International Journal of Approximate Reasoning 83 (2017) 21–42

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Learning Gaussian graphical models with fractional marginal pseudo-likelihood

Janne Leppä-aho [a,*], Johan Pensar [b], Teemu Roos [a], Jukka Corander [c,d]

[a] HIIT/Department of Computer Science, University of Helsinki, Gustaf Hällströmin katu 2, 00560 Helsinki, Finland
[b] Department of Mathematics and Statistics, Åbo Akademi University, Vänrikinkatu 3, 20500 Turku, Finland
[c] Department of Biostatistics, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway
[d] Department of Mathematics and Statistics, University of Helsinki, Gustaf Hällströmin katu 2, 00560 Helsinki, Finland

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 4 January 2017
Accepted 5 January 2017
Available online 11 January 2017

Keywords:
Approximate likelihood
Fractional Bayes factors
Model selection
Structure learning
Gaussian graphical models

ABSTRACT

We propose a Bayesian approximate inference method for learning the dependence structure of a Gaussian graphical model. Using pseudo-likelihood, we derive an analytical expression to approximate the marginal likelihood for an arbitrary graph structure without invoking any assumptions about decomposability. The majority of the existing methods for learning Gaussian graphical models are either restricted to decomposable graphs or require specification of a tuning parameter that may have a substantial impact on learned structures. By combining a simple sparsity inducing prior for the graph structures with a default reference prior for the model parameters, we obtain a fast and easily applicable scoring function that works well for even high-dimensional data. We demonstrate the favourable performance of our approach by large-scale comparisons against the leading methods for learning non-decomposable Gaussian graphical models. A theoretical justification for our method is provided by showing that it yields a consistent estimator of the graph structure.

1. Introduction

1.1. Bayesian learning of Gaussian graphical models

Gaussian graphical models provide a convenient framework for analysing conditional independence in continuous multivariate systems [1–3]. We consider the problem of learning Gaussian graphical models from data using a Bayesian approach. Most of the Bayesian methods for learning Gaussian graphical models make the assumption about the decomposability of the underlying graph [4–6]. Recently, Fitch et al. [7] investigated how Bayesian methods assuming decomposability perform in model selection when the true underlying model is non-decomposable. Bayesian methods that do not assume decomposability have been considered more seldom in the literature, and in particular not in the high-dimensional case [8–14].

A widely used frequentist method for learning Gaussian graphical models is the graphical lasso [15,16]. Graphical lasso (`glasso`) uses l_1-penalized Gaussian log-likelihood to estimate the inverse covariance matrices and does not rely on the assumption of decomposability. Other approaches include a neighbourhood selection (`NBS`) method by Meinshausen and Bühlmann [17] and Sparse Partial Correlation Estimation method (`space`) by Peng et al. [18]. The `NBS`-method estimates

* Corresponding author.
E-mail address: janne.leppa-aho@helsinki.fi (J. Leppä-aho).

http://dx.doi.org/10.1016/j.ijar.2017.01.001
0888-613X/

the graphical structure by performing independent lasso regressions for each variable to find the estimates for the neighbourhoods whereas `space` imposes an l_1-penalty on an objective function corresponding to an l_2-loss of a regression problem in order to estimate the non-zero partial correlations which correspond to edges in the graphical model.

Assuming decomposability in Bayesian methods has been popular, since it enables derivation of a closed form expression for the marginal likelihood under a conjugate prior. In our approach we bypass this restriction by replacing the true likelihood in the marginal likelihood integral by a pseudo-likelihood. This implies a factorization of the marginal pseudo-likelihood into terms that can be evaluated in closed form by using existing results for the marginal likelihoods of Gaussian directed acyclic graphs. The marginal pseudo-likelihood offers further advantages by allowing efficient search algorithms to be used, such that model optimization becomes realistic for even high-dimensional data.

Dobra et al. [19] considered a similar pseudo-likelihood based approach. These two methods involve similar techniques in the first step where a general dependency network is learned using a Bayesian approach. A dependency network [20] is a collection of conditional distributions for each variable given the others which are all fitted separately. In the general case, this network does not define a proper joint distribution for the variables. Dobra et al. use this dependency network in order to define an ordering for the variables before learning a directed acyclic graphical model over the variables. The found directed graph is then moralized in order to produce an undirected graph. In other words, their method does not consider general non-decomposable graphs.

Marginal pseudo-likelihood has been previously used to learn undirected graphical models with discrete variables in Pensar et al. [21]. Our paper can be seen to generalize the ideas developed there to the continuous domain by introducing the required methodology and providing a formal consistency proof under the multivariate normal assumption. Our method utilizes the fractional Bayes factors based approach of Consonni and La Rocca [22] to cope automatically with the difficulty of setting up prior distributions for the models' parameters.

The rest of the paper is organized as follows. After introducing the notation, we briefly review the results by Consonni and La Rocca that are needed in deriving the expression for the marginal pseudo-likelihood. In Section 3 we state our main result by introducing the fractional marginal pseudo-likelihood. The detailed proof of its consistency for Markov blanket estimation is given in Appendix. A score-based search algorithm adopted from Pensar et al. [21] is presented in order to implement the method in practice. In Section 4 we demonstrate the favourable performance of our method by several numerical experiments involving a comparison against `glasso`, `NBS` and `space`.

1.2. Notations and preliminaries

We will start by reviewing some of the basic concepts related to graphical models and the multivariate normal distribution. For a more comprehensive presentation, see for instance [2] and [3].

Consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes (vertices) and $E \subset V \times V$ is the set of edges. There exists an (undirected) edge between the nodes i and j, if and only if $(i, j) \in E$ and $(j, i) \in E$. Each node of the graph corresponds to a random variable, and together they form a p-dimensional random vector $\boldsymbol{x}$. We will use the terms node and variable interchangeably. Absence of an edge in the graph G is a statement of conditional independence between the corresponding elements of $\boldsymbol{x}$. More in detail, $(i, j), (j, i) \notin E$ if and only if x_i and x_j are conditionally independent given the remaining variables $\boldsymbol{x}_{V\setminus\{i,j\}}$. This condition is usually referred as the pairwise Markov property. We let $mb(j)$ denote the Markov blanket of node j. The Markov blanket is defined as the set containing the neighbouring nodes of j, $mb(j) = \{i \in V \mid (i, j) \in E\}$. The local Markov property states that each variable is conditionally independent of all others given its Markov blanket. An undirected graph G is called decomposable or equivalently chordal if each cycle, whose length is greater or equal than 4, contains a chord. By a cycle, we mean a sequence of nodes such that the subsequent nodes are connected by an edge and the starting node equals the last node in the sequence. The length of a cycle equals the number of edges in the cycle. A chord is an edge between two non-subsequent nodes of the cycle.

We will write $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$ to state that a random vector $\boldsymbol{x}$ follows a p-variate normal distribution with a zero mean and precision matrix $\boldsymbol{\Omega}$. We will denote the covariance matrix by $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$. The precision matrix $\boldsymbol{\Omega}$, and also equivalently $\boldsymbol{\Sigma}$, are always assumed to be symmetric and positive definite.

Given an undirected graph G and a random vector $\boldsymbol{x}$, we define a Gaussian graphical model to be the collection of multivariate normal distributions for $\boldsymbol{x}$ that satisfy the conditional independences implied by the graph G. Hence, a Gaussian graphical model consists of all the distributions $N_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega}_{ij} = 0$ if and only if $(i, j) \notin E$, $i \neq j$. Otherwise, the elements of the inverse covariance matrix can be arbitrary, as long as symmetry and positive definiteness hold.

In contrast to the above undirected model, a Gaussian directed acyclic graphical model is a collection of multivariate normal distributions for $\boldsymbol{x}$, whose independence structure can be represented by some directed acyclic graph (DAG) $D = (V, E)$. When considering directed graphs, we use $pa(j)$ to denote the parent set of the node j. The set $pa(j)$ contains nodes i such that $(i, j) \in E$. That is, there exists a directed edge from i to j. Similar Markov assumptions as those characterizing the dependency structure under undirected models, as described above, hold also for directed models, see, for instance, [3]. For each decomposable undirected graph, we can find a DAG which defines the same conditional independence assertions. In general, the assertions representable by DAGs and undirected graphs are different.