Detecting Influential Observations for a Graphical Model

Avner Bar-Hen*
Laboratoire MAP5, Université Paris Descartes, Paris, France Avner.Bar-Hen@mi.parisdescartes.fr

Jean-Michel Poggi
Laboratoire de Mathématiques, Université Paris Sud, Orsay, France and Université Paris Descartes, Paris, France Jean-Michel.Poggi@math.u-psud.fr

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 and obtain distributional results. An application to a gene expression dataset illustrates the proposals and an investigation of clustering with respect to influence is finally presented.

Keywords: Graphical model; Influence; Jackknife.

1. Introduction

Graphical models allow to 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 Lauritzen (1996)) or the more specific one called Bayesian networks (see Ben-Gal (2007)).

Such graphical models have descriptive qualities since they can represent a graph of knowledge about relationships between the variables of interest to model a domain or a problem. In addition, from the computational viewpoint 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. From the applied side, they capture knowledge from multiple experts and experience from knowledge and data. 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. The question of measuring influence of observations on the results obtained with a graphical models is of interest. A key tool is such a direction can be the use of an influence measure which is a classical diagnostic method to measure 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 various criterions to measure the sensitivity of the graphical model using jackknife network. More precisely, to compute the influence of one observation we compare the network based on all observations except the concerned observation with the reference network based on all observations.

An application to gene expression data set is carried out. This dataset provided by Hess et al. (2006) concerns 133 patients with stage I-III breast cancer. 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 contexts.
This dataset allows us to explore some clustering issues using influence-based tools introduced in the first part of the paper.

2. Model and Dataset

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: $\mathrm{cor}(X_a, X_b|\{X_k, k \in \Gamma \setminus \{a, b\}\}) = 0$ 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}. The log-likelihood for μ and $\Theta = \Sigma^{-1}$ based on a random sample $X_1, \ldots, X_n$ of X is

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

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 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

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

over the set of positive definite matrices. Here $t \geq 0$ is the tuning parameter. When $t = \infty$, the solution to 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. 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:

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

where $\lambda \geq 0$ being the tuning parameter, and where $S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$ is the empirical covariance matrix. Then the non-null entries of $\hat{\Theta} = \arg\max \ell_\lambda^S(\Theta)$, the ML estimate of the concentration matrix Σ^{-1}, define the edges of the estimated graphical model.

A gene expression data set provided by Hess et al. (2006) 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. (2006) and Natowicz et al. (2008) developed and tested a multigene predictor for treatment response on this data set. 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. This data set was already considered by Ambroise et al. (2009) 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. (2012).

3. Influence measures

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

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

The IF describes the effect of an infinitesimal contamination at point x_i on the estimator, standardized by the mass of the contamination. The IF 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 IF.
There is a strong connection between empirical influence function and jackknife (see Miller 1974):

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

with $F_{n-1}^{(i)} = \frac{1}{n-1}\sum_{j \neq i} \delta_{x_j}$. Note that $F_n = \frac{n-1}{n}F_{n-1}^{(i)} + \frac{1}{n}\delta_{x_i}$. Under mild assumption, the empirical influence function $IC_{T,F_n}(x_i)$ is a consistent estimator of the influence function $IC_{T,F}(x_i)$.

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:

$$S_{-j} = \frac{n}{n-1}S - \frac{2}{n}(X_j - \overline{X}_{-j})(X_j - \overline{X}_{-j})' \tag{5}$$

Therefore the L_1-penalized maximum likelihood using S_{-j} can be expressed in terms of S:

$$\log\det\Theta - \frac{n}{n-1}\mathrm{tr}(\Theta S) - \frac{1}{n}(x_j - \overline{x}_{-j})'\Theta(x_j - \overline{x}_{-j}) - \lambda||\Theta||_1$$

where the maximum is to be taken over the set of positive definite matrices. In other words:

$$\ell_\lambda^S(\Theta) = \log\det\Theta - \mathrm{tr}(\Theta S) - \lambda||\Theta||_1 \ ; \ \lambda \geq 0 \tag{6}$$

then denoting by

$$\hat{\Theta} = \arg\max \ell_\lambda^S(\Theta)$$

the MLE of Σ^{-1}, the estimated concentration matrix based on X_i, $i = 1, \ldots, n$ and

$$\widehat{\Theta_{-j}} = \arg\max \ell_\lambda^{S_{-j}}(\Theta)$$

the MLE of Σ^{-1} based on X_i, $i \neq j$, the jackknife dataset and considering some kind of adjacency matrix, the following matrix of 0's and 1's: $\underline{\Theta} = \left(1_{\theta_{ij} \neq 0}\right)_{1 \leq i,j \leq n}$.
The first influence measure $I_1(.)$ can be defined as the number of edges affected by the removing observation j:

$$I_1(j) = \frac{1}{2}||\underline{\hat{\Theta}} - \underline{\widehat{\Theta_{-j}}}||_0 \tag{7}$$

Since the graph is undirected, the adjacency matrix is symmetric and the factor 1/2 is necessary.
There are strong links between jackknife and likelihood (influence function as derivative of the statistic) and as it can be seen from the previous derivation of S_{-j} from S, the L_1-penalized log-likelihood of S_{-j} can be expressed in terms of S:

$$\ell_\lambda^{S_{-j}}(\Theta) = \log\det\Theta - \frac{n}{n-1}\text{tr}(\Theta S) - \frac{1}{n}(x_j - \overline{x}_{-j})'\Theta(x_j - \overline{x}_{-j}) - \lambda||\Theta||_1 \tag{8}$$

where the max is taken over the set of positive definite matrices. Then:

$$\ell_\lambda^{S_{-j}}(\Theta) = \ell_\lambda^{S}(\Theta) - \frac{1}{n}(x_j - \overline{x}_{-j})'\Theta(x_j - \overline{x}_{-j}) \tag{9}$$

Note that the effect is to add a L_2 term that taking into account the contribution of x_j to the penalized likelihood. So, a natural definition of influence could be given by $(x_j - \overline{x}_{-j})'\hat{\Theta}(x_j - \overline{x}_{-j})$.
Let $I_2(.)$ be the difference of the likelihoods induced by the removing of one observation:

$$I_2(j) = \ell_\lambda^{S}(\hat{\Theta}) - \ell_\lambda^{S_{-j}}(\widehat{\Theta_{-j}}) = \frac{1}{n}(x_j - \overline{x}_{-j})'\hat{\Theta}(x_j - \overline{x}_{-j}) \tag{10}$$

Various authors studied the asymptotic distribution of the empirical influence function (see Gill (1989), and Cuevas, Romo (1995) for example). They obtain that the asymptotic distribution of $\sqrt{n}\left(I(F_n) - I(F)\right)$ is the same that of $\sqrt{n}\int IC_{I,F}(x)\, F_n(dx) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_{I,F}(x_i)$ which is, by Central Limit Theorem (CLT), asymptotically distributed as a normal law with zero mean and variance $\sigma^2 = \int IC_{I,F}^2(x)\, F(dz)$.
To get this, we need to impose on I more than the mere existence of $IC_{I,F}(x)$. We refer to Hampel (1988) for details and we will work with Hadamard differential.
We obtain a distributional result for I_2. Indeed, it is well known that likelihood and empirical likelihood are Hadamard differential (see Bertail Gautherat (2003) for example) therefore the asymptotic distribution of I_2 is $\sqrt{n}\left(I_2(F_n) - I_2(F)\right) \sim \mathcal{N}(0, \sigma^2)$.

4. Influence in action on the gene expression data set

We computed the lasso estimate of the inverse covariance matrix $\hat{\Theta}$ for the whole gene expression dataset as well as the 133 $\widehat{\Theta_{-j}}$ obtained by removing one observation. And we observe three different aspects. First, the distributions of I_1 are compared by distinguishing the two classes. Except one extreme observation, there is no notable difference between the two distributions. Second, inspecting the values of I_2, the difference between the penalized maximum likelihood computed on the whole set of data and the penalized maximum likelihood computed after removing observation j leads to the conclusion that while most of the observations lead to a moderate variation of the likelihood when removed, few observations lead to a strong perturbation of the maximum likelihood. Third, looking at the relationships between perturbation of the matrix and perturbation of the likelihood when removing one observation, that is $I2$ versus $I1$, leads to conclude that, at least on this example, the fluctuations of maximum likelihood of concentration matrix $I_2(j)$ is not enough to infer stability of adjacency matrix $I_1(j)$.

Let us now look at the dataset using the two groups and looking at the notion of influence with some clustering-like perspective. Let us begin by a remark about the two groups. Recall that the patient response to the treatment is classified as either a pathologic complete response (pCR) 34 individuals or a residual disease (not-pCR) 99 individuals. Inspecting the estimated graph structures for the full dataset on the left, the pCR class and the not-pCR class, leads to the following main conclusion: the structure identified from the pCR class is rich and exhibits some complex structure of conditional dependence, especially when it is compared to the graphs identified from the not-pCR class and (consequently) from the full dataset. This comes from the fact that clearly the pCR class is "really" a class and the structure is meaningful. At the contrary, the not-pCR patients are not at all a class but only the complement of a class and is too heterogeneous.

Then, a natural question is how to assess the groups with respect to an observation? How to inspect using influence measures which class is the less affected by removing or adding an observation? We can use the following idea: quantifying the influence to a given observation i by removing or adding this observation to each of the class of observation and then allocate the observation to the less affected class. More precisely starting from the two classes: pCR/not-pCR and the two associated adjacency matrices $\underline{\Theta^{(1)}}$ and $\underline{\Theta^{(2)}}$, we can note $\underline{\Theta^{(k\vee i)}} = \underline{\Theta^{(k)}}$ if the observation i is from class k and $\underline{\Theta^{(k\vee i)}}$ is the adjacency matrix computed from the set of individuals of class k + individual i. Then we can define by $I_1^k(i)$ the number of edges of $\underline{\Theta^{(k\vee i)}}$ affected by removing of observation i ($k = 1, 2$):

$$I_1^k(i) = \frac{1}{2} || \underline{\widehat{\Theta^{(k\vee i)}}} - \underline{\widehat{\Theta^{(k\vee i)}_{-i}}} ||_0$$

Finally, for each i we can compute $\arg\min_k I_1^k(i)$ and allocate accordingly the observation to the less affected class. Comparing the estimated graph structures the pCR and the not-pCR classes with the graphs obtained from the two groups obtained after one iteration leads to focus on the meaningful comparison between the graph of the class pcR and the one of the corresponding cluster and exhibits the large impact of reallocations on identified structures.

At this stage, it is natural to consider the reallocation sketched in the previous section as a seminal idea for a clustering procedure driven by such influence ideas. We close the paper by some ideas to define a class centroid and to examine stability with respect to the starting point.

References

Ambroise, C., Chiquet, J., and Matias, C. (2009). *Inferring sparse Gaussian graphical models with latent structure.* Electron. J. Stat., 3:205–238.

Bertail P., Gautherat, E. (2013). Generalized empirical likelihood for Hadamard diffferentaible functionals. In: Topics in Non-Parametric Statistics, M.G Akritas, D. Politis, S.N. Lahiri. (Eds.) Springer.

Ben-Gal, I. (2007). *Bayesian networks.* In: Ruggeri F., Faltin F. & Kenett R. (Eds.), Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons.

Campbell, N.A. (1978). *The influence function as an aid in outlier detection in discriminant analysis.*, Appl. Statist., 27, 251–258.

Critchley, F. and Vitiello, C. (1991). *The influence of observations on misclassification probability estimates in linear discriminant analysis.*, Biometrika, 78, 677–690.

Croux, C. and Joossens, K. (2005). *Influence of observations on the misclassification probability in quadratic discriminant analysis.*, Journal of Multivariate Analysis, 96(2), 384–403.

Croux, C., Filzmoser, P., and Joossens, K. (2008). *Classification Efficiencies for Robust Linear Discriminant Analysis*, Statistica Sinica, 18(2), 581–599.

Croux, C., Haesbroeck, G., and Joossens, K. (2008). *Logistic Discrimination using Robust Estimators: an influence function approach*, The Canadian Journal of Statistics, 36(1), 157–174.

Csardi, G., Nepusz, T. (2006). *The igraph software package for complex network research.* InterJournal, Complex Systems, 1695(5).

Cuevas A., Romo J. (1995). *On the estimation of the influence curve.* The Canadian Journal of Statistics, vol. 23, 1–9.

Fellinghauer, B., Bühlmann, P., Ryffel, M., von Rhein, M., Reinhardt, J.D. (2013). *Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables.* Computational Statistics & Data Analysis, vol. 64, 132–152.

Friedman, J., Hastie, T. and Tibshirani R. (2008). *Sparse inverse covariance estimation with the graphical Lasso.* Biostatistics, 9:432–441.

Giraud, C. , Huet, S. and Verzelen, N. (2012). *Graph selection with GGMselect.* SAGMB, Vol. 11 (3) 1544–6115.

Gill R.D. (1989), *Non- and semi-parametric maximum likelihood estimators and the von Mises method (part. 1).* Scand. J. Statist., 16, 97–128).

Hampel, F. R. (1988). *The influence curve and its role in robust estimation.* J. Amer. Statist. Assoc., 69(346):383–393.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (2005), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York,

Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, U., Dempsey, P.J., Rouzier, R., Sneige, N., Ross, J.S., Vidaurre, T., Gomez, H.L., Hortobagyi, G.N., and Pustzai, L. (2006). *Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.* Journal of Clinical Oncology, 24(26):4236-4244.

Huber, P. J. (1981). *Robust Statistics*, Wiley & Sons.

Lauritzen, S. L. (1996). *Graphical models.* Oxford University Press.

Meinshausen N., Bühlmann P. (2006). *High-dimensional graphs and variable selection with the Lasso.* Annals of Statistics, 34:1436–1462.

Miller, R. G. (1974). *The jackknife - a review.* Biometrika, 61, 1–15.

Natowicz, R., Incitti, R., Horta, E.G., Charles, B., Guinot, P., Yan, K., Coutant, C., André F., Pusztai, R., and Rouzier, L. (2008). *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).