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

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: $\mathrm{cor}(X_a, X_b | \{X_k, k \in \Gamma \setminus \{a, b\}\}) = 0$ corresponds to a zero entry in $\Theta = \Sigma^{-1}$.