Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Data and text mining

HeteroGGM: an R package for Gaussian graphical model-based heterogeneity analysis

Mingyang Ren [1, 2], Sanguo Zhang [1, 2], Qingzhao Zhang [3], and Shuangge Ma [4]*

[1]School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [2]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China, [3]MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, Xiamen 361005, China, and [4]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Heterogeneity is a hallmark of many complex human diseases, and unsupervised heterogeneity analysis has been extensively conducted using high-throughput molecular measurements and histopathological imaging features. "Classic" heterogeneity analysis has been based on simple statistics such as mean, variance, and correlation. Network-based analysis takes interconnections as well as individual variable properties into consideration and can be more informative. Several Gaussian graphical model (GGM)-based heterogeneity analysis techniques have been developed, but friendly and portable software is still lacking. To facilitate more extensive usage, we develop the R package HeteroGGM, which conducts GGM-based heterogeneity analysis using the advanced penaliztaion techniques, can provide informative summary and graphical presentation, and is efficient and friendly.
Availability: The package is available at `https://CRAN.R-project.org/package=HeteroGGM`.
Contact: shuangge.ma@yale.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Heterogeneity is a hallmark of cancer, diabetes, and many other complex diseases. It has different definitions under different contexts. Here we focus on the scenario under which samples form subgroups, and (molecular, imaging, etc.) variables have different properties in different subgroups. Unsupervised heterogeneity analysis can assist identifying disease subtypes, provide a deeper understanding of disease biology, and serve as the basis of downstream analysis such as regression. It has been based on high-throughput molecular measurements (gene expression, SNP, methylation, etc.) as well as histopathological imaging features. A few example are provided in Section 1 of the Supplementary Materials. "Classic" heterogeneity analysis has been based on simple statistics, such as mean, variance, and correlation. Network-based analysis can accommodate such information as well as that on the interconnections among variables, take a system perspective, and be more effective. One of the most popular network analysis approaches is Gaussian graphical model (GGM) and has been applied to a variety of molecular, histopathological imaging, and other types of data. With minor revisions, GGM techniques can also be applied to non-normal data. A few examples are described in Section 1 of the Supplementary Material. GGM-based heterogeneity analysis approaches include JGL – joint graphical Lasso (Danaher *et al*., 2014), pGMM – parsimonious Gaussian mixture models (Gao *et al*., 2016), SCAN – Simultaneous Clustering And estimatioN of heterogeneous graphical models (Hao *et al*., 2018), and others, and have led to promising findings. These and some other early studies are limited in that the number of subgroups is assumed to be known *a priori*, which is not realistic. In addition, the accompanying software programs are not sufficiently "friendly", hindering broad utilization.

In a very recent study (Ren *et al*., 2021), a novel approach based on the penalized fusion technique is developed to fully data-dependently determine the number and structure of subgroups in GGM-based heterogeneity analysis. The goal of this study is to develop a user-friendly R package implementing this and the highly relevant approach developed in Zhou *et al*. (2009). Beyond the original approaches, more estimations are added, so that the package is self-contained and more comprehensive. Presentation functions are developed, so that the package can provide "more direct" insights for practitioners.

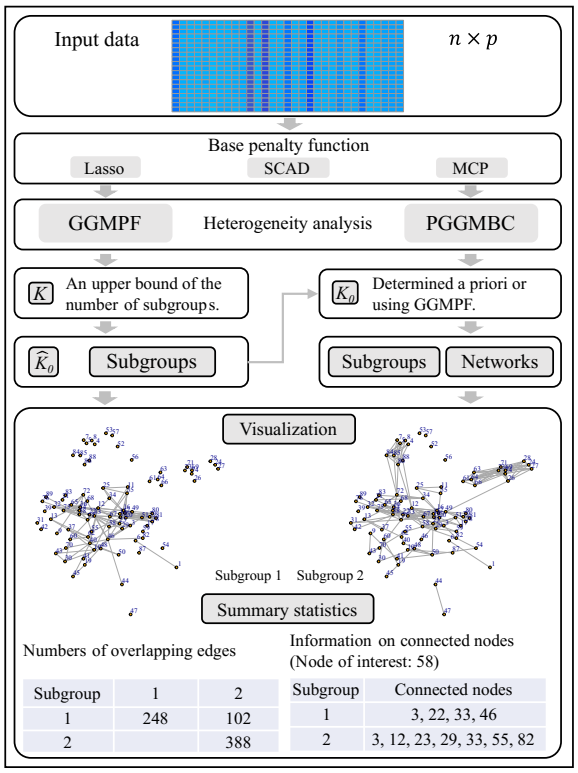

Fig. 1. Workflow of the HeteroGGM package

2 The HeteroGGM package

The main workflow is presented in Figure 1. This package implements two GGM-based heterogeneity analysis methods: (i) the penalized fusion-based method developed in Ren *et al*. (2021), which data-dependently determines the number of sample subgroups. In the original development, the penalty is built on MCP (Minimax Concave Penalty); and (ii) the method developed in Zhou *et al*. (2009), which assumes that the number of subgroups is known *a priori* or determined elsewhere (for example using the approach in Ren *et al*. (2021)). In the original development, the penalty is built on Lasso. In the package, to be more comprehensive, we allow users to choose from Lasso, MCP, and SCAD (Smoothly Clipped Absolute Deviation Penalty) base penalties for both methods. Computation of both methods is realized using EM (expectation maximization), ADMM (Alternating Direction Method of Multipliers), and S-AMA (Sparse Alternating Minimization Algorithm) techniques. Details on the methods and computation are provided in Section 2 of the Supplementary Material. The package has the following key functions:

- *GMMPF*: It applies the method developed in Ren *et al*. (2021). Input includes the data matrix, an upper bound for the number of subgroups, and the choice of base penalty (with MCP being the default). It generates the number of estimated subgroups, subgrouping memberships for samples, and network structures for all subgroups.
- *PGGMBC*: It applies the method developed in Zhou *et al*. (2009). The design of the function is similar to the above. The key difference is that the number of subgroups needs to be specified, either based on prior knowledge or from other analysis (for example using *GMMPF*).
- *summary-network & plot-network*: These two functions summarize the key findings, including the numbers of edges for all subgroups, numbers of overlapping edges, graphical presentation of the networks, and information on nodes that are connected to a specific node of interest.

In Section 3 of the Supplementary Material, we provide demonstrative code for implementing the above functions to a sample dataset.

3 Application examples

We apply the aforementioned functions to two data examples. (i) We analyze the TCGA gene expression data on breast cancer patients. The data contains measurements on 73 genes in the Wnt signaling pathway and 771 subjects with primary solid tumors. The *GGMPF* function identifies three sample subgroups with sizes 156, 331, and 284, respectively, and their networks have 322, 252, and 68 edges, respectively. The *PGGMBC* function, with the number of subgroups set as three, generates subgroups with sizes 172, 320, and 279, whose networks have 402, 302, and 88 edges, respectively. Significant differences are observed, and discussions are provided in the Supplementary Materials. (ii) We analyze the TCGA lung squamous cell carcinoma (LUSC) data. To demonstrate the broad applicability of the GGM-based heterogeneity analysis and the package, we analyze 89 histopathological imaging features extracted using an automated digital signal processing pipeline. Six sample subgroups with significantly different networks are identified. For both datasets, we provide additional numerical and graphical results in Section 4 of the Supplementary Material.

4 Discussion

With the still strong demand for heterogeneity analysis and successes of recent network-based analysis, we expect a significant growth in network-based heterogeneity analysis. The HeteroGGM package can realize the most advanced and recent GGM-based heterogeneity analysis methods and, with its comprehensiveness and user-friendly functions, significantly facilitate routine data analysis. It only demands basic R settings, and its compartmentalized design will also facilitate revision and partial adoption.

Funding

Beijing Natural Science Foundation (Z190004), National Natural Science Foundation of China (11971404, Basic Scientific Project 71988101), 111 Project (B13028), NSF (1916251), NIH (CA241699, CA196530), and a Yale Cancer Center Pilot Award.

Conflict of Interest: none declared.

References

Danaher, P., Wang, P. and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **76**, 373–397.

Gao, C., Zhu, Y., Shen, X. and Pan, W. (2016). Estimation of multiple networks in gaussian mixture models. *Electron. J. Stat.*, **10**, 1133-1154.

Hao, B., Sun, W., Liu, Y. and Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res.*, **18**, 7981–8038.

Ren, M., Zhang, S., Zhang, Q. and Ma, S. (2021). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*, https://doi.org/10.1111/biom.13426.

Zhou, H., Pan, W. and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.*, **3**, 1473-1496.