Journal of Multivariate Analysis 147 (2016) 145–154

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Influence measures and stability for graphical models

Avner Bar-Hen [a,*], Jean-Michel Poggi [b,c]

[a] Laboratoire MAP5, Université Paris Descartes, France
[b] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
[c] Université Paris Descartes, France

ARTICLE INFO

Article history:
Received 9 April 2015
Available online 28 January 2016

AMS 2010 subject classifications:
62-07
62-09
62G09
62G35

Keywords:
Influence measure
Graphical model
Robustness

ABSTRACT

Graphical models allow to represent a set of random variables together with their probabilistic conditional dependencies. Various algorithms have been proposed to estimate such models from data. The focus of this paper is on individual observations diagnosis issues. The use of an influence measure is a classical diagnostic method to measure the perturbation induced by a single element, in other terms we consider stability issue through jackknife. For a given graphical model, we provide tools to perform diagnosis on observations. In a second step we propose a filtering of the dataset to obtain a stable network. All along the paper an application to a gene expression dataset illustrates the proposals.

1. Introduction

Graphical models represent a set of random variables and encode their probabilistic conditional dependencies as a graph in which nodes represent random variables and edges represent conditional dependencies among them. Depending on the non-oriented or oriented feature of the dependencies, we get the general framework of graphical models (see [15]) or the more specific one called Bayesian networks (see [3]). Such graphical models have descriptive qualities and simulation capabilities. Indeed they can represent a graph of knowledge about relationships between the variables of interest to model a domain or a problem as well as they allow to propagate changes in the graph of conditional probabilities of the effects related to the observation of one or more causes, in the case of Bayesian networks. The interest in such models, since the graph can represent the scientific content of a given model, is twofold. On the one hand, from the applied side, they capture knowledge from multiple experts and experience from knowledge and data. On the other hand, from a statistical viewpoint, we are interested to examine issues of stability, sensitivity, scalability to cope with massive data. Therefore graphical models infer probabilistic relationships among variables and conditional dependence probabilities are estimated from data. Various algorithms have been proposed to estimate the topology and we focus on Maximum Likelihood Estimation.

Sensitivity issues are naturally of interest since the topology of the network and new relationships are estimated from data. For example, Cornalba et al. [5] consider in the context of Bayesian networks, sensitivity values as partial derivatives of output probabilities with respect to a given set of varying parameters. Sensitivity to the goodness of fit can be studied generating posterior distributions and applying a sensitivity analysis on posterior distributions, using design of experiments methodology. An alternative approach was proposed by Vogel and Tyler [23], they considered elliptical graphical models as a robust generalization of Gaussian graphical models and derived asymptotic properties of scatter estimators.

* Corresponding author.
E-mail addresses: Avner.Bar-Hen@mi.parisdescartes.fr (A. Bar-Hen), Jean-Michel.Poggi@math.u-psud.fr (J.-M. Poggi).

http://dx.doi.org/10.1016/j.jmva.2016.01.006
0047-259X/

This paper first focuses on a different viewpoint centered on individuals. The question of measuring influence of observations on the results obtained with a graphical model is of interest. A key tool in such a direction can be the use of an influence measure which is a classical diagnostic method to quantify the perturbation induced by a single element, in other terms we examine stability issue through jackknife highlighting influential observations.

To define the influence of individuals on the analysis, we propose a criterion to measure the sensitivity of the reference network defined as the estimated graphical model based on all observations. More precisely, we compare the network based on all observations except the concerned observation with the reference network, and we quantify the influence of one observation by the variation of penalized maximum likelihood. This first step allows to identify influential observations. We define the influential observations as those whose influence values are greater than a threshold. Taking a further step toward robustness, we derive a new network after removing the most influential observations. The dataset used to infer the new network has one observation less that the original dataset and we can compute the influence of each observation of the new dataset on this new network. This is the basic step of a procedure to define a stable network, described more precisely below.

All along the paper an application to gene expression dataset is carried out. This dataset provided by Hess et al. [14] concerns 133 patients with stage I–III breast cancer. The patients were treated with chemotherapy prior to surgery. Patient response to the treatment is classified as either a pathologic complete response (pCR) (34 patients) or a residual disease (not-pCR) (99 patients).

Graphical models are also used in a large variety of fields such as sociology, marketing, etc. In this article we mainly focus on a biological example but our work can be easily adapted to other context. Assuming that the common distribution of genes expressions is Gaussian, conditional independence is equivalent to independence. Graphical Gaussian Models have recently become a popular tool to study gene association networks. The key idea is to use partial correlations to measure dependence between two genes and the non-null entries of the inverse of the covariance matrix between genes allows to reconstruct the dependency network. For a multivariate normal distribution, using L_1-norm penalized likelihood maximization, lasso-type estimators of the concentration matrix of the graph are available. Most of the work focus on the model properties and aim at producing relevant network. On the other hand little attention have been paid on the observations that generate the network. This is the main topic of Section 3 of this article.

Meinshausen and Bühlmann [19] proposed stability selection as a very general technique designed to improve the performance of a variable selection algorithm. They illustrate the interest of the algorithm in the context of selection of stable graphs. The basic idea is that, instead of applying one's favorite algorithm to the whole dataset to determine the selected set of variables, one instead applies it several times to random subsamples of the data of size $\lfloor n/2 \rfloor$ and chooses those conditional dependencies that are selected most frequently on the subsamples. One may notice that it does not give any clue to conclude whether the conclusions are only driven by a few peculiar observations. Although the classical emphasis is to minimize the influence of such observation, another interesting aspect might be to detect them. In other words, are there any observation that drive the network topology, thus inducing changes when deleted? Since we want to characterize the influence of each observation, it is crucial to study them one at a time.

In order to achieve a more robust network, we removed the most influential observations from the analysis. If outliers indeed disrupt the inferred network, we expect that, after discarding enough of them, the inferred network will not be oversensitive to the sample anymore, that is, removing or adding one observation from the analysis will not drastically change it. In order to test this belief, we remove the most influential observation and compute a second network. We then compute the influence of each observation on this second network. The process can be iterated by removing the most influential observation on this second network and compute a third network. Finally we obtain a sequence of networks as well as a sequence of removed observations. The question of choosing the most stable network is addressed in the Section 4 of the paper.

The paper is organized as follows. Section 2 recalls first the basics on Gaussian graphical model. Then it introduces gene expression dataset (see [14]). Section 3 recalls the definition of influence functions and then define an influence measure for graphical models. Section 4 is devoted to the question of defining a stable network.

2. Model and dataset

2.1. Graphical models

Before introducing the framework, let us mention that while relying on sparse estimators, our developments are valid in a classical low-dimensional setting only.

Let $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mu, \Sigma)$ be a p-dimensional multivariate normal distributed random variable. Assuming that covariance matrix Σ is invertible, the conditional independence structure of the distribution can be represented as a graphical model $G = (\Gamma, E)$ where $\Gamma = \{1, \ldots, p\}$ is the set of nodes and E is the set of edges in $\Gamma \times \Gamma$. A pair (a, b) is contained in the set of edges if and only if X_a is dependent on X_b conditionally to the remaining variables $\{X_k, k \in \Gamma \setminus \{a, b\}\}$. Every pair of variables not contained in the edge set is conditionally independent given all remaining variables and corresponds to a zero entry in the inverse covariance matrix, that is: $\text{cor}(X_a, X_b | \{X_k, k \in \Gamma \setminus \{a, b\}\}) = 0$ corresponds to a zero entry in $\Theta = \Sigma^{-1}$.

Thus parameter estimation and model selection in the Gaussian concentration graph model are equivalent to estimating parameters and identifying zeros in the concentration matrix Σ^{-1} [10].

The log-likelihood for μ and $\Theta = \Sigma^{-1}$ based on a random sample $X_1, \ldots, X_n$ of X is

$$\frac{n}{2} \ln \det \Theta - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)' \Theta (X_i - \mu) \tag{1}$$

up to a constant not depending on μ and Θ. The maximum likelihood estimator of (μ, Σ) is $(\bar{X}, S)$ with $S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$, the empirical covariance matrix.

The concentration matrix Θ can be naturally estimated by S^{-1}. However, because of the possibly large number of unknown parameters $(p(p+1)/2)$ to be estimated, S is not a stable estimator of Σ for moderate or large p. In general, the matrix S^{-1} is positive definite when $n \geq p$, but does not lead to sparse graph structure since it typically contains no zero entry. To achieve sparse graph structure and to give a better estimator of the concentration matrix, the lasso idea is used and seek the minimizer

$$\ln \det \Theta - \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' \Theta (X_i - \mu) \quad \text{subject to} \quad \sum_{i \neq j} |\theta_{ij}| \leqslant t \tag{2}$$

over the set of positive definite matrices. Here $t \geq 0$ is the tuning parameter. When $t = \infty$, the solution is the maximum likelihood estimator S^{-1} provided that the inverse exists. On the other hand, if $t = 0$, then the constraint forces Θ to be diagonal, which implies that $X_1, \ldots, X_p$ are mutually independent. It is clear that $\hat{\mu} = \bar{X}$ regardless of t (see [18]).

Since both the objective function and feasible region of (2) are convex, we can equivalently use the Lagrangian form. Therefore, $\hat{\Theta}$ is the positive definite matrix that minimize the L_1-penalized log-likelihood given by:

$$\ell_\lambda^S(\Theta) = \ln \det \Theta - \operatorname{tr}(\Theta S) - \lambda \|\Theta\|_1 \tag{3}$$

where $\lambda \geq 0$ being the tuning parameter.

Let $\hat{\Theta} = \arg\max \ell_\lambda^S(\Theta)$ be the penalized ML estimate of the concentration matrix Σ^{-1} based on X_i, $i = 1, \ldots, n$. The main information within the concentration matrix used to define the graphical model is the adjacency matrix

$$\underline{\Theta} = \left(\mathbb{1}_{\theta_{ij} \neq 0}\right)_{1 \leqslant i,j \leqslant n}$$

which is a matrix of 0's and 1's. Since $\hat{\Theta}$ is sparse, a natural estimate is

$$\widehat{\underline{\Theta}} = \left(\mathbb{1}_{\hat{\theta}_{ij} \neq 0}\right)_{1 \leqslant i,j \leqslant n}.$$

In this paper, to estimate the models, we used the graphical lasso for the graph estimation [10] implemented in the R package `huge` (see [25]). There are various proposals to choose the optimal λ. Since our principal aim is robustness we used stability approach to regularization selection (StARS) [16] implemented in `huge.select`.

2.2. Gene expression dataset

A gene expression dataset provided by Hess et al. [14] and concerning 133 patients with stage I–III breast cancer, is used all along the paper. The patients were treated with chemotherapy prior to surgery. Patient response to the treatment is classified as either a pathologic complete response (pCR) or a residual disease (not-pCR). Hess et al. [14]; Natowicz [21] developed and tested a multigene predictor for treatment response on this dataset. They focused on a set of 26 genes having a high predictive value. We thus consider a total of $n = 133$ cases containing $p = 26$ gene expression levels leading to 133 rows and 26 columns. The kth row gives the expression levels of the 26 identified genes for the kth patient. The p columns are named according to the genes. The simple PCA of this dataset leads to the Fig. 1 showing on the left, the 133 patients classified as pCR in red and not-pCR in black, projected on the first factorial plane and, on the right, the 26 genes on the correlation circle. Let us remark that the patients classified as pCR are located in the upper right corner of factorial plane. The coordinates of a variable on the right part of Fig. 1 correspond to correlations with the two first principal components (therefore is included in the circle of radius 1). First axis leads to two groups of variables.

This dataset was already considered by Ambroise et al. [1] who proposed a method to infer a Gaussian Graphical Model taking into account some hidden structure on the nodes. They simultaneously infer the nodes groups and the graph using an L_1-penalized likelihood criterion. It was also studied by Giraud et al. [12].

As a result, the genes can then be represented in a plane and the estimated network corresponding to the complete dataset is figured by connecting the genes as in the left part of the Fig. 2. This plot has been obtained using the R package `graph` (see [7]) and the network is estimated through the graphical lasso procedure. By inspecting the gene graph representation together with the PCA correlation circle, recalled on the right part of the figure, it can be seen that the two main connected components, of 8 and 3 genes respectively, can be related to the upper left side and the right side of the correlation circle respectively.

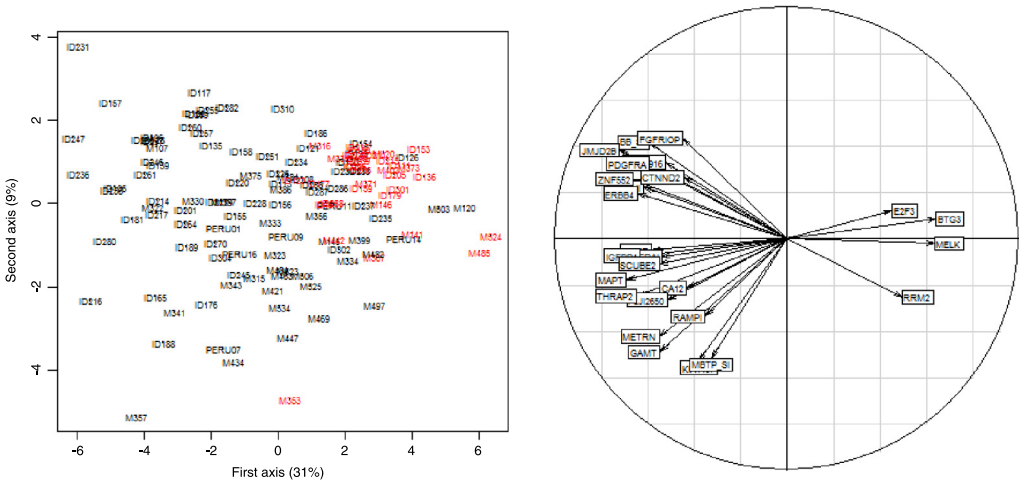

Fig. 1. PCA of the gene expression dataset. On the left, the 133 patients classified as pCR in red and not-pCR in black, projected on the first factorial plane. On the right, the 26 genes on the correlation circle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

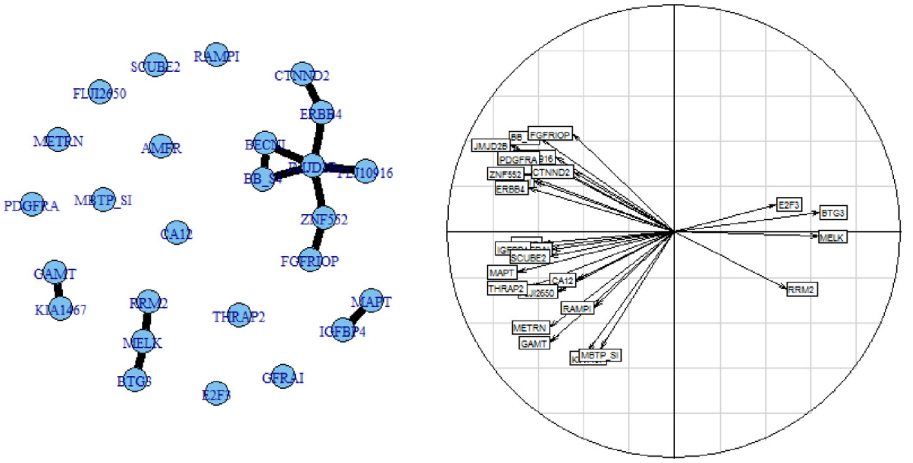

Fig. 2. On the left, gene graph representation. On the right the PCA correlation circle.

3. Influence measures

3.1. Influence function

3.1.1. Definitions and notations

Let $X_1, \ldots, X_n$ be random vectors of common distribution function F on $\mathbb{R}^p$ ($p \geq 1$). Let denote the point mass at the observation x_i by δ_{x_i}.
The influence function (IC) of a statistic T at F is

$$IC_{T,F}(x_i) = \lim_{\epsilon \to 0} \frac{T\big((1-\epsilon)F + \epsilon\, \delta_{x_i}\big) - T\big(F\big)}{\epsilon}. \tag{4}$$

The IC describes the effect of an infinitesimal contamination at point x_i on the estimator, standardized by the mass of the contamination.

The IC is an asymptotic concept and therefore we need a finite sample version. Using the empirical estimator of F:

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

we can define the empirical influence function $IC_{T,F_n}(x_i)$ by suppressing the limit and choosing $\epsilon = -\frac{1}{n-1}$ in the definition of IC.

There is strong connections between empirical influence function and jackknife [17]:

$$\begin{aligned} IC_{T,F_n}(x_i) &\approx \frac{T\big((1-\epsilon)F_n + \epsilon\,\delta x_i\big) - T\big(F_n\big)}{\epsilon} \\ &\approx (n-1)(T(F_n) - T(F_{n-1}^{(i)})) \end{aligned} \tag{5}$$

with $F_{n-1}^{(i)} = \frac{1}{n-1}\sum_{j\neq i}\delta_{x_j}$. Using the relation

$$F_n = \frac{n-1}{n}F_{n-1}^{(i)} + \frac{1}{n}\delta_{x_i}$$

and some mild assumptions, it is possible to prove that the empirical influence function $IC_{T,F_n}(x_i)$ is a consistent estimator of the influence function $IC_{T,F}(x_i)$.

3.1.2. Jackknife concentration matrix

Let first look at the perturbation induced by removing one observation X_j. The empirical covariance matrix then becomes

$$S_{-j} = \frac{1}{n-1}\sum_{i\neq j}(X_i - \bar{X}_{-j})(X_i - \bar{X}_{-j})'$$

where $\bar{X}_{-j}$ is the mean of the X_i for $i \neq j$.
The calculation of S and S_{-j} for $j = 1, \ldots, n$ can be computationally heavy as soon as p or n is large. Therefore it can be useful to express S_{-j} as a function of S.
Let us note at first that

$$\overline{X} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\left(\sum_{i\neq j} X_i + X_j\right) = \frac{n-1}{n}\overline{X}_{-j} + \frac{1}{n}X_j = \overline{X}_{-j} + \frac{1}{n}(X_j - \overline{X}_{-j}).$$

Thus we have

$$\begin{aligned} S &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})' \\ &= \frac{1}{n}\sum_{i\neq j}\left(X_i - \overline{X}_{-j} - \frac{1}{n}(X_j - \overline{X}_{-j})\right)\left(X_i - \overline{X}_{-j} - \frac{1}{n}(X_j - \overline{X}_{-j})\right)' + \frac{n-1}{n^2}(X_j - \overline{X}_{-j})(X_j - \overline{X}_{-j})' \\ &= \frac{n-1}{n}S_{-j} + \frac{n-1}{n^2}(X_j - \overline{X}_{-j})(X_j - \overline{X}_{-j})' \end{aligned}$$

since $\sum_{i\neq j}(X_i - \overline{X}_{-j}) = 0$.
Finally we have

$$S_{-j} = \frac{n}{n-1}S - \frac{1}{n}(X_j - \overline{X}_{-j})(X_j - \overline{X}_{-j})'. \tag{6}$$

Therefore we have the following lemma.

Lemma 1. *Following the previous derivation of S_{-j} from S, L_1-penalized likelihood using S_{-j} can be expressed in terms of S:*

$$\ell_\lambda^{S_{-j}}(\Theta) = \ln\det\Theta - \frac{n}{n-1}tr(\Theta S) - \frac{1}{n}(x_j - \bar{x}_{-j})'\Theta(x_j - \bar{x}_{-j}) - \lambda\|\Theta\|_1. \tag{7}$$

Then:

$$\ell_\lambda^{S_{-j}}(\Theta) = \ell_\lambda^{S}(\Theta) - \frac{1}{n}(x_j - \bar{x}_{-j})'\Theta(x_j - \bar{x}_{-j}) - \frac{1}{n-1}tr(\Theta S). \tag{8}$$

Note that the effect is to add a quadratic term that taking into account the contribution of x_j to the penalized likelihood.

3.2. Influence measure based on jackknife concentration matrix

Let $\widehat{\Theta_{-j}} = \arg\max \ell_\lambda^{S_{-j}}(\Theta)$ be the penalized MLE of Σ^{-1} based on $X_i,\ i \neq j$, the jackknife dataset. The maximum is to be taken over the set of positive definite matrices.

To estimate $n(\ell_\lambda^S(\Theta) - \ell_\lambda^{S_{-j}}(\Theta))$ Eq. (8) gives two candidates: $(x_j - \bar{x}_{-j})'\hat{\Theta}(x_j - \bar{x}_{-j}) - \frac{n}{n-1}\mathrm{tr}(\hat{\Theta}S)$ and $(x_j - \bar{x}_{-j})'\widehat{\Theta_{-j}}(x_j - \bar{x}_{-j}) - \frac{n}{n-1}\mathrm{tr}(\widehat{\Theta_{-j}}S)$.

Alternatively, using Lemma 1, we capture the variability of the penalized likelihood when considering S_{-j} instead of S. Therefore we define:

$$I(j) = \ell_\lambda^S(\hat{\Theta}) - \ell_\lambda^{S_{-j}}(\widehat{\Theta_{-j}}). \tag{9}$$

Taylor expansion is a natural tool to quantify the distance between $I(j)$ and $\ell_\lambda^S(\Theta) - \ell_\lambda^{S_{-j}}(\Theta)$.

Lemma 2. *Let assume the classical hypothesis for convergence of maximum likelihood (see [22] for example) then*

$$\ell_\lambda^S(\Theta) = \ell_\lambda^S(\widehat{\Theta}) + \left\langle (\hat{\Theta} - \Theta), \frac{\partial \ell_\lambda^S(\Theta)}{\partial \Theta} \right\rangle + \frac{1}{2}(\hat{\Theta} - \Theta)' \frac{\partial^2 \ell_\lambda^S(\Theta)}{\partial \Theta^2}(\hat{\Theta} - \Theta) + \varepsilon(\hat{\Theta} - \Theta)$$

where $\langle , \rangle$ denote the canonical scalar product and $\varepsilon(\hat{\Theta} - \Theta) \to 0$ when $\hat{\Theta} \to \Theta$ and is negligible in front of the other terms.

Let us note at first that the same penalty value λ is used for all the jackknifed versions, in order to control the variation of the penalized likelihood evaluated on $\hat{\Theta}$.

The effect of deleting observation X_i can be measured by its influence value $IC_{T,F_n}(X_j)$. Since the penalized log-likelihood is a statistic, Eq. (5) gives:

$$IC_{T,F_n}(X_j) = (n-1)\left(\ell_\lambda^S(\widehat{\Theta}) - \ell_\lambda^{S_{-j}}(\widehat{\Theta_{-j}})\right) \tag{10}$$

$$= (n-1)I(j). \tag{11}$$

The most interesting property of Eq. (10) is the possibility to characterize the influential observations, *i.e.* observations for which $IC_{T,F_n}(X_i)$ is either very positive or very negative. Therefore we are mainly interested by the absolute value of influence index. In real case dataset, and under our assumption that only a few influential observations disrupt the robustness of the inferred graph, we expect to find many observations with small influence value and a few observations with large influence value. Therefore, we focus on observations with very large absolute value of influence index and call them influential observations.

(Graphical) Lasso estimates are known to be biased estimates of the partial correlations but they consistently estimate (under some assumptions) the adjacency matrix of the graph. So an alternative idea would be to focus on the topology of the graphical model and to count the number of edges affected by the removing observation j. This can be written as:

$$\frac{1}{2}\|\underline{\widehat{\Theta}} - \underline{\widehat{\Theta_{-j}}}\|_0. \tag{12}$$

The adjacency matrix is symmetric and the factor 1/2 is necessary. Since self-loops are not allowed in graphical models, the diagonal elements are zero are not involved in Eq. (12).

Distributional results for this index are far from obvious since the indicator function is not continuous and therefore there is no guarantee of convergence of $\underline{\widehat{\Theta}} - \underline{\Theta}$ to zero. This question is well known in robustness literature (see [6] for example). Note that in high-dimension, model selection consistency results for the lasso estimator exist under some assumptions. See [24,20] for example.

3.3. Distributional results

Since I is derived from penalized log-likelihood it is easy to derive the distributional properties of I.

Various authors studied the asymptotic distribution of the empirical influence function (see [11,8] for example). They obtain that the asymptotic distribution of $\sqrt{n}\left(IC_{T,F_n}(x) - IC_{T,F}(x)\right)$ is the same that of

$$\sqrt{n}\int IC_{T,F}(x)\, F_n(dx) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_{T,F_n}(x_i)$$

which is, by Central Limit Theorem (CLT), asymptotically distributed as a normal law with zero mean and variance

$$\sigma^2 = \int IC_{T,F}^2(x)\, F(dx)$$

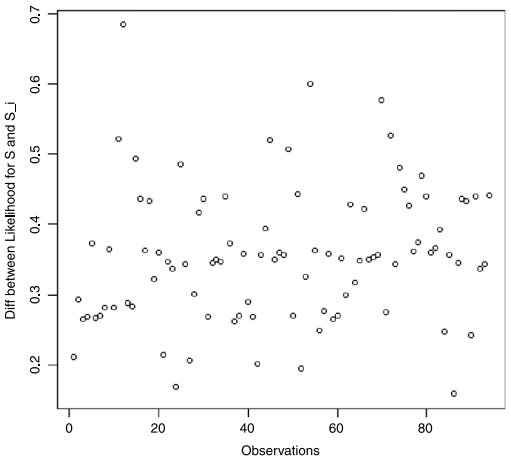

Fig. 3. Absolute value of the variation of the penalized maximum likelihood induced by removing the ith observation.

and a natural estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} IC^2_{T,F_n}(x_i).$$

To get this, we need to impose on T more than the mere existence of $IC_{T,F}(x)$. We refer to Hampel et al. [13] for details and we will work with Hadamard differential.

It is well known that likelihood and empirical likelihood are Hadamard differentiable [4, for example] therefore the asymptotic distribution of I is

$$\sqrt{n}IC_{T,F_n}(X_i) = \sqrt{n}(n-1)\left(\ell^S_\lambda(\hat{\Theta}) - \ell^{S_{-j}}_\lambda(\widehat{\Theta_{-j}})\right) \sim \mathcal{N}(0,\sigma^2).$$

One may notice that this gives the speed of convergence of $\ell^{S_{-j}}_\lambda(\widehat{\Theta_{-j}})$ to $\ell^S_\lambda(\hat{\Theta})$ but does not give any indication on the convergence of the parameters of the graphical model based on $X_1,\ldots,X_n$ to the true parameters of the graphical model.

3.4. Influence on concentration in action

The Fig. 3 shows the absolute values of I the difference between the penalized maximum likelihood computed on the whole set of data and the penalized maximum likelihood computed after removing observation j. While most of the observations lead to a moderate variation of the likelihood when removed, few observations lead to a perturbation of the maximum likelihood. A natural question is to define a threshold to identify outliers. This is the main interest of our distributional results. At the 5% level of confidence (without correction for multiple testing) the three largest differences are significant.

Moreover it is of interest to inspect the edges that appear or disappear when computing the penalized ML graphical model after removing potential outliers.

4. Stable network

Our first interest is to quantify the influential observation but a dual approach to this question is to look at the stability of the network.

4.1. Influence measuring stability of the links through jackknife

The first idea is to study the stability of the graph obtained from the full dataset. In this section we study the influence of a perturbation on the network induced by the deletion of an observation. This leads to define a distance between the estimated concentration matrices. Any distance between matrices is suitable but a natural choice is to count the number of edges altered by the deletion of a single observation.

This simple idea applied to the gene expression dataset leads to the following results. Reference graph is generated from the whole dataset and influence of a perturbation induced by the deletion of an observation can be measured by any distance between $\hat{\Theta}$ and $\widehat{\Theta}_{-i}$.

Table 1
Distribution of index J.

0	1	2	3	36	39	40	42	50	51	52	53	54
285	7	2	1	1	1	1	2	2	4	3	3	13

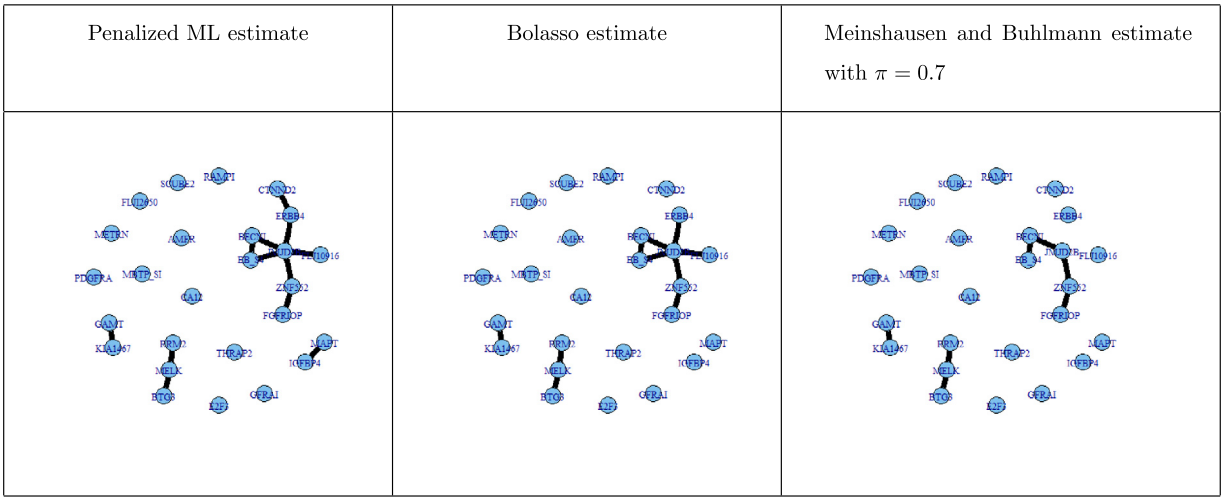

Fig. 4. Estimated graph structures for penalized MLE and Bolasso and Meinshausen and Bühlmann procedure with $\pi = 0.7$.

Let $J(a, b)$ be the number of times that status of edge (a, b) is changed by the removing of one observation:

$$J(a,b) = \sum_{i=1}^{n} \mathbb{1}_{\left|\widehat{\underline{\Theta}(a,b)} - \widehat{\underline{\Theta}_{-i}(a,b)}\right| \neq 0}.$$

Table 1 gives the distribution of the $25 * 26/2 = 325$ possible edges and for each edge the theoretical range of J is between 0 and 133. Most of the links are totally stable (285 edges) or very stable (295 edges are changed less than three times) but a set of thirty links can be considered as very unstable ($J > 35$). This result should be taken into account before deriving biological properties of the network.

Basically this approach gives insight about the stability of the graph inferred from the full dataset but does not provide a stable network.

A similar approach can be found in [16,19]. They proposed methods to choose the regularization parameter such that the resulting graph is sparse and replicable under random sampling. In both cases, they used the stability of graphical network constructed from subsampled datasets. It has the property to minimize the weight of outliers but it would be more satisfactory to remove it.

4.2. Stable graphical model

Since graphical models are very sensitive to outliers it is important to define stable ones. The exhaustive search among all subsets of observations is unfeasible due to computational complexity and a sequential approach is preferable even if it can lead to a local optimum.

In such a direction, Fellinghauer [9] use random forests in combination with stability selection to estimate stable conditional independence graphs with an error control mechanism for false positive selection.

Let $\widehat{\underline{\Theta^{(k)}}} = \left(\mathbb{1}_{\hat{\theta}_{ij}^{(k)} \neq 0}\right)_{1 \leqslant i,j \leqslant n}$ be the adjacency matrix computed from the penalized ML based on $(X_i, i \neq k)$.

Following Bach [2] we can propose a consistent estimate of $\underline{\Theta}$ by computing the lasso estimates $\widehat{\underline{\Theta^{(k)}}}$ of jackknifed samples, $k = 1, \ldots, m$ and intersect them to define a Bolasso estimate of $\underline{\Theta}$ as $\widehat{\underline{\Theta_b}} = \bigcap_{k=1}^{m} \widehat{\underline{\Theta^{(k)}}}$. Practically it means that an entry of $\widehat{\underline{\Theta_b}}$ is non null if and only if there is no jackknifed sample such that this entry is null.

Construction of a Bolasso graph is not a way to define a stable graph. The Bolasso graph is very closed to the graph estimated on the full dataset except for the link between IGFBP4 and MAPT (see Fig. 4). But this link is present in most of the jackknifed graphs.

We also applied Meinshausen and Bühlmann stability procedure [19] to our dataset. We generated 1000 datasets with 66 observations sampled from the original dataset and estimated the graph with the graphical lasso [10]. For each edge

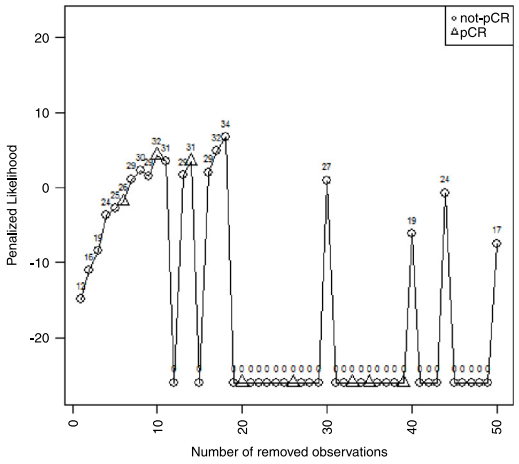

Fig. 5. Penalized likelihood after removing the i highest influential observations. The number of edges of the resulting graph is indicated above the corresponding points.

we can compute a probability of presence, π as the proportion of graph with this given edges. For $\pi = 1/2$ we obtain exactly the penalized maximum likelihood graph and for $\pi = 0.6$, we obtain exactly the Bolasso graph. Fig. 4 also gives the Meinshausen and Bühlmann graph for $\pi = 0.7$. This gives clue for the most stable structure in this graph. One may notice that Bolasso procedure ca be rephrased as Meinshausen and Bühlmann procedure with sample size equal to $n-1$ and $\pi = 1$ for threshold. Our results suggest that it could be interesting to study the relationships between sample size and threshold in Meinshausen and Bühlmann procedure. It is out of the scope of this article.

An idea more closely related with the indicators introduced in the previous sections of this paper could be to start from G the reference graphical model and iteratively remove the most influential observations. More precisely, by considering the n jackknifed graphical models $(G_{-i})_{i=1,\ldots,n}$ coming from G it is possible to detect the most influential observation, say i_0 and to remove it from the reference graph and then iterate this process starting from an updated reference graph $G = G_{-i_0}$. Note that the regularization parameter has to be recalculated. The loop is performed until a terminal condition is reached, for example equality of two successive reference graphs or a given minimal number of observations.

4.3. Applications to the gene dataset

Fig. 5 presents the penalized likelihood along the sequence of $\widehat{\Theta^{-i}}$ $(i = 1, \ldots, 30)$. As expected, removing influential observations leads to an increase of the likelihood. One may notice that after removing 12 observations the graph is empty and one can conclude that there is not enough observations to infer a graph. Meinshausen and Bühlmann [19] observed the same pattern.

We observed also that most of the influential observations are coming from not-PCR group. It sounds logical since this group is intrinsically heterogeneous since the cause of residual disease can be very diverse.

The number of edges of the graph obtained after removing the i highest influential observations is given above the ith point of the Fig. 5. Clearly the number of edges increase as the number of observations decreased. One may notice that, due to the form of the Lasso penalty, the penalized increases as soon as the number of edges increases. From this graph it seems difficult to infer a stable graph and we should be probably very cautious about biological conclusions. This result can be seen as very negative but the idea behind the original experiment is the existence of a biological signature for cancer. The hope of specific biomarkers for cancer is probably lower today than ten years ago.

Acknowledgments

We thank the AE and the two reviewers for very helpful and thorough comments on previous versions of this paper.

References

[1] C. Ambroise, J. Chiquet, C. Matias, Inferring sparse Gaussian graphical models with latent structure, Electron. J. Stat. 3 (2009) 205–238.
[2] F. Bach, Bolasso: model consistent lasso estimation through the bootstrap, in: Proc. 25th Int. Conf. Machine Learning, Association for Computing Machinery, New York, 2008, pp. 33–40.
[3] I. Ben-Gal, Bayesian networks, in: F. Ruggeri, F. Faltin, R. Kenett (Eds.), Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, 2007.
[4] P. Bertail, E. Gautherat, H. Harari-Kermadec, Empirical φ-divergence minimizers for hadamard differentiable functionals, in: M.G. Akritas, D. Politis, S.N. Lahiri (Eds.), Topics in Non-Parametric Statistics, Springer, 2014.
[5] C. Cornalba, R.S. Kenett, P. Giudici, Sensitivity analysis of Bayesian networks with stochastic emulators, in: ENBIS-DEINDE 2007 Conf. on Computer Experiments versus Physical Experiments, University of Turin, 11 April 2007, 2007.

[6] C. Croux, Limit behavior of the empirical influence function of the median, Statist. Probab. Lett. 37 (1998) 331–340.
[7] G. Csardi, T. Nepusz, The igraph software package for complex network research, Int. J. Complex Syst. 1695 (5) (2006).
[8] A. Cuevas, J. Romo, On the estimation of the influence curve, Canad. J. Statist. 23 (1995) 1–9.
[9] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. von Rhein, J.D. Reinhardt, Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables, Comput. Statist. Data Anal. 64 (2013) 132–152.
[10] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical Lasso, Biostatistics 9 (2008) 432–441.
[11] R.D. Gill, Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part. 1), Scand. J. Statist. 16 (1989) 97–128.
[12] C. Giraud, S. Huet, N. Verzelen, Graph selection with GGMselect, SAGMB 11 (3) (2012) 1544–6115.
[13] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, New York, 2005.
[14] K.R. Hess, K. Anderson, W.F. Symmans, V. Valero, N. Ibrahim, J.A. Mejia, D. Booser, R.L. Theriault, U. Buzdar, P.J. Dempsey, R. Rouzier, N. Sneige, J.S. Ross, T. Vidaurre, H.L. Gomez, G.N. Hortobagyi, L. Pustzai, Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer, J. Clin. Oncol. 24 (26) (2006) 4236–4244.
[15] S.L. Lauritzen, Graphical Models, Oxford University Press, 1996.
[16] H. Liu, K. Roeder, L. Wasserman, Stability approach to regularization selection (StARS) for high dimensional graphical models, Adv. Neural Inf. Process. Syst. 23 (2010).
[17] R.G. Miller, The jackknife—a review, Biometrika 61 (1974) 1–15.
[18] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Statist. 34 (2006) 1436–1462.
[19] N. Meinshausen, P. Bühlmann, Stability selection (with discussion), J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (2010) 417–473.
[20] N. Meinshausen, B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, Ann. Statist. (2009) 246–270.
[21] R. Natowicz, R. Incitti, E.G. Horta, B. Charles, P. Guinot, K. Yan, C. Coutant, F. André, R. Pusztai, L. Rouzier, Prediction of the outcome of a preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete response, BMC Bioinformatics 9 (149) (2008).
[22] A.W. Van der Vaart, Asymptotic Statistics, Cambridge University Press, 2000.
[23] D. Vogel, D. Tyler, Robust estimators for nondecomposable elliptical graphical models, Biometrika 101 (4) (2014) 865–882.
[24] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
[25] T. Zhao, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, The huge package for high-dimensional undirected graph estimation in R, J. Mach. Learn. Res. 13 (1) (2012) 1059–1062.