Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202023.0099

REPRODUCIBLE LEARNING IN LARGE-SCALE MULTIPLE GRAPHICAL MODELS

Jia Zhou[1], Guangming Pan*[2], Zeming Zheng[3] and Changchun Tan[1]

[1]*Hefei University of Technology,* [2]*Nanyang Technological University and* [3]*University of Science and Technology of China*

Abstract: Reproducible learning of the underlying structure among large-scale network data is important in many contemporary applications. Despite the fast-growing literature on this subject, the practical issue of data heterogeneity has rarely been addressed. In this paper, we propose a new method called the multiple graphical knockoff filter to efficiently recover the underlying sparse connected structure of a general population from a high-dimensional heterogeneous dataset. We provide theoretical justification on the asymptotic false discovery rate control, and the theory for the power analysis is also established. To the best of our knowledge, this is the first formal theoretical result on the power for the graphical knockoffs procedure. Our new methodology and results are evidenced by numerical studies.

Key words and phrases: False discovery rate, heterogeneity, high-dimensionality, multiple graphical models, power.

1. Introduction

The surge of big data in an unprecedented scale has brought us an enormous amount of information that makes large-scale network analysis increasingly frequent in many contemporary applications, such as biology, economics, and social science (Giudici and Alessanfro, 2016; Shin et al., 2014). It is often of practical interest to uncover the underlying network formed by a large number of individuals that are sparsely related. As a popular choice, Gaussian graphical models provide a flexible way to specify the conditional independence structure among a large number of nodes. There is a growing literature on Gaussian graphical models, mainly focusing on the problem of support recovery and link strength estimation; see for example, Friedman, Hastie and Tibshirani (2008), Fan and Lv (2016), Cheng et al. (2017), and Zhou et al. (2022), among many others.

To obtain a reliable outcome and alleviate reproducibility issues, controlling the false discovery rate (FDR) which is defined as the expected proportion of false discoveries among all the discoveries proposed by Benjamini and Hochberg (1995) has gained much attention recently. There have been several studies proposed

*Corresponding author. E-mail: gmpan@ntu.edu.sg