

VU Research Portal

Testing for pathway (in)activation by using Gaussian graphical models

van Wieringen, Wessel N.; Peeters, Carel F.W.; de Menezes, Renee X.; van de Wiel, Mark A.

published in

Journal of the Royal Statistical Society. Series C: Applied Statistics
2018

DOI (link to publisher)

[10.1111/rssc.12282](https://doi.org/10.1111/rssc.12282)

document version

Publisher's PDF, also known as Version of record

document license

Article 25fa Dutch Copyright Act

[Link to publication in VU Research Portal](#)

citation for published version (APA)

van Wieringen, W. N., Peeters, C. F. W., de Menezes, R. X., & van de Wiel, M. A. (2018). Testing for pathway (in)activation by using Gaussian graphical models. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 67(5), 1419-1436. <https://doi.org/10.1111/rssc.12282>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Appl. Statist. (2018)
67, Part 5, pp. 1419–1436

Testing for pathway (in)activation by using Gaussian graphical models

Wessel N. van Wieringen, Carel F. W. Peeters, Renee X. de Menezes
and Mark A. van de Wiel

Vrije Universiteit Amsterdam, The Netherlands

[Received May 2017. Final revision March 2018]

Summary. Genes work together in sets known as pathways to contribute to cellular processes, such as apoptosis and cell proliferation. Pathway activation, or inactivation, may be reflected in varying partial correlations between the levels of expression of the genes that constitute the pathway. Here we present a method to identify pathway activation status from two-sample studies. By modelling the levels of expression in each group by using a Gaussian graphical model, their partial correlations are proportional, differing by a common multiplier that reflects the activation status. We estimate model parameters by means of penalized maximum likelihood and evaluate the estimation procedure performance in a simulation study. A permutation scheme to test for pathway activation status is proposed. A reanalysis of publicly available data on the hedgehog pathway in normal and cancer prostate tissue shows its activation in the disease group: an indication that this pathway is involved in oncogenesis. Extensive diagnostics employed in the reanalysis complete the methodology proposed.

Keywords: Conditional independence; Multivariate normality; Network; Partial correlations; Penalized estimation

1. Introduction

Gene sets, known as pathways, are involved in many cellular processes, such as apoptosis and cell proliferation. When a pathway is observed to be activated in one group of samples, but not in another, it indicates that pathway activation status may help to explain differences between these two groups. Many pathways, such as deoxyribonucleic acid damage repair and TP53 signalling, are known to be involved in oncogenesis (Weinberg, 2006). Changes in activation status of such pathways between tumour and normal samples are thus indicative of cancer. Although individual study of genes in such pathways is possible, it is more efficient to look at changes in pathways as a whole, as these are representative of the entire process.

Transcription levels of genes in a pathway are often described by a Gaussian graphical model (GGM) (see for example Dobra *et al.* (2004)). Such a model represents interactions between genes in a pathway as conditional dependences. All genes and their interactions amount to a conditional independence graph, in which nodes represent the genes and the presence of edges corresponds to conditional dependences (Whittaker, 1990). Under the assumption of multivariate normality, the conditional dependences between the genes in the pathway can be inferred from the scaled inverse covariance matrix (Whittaker, 1990). This scaled inverse covariance matrix contains the minus partial correlations, i.e. the correlations between a pair of

Address for correspondence: Wessel N. van Wieringen, Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands.

E-mail: w.vanwieringen@vumc.nl

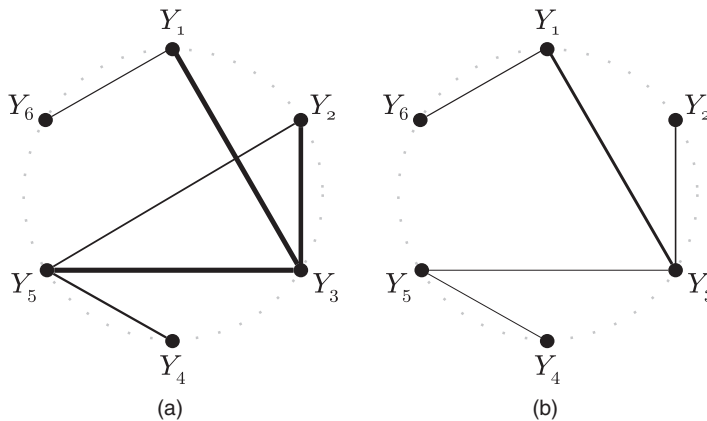


Fig. 1. A pathway in the two conditions (a) 'active' and (b) 'inactive': in both conditions the pathway is represented by its conditional independence graph, in which an edge represents a conditional dependence between a gene pair; the edge width indicates the strength of the conditional dependence, as reflected in the partial correlations absolute value; the global difference, strong (active) *versus* weak (active), between these strengths distinguishes the two conditions

genes conditionally on all other genes in the pathway. The size of the partial correlations reflects the strength of the conditional dependences: the smaller the absolute partial correlation, the weaker the conditional dependence between the associated gene pair. In case a partial correlation is negligible, the two genes can be considered conditionally independent.

The pathway activation status can be captured within a GGM, as the strength of the conditional dependences reflects the partial correlations. Indeed, within an inactive pathway genes typically display weaker conditional dependences compared with when the pathway is active. The conditional dependence structure can be visualized as a graph, where genes are represented by nodes and conditional dependences of a gene pair are symbolized by the presence of an edge between the nodes of this pair. Similarly, conditional independence is represented by the absence of an edge between the genes. Fig. 1 portrays the graph of an active and an inactive pathway, with edge width representing the strength of the conditional dependences. The narrower width of the edges in the inactive pathway indicates weakened conditional dependence compared with its active counterpart.

Here we present a test to compare pathway activation status between two groups of samples. In Section 2 we propose the use of a joint GGM for both samples, in which differential conditional dependence is represented by a single parameter. In Section 3, we describe how the model parameters can be estimated to accommodate various sample-to-pathway sizes. In Section 4, it is outlined how differential conditional dependence is tested within the model proposed. The performance of estimating the differential conditional dependence parameter is investigated by using a simulation study in Section 5. Finally, in Section 6 we use the method to reanalyse *in vivo* prostate gene expression data, corroborating the previously *in-vitro*-found activation of the hedgehog signalling pathway in cancer. The reanalysis introduces several diagnostic techniques that complement the method.

1.1. Related work

Reconstruction of differential GGMs is an active field of research. Published methods can roughly be categorized on the basis of the cross-tabulation of two criteria:

- (a) are network differences identified by estimation or testing?;
- (b) are the to-be-detected differences local (e.g. an edge or subnetwork) or global (e.g. some aggregated property of the whole network)?

Methods that focus on identification of local differences joint precision estimation in penalized fashion can be found in Guo *et al.* (2011), Danaher *et al.* (2014), Zhao *et al.* (2014), Price *et al.* (2015), Bilgrau *et al.* (2015) and Saegusa and Shojaie (2016). Ha *et al.* (2015) and Xia (2017) augmented such approaches by a *post hoc* test procedure for differential edge identification. Alternatively, Peterson *et al.* (2015) and Mitra *et al.* (2016) described Bayesian approaches to estimate groupwise precision matrices jointly and to derive the posterior probability of an edge differing between networks. Testing methods for local differences (under a sparsity assumption) are given in Xia *et al.* (2015) and Städler and Mukherjee (2015). There also exist ‘global plus estimation’ methods that are more heuristic in the sense that these construct the network for each group separately followed by calculation and comparison of various network metrics (Fuller *et al.*, 2007). The method proposed tests for a global difference between two (saturated) networks. Related is the omnibus-type approach of Schott (2007) which tests for a difference in the Frobenius norm of both sample covariance matrices and is a natural competitor to the work that is presented here (see Section 5). For more related methods and an overview of the field refer to the review of Cai (2017).

2. Model

Consider an oncogenomics study with an observational experimental design involving n samples. The transcriptome of each sample in the study has been profiled. The interest lies in the expression levels of a set of p genes that constitute a pathway. Let $Y_{i,j}$ be the expression level of gene j in sample i . The expression profile of sample i of all p genes that belong to the pathway is denoted by $\mathbf{Y}_{i,*}$ and the collection of n expression profiles by $\mathbf{Y}_{*,*}$.

The n samples originate from two groups, which we take to represent two different stages of cancer. One group is thought of as being further advanced in the disease. With the progression of cancer, the pathway under consideration is assumed to be dysregulated, so that the pathway activation status will differ between the groups.

The gene expression data $\mathbf{Y}_{*,*}$ are modelled as follows. Let $\mathbf{Y}_{i,*}^T \sim \mathcal{N}(\mathbf{0}, \Sigma_1)$ for $i = 1, \dots, n_1$ and $\mathbf{Y}_{i,*}^T \sim \mathcal{N}(\mathbf{0}, \Sigma_2)$ for $i = n_1 + 1, \dots, n = n_1 + n_2$. The covariance matrix Σ_1 is parameterized as $\Sigma_1 = \Upsilon_1 \Omega^{-1} \Upsilon_1$ with Υ_1 a diagonal matrix such that Ω is the standardized precision matrix (related to the partial correlation matrix up to the sign of the off-diagonal elements). The covariance matrix Σ_2 is parameterized as $\Sigma_2 = \Upsilon_2 \{(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}\}^{-1} \Upsilon_2$ with Υ_2 a diagonal matrix and $\gamma \in [0, 1]$. If $\gamma = 0$, there is no difference between the two groups. If $\gamma = 1$, all conditional dependences in the second group have vanished. Other values of γ , i.e. $\gamma \in (0, 1)$, represent intermediate degrees of differential conditional dependence. Hence, any $\gamma \in (0, 1]$ corresponds to pathway inactivation in the more advanced stages of cancer. Note that γ represents an aggregation of differential conditional dependence, saying little about the conditional dependence of a particular gene pair. The difference in distribution between the groups when $\gamma > 0$ is illustrated in the on-line supplementary material (SM) section I.

The joint model describing both groups is identifiable. This is obvious for the parameters of the first group’s GGM. However, the GGM of the second group is overparameterized. However, by coupling the standardized precision matrices between both groups, all its parameters become identifiable. As Ω is estimable from the first group and Υ_2 from the marginal variances of \mathbf{S}_2 , there remain $\frac{1}{2}p(p - 1)$ degrees of freedom for the estimation of the scalar γ .

3. Estimation

In this section we describe how the parameters of the two-sample GGM are estimated in a classical context in Section 3.1, and in a penalized context in Section 3.2. Finally, in Section 3.3 we outline the choice of the penalty parameter.

3.1. Basics

The model parameters are estimated by means of maximum likelihood (ML). Below we present the derivation of the ML parameter estimates. The likelihood of the data is

$$L(\mathbf{Y}_{*,*}; \Upsilon_1, \Upsilon_2, \Omega, \gamma) = \prod_{i=1}^{n_1} P(\mathbf{Y}_{i,*}; \Upsilon_1, \Omega) \prod_{i=n_1+1}^{n_1+n_2} P(\mathbf{Y}_{i,*}; \Upsilon_2, \Omega, \gamma).$$

The corresponding log-likelihood (with dependence on the data now running through the group sample covariance matrices) equals

$$\mathcal{L}(\mathbf{S}_1, \mathbf{S}_2; \Upsilon_1, \Upsilon_2, \Omega, \gamma) \propto n_1 \log|\Omega| - 2n_1 \log|\Upsilon_1| - n_1 \text{tr}(\tilde{\mathbf{S}}_1 \Omega) - 2n_2 \log|\Upsilon_2| + n_2 \log|(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}| - n_2 \text{tr}[\tilde{\mathbf{S}}_2 \{(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}\}], \tag{1}$$

where $\tilde{\mathbf{S}}_1 = \Upsilon_1^{-1} \mathbf{S}_1 \Upsilon_1^{-1}$ and $\tilde{\mathbf{S}}_2 = \Upsilon_2^{-1} \mathbf{S}_2 \Upsilon_2^{-1}$, and \mathbf{S}_1 and \mathbf{S}_2 denote the sample covariance matrices of groups 1 and 2. To obtain the ML estimate of Ω , equate the derivative of the log-likelihood with respect to Ω to zero. This yields the following estimating equation:

$$\tilde{\mathbf{S}} = p_1 \Omega^{-1} + p_2 (1 - \gamma) \{(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}\}^{-1}, \tag{2}$$

where $p_1 = n_1 / (n_1 + n_2)$, $p_2 = n_2 / (n_1 + n_2)$ and $\tilde{\mathbf{S}} = p_1 \tilde{\mathbf{S}}_1 + p_2 (1 - \gamma) \tilde{\mathbf{S}}_2$. From this estimation equation we derive two new equations: one through premultiplication by $\Omega^{-1} \{(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}\}$ and the other via post-multiplication by $\{(1 - \gamma)\Omega + \gamma \mathbf{I}_{pp}\} \Omega^{-1}$. Adding the resulting two equations gives

$$2(1 - \gamma) \tilde{\mathbf{S}} + \gamma \Omega^{-1} \tilde{\mathbf{S}}(\gamma) + \gamma \tilde{\mathbf{S}} \Omega^{-1} = 2p_1 \gamma \Omega^{-2} + 2(1 - \gamma) \Omega^{-1}.$$

Using standard matrix algebra we complete the square and reformulate the equation as

$$\{2\gamma p_1 \Omega^{-1} + \Theta(\gamma, \tilde{\mathbf{S}})\}^2 = \Theta(\gamma, \tilde{\mathbf{S}})^2 + 4p_1 \gamma (1 - \gamma) \tilde{\mathbf{S}},$$

where $\Theta(\gamma, \tilde{\mathbf{S}}) = (1 - \gamma) \mathbf{I}_{pp} - \gamma \tilde{\mathbf{S}}$. From this Ω can be solved explicitly, resulting in

$$\hat{\Omega} = 2\gamma p_1 [\{\Theta(\gamma, \tilde{\mathbf{S}})^2 + 4p_1 \gamma (1 - \gamma) \tilde{\mathbf{S}}\}^{1/2} - \Theta(\gamma, \tilde{\mathbf{S}})]^{-1}. \tag{3}$$

It can be verified that for $\gamma = 0$ and $\gamma = 1$ the estimator $\hat{\Omega}$ equals $(p_1 \tilde{\mathbf{S}}_1 + p_2 \tilde{\mathbf{S}}_2)^{-1}$ and $\tilde{\mathbf{S}}_1^{-1}$ respectively. Moreover, $\hat{\Omega}$ is positive definite. This follows from the facts that

- (a) $2\gamma p_1 > 0$,
- (b) the elements of $\Theta(\gamma, \tilde{\mathbf{S}})$ are real,
- (c) $\Theta(\gamma, \tilde{\mathbf{S}})^2$ is semipositive definite,
- (d) $4p_1 \gamma (1 - \gamma) \tilde{\mathbf{S}}$ is positive definite and
- (e) the sum of a positive definite and a semipositive definite matrix is positive definite (see lemma 14.2.4 of Harville (2008)).

Note, however, that the estimator $\hat{\Omega}$ need not have a unit diagonal. To achieve this, $\hat{\Omega}$ is standardized to have a unit diagonal, which warrants the interpretation of $\hat{\Omega}$ as a standardized precision matrix.

The estimating equation of the ML estimate of γ is

$$\text{tr}\{(1 - \gamma)\mathbf{\Omega} + \gamma\mathbf{I}_{pp}\}^{-1}(\mathbf{\Omega} - \mathbf{I}_{pp}) = \text{tr}\{\tilde{\mathbf{S}}_2(\mathbf{\Omega} - \mathbf{I}_{pp})\}. \tag{4}$$

No explicit solution for γ appears to exist, so we resort to standard numerical means to find the root of this equation.

For the ML estimation of Υ_1 consider the eigendecomposition $\mathbf{\Omega} = \mathbf{V}_\omega \mathbf{D}_\omega \mathbf{V}_\omega^T$. The log-likelihood for Υ_1 is then

$$\mathcal{L}(\Upsilon_1) \propto \log|\Upsilon_1^{-1} \mathbf{V}_\omega \mathbf{D}_\omega \mathbf{V}_\omega^T \Upsilon_1^{-1}| - \text{tr}(\mathbf{S}_1 \Upsilon_1^{-1} \mathbf{V}_\omega \mathbf{D}_\omega \mathbf{V}_\omega^T \Upsilon_1^{-1}).$$

Now apply the change of variables $\Upsilon_1 = \mathbf{V}_\omega \mathbf{D}_\omega^{1/2} \mathbf{X} \mathbf{D}_\omega^{1/2} \mathbf{V}_\omega^T$ with \mathbf{X} denoting a symmetric $p \times p$ dimensional matrix. The log-likelihood for Υ_1 becomes (after some manipulation)

$$\mathcal{L}(\Upsilon_1) \propto \log|\mathbf{X}^{-2}| - \text{tr}(\mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \mathbf{S}_1 \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2} \mathbf{X}^{-2}).$$

This has a well-known maximizer: $\mathbf{X}^2 = \mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \mathbf{S}_1 \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2}$. Take the (matrix) square root on both sides and obtain

$$\text{diag}(\hat{\Upsilon}_1) = \text{diag}\{\mathbf{V}_\omega \mathbf{D}_\omega^{1/2} (\mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \mathbf{S}_1 \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2})^{1/2} \mathbf{D}_\omega^{1/2} \mathbf{V}_\omega^T\}.$$

The estimator of Υ_2 is derived similarly, but the role of $\mathbf{\Omega}$ is now played by $(1 - \gamma)\mathbf{\Omega} + \gamma\mathbf{I}_{pp}$.

The estimators of the individual parameters γ , $\mathbf{\Omega}$, Υ_1 and Υ_2 are combined into a single estimation procedure. This gives an iterative procedure in which at each iteration the individual estimators are updated, keeping the others constant. The estimation is terminated if the absolute difference between the (penalized) log-likelihood of two successive iterations is smaller than a user-specified threshold. A full pseudocoded description of the algorithm producing the ML estimates can be found in the on-line SM section II.

3.2. Penalized estimation

In the high dimensional setting, the estimator $\hat{\mathbf{\Omega}}$ is not positive definite. Here we describe how to arrive at a positive definite estimate of $\mathbf{\Omega}$ through regularization. This estimate of $\mathbf{\Omega}$ optimizes a penalized log-likelihood, which is formed from the original log-likelihood (1) augmented with a penalty:

$$\mathcal{L}^{\text{pen}}(\mathbf{S}_1, \mathbf{S}_2; \Upsilon_1, \Upsilon_2, \mathbf{\Omega}, \gamma, \lambda) = \mathcal{L}(\mathbf{S}_1, \mathbf{S}_2; \Upsilon_1, \Upsilon_2, \mathbf{\Omega}, \gamma) + \lambda \mathcal{L}^{\text{simple}}(\mathbf{I}_{pp}; \mathbf{\Omega}), \tag{5}$$

where

$$\mathcal{L}^{\text{simple}}(\mathbf{I}_{pp}; \mathbf{\Omega}) = (n_1 + n_2)\{\log|\mathbf{\Omega}| - \text{tr}(\mathbf{\Omega})\}.$$

The penalty (of the same form as the penalty that was implicitly used by Ledoit and Wolf (2004)) is a simplified form of the log-likelihood (1) with $\gamma = 0$ and $\Upsilon_1 = \Upsilon_2 = \mathbf{S}_1 = \mathbf{S}_2 = \mathbf{I}_{pp}$. Because the penalty is also a log-likelihood, its addition to log-likelihood (1) can be viewed as data augmentation: $(n_1 + n_2)\lambda$ samples drawn from the uninformative standard multivariate normal $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_{pp})$ are added to the original experimental data. From a Bayesian perspective this may be viewed as a historical prior, although based on artificial data. More importantly, inclusion of the penalty ensures that the penalized estimate of $\mathbf{\Omega}$ is

- (a) unique, which follows from the concavity of the two summands that constitute the penalized log-likelihood, and
- (b) shrinks to the identity matrix as $\lambda \rightarrow \infty$.

Instead of the penalty employed we may consider an l_1 -penalty. The resulting estimate would be sparse, i.e. have many off-diagonal elements equaling 0. The sparsity of such an estimate

appeals to the parsimony principle that prefers simple models. However, a simple model should not oversimplify. The l_1 -methods generally assume a severe degree of sparsity to warrant theoretical results. Depending on the context the sparsity required may be too stringent. Here the focus is on molecular biology, where the validity of a stringent sparsity assumption is doubtful (see Boyle *et al.* (2017)). The method proposed, making no assumption on sparsity, seems then to be more appropriate for the application in mind. Moreover, a sparse estimate may hamper the estimation of the differential conditional dependence parameter. It can then only be learned from the few non-zero elements of the sparse standardized precision matrix estimate, as it is unidentifiable from the 0s.

Maximization of the penalized log-likelihood follows the iterative procedure that was discussed previously in Section 3.1. In that procedure only the estimation of Ω needs modification, as the penalty affects the estimating equation of Ω only. As before, an explicit solution for Ω exists. To see this, consider the estimation equation of the penalized estimate of Ω :

$$\tilde{\mathbf{S}} = p_1 \Omega^{-1} + p_2 (1 - \gamma) \{ (1 - \gamma) \Omega + \gamma \mathbf{I}_{pp} \}^{-1} + \lambda \Omega^{-1} - \lambda \mathbf{I}_{pp}.$$

Algebraic manipulations (e.g. the premultiplication and post-multiplication) that are completely analogous to the unpenalized case yield the penalized estimate

$$\hat{\Omega}(\lambda) = 2(p_1 + \lambda) \gamma \{ \tilde{\Theta}(\gamma, \tilde{\mathbf{S}}, \lambda)^2 + 4(p_1 + \lambda) \gamma (1 - \gamma) (\tilde{\mathbf{S}} + \lambda \mathbf{I}_{pp}) \}^{1/2} - \tilde{\Theta}(\gamma, \tilde{\mathbf{S}}, \lambda)^{-1}, \quad (6)$$

where $\tilde{\Theta}(\gamma, \tilde{\mathbf{S}}, \lambda) = (1 + \lambda)(1 - \gamma) \mathbf{I}_{pp} - \gamma(\tilde{\mathbf{S}} + \lambda \mathbf{I}_{pp})$. It is immediate that $\hat{\Omega}(\lambda)$ reduces to the unpenalized estimate when $\lambda = 0$. Moreover, using similar arguments to those for its unpenalized counterpart, $\hat{\Omega}(\lambda)$ is positive definite. It remains to be verified that $\Omega(\lambda)$ converges to \mathbf{I}_{pp} as $\lambda \rightarrow \infty$. Hereto consider the eigendecomposition of $\tilde{\mathbf{S}} = \mathbf{V}_s \mathbf{D}_s \mathbf{V}_s^T$, where the diagonal of \mathbf{D}_s contains the eigenvalues of $\tilde{\mathbf{S}}$ and the corresponding eigenvectors (as columns) form \mathbf{V}_s . Then, as $\mathbf{I}_{pp} = \mathbf{V}_s \mathbf{V}_s^T$ and $\tilde{\Theta}(\gamma, \tilde{\mathbf{S}}, \lambda) = \mathbf{V}_s \{ (1 + \lambda)(1 - \gamma) \mathbf{I}_{pp} - \gamma(\mathbf{D}_s + \lambda \mathbf{I}_{pp}) \} \mathbf{V}_s^T$, the eigendecomposition of $\hat{\Omega}(\lambda)$ is $\mathbf{V}_s \mathbf{D}_\omega \mathbf{V}_s^T$ where the diagonal matrix \mathbf{D}_ω contains the eigenvalues of $\hat{\Omega}(\lambda)$. Algebraic manipulations reveal that the diagonal elements of \mathbf{D}_ω can be expressed as

$$\frac{2\gamma(p_1 + \lambda)}{\{(\lambda + c_1)^2 + c_2\}^{1/2} - (\lambda + c_1) + 2\gamma\lambda + c_3},$$

for suitable constants c_1, c_2 and c_3 that do not depend on λ . As $\{(\lambda + c_1)^2 + c_2\}^{1/2} - (\lambda + c_1)$ converges to 0 as $\lambda \rightarrow \infty$, both numerator and denominator are dominated by the term $2\gamma\lambda$ for sufficiently large λ . Hence, $\lim_{\lambda \rightarrow \infty} \hat{\Omega}(\lambda) = \lim_{\lambda \rightarrow \infty} \mathbf{V}_s \mathbf{D}_\omega \mathbf{V}_s^T = \mathbf{V}_s \mathbf{V}_s^T = \mathbf{I}_{pp}$.

Note that, although the regularization of the estimation of Ω leaves the estimating equation of γ unchanged, it affects the estimate of γ . This is because an increase in either γ or λ reduces the size of the off-diagonal elements of the standardized precision matrix (*pari passu* the partial correlation matrix) of the second group. Consequently, the larger λ is, the closer the penalized estimate of Ω to \mathbf{I}_{pp} , and the smaller the change in the partial correlations that can be assigned to γ . In particular, it is clear from equation (4) that γ cannot be estimated in the limit $\lambda \rightarrow \infty$. In conclusion, for large λ the γ -parameter is likely to be underestimated and thus the differential correlation between the two conditions.

At the other end of the penalty scale, as λ approaches 0, the opposite may happen: the estimate of γ could be biased upwards. To see this, consider the setting in which $\tilde{\mathbf{S}}$ is (close to being) of deficient rank, i.e. $\text{rank}(\tilde{\mathbf{S}}) < p$. When $\lambda \downarrow 0$ the largest eigenvalue of $\hat{\Omega}$ tends to ∞ (because of the rank deficiency of $\tilde{\mathbf{S}}$). Consequently, the right-hand side of γ 's estimating equation (4) is dominated by the contribution of this eigenvalue to the trace. In contrast, on the left-hand side of equation (4) its contribution is $\{(\mathbf{D}_\omega)_{1,1} - 1\} / \{(1 - \gamma)(\mathbf{D}_\omega)_{1,1} + \gamma\}$, where $(\mathbf{D}_\omega)_{1,1}$ corresponds

to the first and thus largest eigenvalue of $\hat{\Omega}$. For small γ and large $(\mathbf{D}_\omega)_{1,1}$ the latter contribution is (relatively) small and does not balance the right-hand side of equation (4). Then, to satisfy its estimating equation, γ must compensate for the ill-conditioned $\hat{\Omega}$, irrespectively of the ‘true’ value of γ . In particular, if the true γ is small, its estimate will shy away from 0. Hence, in undersampled situations too little penalization may bias the estimate of γ upwards.

Penalized estimation procedures have a Bayesian counterpart. To identify the latter for the case at hand assume a Wishart prior $\pi(\Omega), \mathcal{W}(\delta \mathbf{I}_{pp}, \delta^{-1} + p + 1)$ with $\delta^{-1} = 2\lambda(n_1 + n_2)$, on Ω . The Bayesian mode *a posteriori* estimator will then coincide with the proposed penalized estimator. This is immediate after noting that

- (a) the posterior is proportional to the likelihood times the prior,
- (b) the location of the maximum of the posterior is that of the log-posterior and
- (c) substitution of the prior’s parameter in the Wishart density reveals that $\pi(\Omega) \propto |\Omega|^{\lambda(n_1+n_2)} \times \exp\{-\lambda(n_1 + n_2)\text{tr}(\Omega)\}$ and the logarithm of the prior is thus proportional to the employed penalty.

Put together, the log-posterior and the penalized log-likelihood are proportional and share the same maximizer. In particular, the mode *a posteriori* and penalized estimator share the same λ -limiting behaviour: as λ increases $\pi(\Omega)$ becomes more concentrated around zero with the posterior in its trail.

3.3. The penalty parameter

The optimal penalty parameter is chosen by means of conventional leave-one-out cross-validation (LOOCV). LOOCV leaves out the samples one by one, refits the model for the remaining samples and evaluates the performance (here operationalized as the unpenalized log-likelihood) of the refitted model on the left-out sample. The penalty parameter that yields on average (over the left-out samples) the best performance is considered optimal.

The particular form of the penalty provides the unitless penalty parameter with a tangible interpretation. Here note that the penalized estimate of Ω also maximizes

$$\mathcal{L}^{\text{pen}}(\mathbf{S}_1, \mathbf{S}_2; \mathbf{Y}_1, \mathbf{Y}_2, \Omega, \gamma, \lambda) \propto (1 - \nu)\mathcal{L}(\mathbf{S}_1, \mathbf{S}_2; \mathbf{Y}_1, \mathbf{Y}_2, \Omega, \gamma) + \nu\mathcal{L}^{\text{simple}}(\mathbf{I}_{pp}; \Omega),$$

where $\nu = \lambda/(1 + \lambda)$. The penalized log-likelihood of each group can thus be interpreted as the complete log-likelihood of a two-component mixture distribution with the same, known class probability ν for each sample. The mixture leading to this complete log-likelihood assumes that the gene expression profile of sample i from (say) the first group is distributed as

$$\mathbf{Y}_{i,*}^T \sim \nu\phi_{\mu=0_p; \Sigma=\mathbf{I}_{pp}}(\mathbf{Y}_{i,*}^T) + (1 - \nu)\phi_{\mu=0_p; \Sigma=\Sigma_1}(\mathbf{Y}_{i,*}^T),$$

where $\phi_{\mu, \Sigma}(\cdot)$ is the density of the multivariate normal distribution. The penalty parameter (through ν) thus reflects the extent to which the data warrant a more sophisticated description than the uninformative white noise distribution. In the Bayesian interpretation it thus amounts to the value or weight of the historical prior.

4. Testing for pathway activation status

Here we describe how to evaluate whether there is a difference in the ‘average’ edge strength between the groups, represented by the parameter γ : in other words, how to test the null hypothesis $H_0: \gamma = 0$ against its alternative $H_a: \gamma > 0$. A natural test statistic for this is the estimate of γ itself. The null distribution of this test statistic is generated through permutation of the group labels. Denote for permutations $k = 1, \dots, K$ the test statistic, the estimated γ , by T_k . The set of

test statistics that is obtained through permutation forms the null distribution associated with $H_0: \gamma = 0$ from which the p -value can be estimated by $(1/K) \sum_{k=1}^K I\{T_k \geq T_{\text{obs}}\}$. When interest lies only in the p -value exceeding the significance level, the efficient permutation scheme that was proposed by Van Wieringen *et al.* (2008) may be used. Its efficiency gain is due to preliminary termination of the permutation scheme when the probability that the p -value is small is quite low (under the assumption of a binomial process). To decide on this, the 99.9% confidence interval of the current p -value estimate is calculated after a limited number of permutations. When the lower confidence bound does not exceed the significance level, the true p -value is unlikely to be smaller and the permutation scheme may be terminated.

The test statistic employed is estimated in a penalized fashion. Consequently, inference is conditional on the penalty parameter λ . As the choice of λ may affect the inference, we ensure its representativeness under H_0 through two modifications of the scheme that was outlined in Section 3.3:

- (a) fix $\gamma = 0$ (as prescribed by H_0) in the estimation procedure and
- (b) permute the group labels before the selection of the penalty.

The latter warrants that λ is not only chosen under the null model, but also under the ‘null data’. Finally, to eliminate artefacts that are introduced by permutation of the group labels before penalty selection, the selection procedure may be repeated.

The permutation test that was described above is valid only if observations from both groups are exchangeable, i.e. the distribution of both groups is identical under the null hypothesis. This requirement is met when $\Upsilon_1 = \Upsilon_2$. The estimation procedure of Section 3 needs modification to incorporate this requirement, which is described next. Assume that $\Upsilon_1 = \Upsilon = \Upsilon_2$ and (as before) apply the change of variable $\Upsilon = \mathbf{V}_\omega \mathbf{D}_\omega^{1/2} \mathbf{X} \mathbf{D}_\omega^{1/2} \mathbf{V}_\omega^T$. Limiting the log-likelihood to the terms involving Υ gives

$$\begin{aligned} \mathcal{L}(\Upsilon) &\propto p_1 \log |\Upsilon^{-1} \Omega \Upsilon^{-1}| - p_1 \text{tr}(\mathbf{S}_1 \Upsilon^{-1} \Omega \Upsilon^{-1}) \\ &= \log |\mathbf{X}^{-2}| - \text{tr}[\mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \{p_1 \mathbf{S}_1 + p_2(1 - \gamma) \mathbf{S}_2\} \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2} \mathbf{X}^{-2}] \\ &\quad - \text{tr}(p_2 \gamma \mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \mathbf{S}_2 \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2} \mathbf{X}^{-1} \mathbf{D}_\omega^{-1} \mathbf{X}^{-1}). \end{aligned}$$

We now replace \mathbf{X}^{-1} by \mathbf{W} and equate the derivative of the log-likelihood above with respect to \mathbf{W} to zero, yielding

$$\mathbf{0}_{pp} = 2\mathbf{W}^{-1} - \mathbf{A}\mathbf{W} - \mathbf{W}\mathbf{A} - \mathbf{B}\mathbf{W}\mathbf{C} - \mathbf{C}\mathbf{W}\mathbf{B},$$

where $\mathbf{A} = \mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \{p_1 \mathbf{S}_1 + p_2(1 - \gamma) \mathbf{S}_2\} \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2}$, $\mathbf{B} = p_2 \gamma \mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^T \mathbf{S}_2 \mathbf{V}_\omega \mathbf{D}_\omega^{-1/2}$ and $\mathbf{C} = \mathbf{D}_\omega^{-1}$. This estimating equation may be solved by means of Newton’s method (Higham, 2008). Hereto write $\mathbf{W} = \mathbf{Z} + \mathbf{E}$, where \mathbf{Z} is an approximation of the solution and \mathbf{E} the error to approximation \mathbf{Z} . Substitution in the estimating equation gives

$$\mathbf{0}_{pp} = 2(\mathbf{Z} + \mathbf{E})^{-1} - \mathbf{A}(\mathbf{Z} + \mathbf{E}) - (\mathbf{Z} + \mathbf{E})\mathbf{A} - \mathbf{B}(\mathbf{Z} + \mathbf{E})\mathbf{C} - \mathbf{C}(\mathbf{Z} + \mathbf{E})\mathbf{B}.$$

Under the assumption that \mathbf{E} is small compared with \mathbf{Z} we may approximate $(\mathbf{Z} + \mathbf{E})^{-1}$ by $\mathbf{Z}^{-1} - \mathbf{Z}^{-1} \mathbf{E} \mathbf{Z}^{-1}$ to arrive at

$$2\mathbf{Z}^{-1} \mathbf{E} \mathbf{Z}^{-1} + \mathbf{A}\mathbf{E} + \mathbf{E}\mathbf{A} + \mathbf{B}\mathbf{E}\mathbf{C} + \mathbf{C}\mathbf{E}\mathbf{B} = 2\mathbf{Z}^{-1} - \mathbf{A}\mathbf{Z} - \mathbf{Z}\mathbf{A} - \mathbf{B}\mathbf{Z}\mathbf{C} - \mathbf{C}\mathbf{Z}\mathbf{B}.$$

Let \mathbf{H} be an orthonormal matrix that simultaneously diagonalizes the matrices \mathbf{Z} , \mathbf{A} , \mathbf{B} and \mathbf{C} . We can then write $\mathbf{Z} = \mathbf{H}\mathbf{D}_z\mathbf{H}^T$, $\mathbf{A} = \mathbf{H}\mathbf{D}_a\mathbf{H}^T$, $\mathbf{B} = \mathbf{H}\mathbf{D}_b\mathbf{H}^T$ and $\mathbf{C} = \mathbf{H}\mathbf{D}_c\mathbf{H}^T$ with \mathbf{D}_z , \mathbf{D}_a , \mathbf{D}_b and \mathbf{D}_c close to being diagonal in some sense (Flury and Gautschi, 1986; Cardoso and Souloumiac, 1996). Apply the change of variables $\mathbf{E} = \mathbf{H}\tilde{\mathbf{E}}\mathbf{H}^T$, replace \mathbf{Z} , \mathbf{A} , \mathbf{B} and \mathbf{C} by their simultaneous

diagonal decompositions into the last equation and premultiply and post-multiply by \mathbf{H} and \mathbf{H}^T to obtain

$$2\mathbf{D}_z^{-1}\tilde{\mathbf{E}}\mathbf{D}_z^{-1} + \mathbf{D}_a\tilde{\mathbf{E}} + \tilde{\mathbf{E}}\mathbf{D}_a + \mathbf{D}_b\tilde{\mathbf{E}}\mathbf{D}_c + \mathbf{D}_c\tilde{\mathbf{E}}\mathbf{D}_b = 2\mathbf{D}_z^{-1} - \mathbf{D}_a\mathbf{D}_z - \mathbf{D}_z\mathbf{D}_a - \mathbf{D}_b\mathbf{D}_z\mathbf{D}_c - \mathbf{D}_c\mathbf{D}_z\mathbf{D}_b. \tag{7}$$

Now approximate the close-to-diagonal matrices \mathbf{D}_z , \mathbf{D}_a , \mathbf{D}_b and \mathbf{D}_c by diagonal matrices $\tilde{\mathbf{D}}_z$, $\tilde{\mathbf{D}}_a$, $\tilde{\mathbf{D}}_b$ and $\tilde{\mathbf{D}}_c$ obtained from \mathbf{D}_z , \mathbf{D}_a , \mathbf{D}_b and \mathbf{D}_c by setting all off-diagonal elements equal to 0. Substitution of these $\tilde{\mathbf{D}}_z$, $\tilde{\mathbf{D}}_a$, $\tilde{\mathbf{D}}_b$ and $\tilde{\mathbf{D}}_c$ into the estimating equation (7) of $\tilde{\mathbf{E}}$ implies that its solution $\tilde{\mathbf{E}}$ is also diagonal. Thus:

$$\tilde{\mathbf{E}} \approx (\tilde{\mathbf{D}}_z^{-2} + \tilde{\mathbf{D}}_a + \tilde{\mathbf{D}}_b\tilde{\mathbf{D}}_c)^{-1}(\tilde{\mathbf{D}}_z^{-1} - \tilde{\mathbf{D}}_a\tilde{\mathbf{D}}_z - \tilde{\mathbf{D}}_b\tilde{\mathbf{D}}_z\tilde{\mathbf{D}}_c). \tag{8}$$

The solution to the original problem of estimating \mathbf{W} can now be found by an iterative scheme. Iterate between

- (a) diagonalizing simultaneously \mathbf{A} , \mathbf{B} and \mathbf{C} and the current approximation \mathbf{Z} of \mathbf{W} ,
- (b) updating the error \mathbf{E} through formula (8) by using the diagonalized matrices that are acquired in the previous step,
- (c) updating \mathbf{Z} by using the latest estimate of the error \mathbf{E} and
- (d) repeating until convergence.

With an estimate of \mathbf{W} at hand, those of \mathbf{X} and, consequently, \mathbf{Y} are readily acquired.

5. Simulation

We now evaluate the performance of the above estimation and testing procedure by means of a simulation study. Here, several parameters are fixed throughout, whereas a full factorial design is set up for the remaining parameters.

In all simulations the standardized precision Ω is chosen to be symmetric with unit diagonal and three non-zero off-diagonal bands: $(\Omega)_{j,j+1} = 0.5 = (\Omega)_{j+1,j}$ for $j = 1, \dots, p - 1$, $(\Omega)_{j,j+2} = 0.25 = (\Omega)_{j+2,j}$ for $j = 1, \dots, p - 2$ and $(\Omega)_{j,j+3} = 0.1 = (\Omega)_{j+3,j}$ for $j = 1, \dots, p - 3$. Similarly, the diagonal matrices Υ_1 and Υ_2 are set as equal to the identity matrix, $(\Upsilon_1)_{j,j} = 1 = (\Upsilon_2)_{j,j}$ for $j = 1, \dots, p$, and kept constant throughout. While Ω , Υ_1 and Υ_2 are fixed, a full factorial design is built with the remaining free parameters γ , p and $n_1 = n = n_2$. The values of the remaining parameters are set to $\gamma \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, $p \in \{25, 50, 75, 100\}$ and $n \in \{25, 50, 75, 100, 125, 150\}$. The ranges of the last two parameters are motivated from practice. The range of p represents realistically sized and sufficiently specific pathways, whereas that of n covers the typically observed sample sizes of genomics studies. With the parameters at hand, data are drawn in accordance with the model that was outlined in Section 2: $\mathbf{Y}_{i,*} \sim \mathcal{N}(\mathbf{0}, \Upsilon_1\Omega^{-1}\Upsilon_1)$ for $i = 1, \dots, n$ and $\mathbf{Y}_{i,*} \sim \mathcal{N}[\mathbf{0}, \Upsilon_2\{(1 - \gamma)\Omega + \gamma\mathbf{I}_{pp}\}^{-1}\Upsilon_2]$ for $i = n + 1, \dots, 2n$.

To assess the behaviour of the estimation procedure, the model parameters γ , Ω , Υ_1 and Υ_2 are estimated from the generated data by using the estimation procedure in Section 3, for a given choice of the penalty parameter λ in a grid covering the interval $[10^{-2}, 1]$. Each design point of the factorial layout is repeated 100 times. The 100 resulting estimates (or statistics derived thereof) of the parameters are then aggregated (by means of the 10%, 50% and 90% quantiles) and plotted against the penalty parameter λ .

The simulation results on the estimation of γ for all settings (combinations) of the factorial design are displayed in Figs 2(a) and 2(b) and in the on-line SM section III. The obvious conclusion from these plots is that for a reliable and unbiased estimate of γ a reasonable (or large)

number of samples (and little to no penalization) is required. In all other cases, the estimate of γ will be biased. Nevertheless, estimates are concordant with the true value of the parameter, so long as the data are not dramatically undersampled and appropriate penalization is obtained.

To assess the power of the proposed test for $H_0: \gamma = 0$ versus $H_a: \gamma > 0$, parameter choices for γ , Ω , Υ_1 and Υ_2 , as well as values for n and p , were as in the scheme described. Here the combination ($n = 25, p = 100$) is excluded, as the estimates of γ did not reveal a concordant pattern with the true values of γ . Note that the significance α is $k/100$ with $k = 1, \dots, 20$.

The power is assessed as follows. For iteration l ,

- (a) draw from the null model ($H_0: \gamma = 0$),
- (b) determine λ_l for iteration l ,
- (c) obtain the permutation distribution of the test statistic for $\hat{\gamma}$ by using 1000 permutations, given λ_l ,
- (d) find critical value $c_{\alpha,l}$ for γ (using various choices of the level of significance α) and
- (e) draw 1000 times from the null model, estimate γ for each draw and assess whether this estimate exceeds the critical value $c_{\alpha,l}$: the fraction exceeding estimates is the power estimate.

This procedure is repeated for 100 iterations to obtain 100 power estimates. The 100 resulting power estimates are then aggregated by means of the 10%, 50% and 90% quantiles, and plotted against the significance level α . To evaluate the power under non-zero γ s, we also draw from H_a rather than from H_0 .

The simulation results on the power of the test for all combinations of the factorial design are displayed in Figs 2(c) and 2(d) and in the on-line SM section IV. We display the results by fixing three of the parameters and examining the relationship between the remaining quantities on the power. The following conclusions can be drawn from these plots. If $H_0: \gamma = 0$ is true, the power–significance relationship (i.e. $1 - \beta = f(\alpha)$) nicely follows the line $x = y$, indicating that the test exhibits proper behaviour under H_0 . Furthermore, the power of the test increases with both γ and the sample size n . An increase in the dimension p also improves the power. Hence, for larger values of p the test is not hampered by the biased estimation of γ ; rather it still has power to distinguish between different values of γ on the basis of data because of the concordancy of the estimation procedure. These results provide sample size recommendations. For instance, for a study involving a $p = 25$ -sized pathway where it is sought to detect a differential conditional independence of $\gamma = 0.5$ at significance $\alpha = 0.05$ and power $1 - \beta = 0.80$, sample sizes of $n_1 = n_2 = 25$ suffice.

The performance of the proposed testing procedure is evaluated under a misspecified alternative H_a (see the on-line SM section V for details and results). This reveals that the test proposed still has power under a misspecified alternative, but either the sample size should be relatively large, the pathway should be reasonably sized or the proportion of the pathway exhibiting differential conditional dependence should not be too small.

We also compare the proposed testing procedure with the omnibus-type test of Schott (2007). The competing test employs the test statistic $T = \|\hat{\Sigma}_1 - \hat{\Sigma}_2\|_F^2$, where both $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are ridge covariance estimates (Van Wieringen and Peeters, 2016) with a common penalty parameter chosen to be the average of their individual LOOCV penalty parameters. The p -value is then calculated by permutation. The power curves for this test can be found in the on-line SM section VI. The omnibus-type test performs competitively to the proposed test for large sample sizes, strong signals and large levels of significance but is outperformed by our test for smaller sample sizes and weaker signals. For instance, when $n = 75, p = 25$ and $\alpha = 0.05$ the power to

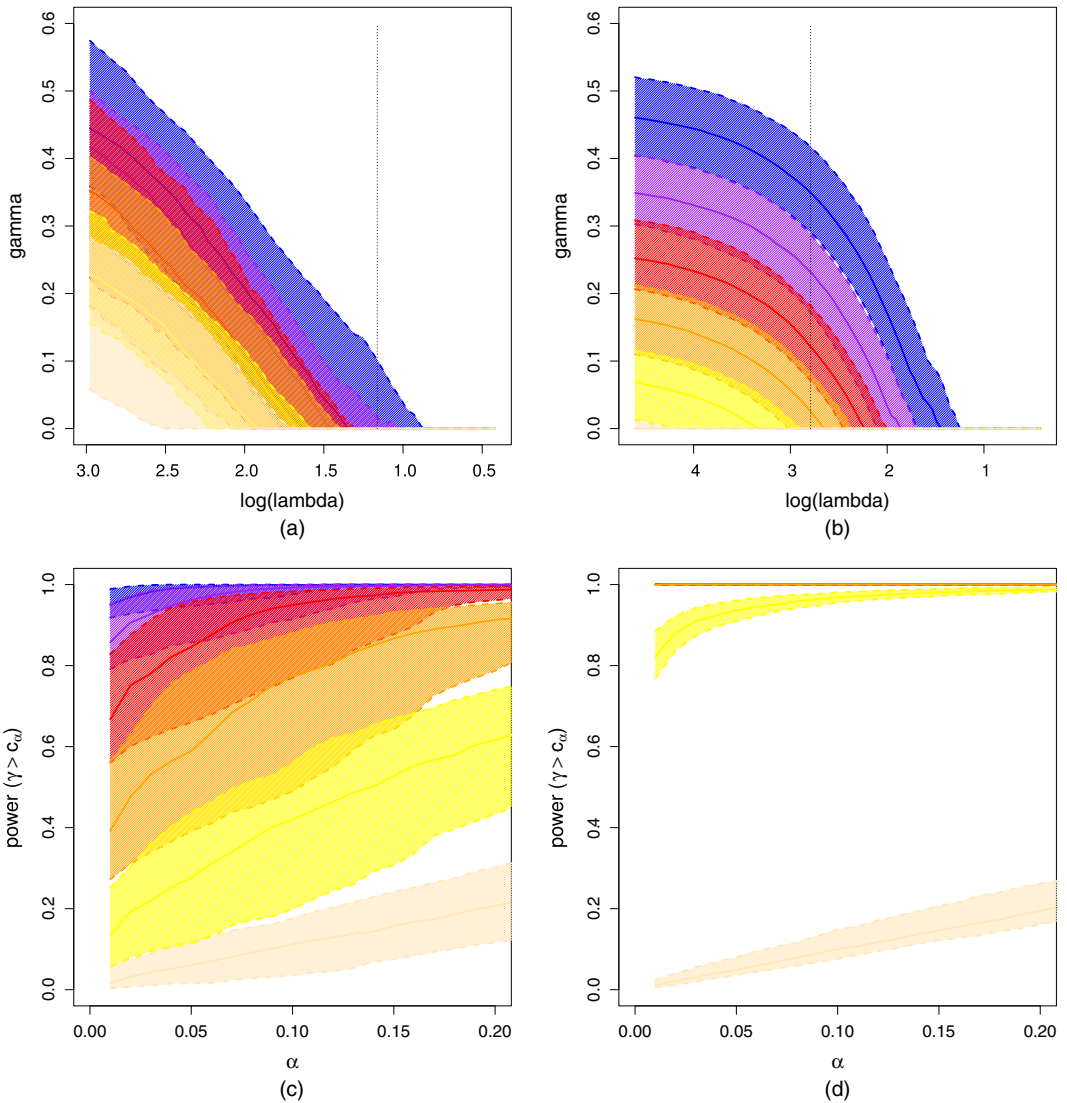


Fig. 2. Simulation results: (a), (b) estimated γ plotted against the (logarithm of the) penalty parameter λ for $\gamma \in \{0.0$ (—), 0.1 (—), 0.2 (—), 0.3 (—), 0.4 (—), 0.5 (—) $\}$ (samples sizes $n = 25$ and $n = 150$ both with a (25×25) -dimensional Ω) (\dagger , penalty parameter at which the estimates of Ω (on average over all 6×100 simulations) have a condition number equal to the true Ω); (c), (d) estimated power functions *versus* significance level α (parameters n and p and γ -colouring are as in (a) and (b))

detect differential conditional independence of $\gamma = 0.2$ is approximately 0.9 and 0.99 for the omnibus-type and the proposed test respectively. Moreover, where a significant result from the omnibus-type test gives no clue about the reason for the difference, the test proposed provides a clear pointer for the difference detected.

6. Application

We applied our proposed test to reanalyse gene expression data from several prostate cancer

studies. The focus is on the hedgehog signalling pathway, which is known to contribute to cellular proliferation in cancer (Evangelista *et al.*, 2006). Both Karhadkar *et al.* (2004) and Sheng *et al.* (2004) have reported the *in vitro* activation of the hedgehog signalling pathway in cancer, based on significant differences in mean expression levels of hedgehog-related biomarkers. This is corroborated by a simple differential expression analysis, which reveals multiple significant gene families in the six *in vivo* data sets (Table 1). Here we use the proposed methodology to investigate whether this activation is also noticeable in the conditional independences among the pathway's genes.

The data originate from six prostate cancer studies that are publicly available from the gene expression omnibus repository (Edgar *et al.*, 2002). Each study comprises normal and cancer samples (Table 1). For five of these studies (studies 1, 2, 3, 4 and 6), preprocessed data were available. A sixth study (study 5) was preprocessed as detailed in the on-line SM section VII. Of each study, probes interrogating genes which entrezID map to the *Kyoto Encyclopedia of Genes and Genomes* definition (Ogata *et al.*, 1999) of the hedgehog signalling pathway were retained. Expression levels of selected probes that contribute to the same family (as defined by the *Encyclopedia* and for completeness given in the SM section VII) are summarized by their median. The number of gene families (see Table 1) may differ between studies because of the employed microarrays interrogating different sets of genes. Finally, the second and third data sets (GSE6919_2 and GSE6919_3) both contain an outlying sample (as can be seen from their principal component plots, which are not shown). These samples are removed, even though results were not qualitatively affected (see the SM section VII for results with outlying samples).

In each data set, we test whether the hedgehog signalling pathway is activated (i.e. has stronger partial correlations) in the cancer group. For this, the optimal penalty parameter is determined by LOOCV under H_0 (see Section 4). Given the penalty parameter selected, the test statistic γ is estimated and its likelihood under $H_0: \gamma = 0$ (versus $H_a: \gamma > 0$) is evaluated by using 1000 permutations.

Table 1 contains the results (i.e. λ_{LOOCV} , $\hat{\gamma}$ and p -value). In each data set, λ_{LOOCV} was rather small. This is due to the reasonable number of samples in these studies in comparison with

Table 1. Results of the test of differential conditional dependence in the hedgehog signalling pathway from six gene expression omnibus prostate cancer data sets (with in both the second and the third data set one outlying sample removed)†

<i>Data set</i>	<i>Number normal</i>	<i>Number cancerous</i>	<i>Number of families</i>	<i>Number significant</i>	λ_{LOOCV}	$\hat{\gamma}$	<i>p-value</i>
GSE6919.1	81	90	13	7	3.36×10^{-8}	0.19659	0.00100
GSE6919.2	76	91	13	4	1.48×10^{-8}	0.57153	0.00100
GSE6919.3	75	89	8	4	1.57×10^{-8}	0.36269	0.00400
GSE29079	48	47	17	14	3.09×10^{-8}	0.01481	0.88212
GSE68907	50	52	13	1	4.45×10^{-8}	0.01977	0.72827
GSE3933	41	71	16	7	1.12×10^{-8}	0.23647	0.00100

† Under the alternative hypothesis the conditional dependences are smaller in the normal group. The columns (from left to right) contain the omnibus data set identifier, the number of normal samples, the number of cancer samples, the number of gene families comprising the pathway, the number of gene families exhibiting differential expression (based on the Wilcoxon test and after Hochberg multiplicity correction), the LOOCV optimal penalty parameter, the estimated differential conditional independence parameter and the p -value.

the pathway size. In four of the six data sets, the estimated γ hinted at stronger conditional dependences in the cancer group and γ significantly different from 0, whereas the other two are virtually 0. Test results on Gaussianized data, obtained through gene familywise application of the probability integral transformation that preserves conditional independences (Liu *et al.*, 2009), confirmed conclusions in Table 1 (the SM section VII). This suggests that removal of the two outlying samples is acceptable.

Considering now the GSE29079 data set, for which the test result was not significant, we noted that it involves many differentially expressed gene families. Therefore, an explanation for its non-significance may be the constitution of the tumour group, comprising two subgroups: TMPRSS2-ERG fusion positive and negative (and several unclassified) tumours (Börno *et al.*, 2012). Heterogeneity often leads to the dilution of partial correlations (Van Wieringen and Van der Vaart, 2015) and, as such, may obscure the activation of the hedgehog signalling pathway. In the case of the GSE68907 data set, also with a not-significant test result, no explanation can be offered except that it also exhibits little differential expression.

The results that are presented in Table 1 are not unanimous about the activation of the hedgehog signalling pathway. We therefore proceeded further exploring analyses results. We started by looking into the possibility of the inactivation of the hedgehog signalling in cancer. For this, the roles of the normal and cancer groups were simply reversed and the analyses repeated. This revealed (see the on-line SM section VII) that all estimated γ s were either 0 or virtually 0, suggesting that inactivation is generally unlikely.

The test results implied strengthened partial correlations in the cancer group in several of the data sets. To verify this, the precision matrices in both groups were estimated by means of the fused ridge ML approach of Bilgrau *et al.* (2015), in which LOOCV-chosen penalty parameters were again rather small (not shown). The partial correlations that were obtained were compared between the groups by using violin plots (Fig. 3(a) for the GSE6919_3 data set—results for the remaining data sets are in the on-line SM section VII). These violin plots show a comparable median partial correlation in both groups, but the larger quartile box and a heavier skewed tail confirm the significant test result for GSE6919_3.

The presence of differential conditional dependence can be assessed visually in various ways. The estimated partial correlations can be plotted against those obtained from the fused ridge estimates of the groupwise precision matrices. To emphasize the salient features, the latter were regressed (in line with the model presented without intercept) on the former. A steeper line indicates strengthened conditional dependences and is indeed observed for most data sets (see Fig. 3(b) for the GSE6919_3 data set, and the on-line SM section VII for the remaining data sets).

Subsequently, we examined the absolute partial correlations distributions to check whether these changed between groups. For this we used the empirical cumulative distribution function which, when shifted to the right (away from zero), indicates stronger conditional dependences. The empirical cumulative distribution functions of both groups in each data set were plotted (see Fig. 3(c) for the GSE6919_3 data set, and the on-line SM section VII for the remaining data sets). For data sets exhibiting significantly differential conditional dependence, these plots revealed a right-shifted empirical cumulative distribution function for the cancer group.

Another important check is that of adequacy of the global likelihood maximization, via the joint loss function of parameters from both groups. Here the inclusion of the differential conditional independence parameter need not lead to a better fit in the normal group. This was assessed by evaluating the Frobenius and quadratic loss of $\hat{\Omega}$ and $(1 - \hat{\gamma})\hat{\Omega} + \hat{\gamma}\mathbf{I}_{pp}$, compared with the fused ridge standardized precision estimate of the normal sample. We found that a precision matrix with globally weakened conditional dependences—i.e. inclusion of a non-zero

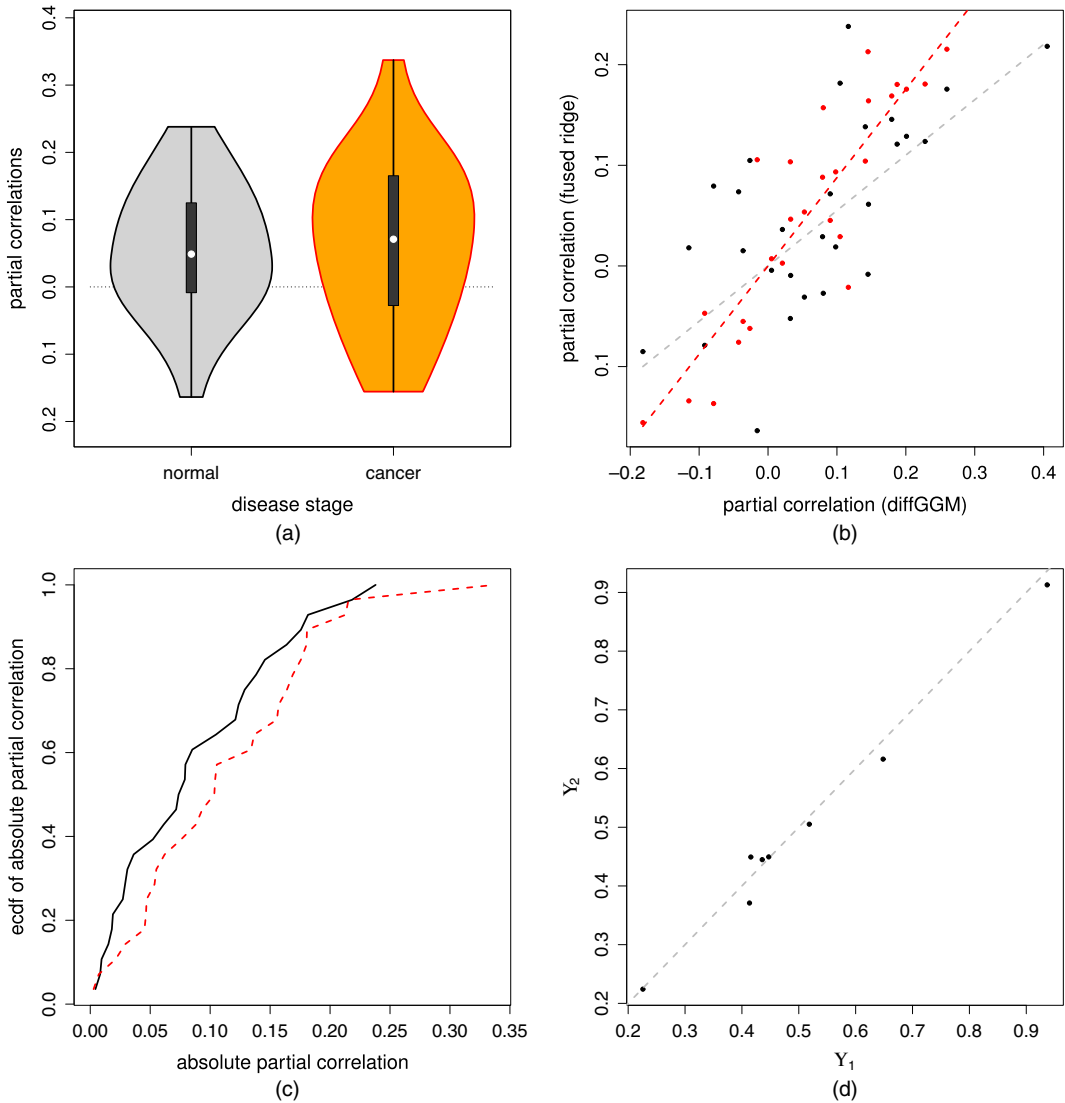


Fig. 3. (a) Violin plot of estimated partial correlations against group, (b) scatter plot of partial correlations from the estimated standardized precision against those obtained from the fused ridge groupwise precision matrix estimates (●, normal; ●, cancer), (c) empirical cumulative distribution functions of the absolute partial correlations from the fused ridged groupwise precision matrix estimates (—, normal; - - -, cancer) and (d) marginal variance estimates ($\hat{\Upsilon}_1$ and $\hat{\Upsilon}_2$) of both groups plotted against each other: all figures relate to the GSE6919.3 data set only

γ —generally leads to an improved description (not shown), in terms of fit criteria other than the (joint) likelihood, of the data from the normal sample.

The equal variance assumption ($\Upsilon_1 = \Upsilon_2$) under which the null hypothesis $H_0: \gamma = 0$ is also examined. For this, the unconstrained model is fitted to the data by using the penalty parameter from the original constrained analysis. For each data set, the resulting $\hat{\Upsilon}_1$ are plotted against the $\hat{\Upsilon}_2$. The plots (see the on-line SM section VII) do not give substantial reason to doubt the equal variance assumption. Finally, the thus—without the equal variance assumption—

re-estimated γ s are compared with those reported in Table 1. The two estimates do not differ substantially.

Once it has been concluded that the hedgehog signalling pathway is implicated in cancer, interest shifts towards identification of its constituents that are most likely to be responsible for this activation. To assess the contribution of each gene family, we evaluated the effect of its ‘removal’ from the pathway (through conditioning) on the differential condition dependence parameter. This was done by replacing a group sample covariance matrix, e.g. \mathbf{S}_1 , by its conditioned counterparts: $(\mathbf{S}_1)_{\setminus j, \setminus j} - (\mathbf{S}_1)_{\setminus j, j} (\mathbf{S}_1)_{j, j}^{-1} (\mathbf{S}_1)_{j, \setminus j}$. The estimated differential conditional dependence parameter is denoted by $\hat{\gamma}_{\setminus j}$, making use of the penalty parameter from the original analysis. Their average approximately equals the original parameter estimate: $\hat{\gamma} \approx (1/p) \sum_{j=1}^p \hat{\gamma}_{\setminus j}$. We observed that the ‘removal’ of gene families with $\hat{\gamma}_{\setminus j}$ positively deviating from $\hat{\gamma}$ resulted in stronger conditional dependences among the remaining families. Similarly, when removing gene families for which $\hat{\gamma}_{\setminus j}$ negatively deviated from $\hat{\gamma}$ resulted in weaker conditional dependences for remaining families. The *WNT*, *PKA* and *Ci* (also called *GLI*) gene families exhibit the largest positive deviations (consistently in significant data sets), whereas the *CKI* (also called *CSNK1*), *Gas1* and *GSK3B* families deviate most consistently in the opposite direction. Strikingly, the first three are known to be implicated in oncogenic processes, whereas the last three are associated with tumour suppressing functionalities (e.g. Del Sal *et al.* (1992) and Fuja *et al.* (2004)). With the effect of their ‘removal’ in mind, it may be suggested that ‘oncogenic families’ have a dysregulating effect on the pathway, reflected in weaker conditional dependences. Conversely, ‘tumour suppressing families’ appear to stimulate cohesion among the pathway’s constituents.

7. Discussion

We have addressed the problem of detecting differences in pathway activation status between two groups. For each group, we posed a GGM with different marginal variances but proportional partial correlations. The proportionality is parameterized by a single parameter representing differential systematic conditional dependence. Parameters of both GGMs are estimated jointly by an iterative procedure maximizing the likelihood. The estimation procedure is extended beyond the realm of low dimensional studies by augmenting the likelihood with a non-informative concave penalty on the standardized precision matrix. The associated penalty parameter is chosen by means of cross-validation and shown to have an interpretation as the mixing parameter between the data and a non-informative sample. The capacity of the proposed estimation procedure to estimate the differential conditional dependence parameter was studied in simulation, revealing good performance in low dimensional studies ($n_1, n_2 \gg p$), yielding useful estimates in medium dimensional studies (up to $n_1 = n_2 \approx p$), but breaking down for higher dimensional studies. Next, a testing procedure for differential systematic conditional dependence by means of a permutation scheme was described. To warrant exchangeability under the null hypothesis, the estimation was modified to accommodate the assumption of equal marginal variances of the groups. Finally, with the method proposed at hand, the *in vitro* activation of the hedgehog signalling pathway was confirmed in various *in vivo* gene expression data sets. The soundness of this conclusion was assessed by a wide range of diagnostics that complement the methodology proposed.

An alternative motivation for a test of pathway inactivation can be found in an emerging paradigm in cancer research (Teschendorff and Severini, 2010; Van Wieringen and Van der Vaart, 2011). Central to this paradigm is the notion of the entropy of a system. Here a system refers to the regulatory network of the pathway. Entropy is then a measure of concentration of the transcription levels of this system and can be interpreted as a measure of dysregulation (see Van

Wieringen and Van der Vaart (2015)). The paradigm suggests that the transcriptomic entropy generally (in most pathways) increases as cancer progresses, with the exception of processes that are essential to the viability of the cancer cell. Previously, in Van Wieringen and Van der Vaart (2015), we have proved (under the assumption of a GGM) that the entropy increase or decrease may be caused by weakening or strengthening of conditional dependences within the system. However, other causes may underlie the surge in transcriptomic entropy (Van Wieringen and Van der Vaart, 2015). A test for pathway (in)activation may enable us to distinguish between causes of entropy increase or decrease.

Future work involves extending the presented work in several directions. Firstly, an alternative procedure for choosing the penalty parameter may exploit the mixture interpretation of the penalized log-likelihood. The mixing proportion could be endowed with a prior, but indirectly to avoid an informative prior. This could be done through a prior on the condition number of the standardized precision matrix (following Won *et al.* (2013)). A large condition number would result from too little penalization (i.e. a small mixing proportion) in the medium-to-high dimensional case. A unit condition number is the result of too much penalization (and a unit mixing proportion). An ideal prior balances between these extremes and is flexible to accommodate the various study dimensionalities. A potential benefit of this approach may be computational gain with respect to the LOOCV procedure described.

Secondly, we could allow the strength of the differential conditional dependence to be associated with a linear covariate. For instance, very small ribonucleic acid (RNA) molecules (micro-RNAs) with expression levels also measured on a continuous scale either contribute to the degradation of messenger RNA or prohibit its translation. As micro-RNAs are believed to play an important role in oncogenesis, it is important to be able to detect micro-RNA-activated pathways.

Thirdly, the approach proposed may need to be robustified. In the application two outlying samples were removed from the analysis. It would be more elegant to leave every sample in and, if the data give rise to it, to downweight their influence, thus robustifying the method against outliers. Pragmatically, one may estimate the groupwise sample covariance matrices—the form in which the data enter the method—robustly by using methodology described in for example Campbell (1980). Preferably, an improved version of the method that is presented here would then complement such pragmatics by in-built safeguards against extreme observations.

A final inroad for improvement would be to strengthen the method's performance beyond the low and medium dimensional domain. This may require the use of a different penalty on the standardized precision matrix, for instance, a combination of the l_1 - and trace norm, that would yield a sparse and low rank matrix which has been shown to preserve the correlation structure reasonably (see Richard *et al.* (2012)). Although the sparsity of such an estimate is not the objective of the method presented, its low rank suggests that it would be stable in the high dimensional case. However, care should be applied as the sparsity of the estimate may hamper the estimation of the differential conditional dependence parameter (see Section 3.2).

References

- Bilgrau, A. E., Peeters, C. F. W., Eriksen, P. S., Bøgsted, M. and Van Wieringen, W. N. (2015) Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *Preprint*. (Available from <http://arxiv.org/pdf/1509.07982.pdf>.)
- Börno, S. T., Fischer, A., Kerick, M., Fälth, M., Laible, M., Brase, J. C., Kuner, R., Dahl, A., Grimm, C., Sayanjali, B., Isau, M., Röhr, C., Wunderlich, A., Timmermann, B., Claus, R., Plass, C., Graefen, M., Simon, R., Demichelis, F., Rubin, M. A., Sauter, G., Schlomm, T., Sültmann, H., Hans Lehrach, H. and

- Schweiger, M. R. (2012) Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov.*, **2**, 1024–1035.
- Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.
- Cai, T. T. (2017) Global testing and large-scale multiple testing for high-dimensional covariance structures. *A. Rev. Statist. Appl.*, **4**, 423–446.
- Campbell, N. A. (1980) Robust procedures in multivariate analysis I: robust covariance estimation. *Appl. Statist.*, **29**, 231–237.
- Cardoso, J.-F. and Souloumiac, A. (1996) Jacobi angles for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.*, **17**, 161–164.
- Danaher, P., Wang, P. and Witten, D. M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B*, **76**, 373–397.
- Del Sal, G., Ruaro, M. E., Philipson, L. and Schneider, C. (1992) The growth arrest-specific gene, Gas1, is involved in growth suppression. *Cell*, **70**, 595–607.
- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004) Sparse graphical models for exploring gene expression data. *J. Multiv. Anal.*, **90**, 196–212.
- Edgar, R., Domrachev, M. and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.*, **30**, 207–210.
- Evangelista, M., Tian, H. and de Sauvage, F. J. (2006) The hedgehog signaling pathway in cancer. *Clin. Cancer Res.*, **12**, 5924–5928.
- Flury, B. N. and Gautschi, W. (1986) An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM J. Scient. Statist. Comput.*, **7**, 169–184.
- Fuja, T. J., Lin, F., Osann, K. E. and Bryant, P. J. (2004) Somatic mutations and altered expression of the candidate tumor suppressors CSNK1 ϵ , DLG1, and EDD/hHYD in mammary ductal carcinoma. *Cancer Res.*, **64**, 942–951.
- Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusk, A. J. and Horvath, S. (2007) Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, **18**, 463–472.
- Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011) Joint estimation of multiple graphical models. *Biometrika*, **98**, 1–15.
- Ha, M. J., Baladandayuthapani, V. and Do, K. A. (2015) DINGO: differential network analysis in genomics. *Bioinformatics*, **31**, 3413–3420.
- Harville, D. A. (2008) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.
- Higham, N. J. (2008) *Functions of Matrices: Theory and Computation*. Philadelphia: Society for Industrial and Applied Mathematics.
- Karhadkar, S. S., Bova, G. S., Abdallah, N., Dhara, S., Gardner, D., Maitra, A., Isaacs, J. T., Berman, D. M. and Beachy, P. A. (2004) Hedgehog signalling in prostate regeneration, neoplasia and metastasis. *Nature*, **431**, 707–712.
- Ledoit, O. and Wolf, M. (2004) A well conditioned estimator for large dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365–411.
- Liu, H., Lafferty, J. and Wasserman, L. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, **10**, 2295–2328.
- Mitra, R., Müller, P. and Ji, Y. (2016) Bayesian graphical models for differential pathways. *Bayesian Anal.*, **11**, 99–124.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) Kegg: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.*, **27**, 29–34.
- Peterson, C., Stingo, F. C. and Vannucci, M. (2015) Bayesian inference of multiple graphical models. *J. Am. Statist. Ass.*, **110**, 159–174.
- Price, B. S., Geyer, C. J. and Rothman, A. J. (2015) Ridge fusion in statistical learning. *J. Computat Graph. Statist.*, **24**, 439–454.
- Richard, E., Savalle, P. A. and Vayatis, N. (2012) Estimation of simultaneously sparse and low rank matrices. In *Proc. 29th Int. Conf. Machine Learning*, pp. 51–58.
- Saegusa, T. and Shojaie, A. (2016) Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Statist.*, **10**, 1341–1392.
- Schott, J. R. (2007) A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computat Statist. Data Anal.*, **51**, 6535–6542.
- Sheng, T., Li, C., Zhang, X., Chi, S., He, N., Chen, K., McCormick, F., Gatalica, Z. and Xie, J. (2004) Activation of the hedgehog pathway in advanced prostate cancer. *Molec. Cancer*, **3**, 29.
- Städler, N. and Mukherjee, S. (2015) Multivariate gene-set testing based on graphical models. *Biostatistics*, **16**, 47–59.
- Teschendorff, A. E. and Severini, S. (2010) Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst. Biol.*, **4**, no. 104, 1–15.
- Van Wieringen, W. N. and Peeters, C. F. W. (2016) Ridge estimation of the inverse covariance matrix from high-dimensional data. *Computat Statist. Data Anal.*, **103**, 284–303.

- Van Wieringen, W. N. and Van der Vaart, A. W. (2011) Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics*, **27**, 556–563.
- Van Wieringen, W. N. and Van der Vaart, A. W. (2015) Transcriptomic heterogeneity in cancer as a consequence of dysregulation of the gene-gene interaction network. *Bull. Math. Biol.*, **77**, 1768–1786.
- Van Wieringen, W. N., Van de Wiel, M. A. and Van der Vaart, A. W. (2008) A test for partial differential expression. *J. Am. Statist. Ass.*, **103**, 1039–1049.
- Weinberg, R. A. (2006) *The Biology of Cancer*. New York: Garland Science.
- Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.
- Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013) Condition-number-regularized covariance estimation. *J. R. Statist. Soc. B*, **75**, 427–450.
- Xia, Y. (2017) Testing and support recovery of multiple high-dimensional covariance matrices with false discovery rate control. *Test*, **26**, 782–801.
- Xia, Y., Cai, T. and Cai, T. T. (2015) Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, **102**, 247–266.
- Zhao, S. D., Cai, T. T. and Li, H. (2014) Direct estimation of differential networks. *Biometrika*, **101**, 253–268.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary Material: Testing for pathway (in)activation using Gaussian graphical models'.