Statistica Sinica **36** (2026), 1-23
doi:https://doi.org/10.5705/ss.202023.0186

COMPONENT-BASED REGRESSION FOR HYBRID DATA

Xiaohu Jiang[1], Xiuli Du*[2], Yenan Ren[2], Jinguan Lin[3]
and The Alzheimer's Disease Neuroimaging Initiative

[1] *Yunnan University,* [2] *Nanjing Normal University*
and [3] *Nanjing Audit University*

Abstract: In recent years, with the deep integration of big data and medical technology, hybrid data with or without block-wise missing arise more commonly in medical care. Efficient dimensionality reduction and extraction of important predictive information for such data have also become a popular research topic. In this article, for hybrid data without missing and with block-wise missing, we proposed a kind of new component-based model based on the unified approach to multi-source principal component analysis and multi-set canonical correlation analysis. After obtaining scores by using the unified framework, component-based regression models are established. Asymptotic properties are established under some mild conditions. Simulations and real data analysis show the proposed method works well.

Key words and phrases: Alzheimer's disease, block-wise imputation, component-based regression, hybrid data, multi-set canonical correlation analysis, multi-source principal component analysis.

1. Introduction

In recent years, with the rapid development of big data and medical technology, hybrid data has become increasingly common in the medical field. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI) data include information from multiple sources such as magnetic resonance imaging (MRI), positron emission tomography (PET), cerebrospinal fluid (CSF), microarray gene expression profile data (GENE) and demographic information. PET and MRI are three-dimensional images, GENE data contain 49,386 gene features, and CSF data have several biomarkers; therefore, ADNI data are typical hybrid data.

Since different data sources can provide complementary information, hybrid data has better predictive performance. However, the high dimensionality of hybrid data can lead to much difficulty in modelling; therefore, it is necessary to reduce the dimensionality of hybrid data.

In recent years, many scholars have proposed component-based regression models based on multi-source principal component analysis (MPCA) for multi-

*Corresponding author. E-mail: duxiuli@njnu.edu.cn