Statistica Sinica **37** (2027), 1-21
doi:https://doi.org/10.5705/ss.202024.0258

# GROUPED HETEROGENEOUS GAUSSIAN GRAPHICAL MODELS FOR HIGH-DIMENSIONAL CLUSTERED DATA

Xin Zeng, Shuangge Ma and Qingzhao Zhang*

*Chongqing Technology and Business University, Yale University and Xiamen University*

*Abstract:* Clustered data-based analysis has been extensively conducted in various studies. Recent research has demonstrated that a network-based heterogeneity analysis, which adopts a system perspective and incorporates the interconnections among variables while considering heterogeneity between components, can provide more informative results compared to approaches based on simpler statistics. Moreover, incorporating grouping strategies in analysis can better delineate the sources of heterogeneity and enable more flexible modeling for clustered data. In this article, we introduce a novel approach called the grouped heterogeneous Gaussian graphical models (Grouped-HGGM) for network analysis of high-dimensional clustered data. Our approach assumes that clusters can be divided into distinct groups, and any heterogeneity across clusters is captured through the cluster-wise mixture probabilities. Unlike most previous approaches that assume that the number of components is known in advance, an appealing feature of our method is the automatic determination of the number of components and sparse estimation using a fusion technique. Consistency properties are rigorously established, and an effective computational algorithm is developed. Extensive simulations demonstrate the practical superiority of the proposed approach over closely related alternatives. In the analysis of breast cancer data, the proposed approach identifies heterogeneity structures different from the alternatives.

*Key words and phrases:* Clustered data, Gaussian graphical models, grouping strategies, heterogeneity analysis.

## 1. Introduction

Consider data that can be divided into $m$ known clusters, which can be based on postal regions, institutions, species, and others. This type of data is referred to as clustered data, which frequently appears in various areas of research, such as neuroscience (Galbraith, Daniel and Vissel, 2010), personalized medicine (Fokkema et al., 2018), and labor economics (Pereda-Fernandez, 2021). For a given cluster $i$ ($i = 1, \ldots, m$), heterogeneity often manifests in real data, indicating the presence of subgroup structures. As a popular analytical strategy, the finite mixture model (McLachlan and Peel, 2000) could be useful to flexibly capture heterogeneity. In the left panel of Figure 1, we show a dataset consisting

*Corresponding author. E-mail: qzzhang@xmu.edu.cn