IEEE Trans. Info. Theory* **55**, 3388–97.

Chiquet, J., Grandvalet, Y. & Ambroise, C. (2011). Inferring multiple graphical structures. *Statist. Comp.* **21**, 537–53.

Danaher, P., Wang, P. & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B **76**, 373–97.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26**, 326–33.

Donoho, D. & Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info. Theory* **47**, 2845–62.

Emmert-Streib, F. & Dehmer, M. (2011). Networks for systems biology: Conceptual connection of data and function. *IET Syst. Biol.* **5**, 185–207.

Friedman, J. H., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

Hauser, A. & Bühlmann, P. (2014). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc*. B. To appear.

Huang, J., Ma, S., Xie, H. & Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339–55.

Hudson, N. J., Reverter, A. & Dalrymple, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comp. Biol.* **5**, e1000382.

IDEKER, T. & KROGAN, N. (2012). Differential network biology. *Molec. Syst. Biol.* **8**, 565.
JOHNSTONE, R. W., FREW, A. J. & SMYTH, M. J. (2008). The TRAIL apoptotic pathway in cancer onset, progression and therapy. *Nature Rev. Cancer* **8**, 782–98.
KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8**, 613–36.
KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M. & TANABE, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–14.
KIPPS, E., TAN, D. S. P. & KAYE, S. B. (2013). Meeting the challenge of ascites in ovarian cancer: New avenues for therapy and research. *Nature Rev. Cancer*, 273–82.
LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.
LI, K.-C., PALOTIE, A., YUAN, S., BRONNIKOV, D., CHEN, D., WEI, X., CHOI, O.-W., SAARELA, J. & PELTONEN, L. (2007). Finding disease candidate genes by liquid association. *Genome Biol.* **8**, R205.
LI, X., ZHAO, T., WANG, L., YUAN, X. & LIU, H. (2013). *flare: Family of Lasso Regression*. R package version 1.0.0.
LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and Dantzig estimators. *Electron. J. Statist.* **2**, 90–102.
MAATHUIS, M. H., KALISCH, M. & BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37**, 3133–64.
MARCHINI, S., FRUSCIO, R., CLIVIO, L., BELTRAME, L., PORCU, L., NERINI, I. F., CAVALIERI, D., CHIORINO, G., CATTORETTI, G., MANGIONI, C., MILANI, R., TORRI, V., ROMUALDI, C., ZAMBELLI, A., ROMANO, M., SIGNORELLI, M., DI GIANDOMENICO, S. & DINCALCI, M. (2013). Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer. *Eur. J. Cancer* **49**, 520–30.
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.
MOHAN, K., CHUNG, M., HAN, S., WITTEN, D., LEE, S.-I. & FAZEL, M. (2012). Structured learning of Gaussian graphical models. In *Adv. Neural Info. Proces. Syst.* 25. New York: Curran Associates, pp. 629–37.
NEWMAN, M. E. (2003). The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.
OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34.
PANG, H., LIU, H. & VANDERBEI, R. (2013). *fastclime: A Fast Solver for Parameterized ℓ_p Problems and Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation*. R package version 1.2.3.
PETRUCCI, E., PASQUINI, L., BERNABEI, M., SAULLE, E., BIFFONI, M., ACCARPIO, F., SIBIO, S., DI GIORGIO, A., DI DONATO, V., CASORELLI, A., BENEDETTI-PANICI, P. & TESTA, U. (2012). A small molecule SMAC mimic LBW242 potentiates TRAIL- and anticancer drug-mediated cell death of ovarian cancer cells. *PloS One* **7**, e35073.
R DEVELOPMENT CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
RAVIKUMAR, P. K., RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Adv. Neural Info. Proces. Syst.* 21. New York: Curran Associates, pp. 1329–36.
TIBSHIRANI, R. J., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B **67**, 91–108.
TOTHILL, R. W., TINKER, A. V., GEORGE, J., BROWN, R., FOX, S. B., LADE, S., JOHNSON, D. S., TRIVETT, M. K., ETEMADMOGHADAM, D., LOCANDRO, B., TRAFICANTE, N., FEREDAY, S., HUNG, J. A., CHIEW, Y.-E., HAVIV, I., AUSTRALIAN OVARIAN CANCER STUDY GROUP, GERTIG, D., DEFAZIO, A. & BOWTELL, D. D. L. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198–208.
VUCIC, D. & FAIRBROTHER, W. J. (2007). The inhibitor of apoptosis proteins as therapeutic targets in cancer. *Clin. Cancer Res.* **13**, 5995–6000.
WANG, S., NAN, B., ZHU, N. & ZHU, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96**, 307–22.
YAGITA, H., TAKEDA, K., HAYAKAWA, Y., SMYTH, M. J. & OKUMURA, K. (2004). TRAIL and its receptors as targets for cancer therapy. *Cancer Sci.* **95**, 777–83.
YUAN, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11**, 2261–86.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

[*Received April* 2013. *Revised February* 2014]