Pattern Recognition Letters 49 (2014) 207–213

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model ☆

Yazhuo Liu, José L. Zayas-Castro, Peter Fabri, Shuai Huang *

University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33613, USA

ARTICLE INFO

Article history:
Received 24 February 2014
Available online 1 August 2014

Keywords:
High-dimensional network learning
Sparse regression
Decision tree model
Nonlinear interactions

ABSTRACT

Network models have been widely used in many domains to characterize relationships between physical entities. Although extensive research efforts have been conducted for learning networks from data, many of them were developed for learning networks with linear relationships. As both linear and nonlinear relationships may appear in many applications, in this paper, we developed a novel graphical model, the sparse tree-embedded graphical model (STGM), which is able to uncover both linear and nonlinear relationships from a large number of variables. We further proposed an efficient regression-based algorithm for learning the STGM from data. We conducted simulation studies that demonstrated the superiority of the STGM over other network learning methods and applied the STGM on a real-world application that demonstrated its efficacy on discovering interesting nonlinear relationships in practice.

1. Introduction

Network models have been widely used in many domains to characterize relationships between physical entities. For example, gene association networks have been used to model how different genes interact in a biological process [11]. Brain connectivity networks have been used to model how different brain regions interact to jointly deliver a brain function such as cognition and emotion [20]. Although the networks are not readily measureable in many applications, recent advancement of sensing technologies have risen the possibility of leaning these networks from the rich amounts of sensing data, such as gene micro-arrays and brain images for the aforementioned networks, respectively.

Extensive research efforts have been conducted for learning networks from data. Many of them focused on one particular type of network model that is called the Gaussian Graphical Model (GGM). A GGM consists of nodes that are random variables following a multivariate normal distribution and undirected arcs that indicate linear relationships between variables. It has been revealed by Dempster [4] that learning a GGM is equivalent to estimating the inverse covariance (IC) of the data, because the undirected arcs in a GGM correspond to nonzero entries in the IC matrix of the data. Existing methods for learning a GGM can be broadly categorized as hypothesis-testing-based methods, likelihood-based methods and regression-based methods.

The hypothesis-testing-based methods employ hypothesis testing techniques to test for each entry of the IC matrix [5,6,38,44]. As the number of entries of an IC matrix grows rapidly with respect to the number of nodes, it is difficult to control the overall type-I error since a large number of hypothesis testing will be conducted. As a remedy, the likelihood-based methods were proposed to identify the zero entries in the IC matrix simultaneously. It takes the advantage of the assumption that the random variables should follow a multivariate normal distribution in a GGM. Penalized maximum likelihood approaches were proposed in Frideman et al., Huang et al., Ravikumar et al., Yuan and Lin [9,20,34,45] that imposed penalties on the entries in the IC matrix, forcing many insignificant entries being zero. Efficient algorithms were proposed by Frideman et al. [9,10] and Sun et al. [40] to implement the penalized maximum likelihood methods, particularly, for high-dimensional problems. Some other methods were proposed, such as a method based on threshold gradient descent regularization developed by Li and Gui [24] and a method for overcoming the ill-conditioned problem of the sample covariance matrix by Schafer and Strimmer [36]. In addition, there are methods dealing with the situations when variables have a natural ordering [2,23]. On the other hand, regression-based methods use regression methods for detecting the network structure. For example, Meinshausen and Buhlmann [29] developed a variable-by-variable approach that used lasso regression to identify the neighborhood for each node in the network. Schafer and Strimmer [36] also developed a joint sparse regression model,

☆ This paper has been recommended for acceptance by Egon L. van den Broek.

* Corresponding author. Tel.: +1 813 974 2090; fax: +1 813 974 5953.
E-mail address: shuaihuang@usf.edu (S. Huang).

http://dx.doi.org/10.1016/j.patrec.2014.07.013
0167-8655/© 2014 Elsevier B.V. All rights reserved.

which simultaneously performs neighborhood selection for all variables. Peng et al. [31] developed a sparse regression technique called SPACE, which is particularly useful in identifying hubs in gene association networks. Friedman et al. [13] also investigated the use of lasso and group lasso for fast approximations to exact penalized maximum likelihood estimation of GGM. Their method leads to sparse network estimation that is not only sparse in edges but also in nodes. Recently, Hsieh et al. [18] and Hsieh et al [19] have developed very efficient algorithms that can remarkably extend the sparse learning of the IC matrix to millions of variables.

Despite the enormous research effort on learning the networks, most of them only focus on linear relationships between variables. For example, a GGM essentially assumes that, the relationship between a variable with the variables that connect with it can be characterized as a linear regression model However, in many applications, both linear and nonlinear relationships will exist between the variables. For example, a particular problem we are studying is the detection of the clinical association networks, which characterize the associations between multiple clinical conditions. Failing to uncover these clinical associations may hinder clinicians detecting important symptoms, potentially leading to inadequate health care such as inappropriate usage of procedures or insufficient treatments. On the other hand, it is very challenging to identify those clinical associations due to their complicate natures [16].

To tackle the challenge of detecting nonlinear relationships in a network, in this paper, we developed a novel graphical model, the sparse tree-embedded graphical model (STGM), which is able to uncover both linear and nonlinear clinical associations from a large number of variables. While the term "nonlinear association" can take many possible forms, we focused on a particular type of nonlinear associations that can be characterized by tree models. The basic idea of our STGM is integrating regression-based methods with decision tree learning, since decision tree has been demonstrated to be a powerful tool for learning nonlinear interactions between variables, with no additional cost of increasing the model complexity due to its nonparametric nature. We further propose an efficient regression-based algorithm for learning the STGM from data.

The paper is organized as follows. In Section 2, we first put our work in perspective by describing related work in learning the networks. Then, we will describe our proposed method and develop the computational algorithm for implementing this method in Section 3, and report the experimental results of applying the proposed method on both simulated and real-world datasets in Section 4. A conclusion will be given in Section 5.

2. Related work

In this section, we will briefly review the related work in the existing methods for learning networks, particularly, in the regression-based methods since our method falls into this category [21,29,31]. We use $\mathbf{X} = \{X_1, \ldots, X_p\}$ to denote the p random variables under study. A graphical model of $\mathbf{X}$ assigns one node for each X_i and connects two nodes if there is association between them. An example of a graphical model of six random variables is shown in Fig. 1(a). The structure of a graphical model can be characterized by a $p \times p$ adjacency matrix $\mathbf{G}$, with entry $\mathbf{G}_{ij} = 1$ representing an arc between X_i to X_j and $\mathbf{G}_{ij} = 0$ otherwise. E.g., the corresponding adjacency matrix for the graphical model is shown in Fig. 1(b).

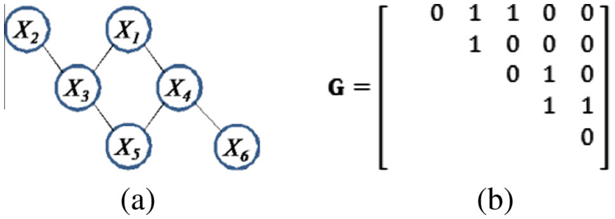

Fig. 1. A graphical model (a) and the corresponding adjacency matrix (b) (only entries at the upper triangle are shown because the matrix is symmetric and the diagonal entries are not used in the graphical model).

The structure learning of the graphical model is equivalent to the identification of the nonzero elements in the adjacency matrix G. Particularly, the regression-based methods [21,29,31] decompose the learning problem into p sub-problems, while each sub-problem concerns the identification of the neighbors of a variable. For example, in GGM, the associations between variable X_i with other variables can be modeled as a linear regression model, such as $X_i = \boldsymbol{\beta}_i^T X_{/i} + \varepsilon_i$, where $X_{/i}$ denotes all the variables except X_i and ε_i denotes the residual term which is modeled as a normal distribution. The regression-based methods repeatedly use some variable selection models for each X_i and identify the non-zero regression coefficients in $\boldsymbol{\beta}_i$ [21,29,31]. The zero regression coefficients in $\boldsymbol{\beta}_i$ correspond to the variables that are not associated with X_i. A general framework for these algorithms is shown in Fig. 2. Here, $f_i(\boldsymbol{\beta}_i)$ could be a loss function that encourages many elements in $\boldsymbol{\beta}_i$ to be zero, i.e., the loss function used in Glasso [9,10]. The regression-based methods can also be applied to other networks rather than GGM. For example, in some Markov graphical models [21] which model discrete variables, the associations between nodes can be modeled as a logistic regression model if the variables are binary, such as $\Pr(X_i = 1) = g(\boldsymbol{\beta}_i^T X_{/i}) + \varepsilon_i$, where g denotes the logit link function and ε_i denotes the residual term which is modeled as a binomial distribution.

The regression-based methods are demonstrated to be computationally efficient and accurate by both theoretical analysis and extensive simulation studies [21,29,31]. However, many of them are limited to the applications where the associations between variables are linear. A few studies have attempted to relax these constraints and extend graphical models to capture nonlinear associations [22,25]. For example, Lafferty et al. [22] proposed two approaches, one made a distributional restriction through the use of copulas as a semiparametric extension of the Gaussian distribution, and another one used kernel density estimation and restricted the underlying graphs to be trees or forests. Apparently, these restrictive assumptions limit their applicability in many real-world cases. This is particularly true in many clinical association studies where the nonlinear associations are usually non-smooth and take a rule-based semantics, while the methods proposed in Lafferty et al., Liu et al. [22,25] restrict the nonlinear associations to be represented as smooth functions.

3. The proposed sparse tree-embedded graphical model (STGM)

3.1. The formulation of the STGM

As mentioned in Section 2, the regression-based methods employ a regression model to characterize the associations

```
Input: sample matrix, $X$; number of variable, $p$;
    regularization parameters, $\{\lambda_i\}_{i=1,2,\ldots,p}$;
For $i = 1,2,\ldots,p$,
    optimize $f_i(\boldsymbol{\beta}_i)$ and get $\boldsymbol{\beta}_i$;
End for
Output: $\boldsymbol{\beta}_i$ for $i = 1,2,\ldots,p$
```

Fig. 2. A general framework for the regression-based methods.