Statistica Sinica Preprint No: SS-2023-0186	
Title	Component-based Regression for Hybrid Data
Manuscript ID	SS-2023-0186
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0186
Complete List of Authors	Xiaohu Jiang, Xiuli Du, Yenan Ren, Jinguan Lin and The Alzheimer's Disease Neuroimaging Initiative
Corresponding Authors	Xiuli Du
E-mails	duxiuli@njnu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

Component-based Regression for Hybrid Data

Xiaohu Jiang, Xiuli Du, Yenan Ren, Jinguan Lin

and The Alzheimer's Disease Neuroimaging Initiative

Yunnan University, Nanjing Normal University and 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: Hybrid data, multi-source principal component analysis, multiple-set canonical correlation analysis, component-based regression, block-wise imputation, Alzheimer's Disease.

*Data used in preparation of this article were obtained from the Alzheimer's Disease

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

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-source high-dimensional data (Bai and Ng (2002); Bair et al.

Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

(2006); Bai and Li (2012); Fan et al. (2017)) and multi-source functional data (Ramsay and Silverman (2005); Berrendero et al. (2011); Chiou et al. (2014); Happ and Greven (2018)) respectively to achieve dimensionality reduction. Another strategy is based on multi-set canonical correlation analysis (MCCA). Correa et al. (2010), Takane et al. (2008) and Tenenhaus et al. (2017) discussed the theory and methods of MCCA for high-dimensional data. Hwang et al. (2012) applied the idea to functional data.

However, MPCA focuses mainly on explaining the variance of the data, whereas MCCA focuses on maximizing the association among different data sources. Hwang et al. (2013) provided a unified framework that combines MCCA and MPCA. Choi et al. (2017) proposed a functional version of the method of Hwang et al. (2013). So far, there is no unified approach to MPCA and MCCA for hybrid data containing functional and high-dimensional data.

Component-based regression is, however, limited in that the components may not be optimal in explaining the variance of the outcome variables because they are extracted only to account for the maximum variance or association of the predictors, without considering their associations with the outcome variables (Choi et al. (2020)). To addressed the issue, extended redundancy analysis (ERA) is proposed to carry out dimension-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

ality reduction and regression as well. Takane and Hwang (2005) first generalized redundancy analysis to ERA that allows analyzing a variety of directional relationships among multiple sets of variables. Hwang et al. (2012) and Tan et al. (2015) proposed a functional version of ERA. Hwang et al. (2015) extended functional ERA into the framework of generalized linear models. Choi et al. (2020) and Park et al. (2020) extended ERA into the Bayesian framework. Kim and Hwang (2021) combined ERA with model-based recursive partitioning in a single framework. Vijayakumar et al. (2022) proposed an NN-ERA model that integrates neural networks algorithms into the framework of ERA. Kim and Hwang (2022) proposed several new model evaluation metrics for ERA that can compute a model's performance on out-of-sample data. Yamashita (2023) proposed an exploratory ERA, which used the dataset directly for estimation and did not require a group structure.

On the other hand, there often exist block-wise missing cases for hybrid data due to a variety of reasons. In recent years, some methods have been proposed to predict the response for multi-source block-wise missing data. A common strategy is to use the Lasso or some other penalized regression methods to implement the regression analysis without imputation (Yuan et al. (2012); Xiang et al. (2014) and Yu et al. (2020)). Another strategy is to

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

impute the missing data firstly by some existing imputation methods, and then implement the regression analysis and variable selection (Wan et al. (2015)). Campos et al. (2015) trained support vector machine and random forest classifiers using all the imputed data for the task of discriminating among different stages of the Alzheimer's disease. For multi-source high-dimensional block-wise missing data, Zhang et al. (2020) proposed a new factor imputed regression model by combining the factor model with block-wise missing data imputation. Xue and Qu (2021) proposed a multiple block-wise imputation (MBI) method to improve the SI method. For a given missing pattern group, MBI uses information from other missing pattern groups in addition to the samples from the complete observed group. For multi-source functional block-wise missing data, Du et al. (2023) proposed the multinomial imputed-component-based Logistic regression model.

In this paper, motivated by Hwang et al. (2013) and Choi et al. (2017), we propose a new component-based regression modelling method for hybrid data containing functional data sources, high-dimensional data sources and multivariate data sources, where components are constructed by the unified approach to MCCA and MPCA. And we also impute the block-wise missing components by the block-wise conditional mean imputation and multiple block-wise imputation methods.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

The main contributions of our methods are as follows. Firstly, we extended the unified approach to MCCA and MPCA proposed by Hwang et al. (2013) and Choi et al. (2017) to hybrid data containing functional data and high-dimensional data. Secondly, we proposed to use the principal component basis of the corresponding univariate functional data as the basis functions to converting functional optimization problem to an approximately equivalent matrix eigen-analysis problem. Besides, the proposed method is suitable not only for hybrid data without missing but also for the data with block-wise missing.

The rest of the article is organized as follows. Section 2 introduces the component model for hybrid data based on the unified approach to MCCA and MPCA. Section 3 discusses the component-based regression model and the idea of component imputation. Section 4 gives some theoretical properties. Numerical simulations are conducted in Section 5. A real data analysis on ADNI data is given in Section 6. Section 7 gives a brief discussion of our method. Imputation methods and proofs can be found in Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

2. Component Model for Hybrid Data

In the following, we assume hybrid data contain multi-source functional data consisting of functions $X^{(1)}, \cdots, X^{(P)}$ $(P \geq 1)$ and multi-source high-dimensional data consisting of vectors $Z^{(1)}, \cdots, Z^{(Q)}$ $(Q \geq 1)$.

2.1 The Unified Approach to MPCA and MCCA

Both MPCA and MCCA can extract the information of original data, where MPCA focuses mainly on explaining the variance of the data and MCCA on maximizing the associations across different data sources. In Hwang et al. (2013) and Choi et al. (2017), they discussed the unified approach to MPCA and MCCA for high-dimensional data and functional data, respectively, by introducing a weight parameter α such that PCA and CCA are integrated into a unified framework. By applying the unified approach, they obtained highly correlated components while also effectively explaining the variance of the data. Inspired by them, we propose a unified approach to MPCA and MCCA for hybrid data. The proposed approach will seek to minimize

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

the following objective function

$$\phi = \frac{\alpha \sum_{i=1}^{N} \sum_{j=1}^{P} ||x_i^{(j)} - \sum_{m=1}^{M} f_{i,m} a_m^{(j)}||^2}{NT} + \frac{\alpha\lambda \sum_{j=1}^{P} \sum_{m=1}^{M} ||\mathrm{D}^2 a_m^{(j)}||^2}{MT} + \frac{(1-\alpha) \sum_{i=1}^{N} \sum_{j=1}^{P} \left|\left|\mathrm{F}_i - \int_{\mathcal{T}_j} x_i^{(j)}(t_j) \mathrm{w}^{(j)}(t_j) \mathrm{d}t_j\right|\right|^2}{NT} + \frac{(1-\alpha)\rho \sum_{j=1}^{P} \sum_{m=1}^{M} ||\mathrm{D}^2 w_m^{(j)}||^2}{MT} + \frac{\alpha \sum_{i=1}^{N} \sum_{j=1}^{Q} ||\mathrm{z}_i^{(j)} - \mathbf{h}^{(j)} \mathrm{F}_i||^2}{NT} + \frac{(1-\alpha) \sum_{i=1}^{N} \sum_{j=1}^{Q} \left|\left|\mathrm{F}_i - \mathbf{v}^{(j)\prime} \mathrm{z}_i^{(j)}\right|\right|^2}{NT} \tag{2.1}$$

to obtain the component scores. In (2.1), $x_i^{(j)}(t_j)$ denotes the ith realization of the jth functional data source $X^{(j)}(t_j)$; $a_m^{(j)}(t_j)$ denotes the jth component of the loading function corresponding to the mth component, $||x_i^{(j)} - \sum_{m=1}^{M} f_{i,m} a_m^{(j)}||^2 = \int_{\mathcal{T}_j} \left(x_i^{(j)}(t_j) - \sum_{m=1}^{M} f_{i,m} a_m^{(j)}(t_j)\right)^2 \mathrm{d}t_j$ denotes the squared norm of function $x_i^{(j)}(t_j) - \sum_{m=1}^{M} f_{i,m} a_m^{(j)}(t_j)$; $f_{i,m}$ denotes the mth component score of subject i, $\mathrm{F}_i = (f_{i,1}, \cdots, f_{i,M})'$; D^2 in $||\mathrm{D}^2 a_m^{(j)}||$ denotes the second-order derivative of function $a_m^{(j)}(t_j)$; $\mathrm{w}^{(j)}(t_j) = (w_1^{(j)}(t_j), \cdots, w_M^{(j)}(t_j))'$ and $\int_{\mathcal{T}_j} x_i^{(j)}(t_j) \mathrm{w}^{(j)}(t_j) \mathrm{d}t_j$ denote the first M canonical weight functions and canonical variates assigned to the jth functional data source, respectively; $\mathrm{z}_i^{(j)} = (z_{i,1}^{(j)}, \cdots, z_{i,T_j}^{(j)})'$ denotes the T_j-dimensional vector which is the ith realization of the jth high-dimensional data source; $\mathbf{h}^{(j)} = (\mathrm{h}_1^{(j)}, \cdots, \mathrm{h}_M^{(j)})$ with $\mathrm{h}_m^{(j)} = (h_{m,1}^{(j)}, \cdots, h_{m,T_j}^{(j)})'$, $\mathbf{v}^{(j)} = (\mathrm{v}_1^{(j)}, \cdots, \mathrm{v}_M^{(j)})$ with $\mathrm{v}_m^{(j)} = (v_{m,1}^{(j)}, \cdots, v_{m,T_j}^{(j)})'$ and $\mathbf{v}^{(j)\prime} \mathrm{z}_i^{(j)}$ denote the first M component loading, canonical weights and canonical variates for the jth high-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

data source, respectively. ||v|| denotes the 2-norm of vector v. λ and ρ are the smooth parameters. α is the weight parameter. N is the number of subjects. $T = P + \sum_{j=1}^{Q} T_j + (P+Q)M$ is the number of squared residuals for each subject. As pointed out by Ramsay and Silverman (2005), the variations in the functional and vector parts of a hybrid observation are almost inevitably not comparable, adding T to each term of the objective function is to make them comparable.

In (2.1), the first 4 terms involve the computation of inner product of functions. One way of reducing the integral to discrete form is to express each function as a linear combination of known basis functions. This paper proposes to using the principal component basis of each univariate functional data source as basis functions. The Karhunen-Loève expansion of univariate functional data $\{x_i^{(j)}(t_j), t_j \in \mathcal{T}_j\}$ is described as follows:

$$x_i^{(j)}(t_j) = \sum_{m=1}^{\infty} \zeta_{i,m}^{(j)} \phi_m^{(j)}(t_j),$$

where $j = 1, \cdots, P$. The univariate principal component score $\zeta_{i,m}^{(j)} =< x_i^{(j)}, \phi_m^{(j)} >$ with $\mathrm{Cov}(\zeta_{i,m}^{(j)}, \zeta_{i,n}^{(j)}) = \lambda_m^{(j)} \delta_{mn}$, and the univariate eigenfunction $\phi_m^{(j)}(t_j) \in \mathbb{R}$ with $< \phi_m^{(j)}, \phi_n^{(j)} >= \delta_{mn}$. In the paper, $< \cdot, \cdot >$ represents the inner product operation.

The set of univariate eigenfunctions $\{\phi_m^{(j)}(t_j), t_j \in \mathcal{T}_j, m = 1, 2, \cdots\}$ is called the principal component basis. However, in practice, there is no way

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

to obtain all principal components based on finite observation, we can only truncate it to take the first M_j terms as basis functions, then expanding $a_m^{(j)}(t_j)$ and $w_m^{(j)}(t_j)$ by which can obtain $a_m^{(j)}(t_j) = \sum_{n=1}^{M_j} s_{m,n}^{(j)} \phi_n^{(j)}(t_j)$ and $w_m^{(j)}(t_j) = \sum_{n=1}^{M_j} b_{m,n}^{(j)} \phi_n^{(j)}(t_j)$.

Let $\boldsymbol{\phi}^{(j)}(t_j) = \left(\phi_1^{(j)}(t_j), \cdots, \phi_{M_j}^{(j)}(t_j)\right)'$, $\mathbf{R}^{(j)} = \int_{\mathcal{T}_j} \mathrm{D}^2\boldsymbol{\phi}^{(j)}(t_j)\mathrm{D}^2\boldsymbol{\phi}^{(j)'}(t_j)\mathrm{d}t_j$; $\boldsymbol{\zeta}^{(j)} = (\boldsymbol{\zeta}_1^{(j)}, \cdots, \boldsymbol{\zeta}_N^{(j)})'$ with $\boldsymbol{\zeta}_i^{(j)} = (\zeta_{i,1}^{(j)}, \cdots, \zeta_{i,M_j}^{(j)})'$; $\mathrm{x}^{(j)}(t_j) = (x_1^{(j)}(t_j), \cdots, x_N^{(j)}(t_j))'$; $\mathbf{s}^{(j)} = (\mathrm{s}_1^{(j)}, \cdots, \mathrm{s}_M^{(j)})$ with $\mathrm{s}_m^{(j)} = (s_{m,1}^{(j)}, \cdots, s_{m,M_j}^{(j)})'$; $\mathbf{b}^{(j)} = (\mathrm{b}_1^{(j)}, \cdots, \mathrm{b}_M^{(j)})$ with $\mathrm{b}_m^{(j)} = (b_{m,1}^{(j)}, \cdots, b_{m,M_j}^{(j)})'$; $\mathbf{Z}^{(j)} = (\mathrm{z}_1^{(j)}, \cdots, \mathrm{z}_N^{(j)})'$; $\mathbf{H} = (\mathbf{h}^{(1)'}, \cdots, \mathbf{h}^{(Q)'})'$; $\mathbf{F} = (\mathrm{F}_1, \cdots, \mathrm{F}_N)'$. (2.1) can be rewritten as follows:

$$
\begin{aligned}
\phi = {} & \alpha(NT)^{-1} \sum_{j=1}^{P} \mathrm{tr} \left(\int_{\mathcal{T}_j} \mathrm{x}^{(j)}(t_j)\mathrm{x}^{(j)'}(t_j)\mathrm{d}t_j - 2\mathbf{s}^{(j)}\mathbf{F}'\boldsymbol{\zeta}^{(j)} + \mathbf{s}^{(j)}\mathbf{F}'\mathbf{F}\mathbf{s}^{(j)'} \right) \\
& + \alpha(NT)^{-1} \sum_{j=1}^{Q} \mathrm{tr} \left(\mathbf{Z}^{(j)'}\mathbf{Z}^{(j)} - 2\mathbf{h}^{(j)}\mathbf{F}'\mathbf{Z}^{(j)} + \mathbf{h}^{(j)}\mathbf{F}'\mathbf{F}\mathbf{h}^{(j)'} \right) \\
& + \alpha(MT)^{-1}\lambda \sum_{j=1}^{P} \mathrm{tr} \left(\mathbf{s}^{(j)'}\mathbf{R}^{(j)}\mathbf{s}^{(j)} \right) \\
& + (1-\alpha)(NT)^{-1} \sum_{j=1}^{P} \mathrm{tr} \left(\mathbf{F}'\mathbf{F} - 2\mathbf{b}^{(j)'}\boldsymbol{\zeta}^{(j)'}\mathbf{F} + \mathbf{b}^{(j)'}\boldsymbol{\zeta}^{(j)'}\boldsymbol{\zeta}^{(j)}\mathbf{b}^{(j)} \right) \\
& + (1-\alpha)(NT)^{-1} \sum_{j=1}^{Q} \mathrm{tr} \left(\mathbf{F}'\mathbf{F} - 2\mathbf{v}^{(j)'}\mathbf{Z}^{(j)'}\mathbf{F} + \mathbf{v}^{(j)'}\mathbf{Z}^{(j)'}\mathbf{Z}^{(j)}\mathbf{v}^{(j)} \right) \\
& + (1-\alpha)(MT)^{-1}\rho \sum_{j=1}^{P} \mathrm{tr} \left(\mathbf{b}^{(j)'}\mathbf{R}^{(j)}\mathbf{b}^{(j)} \right). \qquad (2.2)
\end{aligned}
$$

To obtain the component scores $f_{i,m}$'s, we use the profile least squared algorithm below.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

Step 1. Obtain the estimation of $\mathbf{s}^{(j)}$, $\mathbf{h}^{(j)}$, $\mathbf{b}^{(j)}$ and $\mathbf{v}^{(j)}$.

By solving the equations $\frac{\partial \phi}{\partial \mathbf{s}^{(j)}} = \mathbf{0}$, $\frac{\partial \phi}{\partial \mathbf{h}^{(j)}} = \mathbf{0}$, $\frac{\partial \phi}{\partial \mathbf{b}^{(j)}} = \mathbf{0}$, $\frac{\partial \phi}{\partial \mathbf{v}^{(j)}} = \mathbf{0}$, and under the condition that $\mathbf{F}'\mathbf{F}/N = \mathbf{I}$ which ensures the identifiability of the component model, we can obtain the estimation below:

$$\begin{aligned}
\mathbf{s}^{(j)} &= N^{-1}\left(\mathbf{I} + \lambda M^{-1}\mathbf{R}^{(j)}\right)^{-1}\boldsymbol{\zeta}^{(j)\prime}\mathbf{F}, && j = 1, \cdots, P;\\
\mathbf{h}^{(j)} &= \mathbf{Z}^{(j)\prime}\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} = N^{-1}\mathbf{Z}^{(j)\prime}\mathbf{F}, && j = 1, \cdots, Q;\\
\mathbf{b}^{(j)} &= N^{-1}\left(\boldsymbol{\zeta}^{(j)\prime}\boldsymbol{\zeta}^{(j)}/N + \rho M^{-1}\mathbf{R}^{(j)}\right)^{-1}\boldsymbol{\zeta}^{(j)\prime}\mathbf{F}, && j = 1, \cdots, P;\\
\mathbf{v}^{(j)} &= \left(\mathbf{Z}^{(j)\prime}\mathbf{Z}^{(j)}\right)^{-1}\mathbf{Z}^{(j)\prime}\mathbf{F}, && j = 1, \cdots, Q.
\end{aligned}$$

Step 2. Obtain the estimation of $f_{i,m}$'s.

Substituting the estimation of $\mathbf{s}^{(j)}$, $\mathbf{h}^{(j)}$, $\mathbf{b}^{(j)}$ and $\mathbf{v}^{(j)}$ above into (2.2) can obtain the expression as follows:

$$\begin{aligned}
\phi = (NT)^{-1}\alpha \sum_{j=1}^{P} \mathrm{tr}\left(\int_{\mathcal{T}_j} \mathrm{x}^{(j)}(t_j)\mathrm{x}^{(j)\prime}(t_j)\mathrm{d}t_j\right) + (NT)^{-1}\alpha \sum_{j=1}^{Q} \mathbf{Z}^{(j)\prime}\mathbf{Z}^{(j)}\\
+ T^{-1}(1-\alpha)(P+Q)M\\
- (NT)^{-1}\mathrm{tr}\left\{N^{-1}\mathbf{F}'\left(\sum_{j=1}^{P}\boldsymbol{\zeta}^{(j)}\mathbf{D}_f^{(j)}\boldsymbol{\zeta}^{(j)\prime} + \sum_{j=1}^{Q}\mathbf{Z}^{(j)}\mathbf{D}_h^{(j)}\mathbf{Z}^{(j)\prime}\right)\mathbf{F}\right\},
\end{aligned} \tag{2.3}$$

where $\mathbf{D}_f^{(j)} = \alpha(\mathbf{I} + \lambda M^{-1}\mathbf{R}^{(j)})^{-1} + (1-\alpha)(\boldsymbol{\zeta}^{(j)\prime}\boldsymbol{\zeta}^{(j)}/N + \rho M^{-1}\mathbf{R}^{(j)})^{-1}$ and $\mathbf{D}_h^{(j)} = \alpha\mathbf{I} + (1-\alpha)(\mathbf{Z}^{(j)\prime}\mathbf{Z}^{(j)}/N)^{-1}$.

Minimizing (2.3) with respect to $\mathbf{F}$ ($\mathbf{F}'\mathbf{F}/N = \mathbf{I}$) thus reduces to max-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

imizing $\mathrm{tr}\left\{N^{-1}\mathbf{F}'\left(\sum_{j=1}^{P}\boldsymbol{\zeta}^{(j)}\mathbf{D}_f^{(j)}\boldsymbol{\zeta}^{(j)\prime}+\sum_{j=1}^{Q}\mathbf{Z}^{(j)}\mathbf{D}_h^{(j)}\mathbf{Z}^{(j)\prime}\right)\mathbf{F}\right\}$, which is further equivalent to consider the eigenvalue decomposition problem below:

$$\sum_{j=1}^{P}\boldsymbol{\zeta}^{(j)}\mathbf{D}_f^{(j)}\boldsymbol{\zeta}^{(j)\prime}+\sum_{j=1}^{Q}\mathbf{Z}^{(j)}\mathbf{D}_h^{(j)}\mathbf{Z}^{(j)\prime}=\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}', \tag{2.4}$$

where $\boldsymbol{\Gamma}'\boldsymbol{\Gamma}=\mathbf{I}$, and $\boldsymbol{\Delta}$ is a diagonal matrix consisting of eigenvalues as elements. Therefore, $\mathbf{F}$ is obtained by $\sqrt{N}$ times the first M eigenvectors of $\boldsymbol{\Gamma}$.

2.2 Choice of the Number of Components for Each Data Source

We can see from the analysis above that the number M_j of univariate principal components for each data source needs to be selected, in this paper, we compared six information criteria below to select them.

$$\begin{aligned}
IC_1(M)&=\ln(\mathrm{MSE})+M\frac{N+T_U}{NT_U}\ln(\frac{NT_U}{N+T_U}),\\
AIC_1(M)&=\ln(\mathrm{MSE})+M\frac{2}{NT_U},\\
AIC_2(M)&=\ln(\mathrm{MSE})+M\frac{2}{T_U},\\
BIC_1(M)&=\ln(\mathrm{MSE})+M\frac{\ln(NT_U)}{NT_U},\\
BIC_2(M)&=\ln(\mathrm{MSE})+M\frac{(N+T_U-M)\ln(NT_U)}{NT_U},\\
BIC_3(M)&=\ln(\mathrm{MSE})+M\frac{(N+T_U-M)\ln(\ln(NT_U))}{NT_U},
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

where MSE denotes the mean squared error, N denotes the size of subjects and T_U denotes the number of time points for each subject. IC_1 and AIC_2 come from Li et al. (2013), AIC_1 and BIC_1 from Yao et al. (2005), and BIC_2 and BIC_3 from Bai and Ng (2002).

2.3 Choice of the Smooth Parameters and Weight Parameter

Step 1. Choice of smooth parameters λ and ρ given α.

The multiple-fold cross-validation criterion is used to select the smooth parameters given weight parameter α. We divide the entire data set into S sub-samples. The cross-validation function (Ramsay and Silverman (2005)) is defined as follows:

$$CV(\lambda,\rho)=\sum_{m=1}^{\infty}\sum_{s=1}^{S}CV_m^{(s)}(\lambda,\rho), \tag{2.5}$$

where

$$\begin{aligned}CV_m^{(s)}(\lambda,\rho)=&(N_sT)^{-1}\Bigg\{\alpha\sum_{j=1}^{P}\sum_{i\in I_s}\int_{\mathcal{T}_j}\left(x_i^{(j)^{(s)}}(t_j)-\mathrm{F}_i{}^{(s)\prime}\hat{\mathrm{a}}^{(j)[s]}(t_j)\right)^2\mathrm{d}t_j\\&+\alpha\sum_{j=1}^{Q}\sum_{i\in I_s}\left(\mathrm{z}_i^{(j)^{(s)\prime}}-\mathrm{F}_i{}^{(s)\prime}\hat{\mathbf{h}}^{(j)[s]\prime}\right)\left(\mathrm{z}_i^{(j)^{(s)}}-\hat{\mathbf{h}}^{(j)[s]}\mathrm{F}_i{}^{(s)}\right)\\&+(1-\alpha)\sum_{j=1}^{P}\sum_{i\in I_s}||\mathrm{F}_i{}^{(s)}-\int_{\mathcal{T}_j}x_i^{(j)^{(s)}}(t_j)\hat{\mathrm{w}}^{(j)[s]}(t_j)\mathrm{d}t_j||^2\\&+(1-\alpha)\sum_{j=1}^{Q}\sum_{i\in I_s}\left(\mathrm{F}_i{}^{(s)\prime}-\mathrm{z}_i^{(j)^{(s)\prime}}\hat{\mathbf{v}}^{(j)[s]}\right)\left(\mathrm{F}_i{}^{(s)}-\hat{\mathbf{v}}^{(j)[s]\prime}\mathrm{z}_i^{(j)^{(s)}}\right)\Bigg\},\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

where $x_i^{(j)^{(s)}}$, $\mathrm{z}_i^{(j)^{(s)}}$ and $\mathrm{F}_i^{(s)}$ denote the ith sample observation of the jth functional data source, the jth high-dimensional data source, and the object score in the sth test set, respectively; I_s is the index set of subjects in the sth test set, N_s is the number of subjects in the sth test set; $\hat{\mathrm{a}}^{(j)^{[s]}}(t_j), \hat{\mathrm{w}}^{(j)^{[s]}}(t_j)$ and $\hat{\mathbf{h}}^{(j)^{[s]}}, \hat{\mathbf{v}}^{(j)^{[s]}}$ denote the estimates which are obtained in Section 2.1 on the corresponding training set consisting of the remaining $N - N_s$ sample observations.

Components F_i's in the sth test set are unknown, this is because they changes with the subjects, while $\mathrm{a}^{(j)}(t_j), \mathrm{w}^{(j)}(t_j), \mathbf{h}^{(j)}$ and $\mathbf{v}^{(j)}$ are irrelevant to the subjects. $\mathrm{a}^{(j)^{[s]}}(t_j), \mathrm{w}^{(j)^{[s]}}(t_j), \mathbf{h}^{(j)^{[s]}}$ and $\mathbf{v}^{(j)^{[s]}}$ and those F_i's of training set can be estimated via the method in Section 2.1, but those F_i's in the test set cannot be obtained directly. Therefore, components $\mathrm{F}_i^{(s)}$'s in the sth test set can be computed by minimizing $CV_m^{(s)}(\lambda, \rho)$ with respect to $\mathrm{F}_i^{(s)}$. Based on the estimated $\mathrm{F}_i^{(s)}$, $CV_m^{(s)}(\lambda, \rho)$ can be obtained.

Hence, for given weight parameter α, the smooth parameters λ and ρ can be estimated by minimizing the cross-validation function $CV(\lambda, \rho)$.

Step 2. Choice of the weight parameter α.

First, let $\alpha = 1$, the model is equivalent to the case where only MPCA is considered. The minimum value of the cross-validation criterion function, denoted as CV_{PCA}, is obtained using the method described in Step 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

Secondly, let $\alpha = 0$, the model is equivalent to the case where only MCCA is considered. Using the above method, the minimal value of the cross-validation criterion function, denoted as CV_{CCA}, can be obtained.

Finally, since α and $1-\alpha$ represent the contribution ratio of MPCA and MCCA techniques to the unified compromise solution, respectively, therefore, the weight parameter α can be estimated by the ratio $\frac{CV_{PCA}}{CV_{PCA}+CV_{CCA}}$.

Step 3. Choice of the smooth parameters λ and ρ.

Once the weight parameter α is estimated, repeating Step 1 can obtain the estimator of the smooth parameters λ and ρ.

3. Component-based Regression Model for Hybrid Data and Component Score Imputation

We consider the following component-based regression model:

$$\begin{cases} y_i \sim EF(\theta_i^M, \varphi), \\ \theta_i^M = \sum_{m=1}^{M} f_{i,m}\beta_m, \end{cases} \tag{3.1}$$

where $EF(\theta_i^M, \varphi)$ means y_i follows the exponential family distribution with parameters θ_i^M and φ.

In the paper, the two-stage procedure is used to estimate the parameters $\{\beta_m, m = 1, \cdots, M\}$. The first stage is to extract components $\{f_{i,m}\}$ based on the method in Section 2. After then maximum likeli-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

hood estimation method is used to obtain the estimators of the parameters $\{\beta_m, m = 1, \cdots, M\}$.

In the component-based regression model (3.1), the number of components M needs to be chosen. As mentioned in Yao et al. (2005), AIC criterion is computationally more efficient while the results are similar to those obtained by cross-validation. Therefore, in the paper, we will determine the number of components by minimizing the AIC function below:

$$AIC(k) = -2 \ln L + 2k,$$

where L denotes the Log-likelihood function of the component-based regression model.

The modelling of component-based regression model above is based on the assumption there is no missing in hybrid data. For block-wise missing hybrid data, in order to extract component scores, we first need to impute the missing blocks. However it is obviously not feasible to impute them directly due to their high dimensionality.

Fortunately, from (2.4) we can see that the component scores of block-wise missing hybrid data can be extracted if both the imputation of univariate principal component scores for functional data and high-dimensional data can be achieved, namely, we need to impute univariate principal component scores for both functional and high-dimensional data sources.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

In this paper, we consider two imputation methods: block-wise conditional mean imputation method (CMI) and multiple block-wise imputation method (MBI, Xue and Qu (2021)), the details are included in Supplementary Material.

4. Theoretical Properties

Assumptions A1-A6 used in Theorems 1 and 2 are concluded in the Supplemental Material. Theorem 1 below gives the convergence rate of estimated components when there is no missing in hybrid data.

Theorem 1. *Under assumptions A1-A6, we have*

$$\tilde{\mathrm{F}}_i - \boldsymbol{H}'\mathrm{F}_i = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right), \quad i = 1, \cdots, N,$$

where $\tilde{\mathrm{F}}_i$ *denotes the estimator of* F_i, $\boldsymbol{H}$ *can be found in Lemma 2 of Supplementary Material.*

Theorem 1 indicates that there exists a rotation matrix $\mathbf{H}$ such that $\tilde{\mathrm{F}}_i$ is an estimator of $\mathbf{H}'\mathrm{F}_i$. In regression analysis, using $\tilde{\mathrm{F}}_i$'s as the regressor gives the same predicted value as using $\mathbf{H}'\mathrm{F}_i$'s.

Theorem 2. *When there are data sources with missing blocks, under the*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

assumptions A1-A6 and $N > \left(\frac{p}{1-p}\right)^2$ *(p is the missing rate), we have*

$$\hat{\mathrm{F}}_k - \hat{\boldsymbol{H}}' \mathrm{F}_k = \begin{cases} O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}, \frac{N-N_{(1)}}{N}\frac{T_{max}^{(mis)}}{\sqrt{NT}}\right\}\right), & k \in G_{(1)}, \\ O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}, \frac{T^{(mis,k)}}{\sqrt{NT}}, \frac{N-N_{(1)}}{N}\frac{T_{max}^{(mis)}}{\sqrt{NT}}\right\}\right), & k \notin G_{(1)}, \end{cases}$$

where $T^{(mis,k)}$ *denotes the number of missing variables for the kth subject,* $T_{max}^{(mis)} = \max_{k=1,\cdots,N}\{T^{(mis,k)}\}$*, 'mis' in* $T^{(mis,k)}$ *and* $T_{max}^{(mis)}$ *means 'missing'.* $\hat{\mathrm{F}}_k$ *denotes the estimator of* F_k*,* $G_{(1)}$ *denotes the index set of subjects included in the complete data group,* $N_{(1)} = |G_{(1)}|$*.*

Theorem 2 implies that the convergence rate of the estimator of F_k when there are data sources with missing blocks becomes a little slow comparing with that without missing in hybrid data as described in Theorem 1, which mainly is due to imputed data rather than the original data used. Furthermore, the convergence rate of the estimator of F_k for the complete-data group is faster than that of the missing pattern groups, this is also consistent with general cognition.

5. Numerical Simulations

To evaluate the performance of our proposed methods: CR-COM (component-based regression without missing), CR-CMI and CR-MBI (component-based regression for block-wise missing data corresponding to CMI and MBI im-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

putation methods), we perform numerical simulations based on 3-source hybrid data, where the details of generating 3-source hybrid data can be found in Supplementary Material. All the simulations and the following real data analysis were conducted in R. Assuming that the number of principal components for each data source is 25 and the number of multi-source principal components is 10, and the sample size $N = 300$ and 600. When coefficients in (3.1) are taken as $\boldsymbol{\beta}$ =(0.934, 0.903, 0.815, 0.604, 0.517, 0.447, 0.392, 0.370, 0.345, 0.3)$'$, and the random errors follow $N(0, 0.2^2)$, the response of linear component-based regression model can be generated. When coefficients in (3.1) are taken as $0\boldsymbol{\beta}$, $2\boldsymbol{\beta}$ and $4\boldsymbol{\beta}$, the 3-classification response of Logistic component-based regression model can be generated. For each simulation, we choose 80%/90% of the samples as training data and the remaining 20%/10% of the samples as testing data. We perform 100 simulation runs for each method.

5.1 Results of Linear Component-based Regression Model

For the case that there are block-wise missing, we first construct the block-wise missing data with missing completely at random (MCAR) such that 80% and 40% of the entire samples are completely observed, and the remaining samples are split into 6 different missing patterns with an equal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

probability. Furthermore, we consider the missing not at random (MNAR) case and construct missing data by a variable δ_i to indicate the missing pattern of the ith sample. The variable δ_i is generated from a multinomial distribution with $P(x_i) = (P_1(x_i), P_2(x_i), \cdots, P_7(x_i))$, where $P_j(x_i) = \frac{|g_j(x_i)|}{\sum_{l=1}^{7} |g_l(x_i)|}$ and $g_j(x_i) =< \gamma_j, x_i >$ for $j = 1, \cdots, 7$. Let $\gamma_1 = (1, \cdots, 1)$ and $\gamma_2 = \cdots = \gamma_7 = (\frac{1}{24}, \cdots, \frac{1}{24})$, then the overall missing rate can be guaranteed to be 20%. Let $\gamma_1 = (1, \cdots, 1)$ and $\gamma_2 = \cdots = \gamma_7 = (\frac{1}{4}, \cdots, \frac{1}{4})$, then the overall missing rate can be 60%.

Based on the method in Sections 2 and 3, we obtained the estimated values of the weight parameter, the smooth parameters and the number of components. For all cases, the estimated weight parameter is close to 0.5 (the true weight); the estimated number of components is close to the true value 10; the estimated smooth parameters are both 10. Mean and standard deviation (SD, in parentheses) of mean squared errors (MSE) of estimated component score $\hat{\mathbf{F}}$, estimated regression coefficient $\hat{\boldsymbol{\beta}}$, and estimated outcome $\hat{\mathbf{Y}}$ for different sample sizes, missingness mechanisms and missing rates are shown in Table 1, respectively. We compare also the performance of our proposed CR-COM method with the ERA model. In the ERA model, we performed functional PCA and multivariate PCA for two functional data sources and one high-dimensional data source, respectively,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

5.1 Results of Linear Component-based Regression Model

Table 1: Results of linear component-based regression model.

Missing Mechenism	Estimation Method	N=300 MSE($\hat{\mathbf{F}}$)	N=300 MSE($\hat{\boldsymbol{\beta}}$)	N=300 MSE($\hat{\mathbf{Y}}$)	N=300 AFIT	N=600 MSE($\hat{\mathbf{F}}$)	N=600 MSE($\hat{\boldsymbol{\beta}}$)	N=600 MSE($\hat{\mathbf{Y}}$)	N=600 AFIT
No missing	CR-COM	0.00207	0.00025	0.03858	**0.98613**	0.00209	0.00021	0.03920	**0.98604**
		(0.00029)	(0.00009)	(0.00345)	(0.00117)	(0.00015)	(0.00006)	(0.00191)	(0.00077)
	ERA			**0.03683**	0.98344			**0.03874**	0.98497
				(0.00329)	(0.00145)			(0.00229)	(0.00095)
Missing rate= 20%									
MCAR	CR-ZERO	0.57678	0.13296	0.19738		0.33517	0.05698	0.19727	
		(0.17920)	(0.07617)	(0.03263)		(0.13699)	(0.04069)	(0.02176)	
	CR-KNN	0.11163	0.01468	0.07283		0.05185	0.00476	0.06487	
		(0.06890)	(0.01478)	(0.00942)		(0.02001)	(0.00356)	(0.00639)	
	CR-CMI	0.00217	0.00027	0.03834		0.00216	0.00021	0.03940	
		(0.00031)	(0.00009)	(0.00350)		(0.00013)	(0.00007)	(0.00232)	
	CR-MBI	**0.00217**	**0.00026**	**0.03836**		**0.00215**	**0.00021**	**0.03941**	
		(0.00031)	(0.00009)	(0.00349)		(0.00013)	(0.00007)	(0.00231)	
MNAR	CR-ZERO	0.54895	0.11893	0.20475		0.36109	0.06011	0.21707	
		(0.17440)	(0.06540)	(0.03038)		(0.13521)	(0.03660)	(0.02213)	
	CR-KNN	0.13012	0.01400	0.07966		0.04949	0.00374	0.06639	
		(0.07833)	(0.01060)	(0.01110)		(0.01266)	(0.00253)	(0.00529)	
	CR-CMI	0.00214	0.00027	0.03866		0.00213	0.00020	0.03954	
		(0.00030)	(0.00009)	(0.00358)		(0.00015)	(0.00007)	(0.00228)	
	CR-MBI	**0.00213**	**0.00027**	**0.03866**		**0.00213**	**0.00020**	**0.03954**	
		(0.00030)	(0.00009)	(0.00358)		(0.00016)	(0.00007)	(0.00228)	
Missing rate=60%									
MCAR	CR-ZERO	0.80148	0.23586	0.45651		0.57705	0.16677	0.43895	
		(0.15864)	(0.15154)	(0.03714)		(0.16101)	(0.09783)	(0.02798)	
	CR-KNN	0.17257	0.02229	0.14578		0.10852	0.00703	0.12385	
		(0.04805)	(0.02339)	(0.01795)		(0.01627)	(0.00422)	(0.01289)	
	CR-CMI	0.00229	0.00028	0.03841		0.00223	0.00022	0.03976	
		(0.00041)	(0.00010)	(0.00379)		(0.00025)	(0.00007)	(0.00232)	
	CR-MBI	**0.00228**	**0.00028**	**0.03840**		**0.00223**	**0.00022**	**0.03976**	
		(0.00042)	(0.00010)	(0.00379)		(0.00025)	(0.00007)	(0.00232)	
MNAR	CR-ZERO	0.84743	0.29347	0.46054		0.60803	0.17866	0.44514	
		(0.16459)	(0.18065)	(0.03698)		(0.17960)	(0.12613)	(0.02786)	
	CR-KNN	0.18305	0.02316	0.14228		0.10576	0.00642	0.12283	
		(0.07389)	(0.01840)	(0.01810)		(0.01509)	(0.00358)	(0.00943)	
	CR-CMI	0.00225	0.00029	0.03879		0.00217	0.00022	0.03975	
		(0.00044)	(0.00009)	(0.00361)		(0.00024)	(0.00007)	(0.00234)	
	CR-MBI	**0.00225**	**0.00029**	**0.03883**		**0.00217**	**0.00022**	**0.03975**	
		(0.00044)	(0.00008)	(0.00355)		(0.00024)	(0.00007)	(0.00234)	

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

and each extracted 25 principal components, so that a total of 75+3 parameters were included in the model. The results can also be found in Table 1, where AFIT denotes the adjusted goodness-of-fit.

For the case without missing data, as the sample size increases, the mean of MSE($\hat{\mathbf{F}}$) slightly increases, but the standard deviation significantly decreases; the mean and SD of MSE($\hat{\boldsymbol{\beta}}$) slightly decrease; the mean of MSE($\hat{\mathbf{Y}}$) slightly increases, but SD significantly decreases. All these indicate that the estimators are consistent. From Table 1 we can also see that the MSE ($\hat{\mathbf{Y}}$) of the ERA model is generally smaller than that of the CR-COM method, which implies that the ERA model is slightly superior to the CR (component-based regression) model from a prediction point of view. However, the AFIT indicator values of the ERA model are lower than those of the CR model, the reason is perhaps that the ERA model is much more complex than the CR model (ERA model: 78 parameters; CR model: 10 parameters), which will make it less capable of generalisation.

Next we consider the case with block-wise missing in the data. For fixed sample size, as the missing rate increases, both means and SDs of MSE($\hat{\mathbf{F}}$) and MSE($\hat{\mathbf{Y}}$) increase; means of MSE($\hat{\boldsymbol{\beta}}$) slightly increase, while SDs remain almost unchanged. All these indicates that as the missing rate increases, the estimation will slightly become worse. For fixed missing rate, as the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

sample size increases, means and SDs of both MSE($\hat{\mathbf{F}}$) and MSE($\hat{\boldsymbol{\beta}}$) decrease; means of MSE($\hat{\mathbf{Y}}$) slightly increase, but SDs significantly decreases, these results indicates that the estimates are consistent. In addition, for different missingness mechanisms, there was no significant difference in the performance of all methods. We also consider the zero mean imputation method and the K-nearest neighbors imputation method for comparison, and denote them as CR-ZERO and CR-KNN, respectively. It is clear that in all cases the CR-ZERO method performs the worst, the CR-KNN is slightly better, and the CR-CMI and CR-MBI methods perform the best with the smallest means and standard deviations. This result is in perfect agreement with our perception, as the CMI and MBI imputation methods are essentially regression-based methods that can extract information much more from the data than the zero mean and KNN imputation methods.

5.2 Results of Logistic Component-based Regression Model

We first construct the block-wise missing data with MCAR as in Section 5.1. And, we also consider the MNAR case by a variable δ_i which is generated from a multinomial distribution with $P(x_i|y_i) = (P_1(x_i|y_i), \cdots, P_7(x_i|y_i))$, where $P_j(x_i|y_i) = \frac{|g_j(x_i|y_i)|}{\sum_{l=1}^{7}|g_l(x_i|y_i)|}$ and $g_j(x_i|y_i) =< \gamma_j|y_i, x_i >$ for $j = 1, 2, \cdots, 7$. For $N = 300$ or 600, because the ratio of the three categories of the de-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

5.2 Results of Logistic Component-based Regression Model

Table 2: Results of Logistic component-based regression model.

	MCAR						MNAR					
Method	Accuracy	Precision	Recall	F_1-score	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)	Accuracy	Precision	Recall	F_1-score	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)
(N, Missing rate)=(300, 0%)												
CR-COM	0.8481	0.7501	0.6909	0.6949	0.1503	0.2927						
	(0.0212)	(0.0494)	(0.0292)	(0.0371)	(0.1606)	(0.2882)						
(N, Missing rate)=(300, 20%)												
CR-ZERO	0.8495	0.7496	0.6995	0.7044	1.0163	3.8588	0.8495	0.7531	0.7011	0.7081	1.0308	3.6242
	(0.0211)	(0.0479)	(0.0344)	(0.0431)	(0.5950)	(2.2864)	(0.0211)	(0.0485)	(0.0300)	(0.0361)	(0.7204)	(2.4708)
CR-KNN	0.8415	0.7389	0.6870	0.6897	0.2115	0.5037	0.8460	0.7414	0.6921	0.6959	0.2132	0.5413
	(0.0234)	(0.0576)	(0.0337)	(0.0427)	(0.1631)	(0.3840)	(0.0220)	(0.0514)	(0.0327)	(0.0410)	(0.1780)	(0.4600)
CR-CMI	0.8480	0.7493	0.6954	0.7009	0.1597	0.3351	0.8479	0.7521	0.6940	0.6975	0.1567	0.3427
	(0.0212)	(0.0484)	(0.0312)	(0.0387)	(0.1682)	(0.3836)	(0.0212)	(0.0454)	(0.0308)	(0.0394)	(0.1610)	(0.3913)
CR-MBI	0.8478	0.7486	0.6950	0.7004	0.1593	0.3307	0.8477	0.7512	0.6927	0.6957	0.1557	0.3401
	(0.0215)	(0.0497)	(0.0328)	(0.0401)	(0.1687)	(0.3831)	(0.0214)	(0.0466)	(0.0308)	(0.0395)	(0.1593)	(0.3915)
(N, Missing rate)=(300, 60%)												
CR-ZERO	0.8404	0.7336	0.6877	0.6925	2.0598	7.6143	0.8403	0.7326	0.6911	0.6961	1.9421	7.5053
	(0.0220)	(0.0531)	(0.0322)	(0.0388)	(1.2799)	(4.8052)	(0.0227)	(0.0467)	(0.0321)	(0.0387)	(1.3166)	(4.8080)
CR-KNN	0.8347	0.7228	0.6782	0.6834	0.2527	0.6659	0.8354	0.7366	0.6827	0.6876	0.2457	0.6922
	(0.0213)	(0.0526)	(0.0316)	(0.0372)	(0.2564)	(0.4882)	(0.0226)	(0.0555)	(0.0305)	(0.0391)	(0.1838)	(0.5188)
CR-CMI	0.8474	0.7416	0.6897	0.6927	0.1804	0.3525	0.8472	0.7490	0.6923	0.6961	0.1725	0.3706
	(0.0227)	(0.0479)	(0.0308)	(0.0397)	(0.2434)	(0.5611)	(0.0226)	(0.0551)	(0.0340)	(0.0434)	(0.2377)	(0.5882)
CR-MBI	0.8479	0.7411	0.6892	0.6921	0.1742	0.3770	0.8476	0.7475	0.6928	0.6969	0.1680	0.3693
	(0.0229)	(0.0529)	(0.0312)	(0.0407)	(0.2314)	(0.5781)	(0.0222)	(0.0541)	(0.0332)	(0.0431)	(0.2319)	(0.5846)
(N, Missing rate)=(600, 0%)												
CR-COM	0.8454	0.7170	0.6676	0.6638	0.0564	0.0973						
	(0.0149)	(0.0350)	(0.0193)	(0.0275)	(0.0432)	(0.0819)						
(N, Missing rate)=(600, 20%)												
CR-ZERO	0.8407	0.7151	0.6664	0.6618	0.4170	1.5898	0.8406	0.7166	0.6681	0.6648	0.3856	1.4192
	(0.0146)	(0.0448)	(0.0188)	(0.0264)	(0.2806)	(1.0777)	(0.0167)	(0.0384)	(0.0191)	(0.0269)	(0.2574)	(0.8597)
CR-KNN	0.8350	0.7050	0.6595	0.6535	0.0697	0.1642	0.8378	0.7178	0.6656	0.6621	0.0712	0.1749
	(0.0147)	(0.0401)	(0.0176)	(0.0250)	(0.0616)	(0.1022)	(0.0150)	(0.0351)	(0.0184)	(0.0259)	(0.0483)	(0.1140)
CR-CMI	0.8439	0.7212	0.6690	0.6662	0.0584	0.1050	0.8436	0.7217	0.6694	0.6662	0.0613	0.1081
	(0.0154)	(0.0394)	(0.0188)	(0.0270)	(0.0539)	(0.0954)	(0.0155)	(0.0416)	(0.0193)	(0.0271)	(0.0505)	(0.0957)
CR-MBI	0.8436	0.7213	0.6691	0.6663	0.0588	0.1059	0.8435	0.7218	0.6698	0.6669	0.0614	0.1098
	(0.0156)	(0.0397)	(0.0188)	(0.0272)	(0.0557)	(0.0974)	(0.0156)	(0.0422)	(0.0199)	(0.0280)	(0.0504)	(0.0981)
(N, Missing rate)=(600, 60%)												
CR-ZERO	0.8394	0.7174	0.6679	0.6665	1.0168	4.0761	0.8332	0.7049	0.6648	0.6635	1.0606	4.1531
	(0.0152)	(0.0407)	(0.0205)	(0.0285)	(0.6785)	(2.7304)	(0.0164)	(0.0395)	(0.0200)	(0.0268)	(0.6729)	(2.5781)
CR-KNN	0.8275	0.7021	0.6551	0.6506	0.0868	0.2265	0.8262	0.7050	0.6582	0.6553	0.0881	0.2415
	(0.0146)	(0.0398)	(0.0180)	(0.0258)	(0.0561)	(0.1173)	(0.0160)	(0.0377)	(0.0186)	(0.0263)	(0.0426)	(0.1253)
CR-CMI	0.8433	0.7186	0.6667	0.6621	0.0728	0.1266	0.8430	0.7195	0.6723	0.6700	0.0692	0.1299
	(0.0159)	(0.0346)	(0.0175)	(0.0256)	(0.0689)	(0.1178)	(0.0157)	(0.0447)	(0.0232)	(0.0317)	(0.0695)	(0.1191)
CR-MBI	0.8430	0.7177	0.6660	0.6610	0.0710	0.1269	0.8432	0.7237	0.6727	0.6703	0.0686	0.1331
	(0.0158)	(0.0358)	(0.0179)	(0.0260)	(0.0677)	(0.1199)	(0.0159)	(0.0412)	(0.0220)	(0.0302)	(0.0691)	(0.1189)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

pendent variable is 13:4:13, we can generate the block-wise missing data in each category by using the method of MCAR. Assume that m_1, m_2 and m_3 are the missing rates for each category, respectively. When $m_1 = m_3 = \frac{2.8}{13}, m_2 = 0.1$, for the first and the third categories, let $\gamma_1 = (1, \cdots, 1), \gamma_2 = \cdots = \gamma_7 = (\frac{7}{153}, \cdots, \frac{7}{153})$; for the second category, let $\gamma_1 = (1, \cdots, 1), \gamma_2 = \cdots = \gamma_7 = (\frac{1}{54}, \cdots, \frac{1}{54})$, then the overall missing rate can arrive 20%. When $m_1 = m_3 = \frac{8.4}{13}, m_2 = 0.3$, for the first and the third categories, let $\gamma_1 = (1, \cdots, 1), \gamma_2 = \cdots = \gamma_7 = (\frac{7}{23}, \cdots, \frac{7}{23})$; for the second category, let $\gamma_1 = (1, \cdots, 1), \gamma_2 = \cdots = \gamma_7 = (\frac{1}{14}, \cdots, \frac{1}{14})$, then the overall missing rate can arrive 60%.

Several metrics for evaluating the performance of classification together with the MSEs of $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ (the estimator of regression coefficients) are shown in Table 2.

From Table 2, we can see that for the case without missing data, as the sample size increases, means of various indicators used to evaluate classification performance in 100 simulations decrease slightly, but the SDs (in parentheses) decrease significantly, resulting in a shorter confidence interval; means and SDs of MSE($\hat{\boldsymbol{\beta}}_2$) and MSE($\hat{\boldsymbol{\beta}}_3$) decrease significantly. These results mean that the classification performance becomes better for larger sample size. Overall, the CR-KNN method performed the worst, CR-ZERO

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

slightly better, while the CR-CMI and CR-MBI methods performed the best, which is consistent with the results in Section 5.1.

We next consider the case with block-wise missing data. For fixed sample size, as the missing rate increases, although means and SDs of MSE($\hat{\boldsymbol{\beta}}_2$) and MSE($\hat{\boldsymbol{\beta}}_3$) have slightly increased, but the classification indicators have remained basically unchanged. It can be found that the prediction performance of the imputation methods is almost the same as that of without missing data, which is an advantage of our methods. For fixed missing rate, means of various classification indicators increase slightly as the sample size increases, but SDs decrease significantly. Therefore, the larger the sample size, the better the results. Means and SDs of MSE($\hat{\boldsymbol{\beta}}_2$) and MSE($\hat{\boldsymbol{\beta}}_3$) are significantly smaller as the sample size increases. In addition, for different missingness mechanisms, there is no significant difference in the performance of all methods.

6. Application to ADNI Study

The detailed description of ADNI Data is contained in Supplementary Material. The data in this paper with four data sources (PET, MRI, GENE and CSF) and ten missing patterns consisted of 687 subjects. The complete data group consisting of 190 subjects accounted for 27.7% of the total

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

sample. The total missing rate is approximately 72.3%. The specific data structure is shown in Table 3.

Table 3: Data aggregation. "o": the observed data, "-": the missing data.

Missing pattern	PET	MRI	GENE	CSF	# of subjects
I	o	o	o	o	190
II	o	o	o	-	22
III	o	o	-	o	104
IV	o	-	o	o	147
V	-	o	o	o	60
VI	o	o	-	-	11
VII	o	-	o	-	25
VIII	-	o	o	-	5
IX	o	-	-	o	117
X	-	o	-	o	6
# of subjects	616	398	449	624	687
Missing rate	10.3%	42.1%	34.6%	9.2%	72.3%

In order to establish the component-based regression models, we firstly need to implement univariate fPCA on PET images and MRI images, and univariate PCA on GENE data, respectively. We can choose univariate principal component scores by comparing the six information criteria in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

Section 2.2. We finally choose the optimal number of univariate principle component scores (235 for MRI, 100 for PET and 295 for GENE) by the BIC_3 criterion.

6.1 Prediction and Classification Results

To evaluate the performance of our method, 80% or 90% of the data is layered into a training set, and the remaining data is used as a verification set. The random sampling was repeated 100 times. By applying the methods in Section 2, the estimated weight parameter is 0.41, and the estimated smooth parameters are both 0.01.

We first established linear component-based regression models. The number of components of the model we choose based on AIC is 88. We also compared our proposed methods with the FR-FI method in Zhang et al. (2020). In Zhang et al. (2020), they used features from four high-dimensional data sources: 3 CSF features, 243 PET features, 317 MRI features, and 49,386 gene features. The mean and standard deviation (SD) of MSE of predicted MMSE are shown in Table 4. It can be seen from Table 4 that, for each method, means of MSE at two training rates are relatively close, but SDs become larger as the training rate increases. The mean and SD of MSE by FR-FI method are larger than those proposed by us. The

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

performance of the CR-CMI and CR-MBI methods is similar, while the CR-KNN and CR-ZERO methods are slightly worse, which is consistent with the simulation study.

Table 4: Mean and standard deviation of MSE for ADNI data.

	80% training rate		90% training rate	
Method	Mean	SD	Mean	SD
CR-ZERO	4.9944	1.1072	4.8462	1.6378
CR-KNN	5.0084	1.0740	4.7440	1.4339
CR-CMI	**4.7885**	0.6814	**4.6671**	0.9385
CR-MBI	**4.8141**	0.6364	**4.8008**	1.0826
FR-FI	5.5765	1.0511	5.4931	1.6009

We also established Logistic component-based regression models for 3-classification and 4-classification cases, results are shown in Table 5. From Table 5, we can see that the performance of the CMI method and MBI method is similar, while the CR-KNN and CR-ZERO methods are slightly worse. In addition, regardless of the training rate, the classification performance of 3-classification case is generally better than that of 4-classification case. This may be because it is difficult to distinguish the subjects of the early mild cognitive impairment group (EMCI) and the late mild cognitive

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

impairment group (LMCI), so combining the two groups into a single category of mild cognitive impairment group (MCI) can improve the accuracy of classification. In the actual classification process, we neither want normal

Table 5: 4-classification and 3-classification results of proposed methods.

	80% training rate				90% training rate			
Method	Accuracy	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score
AD/LMCI/EMCI/CN 4-classification results								
CR-ZERO	0.8590	0.8512	0.8601	0.8534	0.9188	0.9170	0.9208	0.9166
CR-KNN	0.8648	0.8582	0.8664	0.8597	0.9261	0.9251	0.9262	0.9234
CR-CMI	**0.8700**	0.8638	0.8683	0.8640	**0.9362**	0.9340	0.9359	0.9330
CR-MBI	**0.8679**	0.8603	0.8641	0.8603	**0.9350**	0.9347	0.9326	0.9315
AD/MCI/CN 3-classification results								
CR-ZERO	0.8528	0.8314	0.8547	0.8398	0.8929	0.8880	0.8963	0.8895
CR-KNN	0.8877	0.8643	0.8849	0.8722	0.9388	0.9254	0.9388	0.9303
CR-CMI	**0.8905**	0.8664	0.8828	0.8724	**0.9469**	0.9349	0.9459	0.9382
CR-MBI	**0.8958**	0.8712	0.8900	0.8784	**0.9473**	0.9373	0.9446	0.9389

people to be classified into patient categories, nor do we want any patient to be missed, so we pay more attention to the F_1-score which comprehensively considers the precision and the recall. We noticed that the precision and recall of the two methods are very close regardless of the training rate, so the F_1-score is relatively stable and close to the accuracy.

We compare further the classification performance of our proposed CR-CMI method with the iMSF method proposed in Yuan et al. (2012) which considered high-dimensional feature including MRI, PET, proteomics and CSF for distinguishing ADNI subjects into 3 diagnostic groups (AD, MCI

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

and CN). We turn the 3-classification task into three pairwise classification problems, namely AD/CN, MCI/AD and CN/MCI, classification results are shown in Table 6. However, the PET data and MRI data we used are infinite-dimensional, so there is no way to directly apply the iMSF method to our data. Therefore, the results of iMSF method in Table 6 are directly extracted from Yuan et al. (2012).

Table 6: 2-classification comparison of CR-CMI method and the iMSF method.

Metrics	Training rate	CR-CMI	iMSF	CR-CMI	iMSF	CR-CMI	iMSF
		AD/CN task		MCI/AD task		CN/MCI task	
Accuracy	50.0%	**0.8965**	0.8658	**0.9151**	0.8278	**0.8828**	0.8872
	66.7%	**0.9580**	0.8890	**0.9629**	0.8335	**0.9494**	0.9033
	75.0%	**0.9752**	0.8848	**0.9777**	0.8401	**0.9671**	0.8927
Sensitivity	50.0%	**0.8568**	0.8552	**0.8474**	0.4339	**0.8596**	0.6228
	66.7%	**0.9390**	0.8706	**0.9330**	0.4424	**0.9372**	0.6922
	75.0%	**0.9692**	0.8667	**0.9648**	0.4514	**0.9547**	0.7162
Specificity	50.0%	**0.9163**	0.879	**0.9333**	0.9628	**0.8952**	0.9907
	66.7%	**0.9673**	0.9142	**0.9708**	0.9643	**0.9560**	0.9934
	75.0%	**0.9782**	0.9102	**0.9813**	0.9670	**0.9738**	0.9949

For AD/CN classification task, the performance of the CR-CMI method is overall better slightly than of the iMSF method. In addition, under this classification task, the sensitivity and specificity of the two methods are relatively balanced. Generally speaking, as the training rate increases, there will be some improvement in classification performance.

For the MCI/AD classification task, CR-CMI method performs better than the iMSF method in classifying and predicting MCI patients and AD

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

patients with similar disease courses. It is also worth noting that under this classification task, although the iMSF method has a high specificity, its sensitivity is particularly low. However, in the field of medicine, sensitivity is more important because the consequence of diagnosing a patient as a healthy person is more serious than that of diagnosing a healthy person as a patient. Therefore, compared to the highly unbalanced sensitivity and specificity of the iMSF method, our method has the advantage of greatly improving sensitivity at the expense of a small portion of specificity.

For CN/MCI task, we can see that when the training rate is 1/2, the accuracy of the CR-CMI method is close to that of the iMSF method; when the training rate increases to 2/3 and 3/4, the accuracy of the CR-CMI method will exceed that of the iMSF method. The sensitivity of the CR-CMI method is superior to that of the iMSF method, while the specificity is slightly inferior to that of the iMSF method.

7. Conclusions

In this paper, we propose a new component-based regression model for hybrid data containing functional data and high-dimensional data, where components are constructed by the unified approach to MCCA and MPCA. We also propose two component imputation methods (CMI and MBI) to

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

impute the block-wise missing univariate principal component scores. Theoretical properties, numerical simulations and real data analysis show that the prediction and classification performance of the new method work well.

The main contributions of our methods are as follows:

(i) Hwang et al. (2013) proposed a unified approach to MCCA and MPCA for multi-source high-dimensional data, Choi et al. (2017) extended it to the case of multi-source functional data. In the paper, we extended further the unified approach to hybrid data.

(ii) Since functional data are contained in hybrid data, the objective function which is used to extract the components involves the integral operation of original functional data, the loading functions and canonical weight functions. In our paper, we proposed to use the principal component basis of the corresponding univariate functional data as the basis functions, thereby converting functional optimization problem to an approximately equivalent matrix eigen-analysis problem.

(iii) Since the new component model is established by univariate principal component scores of each functional data source and high-dimensional data source, and the calculation of univariate principal component scores is not affected by whether the data are balanced or not, sparse or dense, therefore the applicability of our proposed method is very broad.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

(iv) The proposed method is suitable not only for hybrid data without missing but also for the data with block-wise missing. Instead of directly imputing the original data, we first imputed the corresponding univariate principal component scores by block-wise conditional mean imputation and multiple block-wise imputation methods, thereby achieving imputation of components, which reduces the computational complexity greatly.

However, the method we proposed still needs further research. Firstly, our method is theoretically applicable to multiple data sources, but only three or four data sources are considered in simulation research and empirical analysis. In the future, we can introduce more influencing factors, such as clinical data, plasma data, protein data, etc., which may improve the accuracy of prediction further. Secondly, when we select image data, we only select those with the same size, for example, in MRI images, we only select those with a size of $176 \times 240 \times 256$, but in fact there are various other sizes of MRI images. In subsequent research, we can also use registration techniques to obtain more data, thereby improving utilization efficiency. Finally, our approach considers the factor model and regression model separately, which can be seen as a two-stage approach. Since some information may be lost by discussing them separately, the accuracy of the models may be affected to some extent, so we may further consider combining the fac-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

tor and regression models together to achieve dimensionality reduction and regression as well, thus improving the accuracy of the estimation. In the future, we are going to further investigate the Bayesian estimation method and the method in the ERA model for hybrid data.

Supplementary Material

Supplementary materials available in the attached file include generating hybrid data, imputation methods, technical conditions and proofs.

Acknowledgements

This research is supported by the National Social Science Foundation of China (No.21BTJ044).

References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70(1)**, 191–221.

Bai, J. S. and Li, K. P. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40(1)**, 436–465.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

REFERENCES

Berrendero, J. R., Justel, A. and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis* **55(9)**, 2619–2634.

Campos, S., Pizarro, L., Valle, C., Gray, K. R., Rueckert, D. and Allende, H. (2015). Evaluating imputation techniques for missing data in ADNI: a patient classification study. *Iberoamerican Congress on Pattern Recognition. Springer, Cham, Switzerland* **9423**, 3–10.

Chiou, J. M., Chen, Y. T. and Yang, Y. F.(2014). Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica* **24**, 1571–1596.

Choi, J. Y., Kyung, M., Hwang, H. and Park, J-H. (2020). Bayesian extended redundancy analysis: a Bayesian approach to component-based regression with dimension reduction. *Multivariate Behavioral Research* **55(1)**, 30–48.

Choi, J. Y., Hwang, H., Yamamoto, M., Jung, K. and Woodward, T.S. (2017). A unified approach to functional principal component analysis and functional multiple-set canonical correlation. *Psychometrika* **82**, 427–441.

Correa, N. M., Eichele, T., Adali, T., Li, Y. and Calhoun, V. D. (2010). Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. *NeuroImage* **50**, 1438–1445.

Du, X. L., Jiang, X. H., Lin, J. G. and Alzheimer's Disease Neuroimaging Initiative (2023). Multinomial Logistic factor regression for multi-source functional block-wise missing data. *Psychometrika* **88 (3)**, 975–1001.

Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

Econometrics **201**, 292–306.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* **113(522)**, 649–659.

Hwang, H., Jung, K., Takane, Y. and Woodward, T.S.(2012). Functional multiple-set canonical correlation analysis. *Psychometrika* **77**, 48–64.

Hwang, H., Jung, K., Takane, Y. and Woodward, T.S.(2013). A unified approach to multiple-set canonical correlation analysis and principal components analysis. *British Journal of Mathematical & Statistical Psychology* **66(2)**, 308–321.

Hwang, H., Suk, H.W., Lee, J.H., Moskowitz, D.S. and Lim, J.(2012). Functional extended redundancy analysis. *Psychometrika* **77**, 524–542.

Hwang, H., Suk, H.W., Takane, Y., Lee, J.H. and Lim, J.(2015). Generalized functional extended redundancy analysis. *Psychometrika* **80**,101–125.

Kim, S. and Hwang, H. (2021). Model-based recursive partitioning of extended redundancy analysis with an application to nicotine dependence among US adults. *British Journal of Mathematical and Statistical Psychology* **74(3)**, 567–590.

Kim, S. and Hwang, H. (2022). Evaluation of prediction-oriented model selection metrics for extended redundancy analysis. *Frontiers In Psychology* **13**, 821–897.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

functional data. *Journal of the American Statistical Association* **108**, 1284–1294.

Park, J-H., Choi, J. Y., Lee, J., Kyung, M. (2021). Bayesian approach to multivariate component-based Logistic regression: analyzing correlated multivariate ordinal data. *.Multivariate Behavioral Research* **57(4)**, 543–560.

Ramsay, J. O. and Silverman, B. W.(2005). Functional data analysis. Springer, Berlin.

Takane, Y., Hwang, H. and Abdi, H.(2008). Regularized multiple-set canonical correlation analysis. *Psychometrika* **73**, 753–775.

Takane, Y. and Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics & Data Analysis* **49(3)**, 785–808.

Tan, T., Choi, J. Y. and Hwang, H.(2015). Fuzzy clusterwise functional extended redundancy analysis. *Behaviormetrika* **42(1)**, 37–62.

Tenenhaus, M., Tenenhaus, A. and Groenen, P. J. F. (2017). Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika* **82**, 737–777.

Vijayakumar, R., Choi, J. Y. and Jung, E. H. (2022). A unified neural network framework for extended redundancy analysis. *Psychometrika* **87**, 1503–1528.

Wan, Y., Datta, S., Conklin, D. J. and Kong, M. (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation* **85(9)**, 1902–1916.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

REFERENCES

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J. and Alzheimer's Disease Neuroimaging Initiative. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage* **102**, 192–206.

Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association* **116(536)**, 1914–1927.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100(470)**, 577–590.

Yamashita, N. (2023). Exploratory extended redundancy analysis using sparse estimation and oblique rotation of parameter matrices. *Behaviormetrika* **50**, 679–697.

Yu, G., Li, Q., Shen, D. and Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association* **115(531)**, 1406–1419.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J. and Alzheimer's Disease Neuroimaging Initiative (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage* **61(3)**, 622–632.

Zhang, Y., Tang, N. and Qu, A. (2020). Imputed factor regression for high-dimensional block-wise missing data. *Statistica Sinica* **30(2)**, 631–651.

Xiaohu Jiang

Department of Statistics, Ynnnan University, Kunming, China, 650504

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0186

E-mail: (jjjxhmail@163.com)

Xiuli Du

College of Mathematical Sciences, Nanjing Normal University, Nanjing, China, 210023

E-mail: (duxiuli@njnu.edu.cn)

Yenan Ren

College of Mathematical Sciences, Nanjing Normal University, Nanjing, China, 210023

E-mail: (ryn_dream123@163.com)

Jinguan Lin

Institute of Statistics and Data Science, Nanjing Audit University, Nanjing, China, 211815

E-mail: (101010818@seu.edu.cn)

The Alzheimer's Disease Neuroimaging Initiative