Statistica Sinica **33** (2023), 1-27
doi:https://doi.org/10.5705/ss.202020.0145

HIGH-DIMENSIONAL FACTOR REGRESSION FOR HETEROGENEOUS SUBPOPULATIONS

Peiyao Wang[1], Quefeng Li[1], Dinggang Shen[2,3,4] and Yufeng Liu[1]

[1]*University of North Carolina at Chapel Hill,* [2]*ShanghaiTech University,* [3]*Shanghai United Imaging Intelligence Co. and* [4]*Korea University*

Abstract: In modern scientific research, data heterogeneity is commonly observed owing to the abundance of complex data. We propose a factor regression model for data with heterogeneous subpopulations. The proposed model can be represented as a decomposition of heterogeneous and homogeneous terms. The heterogeneous term is driven by latent factors in different subpopulations. The homogeneous term captures common variation in the covariates and shares common regression coefficients across subpopulations. Our proposed model attains a good balance between a global model and a group-specific model. The global model ignores the data heterogeneity, while the group-specific model fits each subgroup separately. We prove the estimation and prediction consistency for our proposed estimators, and show that it has better convergence rates than those of the group-specific and global models. We show that the extra cost of estimating latent factors is asymptotically negligible and the minimax rate is still attainable. We further demonstrate the robustness of our proposed method by studying its prediction error under a misspecified group-specific model. Finally, we conduct simulation studies and analyze a data set from the Alzheimer's Disease Neuroimaging Initiative and an aggregated microarray data set to further demonstrate the competitiveness and interpretability of our proposed factor regression model.

Key words and phrases: Factor models, heterogeneity, penalized regression, prediction.

1. Introduction

Data heterogeneity is an important issue in modern complex data analysis. In practice, data heterogeneity may come from variables or samples. More specifically, multi-modality/source data have heterogeneity among the variables, because they may correspond to different types of measurements. For example, in biomedical imaging, people may acquire both MRI and PET images (Zhang et al. (2011)). In genomics studies, measurements are collected from different sources, such as mRNA and miRNA (Muniategui et al. (2013)). In addition to

Corresponding author: Yufeng Liu, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: yfliu@email.unc.edu.