Statistica Sinica **33** (2023), 1-30
doi:https://doi.org/10.5705/ss.202021.0002

HETEROGENEITY ANALYSIS VIA INTEGRATING MULTI-SOURCES HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO CANCER STUDIES

Tingyan Zhong, Qingzhao Zhang, Jian Huang, Mengyun Wu
and Shuangge Ma

Shanghai Jiao Tong University, Xiamen University, The Hong Kong Polytechnic University, Shanghai University of Finance and Economics and Yale University

Abstract: This study has been motivated by cancer research, in which heterogeneity analysis plays an important role and can be roughly classified as unsupervised or supervised. In supervised heterogeneity analysis, the finite mixture of regression (FMR) technique is used extensively, under which the covariates affect the response differently in subgroups. High-dimensional molecular and, very recently, histopathological imaging features have been analyzed separately and shown to be effective for heterogeneity analysis. For simpler analysis, they have been shown to contain overlapping, but also independent information. In this article, our goal is to conduct the first and more effective FMR-based cancer heterogeneity analysis by integrating high-dimensional molecular and histopathological imaging features. A penalization approach is developed to regularize estimation, select relevant variables, and, equally importantly, promote the identification of independent information. Consistency properties are rigorously established. An effective computational algorithm is developed. A simulation and an analysis of The Cancer Genome Atlas (TCGA) lung cancer data demonstrate the practical effectiveness of the proposed approach. Overall, this study provides a practical and useful new way of conducting supervised cancer heterogeneity analysis.

Key words and phrases: Cancer heterogeneity, data integration, FMR, molecular and imaging features.

1. Introduction

Heterogeneity is a hallmark of cancer, and thus has gained extensive research (Turajlic et al. (2019)). Heterogeneity analysis can be roughly classified as unsupervised or supervised. In unsupervised analysis, outcomes/phenotypes are not involved, and clustering and other techniques are adopted (Wiwie, Baumbach and Röttger (2015)). Unsupervised analysis can be useful, for example, for identifying new disease subtypes, but it is often difficult to associate clinical implications

Corresponding author: Mengyun Wu. E-mail: wu.mengyun@mail.shufe.edu.cn. Shuangge Ma. E-mail: shuangge.ma@yale.edu.