ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Review

Recent developments in high dimensional covariance estimation and its related issues, a review

Younghee Hong, Choongrak Kim *

Department of Statistics, Pusan National University, Pusan, 609-735, Republic of Korea

ARTICLE INFO

Article history:
Received 2 March 2018
Accepted 24 April 2018
Available online xxxx

AMS 2000 subject classifications:
primary 62H12
secondary 62H25

Keywords:
Gaussian graphical model
High dimensional data
Laplacian matrix
Lasso
Precision matrix
Tracy–Widom law

ABSTRACT

In this paper we review some of recent developments in high dimensional data analysis, especially in the estimation of covariance and precision matrix, asymptotic results on the eigenstructure in the principal components analysis, and some relevant issues such as test on the equality of two covariance matrices, determination of the number of principal components, and detection of hubs in a complex network.

Contents

* Corresponding author.
E-mail address: crkim@pusan.ac.kr (C. Kim).

https://doi.org/10.1016/j.jkss.2018.04.005
1226-3192/

1. Introduction

Consider a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, where X_i could be, for example, the ith gene expression in microarray data. We are interested in the dependency of $X_i's$, and the dependencies of all the $X_i's$ can be represented by a network. If p is very large, then the networks will be very complicated, and we call it *complex network*. Examples of complex networks are biological networks, the world-wide web, the social networks, etc. We can figure out the complex networks by a graph which represents the relationship between X_i and X_j, $1 \le i \ne j \le p$.

Consider a graph $G = G(V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes (vertices) and E is the set of edges in $V \times V$. Let $a_{ij},\ i, j = 1, \ldots, p$ denote the *adjacency* between two nodes i and j, and $\boldsymbol{A} = (a_{ij})$ is called an adjacency matrix. In many cases a_{ij} takes real values between 0 and 1, however, it is not necessary. A graph is called directed if $a_{ij} \ne a_{ji}$, and called undirected if $a_{ij} = a_{ji}$. Hence, the correlation matrix, for example, is undirected. The degree of the ith node is denoted as $d_i = \sum_{j=1}^{p} a_{ij}$, $\boldsymbol{D} = diag(d_1, \ldots, d_p)$ is the *degree* matrix, and the *Laplacian* matrix is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$. Note that if the given network consists of k separate groups, then $k + 1$ eigenvalues of the Laplacian matrix are zero and all others are positive.

Graph theory has been developed by mathematicians for a long time, and mathematicians are interested in properties of eigenvalues of adjacency matrix; distribution of eigenvalues (Marcenko & Pastur, 1967; Wigner, 1955) and distribution of the largest eigenvalue (Tracy & Widom, 1996), lower and/or upper bound of eigenvalues, relationship between degree and eigenvalues, and so on. There are numerous books on the graph theory, we suggest Mieghem (2010) among others. On the other hand, statisticians paid attention to the graph theory quite recently, and they are mainly interested in graphical models and estimation of adjacency matrix using available observations.

When a graph is given, our primal interest is investigating the structure of a graph. To be more specific, we want to estimate the dependency of X_i and X_j, the clustering structure of a graph, detection of hubs, and theoretical results for the estimators of interest, etc. There are two types of dependency; marginal and conditional dependence between X_i and X_j. The marginal dependency is the correlation between X_i and X_j, however, the conditional dependency is the conditional correlation between X_i and X_j given all the $X_i's$ except X_i and X_j. Therefore, the marginal dependency can be estimated by the estimation of the covariance matrix, and conditional dependency can be estimated by the estimation of the inverse of the covariance matrix, also called the *precision matrix* (Dempster, 1972).

One of apparent characteristics of modern statistical data is high dimensionality. This phenomenon is often called "small n, large p", while most of the classical statistical tools are for "large n, small p" setting. The "small n, large p" problem starts with the advent of microarray data in genomics in the early 2000. Sample size n is just tens or hundreds of patients (or normal people), while the number of variables p is thousands or tens of thousands of genes. Since then lots of studies are focused on the analyses of high dimensional data because the existing methods cannot be directly applicable to the high dimensional data. Especially, estimation of the covariance matrix based on a standard method is not desirable in many respects.

The most important concept in high dimensional inference is sparsity, and there are many ways of imposing sparsity. In covariance and precision matrix estimation, the naive one is banding or tapering (see Section 2.2 for definitions) in the elements of covariance or precision matrix itself. The other one is using the penalty function like lasso. Another approach is using the principal component analysis (PCA) which is widely used as a dimension reduction tool in the multivariate data. In fact, the sparse PCA is based on the spiked covariance model by assuming that only few eigenvalues of the covariance matrix are much larger than others.

The studies for investigating the structure of a graph, including the estimation of the covariance matrix in the high dimensional setup, are very active and are published a lot recently in major statistical journals. Also, major journals treated high dimensional inference as special issues; For example, *High Dimensional Inference and Random Matrices* in the Annals of Statistics, Vol. 36, No. 6 (2008), and *Special Issue on Large Dimensional Models* in the Econometric Journal, Vol. 19, Issue 1 (2016).

In this paper, we review developments in statistical methods for the analysis of complex networks especially in high dimensional setup. We tried to cover relevant works as many as possible, however, we are limited to provide a selective review on complex networks and covariance estimation because all the relevant studies are too vast to be covered. For the high dimensional covariance estimation, Fan, Liao, and Liu (2016) and Pourahmadi (2013) made good reviews among others.

This paper is different from the previously published review papers in two aspects. First, we tried to minimize presenting theoretical results as possible because this paper is intended to introduce basic concepts rather than theoretical results in high dimensional inference. Second, we tried to provide global approach to high dimensional inference motivated by complex networks which can be presented by a graph. This paper is organized as follows: In Section 2, two types of dependencies (marginal and conditional) are defined and corresponding estimation problems (covariance and precision matrix). Also, results on testing two covariance matrices are introduced. In Section 3, the Tracy–Widom law and some results on the PCA are reviewed. Also, approaches based on the spiked covariance model and the factor model are reviewed. Further, studies on the determination of the number of principal components and recent statistical developments via the Laplacian matrix are introduced. Finally, an illustrative example showing the difference between the conditional and marginal dependencies is given. Concluding remarks are given in Section 4.

Please cite this article in press as: Hong, Y., Kim, C., Recent developments in high dimensional covariance estimation and its related issues, a review. Journal of the Korean Statistical Society (2018), https://doi.org/10.1016/j.jkss.2018.04.005.