Computational Statistics and Data Analysis 70 (2014) 138–152

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Edge detection in sparse Gaussian graphical models

Shan Luo [a,*], Zehua Chen [b,1]

[a] Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China

[b] Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form 23 July 2013
Accepted 3 September 2013
Available online 20 September 2013

Keywords:
Edge detection
Extended Bayesian information criterion
Graphical model
Selection consistency
Sequential selection

ABSTRACT

In this paper, we consider the problem of detecting edges in a Gaussian graphical model. The problem is equivalent to the identification of non-zero entries of the concentration matrix of a normally distributed random vector. Following the methodology initiated in Meinshausen and Bühlmann (2006), we tackle the problem through regression models where each component of the random vector is regressed on the remaining components. We adapt a method called SLasso cum EBIC (sequential LASSO cum extended Bayesian information criterion) recently developed in Luo and Chen (2011) for feature selection in sparse regression models to suit the special nature of the concentration matrix, and propose two approaches, dubbed SR-SLasso and JR-SLasso, for the identification of non-zero entries of the concentration matrix. Comprehensive numerical studies are conducted to compare the proposed approaches with other available competing methods. The numerical studies demonstrate that the proposed approaches are more accurate than the other methods for the identification of non-zero entries of the concentration matrix.

1. Introduction

Ever since the publication of the paper by Dempster (1972), there has been a considerable literature on the identification of non-zero entries in a concentration matrix, the inverse of the covariance matrix of a random vector. A non-zero entry in the concentration matrix corresponds to two components that have a non-zero partial correlation, i.e., they are correlated conditioning on all the other components. The partial correlation is an important aspect of the inter-relationship among the components of a network. The investigation of such inter-relationships is of great scientific importance. For example, in studies of complex diseases, the exploration of the inter-relationship among the responsible genes is crucial for the understanding of the disease pathologies.

A concentration matrix is closely related to an undirected graphical model. An undirected graphical model is specified by a vertex set V and an edge set E, and is denoted by $G = (V, E)$. The vertex set V represents a collection of random variables $\{Y_1, \ldots, Y_p\}$. The edge set E describes the inter-relationship among the random variables: there is an edge connecting vertices Y_i and Y_j if they are dependent conditioning on all the remaining variables. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ follows a multivariate normal distribution with concentration matrix $\Omega = (\Omega_{ij})$. Then, there is an edge between Y_i and Y_j if and only if $\Omega_{ij} = \Omega_{ji} \neq 0$. Thus, the detection of edges of G is equivalent to the identification of non-zero entries of Ω. A more general problem is the estimation of Ω. In this paper, we focus only on the detection of non-zero entries of Ω.

There are two major methodologies for the identification and estimation of the non-zero entries of a concentration matrix. The first methodology, which was initiated in Meinshausen and Bühlmann (2006), is based on the relationship between

* Corresponding author. Tel.: +86 21 54743151 2304.
E-mail addresses: sluo2012@gmail.com (S. Luo), stachenz@nus.edu.sg (Z. Chen).
URL: http://www.stat.nus.edu.sg/~stachenz/ (Z. Chen).
[1] Contributing author.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.002

entries of Ω and the coefficients of p regression models where each component of $\boldsymbol{Y}$ is regressed on the remaining $p-1$ components. A non-zero entry of Ω corresponds to a non-zero regression coefficient in the regression models. The detection and estimation of non-zero entries of Ω are then boiled down to the selection and estimation of non-zero coefficients in p regression models. Various methods for sparse high-dimensional regression models have been used to deal with this problem. The p regression models are handled either separately or simultaneously. The Lasso is used in Meinshausen and Bühlmann (2006) for each of the models, and an edge between two vertices is claimed existent if at least one of the associated coefficients in the two related models is estimated non-zero (or, alternatively, if both of the coefficients are estimated non-zero). The Dantzig selector is applied in Yuan (2010) to each of the models, and then a symmetrization step is called to obtain the estimated concentration matrix as the symmetric matrix closest to the estimated coefficient matrix. The scaled Lasso is used in Sun and Zhang (2012), and a similar symmetrization step is applied to get the final estimate of the concentration matrix. Recognizing that the separate treatment of each regression model might lose useful information, a method called sparse partial correlation estimation (SPACE) was proposed in Peng et al. (2009). SPACE treats the p regression models simultaneously by using an essentially weighted Lasso approach. The second methodology is the direct regularization on the entries of Ω based on their profile likelihood function with various penalty functions. The L_1-penalty is imposed on the entries of Ω in Friedman et al. (2008), and the resultant approach is dubbed graphical Lasso (GLasso). A variant of GLasso is considered in Ravikumar et al. (2011). Instead of the L_1-penalty, the SCAD penalty is used in Fan et al. (2009). We refer to this approach as G-Scad. The adaptive Lasso approach applied to the profile likelihood of Ω is studied in Zhou et al. (2009). Regularization with a general penalty function is studied in Lam and Fan (2009). It is worth noting that a method called Clime, which is different methodologically from the above approaches, is proposed in Cai et al. (2011). Clime estimates the concentration matrix by minimizing $\|\Omega\|_1$ subject to $\|\Sigma_n\Omega - I\|_\infty \le \lambda$, where Σ_n is the sample covariance matrix, λ is a regularization parameter, and $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are element-wise L_1 and L_∞ norms, respectively.

The methods mentioned above all have certain nice theoretical properties under different conditions. The salient one is the so-called oracle property. One aspect of the oracle property is that, asymptotically, the non-zero and zero entries of Ω will be estimated as non-zero and zero entries exactly. This aspect is also referred to as selection consistency. The other aspect is that the estimation can achieve the same accuracy as that if the zero entries were known in advance. However, the realization of this oracle property depends on the proper choice of a regularization parameter λ which is involved in all the methods. In practice, the parameter is usually chosen by evaluating a certain model selection criterion at regularly spaced points in a certain range of λ. The determination of the range and the density of the points affects not only the computation amount but also the accuracy of the choice. There are no theoretical results at all on how such a chosen value of λ is related to the theoretical one required for the oracle property. In other words, in practice, the realization of the oracle property is not guaranteed.

Recently, a sequential approach for sparse high-dimensional regression models, called SLasso cum EBIC, was developed in Luo and Chen (2011). This approach solves a sequence of partially penalized least squares problems. At each step, the parameters of the features already selected in earlier steps are not penalized, and the penalty parameter is tuned to the largest while still allowing some penalized parameters to be estimated non-zero. At each step, the least squares model consisting of all the features selected so far is evaluated by a model selection criterion called EBIC. The EBIC serves as a stopping rule. Whenever the EBIC ceases decreasing, the sequential procedure stops. The EBIC is developed in Chen and Chen (2008) especially for sparse high-dimensional regression models. SLasso cum EBIC possesses the property of selection consistency, and its computation is simple and light. Along the line of the first methodology discussed above, it is natural to apply the SLasso cum EBIC method to the p regression models for identifying the non-zero coefficients. In this paper, we adapt the original SLasso cum EBIC procedure to suit the special nature of the concentration matrix Ω. We propose two versions of the adaption. In the first one, we treat the p regression models separately, and the resultant approach is referred to as SR-SLasso (single regression SLasso). In the second one, we treat all the p regression models simultaneously, and the resultant approach is referred to as JR-SLasso (joint regression SLasso). The selection consistency of the SLasso cum EBIC method carries over to both SR-SLasso and JR-SLasso, which is verified in this paper as well. Comprehensive numerical studies are conducted to compare these new approaches with some representative methods mentioned earlier. The numerical studies demonstrate that the new approaches are more accurate than the other methods in identifying the non-zero entries of Ω.

The rest of this paper is organized as follows. In Section 2, we describe the SR-SLasso and JR-SLasso approaches in detail. In Section 3, we establish the selection consistency of the proposed approaches. In Section 4, we report two numerical studies. In the first study, we compare our approaches with the others in nine commonly assumed settings on the concentration matrix. In the second study, we make the comparison based on a microarray expression data set from a breast cancer study. The technical proofs are provided in the Appendix.

2. Methods

For an undirected graph $G = (V, E)$, let V be modeled as the set of the components of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$. We assume that $\boldsymbol{Y}$ follows a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. Without loss of generality, assume that $\boldsymbol{\mu} = 0$. Let $\boldsymbol{Y}_{i^-}$ be the vector obtained from $\boldsymbol{Y}$ by eliminating component Y_i. Denote by $\Sigma_{i^-i^-}$ the variance–covariance matrix of $\boldsymbol{Y}_{i^-}$, by σ_i^2 the variance of Y_i, and by Σ_{ii^-} the covariance vector between Y_i and $\boldsymbol{Y}_{i^-}$. By the theory of multivariate normal distribution, the conditional distribution of Y_i given $\boldsymbol{Y}_{i^-}$ is still normal, with the following conditional mean and conditional variance:

$$\boldsymbol{E}(Y_i|\boldsymbol{Y}_{i^-}) = \Sigma_{ii^-}\Sigma_{i^-i^-}^{-1}\boldsymbol{Y}_{i^-}, \tag{1}$$