Commun. Math. Res.
doi: 10.4208/cmr.2020-0001

Vol. **36**, No. 1, pp. 68-92
March 2020

Precision Matrix Estimation by Inverse Principal Orthogonal Decomposition

Cheng Yong Tang[1,*], Yingying Fan[2] and Yinfei Kong[3]

[1] *Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, PA 19122, USA.*
[2] *Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA.*
[3] *Department of Information Systems and Decision Sciences, Mihaylo College of Business and Economics, California State University, Fullerton, CA 92831, USA.*

Received 3 January 2020; Accepted 11 February 2020

Abstract. We investigate the structure of a large precision matrix in Gaussian graphical models by decomposing it into a low rank component and a remainder part with sparse precision matrix. Based on the decomposition, we propose to estimate the large precision matrix by inverting a principal orthogonal decomposition (IPOD). The IPOD approach has appealing practical interpretations in conditional graphical models given the low rank component, and it connects to Gaussian graphical models with latent variables. Specifically, we show that the low rank component in the decomposition of the large precision matrix can be viewed as the contribution from the latent variables in a Gaussian graphical model. Compared with existing approaches for latent variable graphical models, the IPOD is conveniently feasible in practice where only inverting a low-dimensional matrix is required. To identify the number of latent variables, which is an objective of its own interest, we investigate and justify an approach by examining the ratios of adjacent eigenvalues of the sample covariance matrix. Theoretical properties, numerical examples, and a real data application demonstrate the merits of the IPOD approach in its convenience, performance, and interpretability.

Key words: High-dimensional data analysis, latent Gaussian graphical model, precision matrix.

*Corresponding author. *Email addresses:* yongtang@temple.edu (C. Y. Tang), fanyingy@marshall.usc.edu (Y. Fan), yikong@fullerton.edu (Y. Kong)

1 Introduction

Exploring how subjects and/or variables are connected to each other in various systems is one of the most common and important problems in practical applications. Examples of such investigations are regularly seen in scenarios including regression analysis, Gaussian graphical models, classification, principal component analysis and many more. Investigations of this kind are encountered even more often in practical applications in recent popular areas such as finance, biological and medical studies, meteorological and astronomical research, among others. Because of the general interest on the connections between individuals, the scale of these investigations can easily grow beyond a practical and manageable scope — for example, considering the complexity of possible associations among human genes. Therefore, parsimonious modeling approaches are critically important for generating practical, feasible, and interpretable statistical analyses when exploring the association structures of the target systems in many contemporary studies.

For studying the connections between subjects/variables, precision matrix, the inverse of a covariance matrix, is a crucial device in many statistical analyses including Gaussian graphical models [12], discriminant analysis, dimension reduction, and investment portfolio analysis. There has been an increasing interest in penalized likelihood approaches for estimating large precision matrices in recent literature; see, for example, [7, 8, 10, 13, 15–17, 19] and references therein. In Gaussian graphical models, the precision matrix has the interpretation that each of its zero elements implies the conditional independence of the corresponding pair of individuals given the information from all other individuals. In the corresponding graph consisting of a vertex set and an edge set, such conditional independence means that there is no edge between the corresponding pair of vertices representing the individuals.

With latent variables, analyzing Gaussian graphical models becomes substantially more difficult; see [4] in which a penalized likelihood approach is investigated. More specifically, the interpretation of the graphical model becomes less clear if the impact of latent variables is not incorporated in the large precision matrix. Additionally, the unknown number of the latent variables also poses new challenges, both computationally in optimizing the penalized likelihood function and practically in developing most appropriate interpretations of the graphical models. A remarkable feature of the Gaussian graphical model with latent variables is that although the underlying the true precision matrix is sparse indicating small number of connected vertices in the corresponding graph, latent variables generally cause a non-sparse observable precision matrix of the variables