ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Tests for Gaussian graphical models

N. Verzelen [a,b], F. Villers [c,*]

[a] Université Paris-Sud, Laboratoire de Mathématique d'Orsay, 91405 Orsay Cedex, France

[b] INRIA Saclay, Equipe SELECT, Université Paris-Sud 91405 Orsay Cedex, France

[c] INRA, Mathématiques et Informatique Appliquées MIA, 78352 Jouy-en-Josas, France

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

Gaussian graphical models are promising tools for analysing genetic networks. In many applications, biologists have some knowledge of the genetic network and may want to assess the quality of their model using gene expression data. This is why one introduces a novel procedure for testing the neighborhoods of a Gaussian graphical model. It is based on the connection between the local Markov property and conditional regression of a Gaussian random variable. Adapting recent results on tests for high-dimensional Gaussian linear models, one proves that the testing procedure inherits appealing theoretical properties. Besides, it applies and is computationally feasible in a high-dimensional setting: the number of nodes may be much larger than the number of observations. A large part of the study is devoted to illustrating and discussing applications to simulated data and to biological data.

1. Introduction

Biological processes regulating the expression of the genes lead to complex high-dimensional systems. Thus, inferring these underlying networks recently became an issue arising in systems biology. More precisely, the challenge at hand is to use gene expression data coming from microarray experiments to estimate or to test the network. In this regard, mathematical tools were developed to provide a suitable framework for modelling complex dependence structures. Among these, Gaussian graphical models (GGMs; see Lauritzen (1996) and Edwards (2000)) have gained a lot of attention and have already been applied in several works (see Kishino and Waddell (2000), To and Horimoto (2002), Wu et al. (2003), Wille et al. (0000), and Schäfer and Strimmer (2005)). However, the number of genes p will typically exceed by far the number n of the samples given by the microarray experiments. In this high-dimensional setting, estimating or assessing a GGM raises difficult statistical and computational issues. For instance, most of the methodologies based on asymptotic statistics do not apply any longer.

In recent years, the problem of graph estimation for massive data sets became a hot-spot in statistics. Most of the emerging methods fall in two categories. On the one hand, some are based on multiple-testing procedures; see for instance Schäfer and Strimmer (2005) or Wille and Bühlmann (2006). On the other hand, other methods are based on variable selection for high-dimensional data. We mention the seminal work of Meinshausen and Bühlmann (2006), who proposed a computationally feasible model selection algorithm using Lasso penalisation (see Tibshirani (1996)). Huang et al. (2006) and Yuan and Lin (2007) extend this method to directly infer the graph by minimising the log-likelihood penalised by the l_1 norm.

In contrast, there are not many results about the problem of hypothesis testing in a high-dimensional setting. We believe that this issue is significant for two reasons. First, when considering a gene regulation network, biologists often have a

* Corresponding address: INRA, unité MIA, domaine de Vilvert, 78352 Jouy-en-Josas, France. Tel.: +33 1 34 65 22 39; fax: +33 1 34 65 22 17.
E-mail address: fanny.villers@jouy.inra.fr (F. Villers).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.022

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

previous knowledge of the graph and may want to test whether the microarray data match with their model. Second, when applying an estimation method in a high-dimensional setting, it could be useful to test the estimated graph as some of these methods are revealed as too conservative. Admittedly, some of the previously mentioned estimation methods are based on multiple testing. However, as they are constructed for an estimation purpose, most of them do not take into account some previous knowledge about the graph. This is for instance the case for the approaches of Drton and Perlman (2007) and Schäfer and Strimmer (2005). Some of the other existing procedures cannot be applied in a high-dimensional setting (e.g. Drton and Perlman (2008)). Finally, most of them lack theoretical justifications given in a non-asymptotic way. This is why we propose a testing procedure to assess whether some connections are missing in a graph. The procedure starts from a minimal graph, minimal in the sense that all edges are assumed to be relevant: typically this graph is provided by the biologists thanks to their previous knowledge. The aim of the procedure is to test whether microarray data match with this minimal graph or whether there are missing edges. The interest of this test is first for biologists assessing the quality of their knowledge. Second, when the test is rejected, it suggests potential connections between genes that steer biologists towards new experimentations.

Let us make precise our objective: consider $X = (X_1, \ldots, X_p)^t$, a random vector distributed as a multivariate Gaussian $\mathcal{N}(0, \Sigma)$. Throughout this paper, we assume that the matrix Σ is non-singular. The conditional independence structure of this distribution can be represented by an undirected graph $\mathcal{G} = (\Gamma, E)$ where $\Gamma = \{1, \ldots, p\}$ is the set of nodes and E the set of edges. There is an edge between nodes a and b if and only if the random variables X_a and X_b are conditionally dependent, given all remaining variables $X_{-\{a,b\}} = \{X_i, i \in \Gamma \setminus \{a, b\}\}$. The random vector X is then said to be a Gaussian graphical model with respect to the graph $\mathcal{G}$. Given a node $a \in \Gamma$, we define its neighborhood $ne(a)$ as the set of nodes $b \in \Gamma \setminus \{a\}$ such that $(a, b) \in E$. We say that X follows the local Markov property at node a with respect to the graph $\mathcal{G}$ if X_a is independent from $\{X_i, i \in \Gamma \setminus (ne(a) \cup \{a\})\}$ given $\{X_i, i \in ne(a)\}$. Lauritzen (1996) shows that X is a Gaussian graphical model with respect to $\mathcal{G}$ if and only if it follows the local Markov property at each node $a \in \Gamma$.

Suppose we are given an n-sample of the vector X and an undirected graph $\mathcal{G} = (\Gamma, E)$. In the present paper, we construct procedures for testing the hypothesis "X follows the local Markov property at the node a with respect to the graph $\mathcal{G}$" against the hypothesis that it does not. In the following, we refer to such a test as a *test of neighborhood*. We deduce procedures for testing the hypothesis "X is a Gaussian graphical model with respect to the graph $\mathcal{G}$" against the hypothesis that it is not. We call such a test a *test of graph*. Our test of neighborhood applies and is computationally feasible in a high-dimensional setting as long as the graph $\mathcal{G}$ is sparse. Besides, it inherits the appealing theoretical properties shown in a previous paper (see Verzelen and Villers (2007)): we are able to compute non-asymptotic bounds of its power and we show its optimality in the minimax sense.

In Section 2.1.1 we highlight the connection between tests of neighborhood and tests in Gaussian linear regression in a random Gaussian design. Thus, we construct procedures based on tests of linear hypothesis in this regression framework introduced in Verzelen and Villers (2007). They are feasible in a high-dimensional setting and we control exactly their familywise error rate. Then, we exhibit non-asymptotic results on their power in Section 2.2. Finally, we apply our procedures to simulated data in Section 3 and to real data sets in Section 4. In the sequel, we write $\overline{ne}(a) := ne(a) \cup \{a\}$ for any node $a \in \Gamma$.

2. Description of the testing procedures

2.1. Test of neighborhood

2.1.1. Connection with conditional Gaussian regression

In this part, we highlight the connection between the local Markov property and conditional regression of a Gaussian random variable. We define the testing procedure precisely in the next part, following the approach introduced in Verzelen and Villers (2007).

Let $\mathcal{G} = (\Gamma, E)$ be an undirected graph and $a \in \Gamma$ be a node of this graph. We want to test the hypothesis "X_a is independent from $X_{\Gamma \setminus \overline{ne}(a)}$ conditionally on $X_{ne(a)}$" against the general alternative that it is not. This hypothesis corresponds to the local Markov property defined in Lauritzen (1996) of X at the node a. In order to perform this test, we use a different characterisation of conditional independence.

Let us consider the conditional distribution of X_a given all remaining variables $X_{-a} = \{X_b, b \in \Gamma \setminus \{a\}\}$. Using standard Gaussian properties (see for instance Lauritzen (1996) appendix C), we know that this conditional distribution is a Gaussian distribution whose mean is a linear combination of elements in X_{-a} and whose variance does not depend on X_{-a}. Hence, we can decompose X_a as

$$X_a = \sum_{b \in \Gamma \setminus a} \theta_b^a X_b + \epsilon_a, \tag{1}$$

where θ^a is a vector of coefficients in $\mathbb{R}^{p-1}$ and ϵ_a is a zero-mean Gaussian random variable independent from X_{-a} whose variance equals the conditional variance of X_a given X_{-a}, $\mathrm{var}(X_a|X_{-a})$. The vector θ^a is determined by the inverse covariance matrix K of X (see Edwards (2000)). More precisely, $\theta_b^a = -K[a, b]/K[a, a]$ for any $b \neq a$ and $\mathrm{var}(X_a|X_{-a}) = 1/K[a, a]$. As a consequence, the set of non-zero coefficients of θ^a corresponds to the non-zero components of the a-th row of K.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

Equivalently, there is an edge between the nodes a and b in the graph if the quantity $K[a, b]$ is not zero. For any set $V \subset \Gamma\backslash\{a\}$, θ^a_V denotes the sequence $(\theta^a_b)_{b\in V}$.

Testing the null hypothesis "X_a is independent from $X_{\Gamma\backslash\overline{ne}(a)}$ conditionally to $X_{ne(a)}$" against the general alternative is therefore equivalent to testing the null hypothesis $H_{0,a}$: "$\theta^a_{\Gamma\backslash\overline{ne}(a)} = 0$" against the general alternative $H_{1,a}$: "$\theta^a_{\Gamma\backslash\overline{ne}(a)} \neq 0$". Consequently, the test of neighborhood amounts to goodness-of-fit tests for Gaussian regression with random Gaussian covariates as considered in Verzelen and Villers (2007).

2.1.2. Description of the procedure

In this part, we adapt the test introduced in Verzelen and Villers (2007) to our statistical context. We are given n observations of the vector $X = (X_1, \ldots, X_p)^t$. For any $a \in \Gamma$, let us denote as $\mathbf{X}_a$ the n-vector of observations of X_a and as $\mathbf{X}_{-a}$ the set of vectors $\mathbf{X}_b$ where b belongs to $\Gamma \setminus \{a\}$. The joint distribution of (X_a, X_{-a}) is uniquely defined by the vector θ^a, the covariance matrix of X_{-a} denoted as Σ_{-a}, and $\mathrm{var}(X_a|X_{-a})$, the conditional variance of X_a. In the sequel, $\mathbb{P}_{\theta^a}$ refers to the joint distribution of $(\mathbf{X}_a, \mathbf{X}_{-a})$. For the sake of simplicity, we do not emphasise the dependence of $\mathbb{P}_{\theta^a}$ on Σ_{-a} and $\mathrm{var}(X_a|X_{-a})$.

Let us first fix some level $\alpha \in]0, 1[$ and let m be a subset of $\Gamma \setminus \overline{ne}(a)$. In the sequel d_a and D_m denote the cardinalities of $ne(a)$ and m, and we define N_m as $n - d_a - D_m$. We assume that $n \geq d_a + 2$. We define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) := \frac{N_m \|\Pi_{ne(a)\cup m}\mathbf{X}_a - \Pi_{ne(a)}\mathbf{X}_a\|_n^2}{D_m \|\mathbf{X}_a - \Pi_{ne(a)\cup m}\mathbf{X}_a\|_n^2}, \tag{2}$$

where $\|.\|_n$ is the canonical norm in $\mathbb{R}^n$, and $\Pi_{ne(a)}$ and $\Pi_{ne(a)\cup m}$ respectively refer to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b\in ne(a)}$ and to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b\in ne(a)\cup m}$. Then, ϕ_m corresponds to the statistic of the Fisher test of the null hypothesis

$$\begin{aligned} &H_{0,a} : \theta_{\Gamma\backslash\overline{ne}(a)} = 0 \quad \text{against the alternative} \\ &H_{1,a,m} : \theta_{\Gamma\backslash\overline{ne}(a)} \neq 0 \quad \text{and} \quad \theta_{\Gamma\backslash(\overline{ne}(a)\cup m)} = 0. \end{aligned} \tag{3}$$

In the sequel, $\Pi_{ne(a)^\perp}$ stands for the orthogonal projection along the space generated by $(\mathbf{X}_b)$ with b belonging to $ne(a)$. Let us consider a finite collection $\mathcal{M}_a$ of non-empty subsets of $\Gamma \setminus \overline{ne}(a)$. For all $m \in \mathcal{M}_a$, the cardinality D_m must be smaller than $n - d_a$. We define $\{\alpha_m, m \in \mathcal{M}_a\}$, a suitable collection of numbers in $]0, 1[$ (which possibly depend on $\mathbf{X}_{-a}$). Our testing procedure consists in carrying out for each $m \in \mathcal{M}_a$ the Fisher test based on the statistic ϕ_m defined in Eq. (2) at level α_m and rejecting the null hypothesis $H_{0,a}$ if one of those tests does. More precisely, we define the test T_α as

$$T_\alpha := \sup_{m\in\mathcal{M}_a} \left\{\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m,N_m}(\alpha_m(\mathbf{X}_{-a}))\right\}, \tag{4}$$

where for any $u \in \mathbb{R}$, $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. We therefore reject the null hypothesis when T_α is positive. The main difference between this procedure and the one defined in Verzelen and Villers (2007) lies in the fact that we now deal with possibly random collection of models.

In order to ensure that the level T_α is less than α, the collection of weights $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in $]0, 1[$ must satisfy the property: for all $\theta \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma\backslash\overline{ne}(a)} = 0$, then $\mathbb{P}_\theta(T_\alpha > 0) \leq \alpha$. We choose the collection $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in accordance with one of the two following procedures:

- P_1: The α_m 's do not depend on $\mathbf{X}_{-a}$ and satisfy the equality

$$\sum_{m\in\mathcal{M}_a} \alpha_m = \alpha. \tag{5}$$

- P_2: For all $m \in \mathcal{M}_a$, $\alpha_m(\mathbf{X}_{-a}) = q_{\mathbf{X}_{-a},\alpha}$, where $q_{\mathbf{X}_{-a},\alpha}$ is defined conditionally to $\mathbf{X}_{-a}$ as the α-quantile of the distribution of the random variable

$$\inf_{m\in\mathcal{M}_a} \bar{F}_{D_m,N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right). \tag{6}$$

Note that this last distribution does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we can work out $q_{\mathbf{X}_{-a},\alpha}$ using a Monte Carlo method.

2.1.3. Comparison of Procedures P_1 and P_2

If the collection of models is not random, one can use either Procedure P_1 or Procedure P_2. In Verzelen and Villers (2007), Section 2.2, we show that the test T_α with Procedure P_1 has a size less than α, whereas the size of T_α with Procedure P_2 is exactly α. We deduce from this fact that the test T_α with procedure P_2 is more powerful than the corresponding test defined with Procedure P_1 with weights $\alpha_m = \alpha/|\mathcal{M}_a|$ (see Verzelen and Villers (2007), Section 2.3).

On the one hand the choice of Procedure P_1 allows us to avoid the computation of the quantile $q_{\mathbf{X}_{-a},\alpha}$ and possibly permits us to give a Bayesian flavour to the choice of the weights. On the other hand, Procedure P_1 becomes too conservative when the collection of models $\mathcal{M}_a$ is large. This is often the case when the number p of nodes in the graph is large. That is why we advise using Procedure P_2 when considering large graphs. We compare the two Procedures in practice in Verzelen and Villers (2007), Section 6, and in Section 3 of this paper.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

2.1.4. Collection of models $\mathcal{M}_a$

The main advantage of our procedure is that it is very flexible in the choices of the models $m \in \mathcal{M}_a$. If we choose suitable collections $\mathcal{M}_a$, the test is powerful over a large class of alternatives as shown in Verzelen and Villers (2007) for non-random collections. In this part, we propose two relevant classes of models $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ for our issue of test of neighborhood.

The collection $\mathcal{M}_a^1$ is defined as $\mathcal{M}_a^1 := \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}$ and consists in taking each node in $\Gamma \setminus \overline{ne}(a)$ in turn. In Section 2.2, we present theoretical results for the power of T_α with collection $\mathcal{M}_a^1$ and Procedure P_1. This collection presents the advantage of being relatively small compared to other possible collections and the procedure obtained is consequently computationally attractive.

We have shown in Verzelen and Villers (2007), and this will be illustrated again in Section 3, that if there are several non-zero coefficients in $\theta^a_{\Gamma\setminus\overline{ne}(a)}$, considering models of larger dimensions can improve the performance of the test. For instance, if we are given an order on the nodes and if the vector θ^a belongs to an ellipsoid relative to this order, one should choose the collection of nested models defined by this order (see Verzelen and Villers (2007), Section 5). There is not such an order in our context as we do not know in principle which nodes are more relevant to test. That is why we propose to use the LARS (least angle regression) algorithm introduced by Efron et al. (2004). This model selection algorithm provides an order of relevance of the covariates in linear regression. Besides, one of its main advantages lies in its computational attractiveness. The collection of models $\mathcal{M}_a^2$ is built as follows. We first choose an integer J which corresponds to the maximal size of the models that we want to consider. We advise taking J smaller than $n/2$. Then, we apply the LARS algorithm to the response $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in \Gamma \setminus \overline{ne}(a)$ and we obtain the sequence $s_{LARS} = (j_1, \ldots, j_J)$. Finally we define the collection $\mathcal{M}_a^2$ as

$$\mathcal{M}_a^2 := \{\{j_1, \ldots, j_k\}, 1 \le k \le J\}.$$

As the collection of models $\mathcal{M}_a^2$ given by the LARS algorithm now depends on the data, we need to define a new procedure to handle random collections.

If we are given a random collection of models $\mathcal{M}_a$ which only depends on

$$\Psi(\mathbf{X}_a, \mathbf{X}_{-a}) := \left(\frac{\Pi_{ne(a)^\perp}\mathbf{X}_a}{\|\Pi_{ne(a)^\perp}\mathbf{X}_a\|_n}, \mathbf{X}_{-a}\right), \tag{7}$$

then we shall use the test statistic (4) with weights given by the procedure P_3 defined as follows:

- P_3: For all $m \in \mathcal{M}_a[\Psi(\mathbf{X}_a, \mathbf{X}_{-a})]$, $\alpha_m(\mathbf{X}_{-a}) = q'_{\mathbf{X}_{-a},\alpha}$, the α-quantile of the distribution of the random variable is

$$\inf_{m\in\mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a,\mathbf{X}_{-a})]} \bar{F}_{D_m,N_m}(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})), \tag{8}$$

  conditionally to $\mathbf{X}_{-a}$. As for the procedure P_2, the distribution of (8) does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we are able to compute $q'_{\mathbf{X}_{-a},\alpha}$ using a Monte Carlo method.

Clearly, if the collection of models is not random, Procedures P_2 and P_3 lead to the same weights. As with Procedure P_2, the size of T_α with Procedure P_3 is exactly α. More precisely, for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta^a_{\Gamma\setminus\overline{ne}(a)} = 0$, we have that

$$\mathbb{P}_{\theta^a}(T_\alpha \,|\, \mathbf{X}_{-a}) = \alpha \quad \mathbf{X}_{-a} \quad \text{a.s.}$$

The result follows from the fact that $q'_{\mathbf{X}_{-a},\alpha}$ satisfies

$$\mathbb{P}_{\theta^a}\left(\sup_{m\in\mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a,\mathbf{X}_{-a})]}\left\{\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m,N_m}\left(q'_{\mathbf{X}_{-a},\alpha}\right)\right\} > 0 \,\middle|\, \mathbf{X}_{-a}\right) = \alpha,$$

and for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma\setminus\overline{ne}(a)} = 0$,

$$\Pi_{ne(a)\cup m}\mathbf{X}_a - \Pi_{ne(a)}\mathbf{X}_a = \Pi_{ne(a)\cup m}\boldsymbol{\epsilon}_a - \Pi_{ne(a)}\boldsymbol{\epsilon}_a,$$

and

$$\mathbf{X}_a - \Pi_{ne(a)\cup m}\mathbf{X}_a = \boldsymbol{\epsilon}_a - \Pi_{ne(a)\cup m}\boldsymbol{\epsilon}_a.$$

As the sequence of relevant variables given by the LARS algorithm does not depend on the norm of the response, the collection $\mathcal{M}_a^2$ only depends on $\Psi(\mathbf{X}_a, \mathbf{X}_{-a})$ and thus we are able to apply Procedure P_3.

The size of these two collections $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ is smaller than the number of nodes p. Consequently, the computational complexity of our procedure is at most linear with respect to p when considering the collection $\mathcal{M}_a^1$ and is of the same order as the complexity of the LARS algorithm when considering $\mathcal{M}_a^2$.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

2.2. Properties of the test of neighborhood with collection $\mathcal{M}_a^1$

For the convenience of the reader, we recall in this part some of the theoretical results established in Verzelen and Villers (2007). First, we give a proposition which characterises the set of vectors θ^a over which the test T_α with the collection $\mathcal{M}_a^1$ and weights $\alpha_m = \alpha/|\mathcal{M}_a^1|$ is powerful. We shall then discuss the optimality of this test.

Proposition 1. *Let us assume that n satisfies*

$$n - d_a - 1 \geq \left[10 \log\left(\frac{p - d_a - 1}{\alpha} \right) \vee 21 \log(1/\delta) \right].$$

Let us set the quantity

$$\rho^2_{n-d_a,p-d_a} := \frac{C_1}{n - d_a} \log\left(\frac{p - d_a - 1}{\alpha\delta} \right), \tag{9}$$

where C_1 is a universal constant. For any θ^a in $\mathbb{R}^{\Gamma\setminus\{a\}}$, $\mathbb{P}_\theta\,(T_\alpha > 0) \geq 1 - \delta$ if there exists $b \in \Gamma \setminus \overline{ne}(a)$ such that

$$\frac{\mathrm{var}_{\theta^a}(X_a|X_{ne(a)}) - \mathrm{var}_{\theta^a}(X_a|X_{ne(a)\cup\{b\}})}{\mathrm{var}_{\theta^a}(X_a|X_{ne(a)\cup\{b\}})} \geq \rho^2_{n-d_a,p-d_a}. \tag{10}$$

This proposition is a straightforward corollary of Theorem 1 in Verzelen and Villers (2007). One interprets the quantity appearing in (10) as follows: the quotient of conditional variances measures the ratio of the quantity of information brought by X_b for the prediction of X_a to the part of X_a not explained by $X_{ne(a)\cup\{b\}}$. In other words, the test T_α has a power larger than δ for vectors θ^a such that there exists a node $b \in \Gamma \setminus \overline{ne}(a)$ which improves the prediction of X_a enough.

This test is optimal in the minimax sense if we test against the alternative "$\theta^a_{\Gamma\setminus\overline{ne}(a)}$ has only one non-zero component" and if the covariates are independent (see Verzelen and Villers (2007), Section 4.2). The condition of independence for covariates is unrealistic in a Gaussian graphical context, but it is nevertheless relevant as the independent case is an important benchmark from the minimax point of view (see Verzelen and Villers (2007), Section 4.2 for more details). When the covariates are correlated we know from a simulation study (Verzelen and Villers (2007), Section 6) that using Procedure P_2 slightly improves the power of the test T_α.

2.3. Test of graph

From the test of neighborhood we define a procedure to test a graph. More precisely, we test the null hypothesis H_0 that "X is a Gaussian graphical model with respect to $\mathcal{G}$" against the alternative that it is not. Let $\{\alpha_a, a \in \Gamma\}$ be a collection of numbers in]0, 1[. For each node $a \in \Gamma$, we test at level α_a the neighborhood of the node a with one of the procedures explained in Section 2.1.2. We decide to reject the null hypothesis H_0 as soon as one of the test $T^a_{\alpha_a}$ is rejected. We obtain a test of level α of the graph $\mathcal{G}$ if we take $\{\alpha_a, a \in \Gamma\}$ such that $\sum_{a\in\Gamma} \alpha_a = \alpha$. In the sequel we choose $\alpha_a = \alpha/p$ for each $a \in \Gamma$.

This procedure corresponds to a Bonferroni choice of the weights. As a consequence, if the number p of nodes is very large, our test may suffer a loss of its size. This restricts us to considering tests of graph only for relatively small graphs, or for subgraphs of a large graph. Let us recall that when we apply the test of neighborhood to one node, the number p of nodes can be arbitrarily large without any loss in the size of the test, provided that we use Procedure P_2 or P_3.

3. Simulations

In this section we present two simulation studies. First, we study the test of graph when the number of nodes is small. On the one hand we compare the efficiency of Procedures P_1 and P_2 and on the other hand we show the influence of the percentage of edges in the graph on the power of the test. Second, we study the test of neighborhood when p is large, illustrating the power of our procedure in a high-dimensional setting. Besides, we compare the efficiency of the tests based on the collections of models $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ defined in Section 2.1.4.

3.1. Simulation of a GGM

3.1.1. Simulation of a graph

In our simulations we use two different methods to generate random graphs. The first one allows us to control the number of nodes p and the percentages of edges η in the graph. It consists in choosing uniformly and independently the positions of the $\eta \times p(p-1)/2$ edges. We use this method in the simulation experiment on the test of graph, with different values of η to measure the influence of the percentage of edges on the test.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

However, the vertices of real-world networks are often structured in clusters, i.e. groups of proteins functionally related, with different connectivity properties. That is why Daudin et al. (2006) proposed a model called ERMG, for Erdös–Rényi Mixtures for Graphs, which describes the way edges connect nodes, accounting for some groups of nodes, and some preferential connections between the groups. The ERMG model assumes that the nodes are spread into Q clusters with probabilities $\{p_1, \ldots, p_Q\}$. We are given a connectivity matrix C of size $Q \times Q$ which specifies the probability of connection between two nodes according to the clusters that they belong to. More precisely, the probability that two nodes belonging to the clusters i and j share an edge equals $C[i, j]$. We use this method to generate a graph in the simulation experiment on the test of neighborhood, with the following parameters provided by Daudin et al. (2006): $p = 199$ nodes, $Q = 7$ clusters, the probabilities $(p_1, \ldots, p_Q)$ and the connectivity matrix C equal to

$$(p_1, \ldots, p_Q) = \begin{pmatrix} 0.038 & 0.052 & 0.060 & 0.082 & 0.083 & 0.125 & 0.560 \end{pmatrix}, \tag{11}$$

$$C = \begin{pmatrix} 0.999 & 0.319 & 1e-06 & 0.116 & 1e-06 & 1e-06 & 0.007 \\ 0.319 & 0.869 & 1e-06 & 1e-06 & 0.140 & 0.004 & 0.002 \\ 1e-06 & 1e-06 & 0.467 & 0.0155 & 0.005 & 0.014 & 0.004 \\ 0.116 & 1e-06 & 0.016 & 0.216 & 1e-06 & 0.017 & 0.005 \\ 1e-06 & 0.140 & 0.005 & 1e-06 & 0.229 & 1e-06 & 0.004 \\ 1e-06 & 0.004 & 0.014 & 0.017 & 1e-06 & 0.239 & 0.013 \\ 0.007 & 0.002 & 0.004 & 0.005 & 0.0041 & 0.0129 & 0.0163 \end{pmatrix}. \tag{12}$$

Using these parameters, the percentage of edges η in the graph equals 2.5%.

3.1.2. Simulation of the data

Given a graph, we generate random vectors whose conditional independence structure is represented by the graph.

First, we generate the partial correlation matrix Π as follows. To a graph with p nodes we associate a symmetric $p \times p$ matrix U such that for any $(i, j) \in \{1, \ldots, p\}^2$, $U[i, j]$ is drawn from the uniform distribution between -1 and 1 if there is an edge between the nodes i and j and $U[i, j]$ is set to 0 in the other case. We then compute columnwise sums of the absolute values of the matrix U entries, and set the corresponding diagonal element equal to this sum plus a small constant. This ensures that the resulting matrix is diagonally dominant and thus positive definite. Finally, we standardise the matrix so that the diagonal entries all equal 1 to obtain the simulated partial correlation matrix Π.

Second, we simulate data of size n. We generate n independent samples from the multivariate normal distribution with mean zero, unit variance, and correlation structure associated with the partial correlation matrix Π. In the sequel, we denote as $\mathbf{X}$ the $n \times p$ associated data matrix.

3.2. Simulation setup

3.2.1. Simulation study of the test of graph

We evaluate the performance of the test of graph, first with simulations on randomly generated graphs, and secondly on a network coming from the data base KEGG.

(1) First simulation experiment: We estimate the level and the power of the test of graph with 1000 simulations. For fixed parameters (p, η, n), we generate 1000 graphs by using the first method described in Section 3.1.1 and 1000 data matrices as described in Section 3.1.2. Let $\mathcal{G}^s$ and $\mathbf{X}^s$ for $s = 1, \ldots, 1000$ denote the graphs and the data matrices for the 1000 simulations. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s$". We thus estimate the level of the test by dividing the number of simulations for which we reject the null hypothesis by 1000. Let q be a number in]0, 1[. For each simulation s, let $\mathcal{G}^s_{-q}$ be the graph built from the graph $\mathcal{G}^s$ in which we delete randomly $q\frac{p(p-1)}{2}\eta$ edges. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s_{-q}$". We estimate the power of the test by dividing the number of simulations for which we reject the null hypotheses by 1000.

The number of variables p is set to 15, whereas the number of observations n is taken equal to 10, 15 and 30 to study the effect of the sample size. We examine the influence of the percentage of edges in the graph, by taking $\eta = 0.1$ and 0.15. Besides this, we show the effect of the percentage q of missing edges on the power, by presenting the results for q equal to 10%, 40% and 100%.

(2) Second simulation experiment: This simulation is based on the cell cycle of yeast (*Saccharomyces cerevisiae*). This experiment aims at showing the performance of our procedure with simulations on a real biological network. The graph corresponding to the cell cycle of yeast is available in the database KEGG from the following website: http://www.genome.jp/kegg/pathway/sce/sce04111.html. We focus on a part of this pathway involving 16 proteins and 18 interactions. The graph, denoted in the sequel as $\mathcal{G}_{cellcycle}$, is shown in Fig. 1. We estimate the level and the power of the test by simulating 1000 data matrices $(\mathbf{X}^s)_{s=1,\ldots,1000}$ from the graph $\mathcal{G}_{cellcycle}$ as described in Section 3.1.2. We first estimate the level of the test by testing for each simulation s the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_{cellcycle}$". Then, we delete the three edges involving the protein complex *SCF Cdc4* in $\mathcal{G}_{cellcycle}$ in order

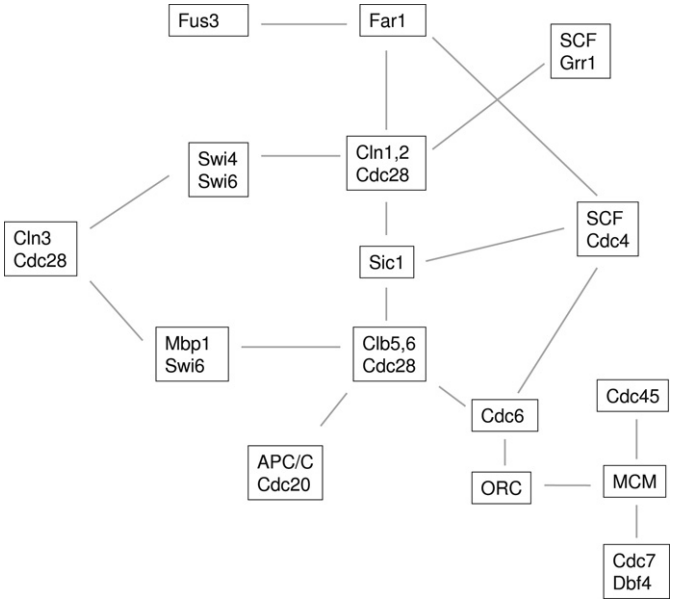

Fig. 1. $\mathcal{G}_{cellcycle}$.

to define the graph $\mathcal{G}_{cellcycle}^{-Cdc4}$. This protein complex *SCF Cdc*4 participates in cell death. We estimate the power of the test by testing for each simulation s the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_{cellcycle}^{-Cdc4}$. In other words we evaluate the ability of our procedure to detect the link of the protein complex *SCF Cdc*4 with the cell cycle.

3.2.2. Simulation study of the test of neighborhood

We first simulate a graph $\mathcal{G}$ according to the ERMG model described in Section 3.1.1 with $p = 199$ nodes, $Q = 7$ clusters, and the parameters $(p_1, \ldots, p_Q)$ and the matrix C defined in Eqs. (11) and (12). We then focus on a node a of this graph, chosen such that it has several neighbors. In our simulation this node has six neighbors. Let us denote as $ne(a)$ its neighborhood given by the graph $\mathcal{G}$. We simulate 1000 data matrices as described in Section 3.1.2, from the graph $\mathcal{G}$, and estimate the level of the test by testing the null hypothesis that the node a has no neighbor other than the set $ne(a)$, and the power by testing the null hypothesis that the node a has no neighbor. We present results when the sample size n is equal to 50, 100, and 200.

3.2.3. Collections of models $\mathcal{M}_a$ and collections $\{\alpha_m, m \in \mathcal{M}_a\}$

For each node a, we use the testing procedure defined in (4) with different collections $\mathcal{M}_a$ and different choices of the weights $\{\alpha_m, m \in \mathcal{M}_a\}$. Let us recall that $ne(a)$ denotes the neighborhood of the node a under the null hypothesis and α_a the level of the test of neighborhood for the node a. For the test of graph we choose $\alpha_a = \alpha/p$ and for the test of neighborhood α_a equals α.

The collections $\mathcal{M}_a$: we consider the two collections defined in Section 2.1.4:

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}$$

and

$$\mathcal{M}_a^2 = \{\{j_1, \ldots, j_k\}, 1 \leq k \leq J\}.$$

where $S_{Lars}[\Psi(\mathbf{X}_a, \mathbf{X}_{-a})] = \{j_1, j_2, \ldots, j_J\}$ is the sequence given by the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in \Gamma \setminus \overline{ne}(a)$. The maximum number of steps J is taken equal to 10. We evaluate the performance of our testing procedure with $\mathcal{M}_a^1$ in the simulation experiment on the test of graph, and we compare collections $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ in the simulation experiment on the test of neighborhood. Indeed, in the second simulation experiment, p and, thus, the collection $\mathcal{M}_a^1$ are large. It is therefore interesting to compare their respective computational costs.

The collection $\{\alpha_m, m \in \mathcal{M}_a\}$: When we consider the collection of models $\mathcal{M}_a^1$ we use either Procedure P_1 or Procedure P_2 defined in Section 2.1.2. For Procedure P_1 the α_m's are taken equal to $\alpha_a/|\mathcal{M}_a|$. The quantity $q_{\mathbf{X}_{-a},\alpha_a}$ occurring in Procedure P_2 is evaluated by simulation. Let Z be a standard Gaussian random vector of size n independent from $\mathbf{X}_{-a}$. As $\boldsymbol{\epsilon}_a$ is independent from $\mathbf{X}_{-a}$, the distribution of (6) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m \subset \mathcal{M}_a} \bar{F}_{D_m,N_m} \frac{\|\Pi_{ne(a)\cup m}(Z) - \Pi_{ne(a)}(Z)\|^2/D_m}{\|Z - \Pi_{ne(a)\cup m}(Z)\|^2/N_m},$$

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

Table 1
Test of graph, first simulation. $\eta = 0.1$. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels								
n				$T_{\mathcal{M}^1,P_1}$				$T_{\mathcal{M}^1,P_2}$
10				0.028				0.046
15				0.035				0.061
30				0.033				0.054
Estimated powers								
$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.73	0.75	10	0.94	0.94	10	0.99	0.99
15	0.83	0.84	15	0.97	0.98	15	1	1
30	0.95	0.95	30	1	1	30	1	1

Table 2
Test of graph, first simulation. $\eta = 0.15$. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels								
n				$T_{\mathcal{M}^1,P_1}$				$T_{\mathcal{M}^1,P_2}$
10				0.031				0.050
15				0.044				0.053
30				0.041				0.058
Estimated powers								
$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.28	0.32	10	0.70	0.72	10	0.90	0.91
15	0.44	0.46	15	0.87	0.88	15	0.99	0.99
30	0.73	0.75	30	0.99	0.99	30	1	1

conditionally to $\mathbf{X}_{-a}$. Consequently, we estimate the quantile $q_{\mathbf{X}_{-a},\alpha_a}$ by a Monte Carlo method with 1000 samples. When we use the collection $\mathcal{M}_a^2$ we apply Procedure P_3. The quantile $q'_{\mathbf{X}_{-a},\alpha_a}$ is again computed by a Monte Carlo method with 1000 simulations. The difference from the simulation of $q_{\mathbf{X}_{-a},\alpha_a}$ lies in the fact that the collection $\mathcal{M}_a^2$ is random and depends on ϵ_a. For each simulation, let Z be a standard Gaussian random vector of size n independent from $\mathbf{X}_{-a}$. We apply the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp} Z$ with the set of covariates $\Pi_{ne(a)^\perp} \mathbf{X}_b$ where $b \in \Gamma_{-a} \setminus ne(a)$. We obtain the sequence $S_{Lars}[\Psi(Z, \mathbf{X}_{-a})]$ which leads to the collection of models $\mathcal{M}_a^2[\Psi(Z, \mathbf{X}_{-a})]$. The Ψ function is defined in (7). As $\boldsymbol{\epsilon}_a$ is independent from $\mathbf{X}_{-a}$, the distribution of (8) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m \in \mathcal{M}_a[\Psi(Z,\mathbf{X}_{-a})]} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{ne(a)\cup m} Z - \Pi_{ne(a)} Z\|_n^2 / D_m}{\|Z - \Pi_{ne(a)\cup m} Z\|_n^2 / N_m}\right),$$

conditionally to $\mathbf{X}_{-a}$ and we therefore estimate the quantile $q'_{\mathbf{X}_{-a},\alpha_a}$. In the sequel, we write as $T_{\mathcal{M}_a^i,P_j}$ the test (4) with collection $\mathcal{M}_a^i$ and Procedure P_j.

3.3. The results

In Tables 1 and 2 we present results of the first simulation experiment on the test of graph respectively for $\eta = 0.1$ and $\eta = 0.15$. As expected, the power of the tests increases with the number of observations n. Besides, the power of the tests increases also with the percentage of missing edges q, the tests indeed being more powerful when the graphs under the null and the alternative hypotheses are more different. As expected, the tests based on Procedure P_2 are more powerful than the corresponding tests based on Procedure P_1. However because p is small, the difference between the two procedures is not really significant. Nevertheless, Procedure P_1 may become too conservative when p is large. As expected, its implementation is faster: for $p = 15$ and $n = 10$ a single simulation using Procedure P_1 takes approximately a tenth of a second whereas a single simulation using Procedure P_2 takes approximately 9 s. For p small, Procedure P_1 is therefore a good compromise in practice, Procedure P_2 being rather recommended when considering large graphs. Let us now compare the influence of η on the power of the test. When the percentage of edges η in the graph increases, the tests are less powerful. This is especially significant for $q = 10\%$. In fact, when η increases, the average number of neighbors for each node increases as well. In practice, the test of neighborhood is less powerful for a node which already has several neighbors under the null hypothesis. Consequently, the issue of testing the graph is more difficult when η is large.

In Table 3 we give the results of the second experiment for the test of graph. The percentage of edges in the graph $\mathcal{G}_{cellcycle}$ equals 15%, whereas the ratio of missing edges is $q = 1/6$ as we delete 3 edges among 18 in $\mathcal{G}_{cellcycle}$. In fact, as q is between

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

Table 3
Test of graph, second simulation experiment. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels			Estimated powers		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.040	0.055	10	0.43	0.46
20	0.046	0.063	20	0.76	0.79
30	0.040	0.058	30	0.89	0.90

Table 4
Test of neighborhood for the simulation experiment described in Section 3.2.2. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels			Estimated powers		
n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$	n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$
50	0.056	0.052	50	0.19	0.15
100	0.044	0.054	100	0.47	0.41
200	0.041	0.043	200	0.85	0.86

10% and 40% the powers of the tests in this setting are comparable to the results in Table 2. For $n = 20$ observations the test is powerful and detects the relation between the protein complex *SCFCdc*4 and the cell cycle with large probability. Even when n is smaller than p, the test detects the relation with a moderate probability.

In Table 4 we give the results of the experiment on the test of neighborhood. For $n = 50$ and 100 the test is more powerful when using the collection of models $\mathcal{M}_a^1$ whereas when n is larger the two procedures exhibit a comparable power. This comes from the fact that the test with collection $\mathcal{M}_a^2$ is performed in two steps, first, the selection of the relevant covariates using LARS and second, the test (4) itself. When n is small, LARS makes mistakes and possibly selects irrelevant covariates. In this case, the collection of models is bad and the test seldom rejects. When n is large, LARS often selects the relevant variables and the test $T_{\mathcal{M}^2,P_3}$ therefore takes advantage of exploiting models of several dimensions. However, its performances are not much better than those of $T_{\mathcal{M}^1,P_2}$ even when n is large. Let us now compare the computational efficiency of these two procedures. For $p = 200$ and $n = 100$ a single simulation using collection $\mathcal{M}_a^1$ is almost three times longer than one using collection $\mathcal{M}_a^2$. It seems natural to exploit models of several dimensions especially when we consider the test of neighborhood for a node which has several missing neighbors. However, the LARS algorithm does not really improve the performance of the procedure. Nevertheless, using collection $\mathcal{M}_a^2$ is computationally more attractive than using collection $\mathcal{M}_a^1$.

4. Application to biological data

In this section, we apply the test of graph to the multivariate flow cytometry data produced by Sachs et al. (2005). These data concern a human T cell signaling pathway whose deregulation may lead to carcinogenesis. Therefore, this pathway was extensively studied in the literature and a network involving 11 proteins and 16 interactions was conventionally accepted (see Sachs et al. (2005)). See Fig. 2 for a representation of this network. The data from Sachs consist of quantitative amounts of these 11 proteins, simultaneously measured from single cells under perturbation conditions. In the sequel, we focus on one general perturbation (anti-CD3/CD28 + ICAM-2) that overall stimulates the cellular signaling network. For this condition the quantities of the 11 proteins are measured in 902 cells. Let denote as D this data set constituted of $p = 11$ variables and $n = 902$ observations. In contrast to most of postgenomic data, flow cytometry data provide a large sample of observations that allow us to measure the influence of the sample size on the power. From this data set we infer the network using three methods and we apply our test of graph as a tool to validate these estimations. As such an abundance of data are rarely available in the postgenomic case, we secondly carry out a simulation study to determine the influence of the number of observations on the test. From the empirical covariance matrix obtained with the whole data set D, we generate data of different sample sizes and we evaluate the performance of the test with respect to the sample size.

We use the methods proposed by Drton and Perlman (2008), Wille and Bühlmann (2006), and Meinshausen and Bühlmann (2006) to infer the network. Let us briefly describe them. The SINful approach introduced by Drton and Perlman is a model selection algorithm based on multiple testing. For any couple of nodes they perform a test of existence of an edge between these two nodes and select the graph by computing the simultaneous p-values of these tests. This method assumes that the number of observations n is larger than the number of variables p. Two other methods have been recently proposed to deal with the usual fact in genomics of p large and n small. Wille and Bühlmann (2006) estimate a lower-order conditional independence graph instead of the concentration graph, while Meinshausen and Bühlmann (2006) estimate the neighborhood of any node with the lasso method. We represent the three estimated graphs in Fig. 3.

Let us define the graph $\mathcal{G}_\cap$ as the intersection of the graph estimated by these three methods and the graph with the connections well established in the literature. This graph $\mathcal{G}_\cap$ is represented in Fig. 4. We test with our procedure the null hypothesis $H_{\mathcal{G}_\cap}$: "the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_\cap$". We use for each node a of the graph the collection of models $\mathcal{M}_a^1$ defined in Section 2.1.4 and the procedure P_1. As p is small,

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

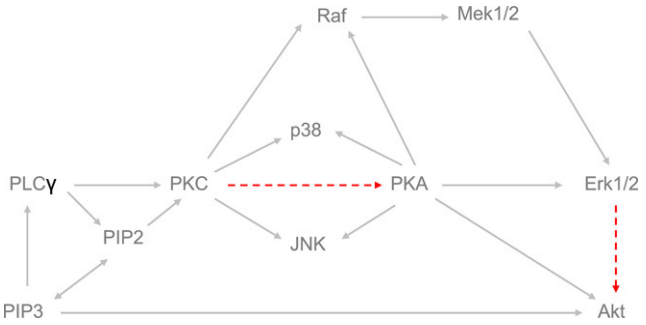

Fig. 2. Classic signaling network of the human T cell pathway. The connections well established in the literature are in grey and the connections cited at least once in the literature are represented by red dotted lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

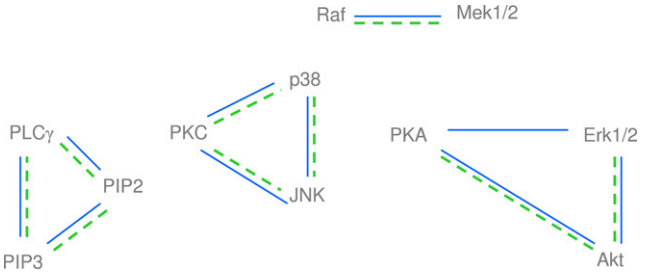

Fig. 3. Inferred graphs. The graphs estimated with the methods of Drton and Perlman and of Wille and Bühlmann are identical and represented in blue. The graph estimated with the method of Meinshausen and Bühlmann is shown by a green dotted line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

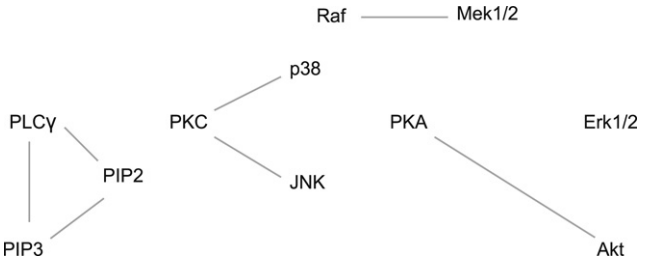

Fig. 4. Graph $\mathcal{G}_\cap$.

Table 5
Rejection of $H_{\mathcal{G}_\cap}$.

Rejection of the neighborhood of	
Node	Because of node(s)
*Erk*1/2	*Akt*, *PKA*
Akt	*Erk*1/2
PKA	*Erk*1/2
*p*38	*JNK*
JNK	*p*38

the difference between Procedures P_2 and P_1 is indeed not significant and the implementation of P_1 is faster. If we apply our procedure at level $\alpha = 5\%$, we reject the null hypothesis $H_{\mathcal{G}_\cap}$. In fact the p-value of the test is smaller than 10^{-10}. As our procedure consists in testing the neighborhood of each node, it is interesting to look for the nodes for which the test of neighborhood is rejected. For any of these rejected neighborhood tests, we then look for the alternatives leading to this rejection. In Table 5 we enumerate the nodes for which the test of neighborhood is rejected and the alternatives which lead to this decision.

As the connection *PKA*–*Erk*1/2 is well established and the connection *Erk*1/2–*Akt* is cited at least once in the literature, we decide to add those two edges in the graph $\mathcal{G}_\cap$, defining thus a new graph $\mathcal{G}_2$ shown in Fig. 5. The test of the null hypothesis $H_{\mathcal{G}_2}$ at level $\alpha = 5\%$: "the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_2$"

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

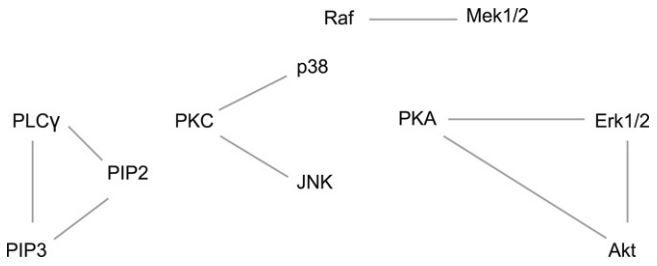

Fig. 5. Graph $\mathcal{G}_2$.

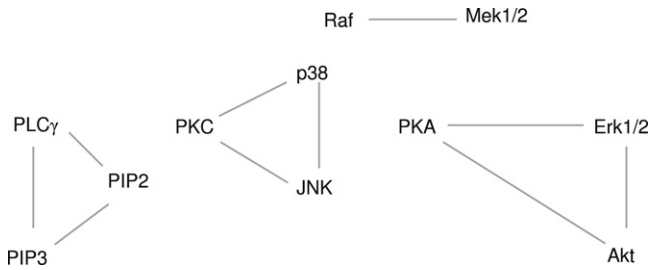

Fig. 6. Graph $\mathcal{G}_T$.

Table 6
Sachs data. Estimated levels and powers.

Estimated levels		Estimated powers	
n	$T_{\mathcal{M}^1,P_1}$	n	$T_{\mathcal{M}^1,P_1}$
10	0.032	10	0.49
15	0.036	15	0.86
20	0.033	20	0.97

is rejected, the p-value of the test being smaller than 10^{-10}. The reason is that the tests concerning respectively nodes $p38$ and *JNK* are rejected when we consider in the alternative respectively nodes *JNK* and $p38$.

We therefore define a new graph $\mathcal{G}_T$ by adding the connection $p38$–*JNK*, even if this connection is not well established in the literature. Let us note that the graph $\mathcal{G}_T$ is the same as the network inferred by Sachs et al. (2005) with approximately the same data set by using a Bayesian approach. We apply our test of graph and we accept the hypothesis that the data set D is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$ at the level $\alpha = 5\%$. In fact, the p-value of the test equals 8%. As n is large we use the result of the test with confidence and assume that the graph $\mathcal{G}_T$ (Fig. 6) represents the conditional independence structure of the data set D.

We now carry out a simulation study using this data set to determine the influence of the number of observations n on the power of our procedure. From the empirical covariance matrix obtained with the data set D, we generate 1000 simulated data $(\mathbf{X}^s)_{s=1,\ldots,1000}$ of different sample sizes n whose conditional independence structure is represented by the graph $\mathcal{G}_T$. First, we estimate the level of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$. Second, we delete the two edges involving protein *PKC* in $\mathcal{G}_T$ in order to define $\mathcal{G}_T^-$. We estimate the power of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T^-$.

The results of the simulation study using the selected data of Sachs are presented in Table 6. We recall that the graph involves $p = 11$ proteins and we take for the sample size n the values 10, 15, and 20. As expected, the power of the test increases with the number of observations n. However, the number of observations does not have to be very large to obtain a powerful test. For $n = 15$ observations, the test is able to recover that the protein *PKC* is not independent from the proteins $p38$ and *JNK* with large probability.

5. Conclusion

In this paper, we propose a multiple-testing procedure to assess whether some connections are missing in a minimal graph derived from experimental knowledge. Besides, when the procedure is rejected the different p-values of the tests suggest potential connections between genes/proteins that steer biologists towards new experimentations.

Our procedure is feasible in a high-dimensional setting. Hence, we advise using it to analyse microarray data for which the number of genes p typically exceeds the number of samples. Of course, when p becomes very large, the power of the procedure decreases, but this is intrinsic to the statistical problem.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022

Acknowledgements

We gratefully thank Sylvie Huet and Pascal Massart for many fruitful discussions.

References

Daudin, J.J., Picard, F., Robin, S., 2006. A mixture model for random graphs, Tech. Rep. RR-5840, INRIA.
Drton, M., Perlman, M., 2007. Multiple testing and error control in Gaussian graphical model selection. Statistical Science 22 (3), 430–449.
Drton, M., Perlman, M., 2008. A SINful approach to Gaussian graphical model selection. Journal of Statistical Planning and Inference 138 (4), 1179–1200.
Edwards, D.M., 2000. Introduction to Graphical Modelling, 2nd edition. Springer-Verlag, New York.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32 (2), 407–499.
Huang, J., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93 (1), 85–98.
Kishino, H., Waddell, P., 2000. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. Genome Informatics 11, 83–95.
Lauritzen, S.L., 1996. Graphical Models. Oxford University Press, New York.
Meinshausen, N., Bühlmann, P., 2006. High dimensional graphs and variable selection with the Lasso. The Annals of Statistics 34 (3), 1436–1462.
Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science 308, 523–529.
Schäfer, J., Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21, 754–764.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, Statistical Methodology 58, 267–288.
To, H., Horimoto, K., 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modelling. Bioinformatics 18, 287–297.
Verzelen, N., Villers, F., 2007. Goodness-of-fit tests for high-dimensional gaussian linear models. arxiv:math.ST/0711.2119.
Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., Bühlmann, P., Sparse graphical Gaussian modelling of the isoprenoid gene network in *arabidopsis thaliana*, Genome Biology 5.
Wille, A., Bühlmann, P., 2006. Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology 5.
Wu, X., Ye, Y., Subramanian, K., 2003. Interactive analysis of gene interactions using graphical gaussian model. In: Proceedings of the ACM SIGKDD Workshop on Data Mining in Bioinformatics. vol. 3, pp. 63–69.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94, 19–35.

Please cite this article in press as: Verzelen, N., Villers, F., Tests for Gaussian graphical models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.09.022