International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

UNCORRECTED PROOF

An empirical comparison of popular structure learning algorithms with a view to gene network inference[☆]

Vera Djordjilović[a,*], Monica Chiogna[a], Jiří Vomlel[b]

[a] Department of Statistical Sciences, University of Padova, Italy
[b] Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

ARTICLE INFO

Article history:
Received 28 March 2016
Received in revised form 10 October 2016
Accepted 20 December 2016
Available online xxxx

Keywords:
Bayesian networks
Structure learning
Reverse engineering
DAGs
Gene networks
Prediction

ABSTRACT

In this work, we study the performance of different structure learning algorithms in the context of inferring gene networks from transcription data. We consider representatives of different structure learning approaches, some of which perform unrestricted searches, such as the PC algorithm and the Gobnilp method, and some of which introduce prior information on the structure, such as the K2 algorithm. Competing methods are evaluated both in terms of their predictive accuracy and their ability to reconstruct the true underlying network. A real data application based on an experiment performed by the University of Padova is also considered.

© 2016 Published by Elsevier Inc.

1. Introduction

Genes and proteins do not act in isolation – it is the interactions between them that underlie all major functions of a living cell, from cell differentiation and signal transduction to metabolic processes and cell division. Better understanding of gene networks is one of the central aims of systems biology. Initial approaches to modeling regulatory mechanisms relied on systems of differential equations [5], which are, although very refined, unfortunately limited to small systems about which we already have a hypothesized theory. However, the invention of the technology for measuring abundance of gene transcripts (gene expression), that serve as a proxy for protein abundance, prompted interest in reconstructing the regulatory network from observational data. The idea is to reason backwards: deduce the structure of a complex system from observations of its behavior. This problem has received much attention in the computational biology literature, resulting in a plethora of different models and methods – we refer the interested reader to [2,12,16] for a comprehensive review of the field. Here, we focus on Bayesian networks, a special class of probabilistic graphical models.

A Bayesian network is a statistical model consisting of a Acyclic Directed Graph (DAG) and a family $\mathcal{F}$ of distributions over a set of variables of interest. The graphical structure consists of a set of nodes V and a set of directed edges E. The nodes represent random variables, while edges imply the absence of conditional independence. If there is a directed edge from u to v, we say that u is a parent of v, and denote by $\mathrm{pa}(v)$ a set of parents of v. If we denote by $\boldsymbol{X} = (X_1, X_2, \dots, X_p)^\top$ the variables of the model, the structure of a DAG is associated with their joint distribution through a factorization property

---

[☆] This work was supported by the Czech Science Foundation through projects 13-20012S and 16-12010S.
[*] Corresponding author.
E-mail address: djordjilovic@stat.unipd.it (V. Djordjilović).

http://dx.doi.org/10.1016/j.ijar.2016.12.012
0888-613X/© 2016 Published by Elsevier Inc.

Please cite this article in press as: V. Djordjilović et al., An empirical comparison of popular structure learning algorithms with a view to gene network inference, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.12.012

$$f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} f\left[x_i \mid \mathrm{pa}(x_i)\right],$$

where f stands for a generic probability distribution function. When X_i, $i = 1, 2, \ldots, p$ are discrete, we usually assume that all conditional distributions on the right hand side are multinomial, giving rise to a joint multivariate multinomial distribution. When they are continuous, we usually assume that all conditional distributions on the right hand side are normal, giving rise to a joint multivariate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ represents the mean vector and $\boldsymbol{\Sigma} = [\sigma_{rs}]$, $r, s = 1, \ldots, p$, is the covariance matrix encoding the structure of the network. In fact, the model can be equivalently represented in the recursive form

$$\boldsymbol{X} = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ is the base level, $\boldsymbol{B} = [b_{rs}]$, $r, s = 1, \ldots p$ is a matrix of regression coefficients and $\boldsymbol{\epsilon} \sim \mathrm{N}_p(0, \mathrm{diag}\left(\theta_1, \ldots, \theta_p\right))$ is the random disturbance. If variables are topologically ordered with respect to the DAG, $\boldsymbol{B}$ will be strictly upper triangular and there is a simple relation between the two parameterizations being $\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\mathrm{diag}\left(\theta_1, \ldots, \theta_p\right)[(\boldsymbol{I} - \boldsymbol{B})^{-1}]^\mathsf{T}$, for more details see [24].

The problem of learning the structure of a Bayesian network from realizations of the random vector $\boldsymbol{X}$ is conceptually simple but computationally very complex. Many algorithms, with optimal asymptotic properties, have been proposed [see for instance [6,28]]. The problem encountered when applying these algorithms to gene networks is the small sample size – very often the number of considered genes p exceeds the number of statistical units n. This typical property of biological datasets poses serious limitations for structure learning, explored in detail in [18]. To attenuate this problem, several authors exploited the expected sparsity of biological networks [4,26]. Another possible remedy is to use other sources of information and include them in the learning process. This seems reasonable, since technological advances seen in the last two decades drastically reduced experimental costs and made measurements of biological activity more readily available.

Much of the experimentally obtained knowledge is stored in online public databases. One instance is represented by pathway diagrams, which are elaborate diagrams featuring genes, proteins and other small molecules, showing how they work together to achieve a particular biological effect. From a technical point of view, they are networks and can be represented through a graph where genes and their connections are, respectively, nodes and edges. We will study the effect of including some of the information they carry into the learning process. More specifically, for a set of genes of interest we will specify a topological ordering according to the pathway information, and then pass this ordering to the algorithms that use prior information.

In this empirical comparison, we consider representatives of different structure learning approaches, such as the constraint based PC algorithm [28], the exact Gobnilp method [8] and the score based K2 algorithm [6]. The software used is a combination of tools available in Hugin [21], Gobnilp [8] and R [23]. We perform an extensive simulation study, considering two data generating mechanisms, with the aim of verifying whether the approaches that include prior information, such as K2, perform better than those that rely on data only. In addition to simulated data, we also consider real data from the *Drosophila melanogaster* experiment. In this experiment, performed by the University of Padova [11], researchers measured the expression of genes participating in the WNT signaling pathway in a fruit fly.

The outline of the paper is as follows. In Section 2, we introduce algorithms considered in this work, we propose the data driven categorization procedure for continuous gene expression measurements, and describe the method we used to evaluate the performance of algorithms. Details and results of the simulation studies are given in Section 3, while the evaluation on the real dataset is given in Section 4. Conclusions, some limitations of the present work and future perspectives are given in Section 5.

2. Structure learning algorithms

Structure learning algorithms can be roughly divided into two groups: (1) the score based approaches that search the space of possible structures to find the one that maximizes the score (reflecting the goodness of fit and model parsimony), and (2) the constraint based approaches, that test a large number of conditional independencies between variables and then find the structure that best fits the set of inferred relations. Computational complexity of both approaches is prohibitive for all but small problems, and so some strategy for restricting the search space and reducing the number of performed tests is usually adopted. In that case, one does not have a guarantee that the inferred structure will be globally optimal, but must rely on proven asymptotic validity. Recently, a number of approaches to structure learning that do not rely on a search strategy, but explore the entire space of DAGs, have been proposed. Two major directions are dynamic programming and integer linear programming. For the former, see for instance [20]. We consider a representative of the latter: the approach introduced in [7], where the problem of structure learning is translated into an optimization problem with a linear objective function and a set of linear constraints (as well as integrality constraints on the variables).

As already stated, in this empirical study we consider representatives of different structure learning strategies: a number of variants of the PC algorithm [28], the K2 algorithm [6] and the exact Gobnilp method [8]. Of the examined approaches, the K2 algorithm and all modifications of the K2 algorithm here considered, include the prior information. The prior information is in the form of the topological ordering of the studied genes. In the simulation study, we specify the topological ordering according to the true underlying graph. In the real study, we relied on public databases of biological knowledge. In particular,

UNCORRECTED PROOF

Please cite this article in press as: V. Djordjilović et al., An empirical comparison of popular structure learning algorithms with a view to gene network inference, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.12.012

Download English Version:

https://daneshyari.com/en/article/4945282

Download Persian Version:

https://daneshyari.com/article/4945282

Daneshyari.com