Journal of Multivariate Analysis 150 (2016) 55–74

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data

T. Tony Cai [a], Anru Zhang [b,*]

[a] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, United States
[b] University of Wisconsin-Madison, Madison, WI, United states

ARTICLE INFO

Article history:
Received 3 August 2015
Available online 19 May 2016

AMS subject classifications:
primary 62H12
62J10
secondary 62C20

Keywords:
Adaptive thresholding
Bandable covariance matrix
Generalized sample covariance matrix
Missing data
Optimal rate of convergence
Sparse covariance matrix
Thresholding

ABSTRACT

Missing data occur frequently in a wide range of applications. In this paper, we consider estimation of high-dimensional covariance matrices in the presence of missing observations under a general missing completely at random model in the sense that the missingness is not dependent on the values of the data. Based on incomplete data, estimators for bandable and sparse covariance matrices are proposed and their theoretical and numerical properties are investigated.

Minimax rates of convergence are established under the spectral norm loss and the proposed estimators are shown to be rate-optimal under mild regularity conditions. Simulation studies demonstrate that the estimators perform well numerically. The methods are also illustrated through an application to data from four ovarian cancer studies. The key technical tools developed in this paper are of independent interest and potentially useful for a range of related problems in high-dimensional statistical inference with missing data.

1. Introduction

The problem of missing data arises frequently in a wide range of fields, including biomedical studies, social science, engineering, economics, and computer science. Statistical inference in the presence of missing observations has been well studied in classical statistics. See, e.g., Ibrahim and Molenberghs [18] for a review of missing data methods in longitudinal studies and Schafer [26] for literature on handling multivariate data with missing observations. See Little and Rubin [20] and the references therein for a comprehensive treatment of missing data problems.

Missing data also occurs in contemporary high-dimensional inference problems, whose dimension p can be comparable to or even much larger than the sample size n. For example, in large-scale genome-wide association studies (GWAS), it is common for many subjects to have missing values on some genetic markers due to various reasons, including insufficient resolution, image corruption, and experimental error during the laboratory process. Also, different studies may have different volumes of genomic data available by design. For instance, the four genomic ovarian cancer studies discussed in Section 4 have throughput measurements of mRNA gene expression levels, but only one of these also has microRNA measurements (Cancer Genome Atlas Research Network [11], Bonome et al. [4], Tothill et al. [27] and Dressman et al. [15]). Discarding samples with any missingness is highly inefficient and could induce bias due to non-random missingness. It is of significant interest to integrate multiple high-throughput studies of the same disease, not only to boost statistical power

* Corresponding author.
E-mail addresses: tcai@wharton.upenn.edu (T.T. Cai), anruzhang@stat.wisc.edu (A. Zhang).

http://dx.doi.org/10.1016/j.jmva.2016.05.002
0047-259X/

but also to improve the biological interpretability. However, considerable challenges arise when integrating such studies due to missing data.

Although there have been significant recent efforts to develop methodologies and theories for high dimensional data analysis, there is a paucity of methods with theoretical guarantees for statistical inference with missing data in the high-dimensional setting. Under the assumption that the components are missing uniformly and completely at random (*MUCR*), Loh and Wainwright [21] proposed a non-convex optimization approach to high-dimensional linear regression, Lounici [23] introduced a method for estimating a low-rank covariance matrix and Lounici [22] considered sparse principal component analysis. In these papers, theoretical properties of the procedures were analyzed. These methods and theoretical results critically depend on the MUCR assumption.

Covariance structures play a fundamental role in high-dimensional statistics. It is of direct interest in a wide range of applications including genomic data analysis, particularly for hypothesis generation. Knowledge of the covariance structure is critical to many statistical methods, including discriminant analysis, principal component analysis, clustering analysis, and regression analysis. In the high-dimensional setting with complete data, inference on the covariance structure has been actively studied in recent years. See Cai, Ren and Zhou [7] for a survey of recent results on minimax and adaptive estimation of high-dimensional covariance and precision matrices under various structural assumptions. Estimation of high-dimensional covariance matrices in the presence of missing data also has wide applications in biomedical studies, particularly in integrative genomic analysis which holds great potential in providing a global view of genome function (see Hawkins et al. [17]).

In this paper, we consider estimation of high-dimensional covariance matrices in the presence of missing observations under a general missing completely at random (*MCR*) model in the sense that the missingness is not dependent on the values of the data. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n independent copies of a p dimensional random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Instead of observing the complete sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, one observes the sample with missing values, where the observed coordinates of $\mathbf{X}_k$ are indicated by a vector $\mathbf{S}_k \in \{0, 1\}^p$, $k = 1, \ldots, n$. That is,

$$X_{jk} \text{ is observed if } S_{jk} = 1 \text{ and } X_{jk} \text{ is missing if } S_{jk} = 0. \tag{1}$$

Here X_{jk} and S_{jk} are respectively the jth coordinate of the vectors $\mathbf{X}_k$ and $\mathbf{S}_k$. We denote the incomplete sample with missing values by $\mathbf{X}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$. The major goal of the present paper is to estimate $\boldsymbol{\Sigma}$, the covariance matrix of $\mathbf{X}$, with theoretical guarantees based on the incomplete data $\mathbf{X}^*$ in the high-dimensional setting where p can be much larger than n.

This paper focuses on estimation of high-dimensional bandable covariance matrices and sparse covariance matrices in the presence of missing data. These two classes of covariance matrices arise frequently in many applications, including genomics, econometrics, signal processing, temporal and spatial data analyses, and chemometrics. Estimation of these high-dimensional structured covariance matrices have been well studied in the setting of complete data in a number of recent papers, e.g., Bickel and Levina [3,2], Karoui [16], Rothman et al. [24], Cai and Zhou [10], Cai and Liu [5], Cai et al. [6,9] and Cai and Yuan [8]. Given an incomplete sample $\mathbf{X}^*$ with missing values, we introduced a "generalized" sample covariance matrix, which can be viewed as an analog of the usual sample covariance matrix in the case of complete data. For estimation of bandable covariance matrices, where the entries of the matrix decay as they move away from the diagonal, a blockwise tridiagonal estimator is introduced and is shown to be rate-optimal. We then consider estimation of sparse covariance matrices. An adaptive thresholding estimator based on the generalized sample covariance matrix is proposed. The estimator is shown to achieve the optimal rate of convergence over a large class of approximately sparse covariance matrices under mild conditions.

The technical analysis for the case of missing data is much more challenging than that for the complete data, although some of the basic ideas are similar. To facilitate the theoretical analysis of the proposed estimators, we establish two key technical results, first, a large deviation result for a sub-matrix of the generalized sample covariance matrix and second, a large deviation bound for the self-normalized entries of the generalized sample covariance matrix. These technical tools are not only important for the present paper, but also useful for other related problems in high-dimensional statistical inference with missing data.

A simulation study is carried out to examine the numerical performance of the proposed estimation procedures. The results show that the proposed estimators perform well numerically. Even in the MUCR setting, our proposed procedures for estimating bandable, sparse covariance matrices, which do not rely on the information of the missingness mechanism, outperform the ones specifically designed for MUCR. The advantages are more significant under the setting of missing completely at random but not uniformly. We also illustrate our procedure with an application to data from four ovarian cancer studies that have different volumes of genomic data by design. The proposed estimators enable us to estimate the covariance matrix by integrating the data from all four studies and lead to a more accurate estimator. Such high-dimensional covariance matrix estimation with missing data is also useful for other types of data integration. See further discussions in Section 4.4.

The rest of the paper is organized as follows. Section 2 considers estimation of bandable covariance matrices with incomplete data. The minimax rate of convergence is established for the spectral norm loss under regularity conditions. Section 3 focuses on estimation of high-dimensional sparse covariance matrices and introduces an adaptive thresholding estimator in the presence of missing observations. Asymptotic properties of the estimator under the spectral norm loss is also studied. Numerical performance of the proposed methods is investigated in Section 4 through both simulation studies