https://doi.org/10.1093/biomet/asac007

Functional hybrid factor regression model for handling heterogeneity in imaging studies

By C. HUANG

Department of Statistics, Florida State University,
117 N. Woodward Ave., Tallahassee, Florida 32304, U.S.A.
chaohuang@stat.fsu.edu

and H. ZHU

Department of Biostatistics, The University of North Carolina at Chapel Hill,
135 Dauer Drive, Chapel Hill, North Carolina 27599, U.S.A.
htzhu@email.unc.edu

Summary

This paper develops a functional hybrid factor regression modelling framework to handle the heterogeneity of many large-scale imaging studies, such as the Alzheimer's disease neuroimaging initiative study. Despite the numerous successes of those imaging studies, such heterogeneity may be caused by the differences in study environment, population, design, protocols or other hidden factors, and it has posed major challenges in integrative analysis of imaging data collected from multicentres or multistudies. We propose both estimation and inference procedures for estimating unknown parameters and detecting unknown factors under our new model. The asymptotic properties of both estimation and inference procedures are systematically investigated. The finite-sample performance of our proposed procedures is assessed by using Monte Carlo simulations and a real data example on hippocampal surface data from the Alzheimer's disease study.

Some key words: Alzheimer's disease; Functional hybrid factor regression model; Hippocampal surface; Imaging heterogeneity; Surrogate variable analysis.

1. Introduction

With the rapid growth of modern technology, many large-scale imaging studies, such as the Alzheimer's disease neuroimaging initiative, ADNI, study (Mueller et al., 2005), the Human Connectome Project (Van Essen et al., 2013) and the UK Biobank study (Sudlow et al., 2015), have been conducted to collect massive datasets with large volumes of complex information from increasingly large cohorts for unravelling the etiology of different diseases, such as Alzheimer's disease. For example, the ADNI study is a multi-phase study that aims to discover the progression of Alzheimer's disease and improve clinical trials for the prevention and treatment of Alzheimer's disease. However, such integrative data analysis is challenging largely due to the heterogeneity in those imaging studies, since the datasets are often collected from different centres and/or phases and need to be rigorously integrated (Lock et al., 2013; Yu et al., 2017). The potential heterogeneity may be caused by the differences in study environment, population (e.g., race), design, protocols (e.g., imaging acquisition protocol) and some other (unknown) hidden factors in multiple centres and/or phases (Leek & Storey, 2007; Mirzaalian et al., 2016; Fortin et al.,

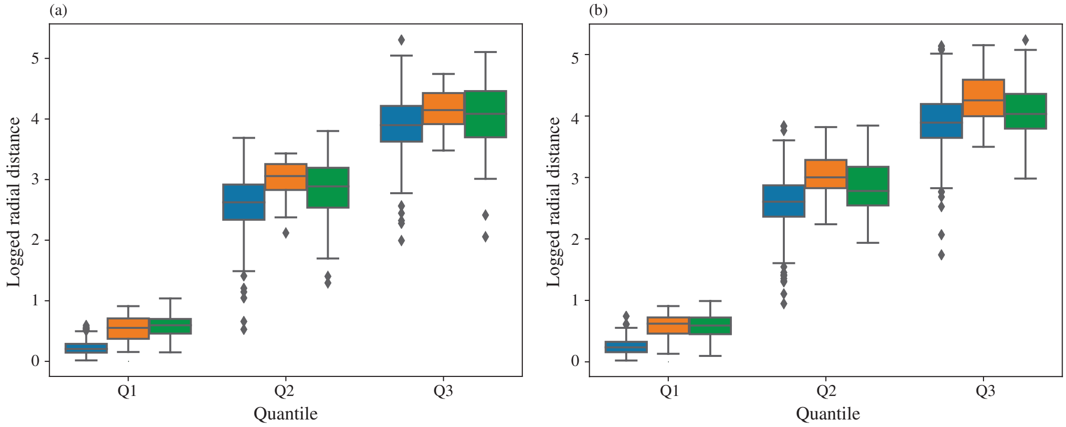


Fig. 1. Heterogeneity in the ADNI hippocampal surface dataset: (a) three quantiles of the logged radial distances across all the vertices on the left hippocampal surface and (b) those on the right hippocampal surface for all subjects obtained from ADNI-1 (blue), ADNI-GO (orange) and ADNI-2 (green).

2017). As an illustration, we consider a hippocampal surface dataset obtained from the three different phases, ADNI-1, ADNI-GO and ADNI-2, of the ADNI study. Fig. 1 presents the three quantiles of loged radial distances across all the vertices on the left and right hippocampal surfaces. More details on the calculation of radial distances will be discussed in § 5. We observe different patterns in the quantile plots across the three phases, especially between ADNI-1 and the other two phases, indicating that the imaging heterogeneity does exist in the ADNI hippocampal surface data. Thus, appropriately handling the imaging heterogeneity can be critically important for understanding the role of imaging biomarkers in the etiological mechanism of Alzheimer's disease. Another example on diffusion tensor imaging also illustrates the heterogeneity in different imaging datasets and can be found in the Supplementary Material.

Currently, there are two approaches to tackling heterogeneity in imaging studies. The first one is image-based meta analysis, in which study-specific statistical analyses are performed first, e.g., Fisher's combined probability test and Stouffer's z-transformation test, and the results are combined afterwards (Salimi-Khorshidi et al., 2009). Although it has shown great promise for some studies with a large number of participants at each phase, or site (Kochunov et al., 2014), this technique still suffers from at least two major limitations: (i) the study-specific population might not be large enough to estimate the true biological variability in the entire population (Mirzaalian et al., 2016); and (ii) computing study-specific summary statistics can be affected by unbalanced data. For instance, the variance in the z-score is highly dependent on the ratio of cases over controls in each individual study, and can lead to inaccurate statistical inferences (Fortin et al., 2017). The second approach is to apply either fixed-effect or mixed-effect models to capture the heterogeneity. These methods estimate primary effects, while adjusting for study related known covariates and unknown hidden factors. To identify those unknown factors, surrogate variable analysis has been developed in various genomic studies (Johnson et al., 2007; Leek & Storey, 2007, 2008; Sun et al., 2012; Lee et al., 2017; Wang et al., 2017), and recently adapted to imaging data analysis (Guillaume et al., 2018). Since surrogate variable analysis assumes that massive univariate regression models share a common set of unknown factors, imaging measures are usually treated as multivariate phenotypes. However, image measures across different voxels, or grid points, are more likely to be treated as functional responses, so it is natural to use functional data analysis tools, which can explicitly account for the three key features of imaging data: spatial smoothness, spatial correlation and low-dimensional representation (Zhu et al., 2012).

Furthermore, by applying some smoothing techniques, the noise component of image measures can be reduced and the estimates of primary effects outperform those under mass-univariate analysis in terms of estimation precision (Ramsay & Silverman, 2002). Therefore, it is greatly important to address the hidden factor issue in functional regression models by borrowing some ideas from surrogate variable analysis.

The aim of this paper is to develop a functional hybrid factor regression modelling framework to investigate the relationship between functional responses and primary covariates, while adjusting for hidden factors. Compared to existing surrogate variable analysis methods, our proposed method is the first designed for functional data. In contrast, although some functional models also consider recovering hidden factors via functional principal component analysis (Zhu et al., 2012), they are inefficient for handling the imaging heterogeneity, since the hidden factors and observed covariates are assumed to be uncorrelated. We develop a three-step estimation procedure to estimate unknown quantities in our proposed model. In addition to the estimation procedure, a global Wald-type test and a simultaneous confidence band are also constructed for coefficient functions. We also systematically investigate the asymptotic properties of estimated coefficient functions, detected hidden factors and test statistics. Furthermore, both simulation studies and real data analysis show that our proposed method outperforms competing methods in terms of both estimation accuracy and robustness.

2. Methods

2.1. *Functional hybrid factor regression model*

Suppose that we observe both imaging data and some covariates from n unrelated subjects. Assume that all the images have been registered to a common template, denoted as $\mathcal{S} \subset \mathbb{R}^d$. The template $\mathcal{S}$ includes n_v grid points, denoted as $s_1, \ldots, s_{n_v}$, which have a common density $p(s)$ with bounded support $\text{supp}(p) \subset \mathcal{S}$. For each registered image, it is assumed that J imaging measurements, or features, are derived at each point such that $y(s_k) = \{y_{\cdot 1}(s_k), \ldots, y_{\cdot J}(s_k)\}$ is an $n \times J$ matrix of J features at s_k across n subjects. Let X be an $n \times p$ full column rank matrix of observed covariates including the intercept, and let Z be an $n \times q$ full column rank matrix of hidden factors, where the number of latent factors, q, is unknown. Let $\mathbb{C}^2(\mathcal{S})$ denote a class of functions whose second-order partial derivatives exist and are continuous everywhere in $\mathcal{S}$.

In this paper, to build up the relationship between imaging responses and both observed covariates and hidden factors, a functional hybrid factor regression model is described as

$$y_{\cdot j}(s) = X\beta_j(s) + Z\gamma_j(s) + \eta_{\cdot j}(s) + \epsilon_{\cdot j}(s) \quad (j = 1, \ldots, J) \tag{1}$$

where $\beta_j(s)$ is a $p \times 1$ vector with entries $\{\beta_{tj}(s) \in \mathbb{C}^2(\mathcal{S})\}_{t=1}^{p}$ representing the primary effect related to X on $y_{,j}(s)$, and $\gamma_j(s)$ is a $q \times 1$ vector with entries $\{\gamma_{lj}(s) \in \mathbb{C}^2(\mathcal{S})\}_{l=1}^{q}$ representing the effect on $y_{.j}(s)$ caused by the hidden factors Z. Moreover, let $\eta(s) = \{\eta_{\cdot 1}(s), \ldots, \eta_{\cdot J}(s)\}$ be an $n \times J$ matrix that characterizes both subject-specific and location-specific spatial variability, and let $\epsilon(s) = \{\epsilon_{\cdot 1}(s), \ldots, \epsilon_{\cdot J}(s)\}$ be measurement errors. It is also assumed that each row in $\eta(s)$ and that in $\epsilon(s)$ are mutually independent and identical copies of $\text{SP}(0, \Sigma_\eta)$ and $\text{SP}(0, \Sigma_\epsilon)$, respectively, where $\text{SP}(\mu, \Sigma)$ denotes a stochastic process vector with mean function $\mu(s)$ and covariance function $\Sigma(s, s')$. Moreover, $\Sigma_\epsilon(s, s')$ takes the form of $\Omega_\epsilon(s)\mathbb{1}(s = s')$, where $\Omega_\epsilon(s)$ is a diagonal matrix and $\mathbb{1}(\cdot)$ is the indicator function. As a comparison, we also consider multivariate varying coefficient models (Zhu et al., 2012) given by

$$y_{\cdot j}(s) = X\beta_j(s) + \eta_{\cdot j}(s) + \epsilon_{\cdot j}(s), \quad (j = 1, \ldots, J). \tag{2}$$

Here models (1) and (2) share several common features. First, both models account for the spatial smoothness, spatial correlation and the low-dimensional representation of functional responses (Zhu et al., 2012). Second, both models are feasible to investigate the relationship between multivariate functional responses and some observed covariates of interest. Third, the individual function variations are considered through $\eta(s)$ in both models (Zhu et al., 2012). Fourth, the detection and adjustment of hidden factors are possible in both models.

However, models (1) and (2) use different strategies to handle the hidden factors. In Zhu et al. (2012), the hidden factors can be captured by the individual functions $\eta(s)$ based on the functional principal component analysis (Wang et al., 2016), where all the principal component scores can be used to recover the structure of hidden factors. A major issue associated with this strategy is that it cannot appropriately handle the case that observed covariates and hidden factors are correlated to each other. Specifically, in Zhu et al. (2012), the observed covariates X are assumed to be uncorrelated with the hidden factors in individual functions $\eta(s)$. However, such an assumption may be questionable in some applications (Helmer et al., 1999; Sundström et al., 2016; Sommerlad et al., 2018) and, thus, model (2) can be problematic for appropriately detecting and adjusting the hidden factors. In contrast, in model (1), the individual functions $\eta(s)$ are assumed to be uncorrelated with both observed covariates X and hidden factors Z, while no assumptions are made for the correlation between X and Z. Therefore, model (1) can handle hidden factors even when they are correlated with the observed covariates.

2.2. *Estimation procedure*

We present the estimation procedure for coefficient functions and hidden factors in three steps.

Step 1. By applying the orthogonal decomposition of the matrix Z onto the column space of X, we reparameterize model (1) as

$$y_{\cdot j}(s) = X\beta_j^*(s) + Z^*\gamma_j(s) + \eta_{\cdot j}(s) + \epsilon_{\cdot j}(s) \quad (j = 1, \ldots, J), \tag{3}$$

where $\beta_j^*(s) = \beta_j(s) + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z\gamma_j(s)$, $Z^* = (I_n - P_X)Z$ and $P_X = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Obviously, the columns of X are orthogonal to those of Z^*. Then, given that $\{y_{\cdot j}(s)\}_{j=1}^J$ and X are observed, the multivariate local linear kernel smoothing technique (Ruppert & Wand, 1994; Fan & Gijbels, 1996) is then used to derive the weighted least squares estimator of $\beta_j^*(s)$ in (3). Let $e^{\otimes 2} = ee^{\mathrm{T}}$ for any vector e, and let $C \otimes D$ be the Kronecker product of two matrices C and D. In addition, define $K_{H_\beta}(s) = |H_\beta|^{-1}K(H_\beta^{-1}s)$ and $z_{H_\beta}(s_k - s) = \{1, (s_k - s)^{\mathrm{T}}H_\beta^{-1}\}^{\mathrm{T}}$, where $K(\cdot)$ is the kernel function, and H_β is the positive definite bandwidth matrix and $|H_\beta|$ is its determinant. For each j and fixed H_β, the estimator of $\beta_j^*(s)$ is derived as

$$\hat{\beta}_j^*(s) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sum_{k=1}^{n_v} \varrho_k(H_\beta, s)y_{,j}(s_k), \tag{4}$$

where $\varrho_k(H_\beta, s) = (1, 0_{1\times d})\{\sum_{k=1}^{n_v} K_{H_\beta}(s_k - s)z_{H_\beta}^{\otimes 2}(s_k - s)\}^{-1}K_{H_\beta}(s_k - s)z_{H_\beta}(s_k - s)$. Since there is no linearity assumption on the coefficient function $\beta_j^*(s)$, the local linear smoother in (4) is a biased estimator (Fan & Gijbels, 1996). To overcome this issue, a standard technique considered here is bias correction. Following the pre-asymptotic substitution method in

Fan & Gijbels (1996), the bias term can be obtained by using local cubic fit with a pilot bandwidth selected in (4). Furthermore, according to the definition of $\beta_j^*(s)$, the aim of the following two steps is to seek an estimate of $Z\gamma_j(s)$. Then the estimate of $\beta_j(s)$ can be derived by subtracting the term $(X^{\mathrm T}X)^{-1}X^{\mathrm T}\hat{Z}\gamma_j(s)$ from $\hat{\beta}_j^*(s)$.

Step 2. The residual term from the previous step is first defined as $R_{\cdot j}(s) = y_{\cdot j}(s) - X\widetilde{\beta}_j^*(s)$, where $\widetilde{\beta}_j^*(s)$ is the refined version of $\hat{\beta}_j^*(s)$ after correcting the bias using the local linear kernel smoothing technique. Next, we construct an $n \times Jn_v$ extended residual matrix written as $\bar{R} = \{R_{\cdot 1}(s_1), \ldots, R_{\cdot 1}(s_{n_v}), \ldots, R_{\cdot J}(s_1), \ldots, R_{\cdot J}(s_{n_v})\}$. Then, given $\mathcal{S}, X$ and Z, the conditional expectation of the extended residual matrix can be derived as (Ruppert & Wand, 1994)

$$E(\bar{R} \mid \mathcal{S}, X, Z) = Z^*\bar{\Gamma} + o_p\{\mathrm{tr}(H_\beta^2)\}, \tag{5}$$

where $\bar{\Gamma} = \{\gamma_{\cdot 1}(s_1), \ldots, \gamma_{\cdot 1}(s_{n_v}), \ldots, \gamma_{\cdot J}(s_1), \ldots, \gamma_{\cdot J}(s_{n_v})\}$ and $\mathrm{tr}(\cdot)$ is the trace of a given matrix. To estimate the primary term Z^* in (5), the singular value decomposition technique is first performed on $\bar{R}$, i.e., $\bar{R} = U\Lambda V^{\mathrm T}$, where the columns of U and V consist of the left and right singular vectors, respectively, and Λ is a diagonal matrix whose diagonal entries are the ordered singular values of $\bar{R}$. Specifically, the first q columns in U, denoted as $U_{1:q}$, can be treated as an estimator of linear combinations of the columns of Z^*; see the Supplementary Material. Then, there exists a $q \times q$ orthonormal matrix Q such that $U_{1:q} = Z^*Q + o_p(1)$ and $Z^*\gamma_j(s) = U_{1:q}\alpha_j(s)$, where $\alpha_j(s) = Q^{\mathrm T}\gamma_j(s)$ $(j = 1, \ldots, J)$.

Step 3. To derive the estimate of $\alpha_j(s)$, the residual terms in the previous steps are treated as functional responses. Then, a new varying coefficient model is constructed via substituting the singular value decomposition results:

$$R_{\cdot j}(s) = U_{1:q}\alpha_j(s) + \tilde{\eta}_{\cdot j}(s) + \tilde{\epsilon}_{\cdot j}(s) \quad (j = 1, \ldots, J)$$

with $\tilde{\eta}_{\cdot\, j}(s)$ and $\tilde{\epsilon}_{,\, j}(s)$ similarly defined as $\eta_{,\, j}(s)$ and $\epsilon_{,\, j}(s)$, respectively. For the fixed H_α, the estimator of $\alpha_j(s)$ can be derived as $U_{1:q}^{\mathrm T}\sum_{k=1}^{n_v}\varrho_k(H_\alpha, s)R_{\cdot\, j}(s_k)$, and $\hat{\alpha}_j(s)$ is denoted as the corresponding bias corrected version. Then, an estimation equation can be constructed as

$$X\tilde{B}^*(s) + U_{1:q}\hat{A}(s) = XB(s) + G\hat{A}(s),$$

where $\tilde{B}^*(s) = \{\tilde{\beta}_1^*(s), \ldots, \tilde{\beta}_J^*(s)\}$, $G = ZQ$ and $\hat{A}(s) = \{\hat{\alpha}_1(s), \ldots, \hat{\alpha}_J(s)\}$. With an additional assumption that the row vectors of $B(s) = \{\beta_1(s), \ldots, \beta_j(s)\}$ and the row vectors of $\Gamma(s) = \{\gamma_1(s), \ldots, \gamma_J(s)\}$ are orthogonal with respect to $p(s)$ on $\mathcal{S}$ after mean centring, we can derive the estimator of G as

$$\hat{G} = U_{1:q} + X\int_{\mathcal{S}} \tilde{B}^*(s)(I_J - P_J)\hat{A}^{\mathrm T}(s)p(s)\,\mathrm{d}s\Omega^{-1},$$

where $\Omega = \int_{\mathcal{S}} \hat{A}(s)(I_J - P_J)\hat{A}^{\mathrm T}(s)p(s)\,\mathrm{d}s$ and $P_J = \mathbb{1}_J(\mathbb{1}^{\mathrm T}\mathbb{1}_J)^{-1}\mathbb{1}_J^{\mathrm T}$, in which $\mathbb{1}_J$ is a $J \times 1$ vector of 1s. Since $Z\gamma_j(s) = G\alpha_j(s)$ for $j = 1, \ldots, J$, the estimator of $B(s)$ is given by

$$\hat{B}(s) = \tilde{B}^*(s) - (X^{\mathrm T}X)^{-1}X^{\mathrm T}\hat{G}\hat{A}(s).$$

2.3. *Other issues in the estimation procedure*

First, according to the definitions of G and $\{\alpha_j(s)\}$, they are not identifiable due to the scaling issue. To address this issue, we impose the constraint $(nq)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{q} G_{i,j}^2 = 1$, where $G_{i,j}$ is the (i,j)th element of G and the estimated $\hat{G}$ is adjusted to satisfy this constraint. Thus, G and $\{\alpha_j(s)\}$ are identifiable up to an orthonormal transformation only.

Second, by using the smoothing method in Ruppert & Wand (1994), we smooth the individual functions of $\eta(s)$ based on the updated residual matrix as

$$\hat{\eta}(s) = \sum_{k=1}^{n_v} \varrho_k(H_\eta, s)\{y(s) - X\hat{B}(s) - \hat{G}\hat{A}(s)\},$$

where H_η is the fixed bandwidth matrix. Furthermore, we use the empirical covariance $\hat{\Sigma}_\eta(s,s') = (n-p-q)^{-1}\sum_{i=1}^{n}\hat{\eta}_{i.}(s)\hat{\eta}_{i.}^{\mathrm{T}}(s')$ to estimate $\Sigma_\eta(s,s')$.

Third, to select the optimal bandwidth in $\hat{B}(s)$ and $\hat{A}(s)$, we use leave-one-curve-out cross-validation, whereas for the optimal bandwidth in $\hat{\eta}(s)$, we use the generalized cross-validation score method (Zhang & Chen, 2007; Zhu et al., 2012). Moreover, we standardize all covariates to have mean zero and standard deviation one, as well as all functional features. Finally, we choose a common bandwidth for all covariates and features. More details can be found in the Supplementary Material.

Fourth, since the number of latent factors, q, is unknown, we consider four different methods: a permutation version of the analytical-asymptotic approach (Johnstone, 2001), parallel analysis (Buja & Eyuboglu, 1992), the eigenvalue difference method (Onatski, 2010) and the bicross-validation method (Owen & Wang, 2016). We compare the four different methods in terms of high detection accuracy and computation time in the simulation studies, and select the one with the best performance in the rest of our data analyses.

2.4. *Inference procedure*

We consider the following linear hypotheses on $B(s)$:

$$\begin{aligned} &\mathrm{H}_0: C\mathrm{vec}\{B(s)\} = b_0(s) \quad \text{for all } s \in \mathcal{S} \\ \text{versus}\quad &\mathrm{H}_1: C\mathrm{vec}\{B(s)\} \neq b_0(s) \quad \text{for some } s \in \mathcal{S}. \end{aligned} \tag{6}$$

Here C is an $r \times Jp$ matrix with rank r, $\mathrm{vec}(\cdot)$ denotes the vectorization of a given matrix and $b_0(s)$ is an $r \times 1$ vector of functions. The global test statistic T_n for (6) is defined as

$$T_n = \int_{\mathcal{S}} T_n(s)p(s)\mathrm{d}s \quad \text{with} \quad T_n(s) = \zeta^{\mathrm{T}}(s)[C\{\hat{\Sigma}_\eta(s,s) \otimes (\hat{M}\hat{M}^{\mathrm{T}})\}C^{\mathrm{T}}]^{-1}\zeta(s), \tag{7}$$

where $\zeta(s) = C\mathrm{vec}\{\hat{B}(s)\} - b_0(s)$, $\hat{M} = (I_p, 0_{q\times q})(\hat{W}^{\mathrm{T}}\hat{W})^{-1}\hat{W}^{\mathrm{T}}$ and $\hat{W} = (X, \hat{G})$.

As the asymptotic distribution of T_n under H_0 is quite complicated, it is difficult to derive the percentiles of T_n directly from the corresponding asymptotic results. To address this issue, the wild bootstrap method is developed (Zhu et al., 2012) consisting of the following four steps.

Step 1. Fit model (1) under H_0 on X and $\{y(s_k)\}_{k=1}^{n_v}$, yielding $\hat{G}, \hat{A}(s), \hat{B}(s), \hat{\eta}(s), \hat{\epsilon}(s)$ and the global test statistic T_n.

Step 2. Generate random vectors $\tau_i^{(m)}$ and $\tau_i^{(m)}(s_k)$ independently from the standard normal distribution $N(0, I_n)$ for $k = 1, \ldots, n_v$, and then construct

$$y^{(m)}(s_k) = X\hat{B}(s_k) + \hat{G}\hat{A}(s_k) + \mathrm{diag}(\tau_i^{(m)})\hat{\eta}(s_k) + \mathrm{diag}\{\tau_i^{(m)}(s_k)\}\hat{\epsilon}(s_k),$$

where $\mathrm{diag}(\tau)$ denotes a diagonal matrix with the elements of τ lying along the diagonal.

Step 3. Based on X and $\{y^{(m)}(s_k)\}_{k=1}^{n_v}$, recalculate $\hat{B}^{(m)}(s)$ and the global test statistic $T_n^{(m)}$.

Step 4. Repeat the previous two steps M times to obtain $\{T_n^{(1)}, \ldots, T_n^{(M)}\}$, which yields the empirical p-value as $p = \sum_{m=1}^{M} \mathbb{1}(T_n^{(m)} > T_n)/M$.

Construction of simultaneous confidence bands for coefficient functions is also of great interest in statistical inference for our proposed model. For a given confidence level ϑ, we construct the $1 - \vartheta$ simultaneous confidence band for $\beta_{tj}(s)$,

$$\{\hat{\beta}_{tj}(s) - n^{-1/2}C_{tj}(\vartheta), \hat{\beta}_{tj}(s) + n^{-1/2}C_{tj}(\vartheta)\}, \quad 1 \leqslant t \leqslant p,\ 1 \leqslant j \leqslant J,$$

where $C_{tj}(\vartheta)$ is a scalar, which is to be determined. Here an efficient resampling method (Kosorok, 2003; Zhu et al., 2007, 2012) is developed to approximate $C_{tj}(\vartheta)$ as follows.

Step 1. Fit model (1) on X and $\{y(s_k)\}_{k=1}^{n_v}$, yielding the residuals $\nu_{.j}(s) = y(s) - X\hat{\beta}(s) + \hat{G}\hat{\alpha}(s)$.

*Step 2.*Generate the random vector $\tau_i^{(m)}$ from the standard normal distribution $N(0, I_n)$, and then construct $\omega_{tj}^{(m)}(s) = n^{1/2} e_t^{\mathrm{T}} \hat{M} \mathrm{diag}(\tau_i^{(m)}) \sum_{k=1}^{n_v} \varrho_k(H, s)\nu_{.j}(s_k)$, where e_t is a $p \times 1$ vector with the tth element being 1 and 0 otherwise.

*Step 3.*Repeat the second step M times to obtain $\{\sup_s |\omega_{tj}^{(1)}(s)|, \ldots, \sup_s |\omega_{tj}^{(M)}(s)|\}$, and use their $1 - \vartheta$ empirical percentile to estimate $C_{tj}(\vartheta)$.

3. Asymptotic properties

We systematically investigate the asymptotic properties of all estimators proposed in § 2.2 and inference procedures in § 2.4. Assumptions used to facilitate the technical details can be found in the Supplementary Material.

The following theorem tackles the theoretical properties of $\hat{B}(s)$ and $\hat{G}$. The detailed proof can be found in the Supplementary Material.

Theorem 1. *Under Assumptions A1–A7 in the Supplementary Material, we have the following results.*

(i) *The columns of* $\hat{G}$ *span the same column space as the columns of* Z *in probability.*
(ii) *It holds that* $n^{1/2}[\{I_J \otimes (\hat{M}\hat{M}^{\mathrm{T}})^{-1/2}\}\mathrm{vec}(\hat{B}(s) - E[\hat{B}(s)]) \mid s \in \mathcal{S}]$ *weakly converges to a centred Gaussian process with covariance matrix* $\Sigma_\eta(s, s) \otimes I_p$.

The following theorem derives the asymptotic distribution of global test statistic T_n in (7) under the null hypothesis and its asymptotic power under local alternative hypotheses. The detailed proof can be found in the Supplementary Material.

Theorem 2. *Under Assumptions A1–A9 in the Supplementary Material, we have the following results.*

(i) *It holds that* $T_n \to \int_{\mathcal{S}} \xi(s)^{\mathrm{T}}\xi(s)\, \mathrm{d}s$ *as* $n \to \infty$, *where* $\xi(s)$ *is a centred Gaussian process.*
(ii) *It holds that* $P\{T_n > T_{n,\vartheta} \mid \mathrm{H}_{1\mathrm{n}}\} \to 1$ *as* $n \to \infty$ *for a sequence of local alternatives* $\mathrm{H}_{1n}\colon C\mathrm{vec}(B(s)) - b_0(s) = n^{-\tau/2}\zeta(s)$, *where* τ *is any scalar in* $[0, 1)$, $T_{n,\vartheta}$ *is the upper* 100ϑ *percentile of* T_n *under* H_0 *and* $0 < \int_{\mathcal{S}} \|\zeta(s)\|^2\, \mathrm{d}s < \infty$.

4. Simulation studies

To examine the proposed methods, we generated synthetic curves from the model

$$y_{ij}(s_k) = x_i^{\mathrm{T}}\beta_j(s_k) + z_i\gamma_j(s_k) + \eta_{ij}(s_k) + \epsilon_{ij}(s_k), \quad j = 1, 2,$$

where $s_1 = 0 \leqslant s_2 \leqslant \cdots \leqslant s_{n_v} = 1$, in which we independently simulated $\tilde{s}_k \sim U(0,1)$ for $k = 2, \ldots, n_v - 1$ and sorted them to obtain $\{s_k : k = 2, \ldots, n_v - 1\}$. We set $x_i = (1, x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$, in which we independently simulated $x_{i1} \sim \mathrm{Ber}(0.5)$, $x_{i2} \sim N(0,1)$ and $x_{i3} \sim N(0,1)$ for $i = 1, \ldots, n$. We simulated z_i as

$$z_i = x_i^{\mathrm{T}}\varphi + \omega_i, \; \omega_i \sim N(0,1) \quad (i = 1, \ldots, n),$$

where $\varphi = \{u_1(2b_1-1), u_2(2b_2-1), u_3(2b_3-1), u_4(2b_4-1)\}^{\mathrm{T}}$ with b_l being independently generated from Ber(0.5). We independently simulated u_l for all l and consider four different simulation scenarios on u_l: (i) $u_l = 0$; (ii) $u_l \sim U(0, 0.2)$; (iii) $u_l \sim U(0.2, 0.5)$ and (iv) $u_l \sim U(0.5, 1)$. Those scenarios correspond to hidden factors Z being (i) independent of X, (ii) weakly correlated with X, (iii) moderately correlated with X and (iv) highly correlated with X, respectively. The $\eta_{ij}(s)$ admits the Karhunen–Loeve expansion as $\eta_{ij} = \xi_{ij1}\psi_{j1}(s) + \xi_{ij2}\psi_{j2}(s)$, where the $\psi_{jl}(s)$ are the eigenfunctions and $\xi_{ijl} \sim N(0, 0.5)$ for $j = 1, 2$ and $l = 1, 2$. We simulated $(\epsilon_{i,1}, \epsilon_{i,2})^{\mathrm{T}} \sim N\{(0,0)^{\mathrm{T}}, 0.5\,\mathrm{diag}(\sigma_1^2, \sigma_2^2)\}$, where $\sigma_l^2 \sim$ Inverse-Gamma(10, 9) for $l = 1, 2$. Also, we set the following functions:

$$\beta_1(s) = \{3s^2, 3(1-s)^2, 6\sqrt{s(1-s)}, -s^2\}^{\mathrm{T}}, \quad \gamma_1(s) = -\sqrt{2}\sin(\pi s),$$
$$\beta_2(s) = \{12(s-0.5)^2, 1.5\sqrt{s}, 3s^2, -2s/3\}^{\mathrm{T}}, \quad \gamma_2(s) = \sqrt{2}\cos(2\pi s),$$
$$\psi_{11}(s) = 0.5, \quad \psi_{12}(s) = s - 0.5, \quad \psi_{21}(s) = 2s - 1 \quad \text{and} \quad \psi_{22}(s) = 1.$$

Throughout the simulation studies, we set $n = 50$ and $n_v = 2000$. Finally, we generated 200 datasets for each simulation scenario.

We compare our method with two other methods: the multivariate varying coefficient model of Zhu et al. (2012) and the confounder adjustment method of Wang et al. (2017). For the method in Wang et al. (2017), the curved data is treated as multivariate responses. To evaluate the finite-sample performance of each method, we consider the integrated square error, i.e., $\sum_{l=1}^{2}\int_0^1 ||\hat{\beta}_l(s) - \beta_l(s)||^2\,\mathrm{d}s$, where $\hat{\beta}_j(s)$ is any estimator of $\beta_j(s)$. For both the method in Wang et al. (2017) and our method, the eigenvalue difference method (Onatski, 2010) is used to estimate the number of factors.

Fig. 2 presents the comparison results for the three methods in all four scenarios. Inspecting Fig. 2 reveals the following results. First, compared with the method in Zhu et al. (2012), both the method in Wang et al. (2017) and our method are very stable and robust to the correlation between X and Z. Second, our method outperforms that in Wang et al. (2017) for all scenarios, indicating that it is critically important to use the functional data analysis tools. Third, when Z and X are independent, the difference between our method and that of Zhu et al. (2012) is very small. Fourth, when the correlation between X and Z is high in scenario (iv), the integrated square errors based on the method in Zhu et al. (2012) dramatically increase. In contrast, those of our method are much smaller even though there are a few outliers, which are caused by the uncertainty of estimating q, as detailed below.

We compare four estimation methods for the number of hidden factors, including the analytical-asymptotic approach in Johnstone (2001), the permutation version of the parallel analysis in

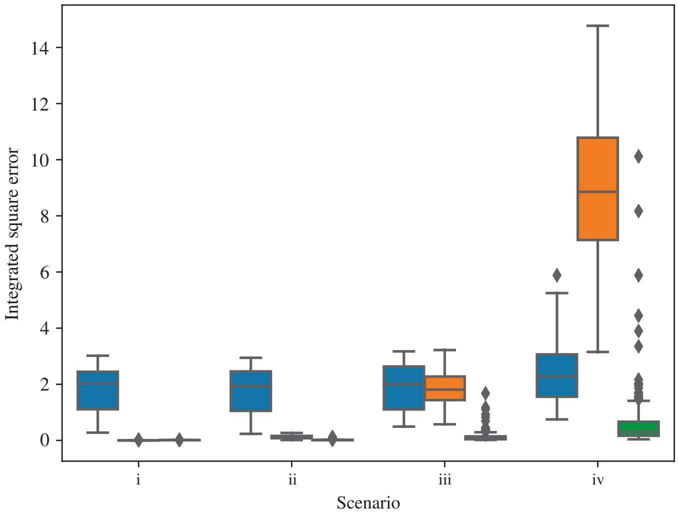


Fig. 2. Simulation results for comparisons of the proposed and competing methods on synthetic curve data in terms of the integrated square error. Four scenarios were considered: the hidden factors Z are (i) independent of X, (ii) weakly correlated with X, (iii) moderately correlated with X and (iv) highly correlated with X. The methods of Wang et al. (2017) (blue) and Zhu et al. (2012) (orange) are compared with our method (green).

Table 1. *Comparison of four methods for estimating the number of hidden factors with $q = 1$. The average computation time for each method is reported as well. In the four scenarios considered Z is (i) independent of X, (ii) weakly correlated with X, (iii) moderately correlated with X and (iv) highly correlated with X*

Method	Scenario				Average computation time
	(i)	(ii)	(iii)	(iv)	(seconds per dataset)
Johnstone (2001)	62/200	65/200	64/200	64/200	0.1
Buja & Eyuboglu (1992)	190/200	191/200	192/200	191/200	70.2
Onatski (2010)	200/200	200/200	198/200	198/200	0.8
Owen & Wang (2016)	200/200	196/200	196/200	196/200	9.7

Buja & Eyuboglu (1992), the eigenvalue difference method in Onatski (2010) and the bi-cross-validation method in Owen & Wang (2016). Table 1 reports the estimation results for the four methods. We observe that the last three methods can achieve almost 100% estimation accuracy, while outperforming the analytical-asymptotic approach with low estimation accuracy around 30%. In addition, in terms of average computation time, the eigenvalue difference method (Onatski, 2010) is much more efficient than the bi-cross-validation method (Owen & Wang, 2016) and the parallel analysis approach (Buja & Eyuboglu, 1992). Thus, the eigenvalue difference method is used in subsequent analyses.

We investigate the sensitivity of our method with respect to the misspecification of q under the four scenarios since there are some outliers in Fig. 2 for our method when Z and X are highly correlated. We also consider three choices of q including $q = 0, 1$ and 2, which represent the underestimated, true and overestimated values, respectively. Fig. 3 presents the box plots of integrated square errors for all q values under the four scenarios. There are three major findings. First, when the hidden factor Z is independent of or weakly correlated with X, the integrated square errors are relatively stable even when q is misspecified. Second, when Z is moderately

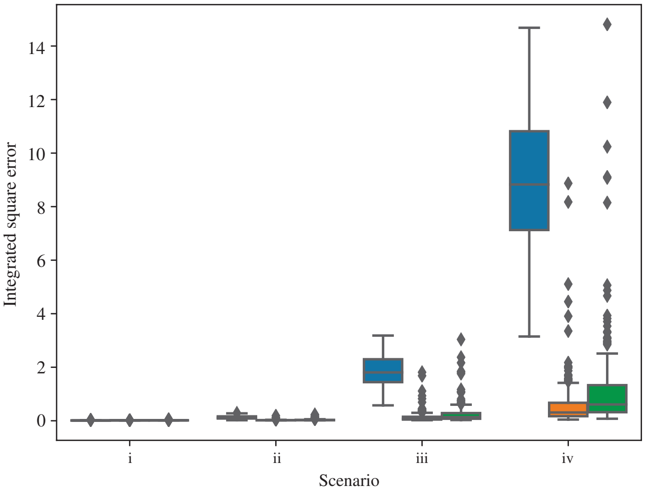


Fig. 3. Simulation results for the sensitivity analysis of our method under the three choices of q, $q = 0$ (blue), $q = 1$ (orange) and $q = 2$ (green) in the four scenarios in which Z is (i) independent of X, (ii) weakly correlated with X, (iii) moderately correlated with X and (iv) highly correlated with X.

or highly correlated with X, the integrated square errors dramatically increase for misspecified q values. Third, the underestimated $q = 0$ has much larger effects on integrated square errors than the overestimated $q = 2$.

We examine the correlation between the space spanned by the columns of detected latent factors with that spanned by the columns of true Z. Fig. 4 presents simulation results in the four scenarios with the absolute values of Pearson correlation between $\hat{G}$ and Z being greater than 0.90, indicating their consistency with each other. Moreover, when the correlation between Z and X gets higher, the absolute values of the Pearson correlation coefficient are closer to 1.

We examine the Type I and Type II error rates of T_n. For the sake of space, we only consider the third scenario (iii), in which $\varphi = (u_1, -u_2, u_3, -u_4)^{\mathrm{T}}$ independently simulating u_l from $U(0.2, 0.5)$ for all $l = 1, 2, 3, 4$. Moreover, we fix all other parameters at their values specified above except that we set $\beta_{14}(s) = -cs^2$ and $\beta_{24}(s) = -2cs/3$, where c is a scalar specified below. We want to test the following hypotheses:

$$\begin{aligned} &\mathrm{H}_0: \beta_{14}(s) = \beta_{24}(s) = 0 \quad \text{for all } s \\ \text{versus} \quad &\mathrm{H}_1: \beta_{14}(s) \neq 0 \quad \text{or} \quad \beta_{24}(s) \neq 0 \quad \text{for at least one } s. \end{aligned} \tag{8}$$

We set $c = 0$ to assess the Type I error rates for T_n, and set $c = 0.1, 0.2, 0.3, 0.4$ and 0.5 to examine the power of T_n. We set the sample size to $n = 100$ and 200. For each case, 500 bootstrap replications were generated to construct the empirical distribution of T_n under H_0. Fig. 5 presents the power curves at the significance levels $\alpha = 0.05$ and 0.01. The rejection rates for T_n based on the wild bootstrap method are accurate for moderate sample sizes with $n = 100$ and 200 at both significance levels $\alpha = 0.01$ and 0.05. As expected, the power increases with the sample size.

Finally, we investigate the coverage probabilities of simultaneous confidence bands for the functional coefficients in $B(s)$ based on the resampling method. We only consider the third scenario (iii). We fix all parameters specified above except that we set $n = 200$ and the number of grid points $n_v = 200$ and 2000. We calculated the simultaneous confidence bands for each component in $B(s)$ based on 200 replications. Table 2 summarizes the empirical coverage probabilities at

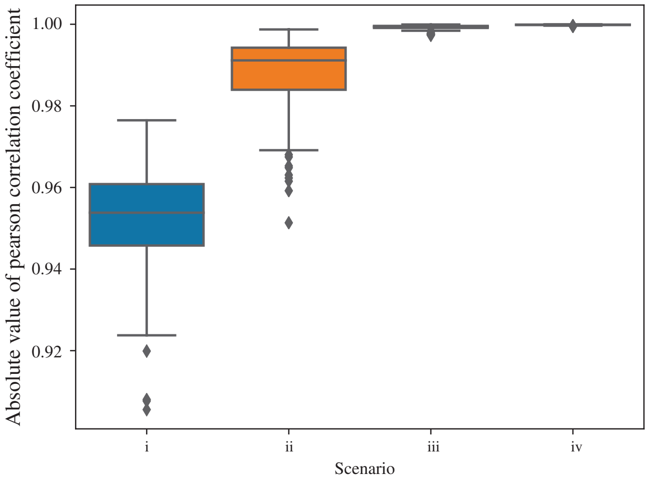


Fig. 4. Simulation results for the absolute values of the Pearson correlation between $\hat{G}$ and Z in the four scenarios in which Z is (i) independent of X, (ii) weakly correlated with X, (iii) moderately correlated with X and (iv) highly correlated with X.

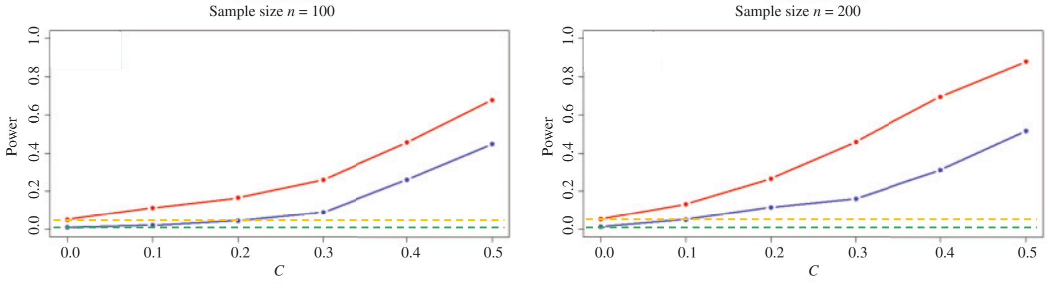


Fig. 5. Power curves for the hypothesis testing problem (8) based on our method with different choices of c and levels of α: $\alpha = 0.05$ (red) and $\alpha = 0.01$ (blue). Two horizontal dashed lines are added to indicate the levels $\alpha = 0.05$ (orange) and $\alpha = 0.01$ (green).

Table 2. *Empirical coverage probabilities of* $1 - \alpha$ *simultaneous confidence bands*

α	n_v	β_{11}	β_{12}	β_{13}	β_{14}	β_{21}	β_{22}	β_{23}	β_{24}
0.05	200	0.935	0.920	0.925	0.920	0.915	0.915	0.930	0.940
	2000	0.945	0.950	0.950	0.950	0.945	0.945	0.955	0.950
0.01	200	0.985	0.990	0.995	0.980	0.980	0.995	0.990	0.990
	2000	0.990	0.995	0.990	0.995	0.995	0.995	0.990	0.995

$\alpha = 0.05$ and 0.01. As expected, the coverage probabilities improve as the number of grid points n_v increases.

5. Real data analysis

5.1. *Data processing*

In this data analysis, we consider 936 MRI scans from normal controls and individuals with mild cognitive impairment or Alzheimer's disease from the three phases ADNI-1, ADNI-GO and ADNI-2. Table 3 summarizes the demographic information of all the subjects.

Table 3. *Hippocampal surface data: demographic information of 936 subjects*

Phase	ADNI-1	ADNI-GO	ADNI-2	Total
Size	800	24	112	936
Gender (F/M)	465/335	13/11	61/51	539/397
Handedness (R/L)	738/62	20/4	9/103	861/75
Age range (years)	[58, 95]	[55, 84]	[53, 87]	[53, 95]
Education length range (years)	[4, 20]	[12, 20]	[8, 20]	[4, 20]
Disease (NC/MCI/AD)	224/389/187	0/24/0	29/58/25	253/471/212

NC, normal control; MCI, mild cognitive impairment; AD, Alzheimer's disease.

We processed the MRI data by using standard steps and generated one-to-one hippocampal surface registration in Shi et al. (2013). Then, we computed the various surface statistics on the registered surface, such as multivariate tensor-based morphometry statistics, which retain the full tensor information of the deformation Jacobian matrix, together with the radial distance, which retains information on the deformation along the surface normal direction. More detailed image data processing procedures can be found in the Supplementary Material.

5.2. *Data analysis*

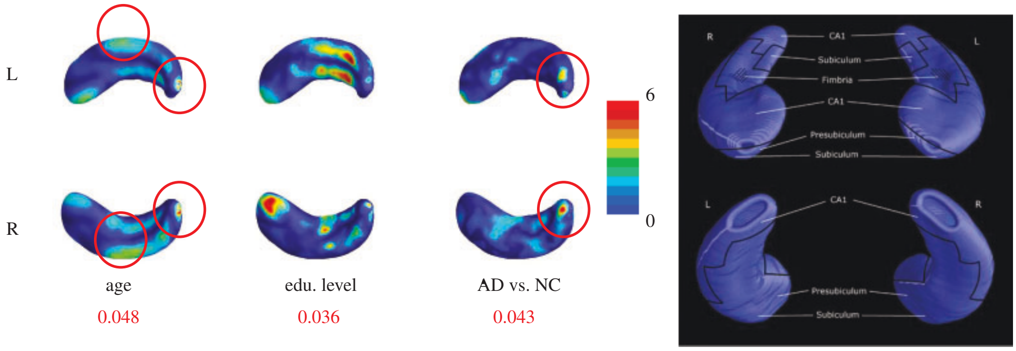
The hippocampus is believed to be involved in memory, spatial navigation and memory, and behavioural inhibition. In Alzheimer's disease, the hippocampus is one of the first regions of the brain to be affected, leading to the confusion and loss of memory so commonly seen in the early stages of the disease (Kong et al., 2019). The objective of this data analysis is to integrate the data from three different data phases, i.e., ADNI-1, ADNI-GO and ADNI-2, and examine the effects of clinical variables and demographic variables on either the left or right hippocampus. Moreover, the hidden factors are expected to be recovered and discussed. Before conducting this analysis, we would like to check if there is any heterogeneity caused by phases. According to Fig. 1 and the related discussion in § 1, this study-level heterogeneity does exist in the ADNI hippocampal surface data. Therefore, the phase information should be included as predictors in the data analysis.

We applied our new method with either the left or right hippocampal surface data as the functional responses. The method in Zhu et al. (2012) was used for comparison. Specifically, we consider four imaging measurements: the logged radial distance and three tensor-based morphometry statistics measured over 7500 vertices on the hippocampal surface (3750 on each side). In this case, we have $J = 4$. Moreover, we included an intercept, gender, handedness, education length, age, diagnostic information and phase information as predictors in X. The corresponding coefficients are considered as functions on the cerebral cortex, and the Gaussian kernel function is adopted in the estimation procedure. Subsequently, we test the effects of all the primary variables on the four functional responses on hippocampal surfaces. We calculated the global test statistic for each predictor and used 500 replications in the wild bootstrap approach. Table 4 summarizes the corresponding p-values, where p-values less than 5% are in red. Given the significant level 0.05, both disease, Alzheimer's disease versus Normal control, and age are found to be significant on the left hippocampal surface based on the method in Zhu et al. (2012). In contrast, more predictors are found to be significant based on our method. For example, significant age effect is found on the left hippocampal surface, while both education length effect and disease effect, Alzheimer's disease versus normal control, are significant on left and right hippocampal surfaces. Among all these variables, education length is found to be significant in our method, but not in

Table 4. *Hippocampal surface data: comparison of p-values for primary variables*

Variable	*p*-value			
	Left hippocampus		Right hippocampus	
	Zhu et al. (2012)	Our method	Zhu et al. (2012)	Our method
Gender	0.212	0.092	0.234	0.116
Handedness	0.652	0.102	0.704	0.082
Education length	0.132	0.036	0.244	0.048
Age	0.048	0.048	0.096	0.052
MCI versus NC	0.156	0.066	0.082	0.064
AD versus NC	0.046	0.034	0.054	0.040
ADNI-GO versus ADNI-1	0.134	0.112	0.136	0.120
ADNI-2 versus ADNI-1	0.118	0.106	0.112	0.114

NC, normal control; MCI, mild cognitive impairment; AD, Alzheimer's disease.



Frisoni et al. (2008) *Brain*

Fig. 6. Hippocampal surface data: adjusted $-\log_{10}(p)$-value maps corresponding to three covariates of interest: age, education level, and diagnosis status.

the competing method. Education length is an important factor for the changes of hippocampus structure in the literature (Arenaza-Urquijo et al., 2013).

Furthermore, we are also interested in detecting significant subregions by using the local test statistic and the false discovery rate (Benjamini & Yekutieli, 2001). Fig. 6 presents the false discovery rate adjusted $-\log_{10}(p)$-value maps. To better understand the significant subregions, we consider the cytoarchitectonic subregions mapped on blank MR-based models at 3 T of the hippocampal formation (Frisoni et al., 2008, Fig. 2). All the significant subregions associated with age and disease circled in red are found in the CA1 subfield. Similar hippocampal subregions were found to be affected by Alzheimer's disease (Frisoni et al., 2008), indicating that our findings are in agreement with those in the literature.

We investigate the potential hidden factors estimated by our method. Applying the eigenvalue difference method yields three hidden factors. Table 5 presents the correlation between primary variables and detected hidden factors, where *p*-values less than 5% are in red. Specifically, we calculated the Pearson correlation between two continuous variables and the polyserial correlation between a continuous variable and a discrete one. Inspecting Table 5 reveals that on both left and right hippocampal surfaces, the detected factors are highly related to education length, age, disease status and phase information. In contrast, for the method in Zhu et al. (2012), the key assumption that the hidden factors and primary variables are uncorrelated is inappropriate.

Finally, we investigate whether there are any other variables not included in our current analysis that may be strongly correlated with the latent factors. We consider seven new variables in the three categories of ethnic group information (three dummy variables were introduced to represent

Table 5. *Hippocampal surface data: correlations between primary variables and detected hidden factors and their associated p-values in parentheses*

Primary variable	Hidden factor					
	Left hippocampus			Right hippocampus		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Gender	−0.038	0.015	−0.048	0.006	0.023	−0.045
	(0.358)	(0.724)	(0.239)	(0.883)	(0.582)	(0.278)
Handedness	−0.013	−0.041	0.076	0.041	−0.055	0.047
	(0.835)	(0.517)	(0.209)	(0.494)	(0.382)	(0.435)
Education length	−0.021	0.024	0.090	0.058	0.014	0.074
	(0.531)	(0.466)	(0.006)	(0.078)	(0.665)	(0.025)
Age	0.120	0.089	−0.079	−0.163	0.071	−0.131
	(<0.001)	(0.007)	(0.015)	(<0.001)	(0.030)	(<0.001)
MCI versus NC	−0.045	0.061	0.020	0.064	0.003	0.062
	(0.272)	(0.144)	(0.617)	(0.119)	(0.944)	(0.131)
AD versus NC	0.087	−0.058	0.061	−0.094	−0.029	−0.008
	(0.041)	(0.228)	(0.507)	(0.039)	(0.530)	(0.853)
ADNI-GO versus ADNI-1	−0.305	0.392	0.215	0.440	−0.176	0.403
	(<0.001)	(<0.001)	(0.011)	(<0.001)	(0.064)	(<0.001)
ADNI-2 versus ADNI-1	−0.221	−0.318	0.213	0.271	−0.469	0.466
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)

NC, normal control; MCI, mild cognitive impairment; AD, Alzheimer's disease.

Asian, African American and White), marital status (three dummy variables were introduced to represent widow, divorce and not married) and retirement status. There are several reasons that we do not include the new regressors in the main model at the beginning. First, we only include a standard set of covariates, which have been widely considered in the existing literature (Kong et al., 2019), in the main model. Second, we apply our proposed method to detect some hidden factors that cannot be explained by the existing covariates. Third, we correlate the hidden factors with a set of new regressors and find that these regressors can partially explain these factors. This process also illustrates the importance of our functional hybrid factor regression model. Another reason is that there are many missing data in these new regressors. Specifically, the missing data rates for the new regressors in the three categories are 9.8% for ethnic group information, 10.9% for marital status and 9.4% for retirement status. We observe that on the left hippocampal surface, the detected hidden factors are strongly correlated with all of them, whereas on the right hippocampal surface, the detected hidden factors are only correlated with the marital status. More detailed results can be found in the Supplementary Material.

6. Discussion

The key assumption of our method is Assumption A6 in the Supplementary Material, which requires that the row vectors of $B(s)$ and the row vectors of $\Gamma(s)$ are orthogonal with respect to the underlying density function $p(s)$ after mean centring. Similar assumptions for model identification can be found in some existing methods (Sun et al., 2012; Lee et al., 2017). This assumption is reasonable in many imaging studies. For example, in neuroimage data analysis, batch effects are usually caused by the heterogeneity in imaging acquisition protocols. Their effect sizes would not be correlated with those of population differences or diagnostic status (Lee et al., 2017). Also, our simulation studies show that our method is robust even when this assumption is violated. Specifically, in our simulation settings, when $\| \int_S B(s)(I_J - P_J)\Gamma^{\mathrm{T}}(s)p(s)\,\mathrm{d}s\|_1 = 3.544$, indicating that this assumption does not hold, our method still outperforms the two competing methods.

Besides the assumption on functional coefficients, modelling of latent factors Z is also a key term in our method. In this paper, we treat the latent factors as fixed. However, to account for the imaging heterogeneity, it will be more flexible to assume that the latent factors are random. For example, Wang et al. (2017) modelled the latent factors Z through a linear model on primary variables X, i.e., $Z = X\alpha^{\mathrm{T}} + W$ and W is normally distributed. Therefore, it is important to extend our model in this paper to handle the random setting of latent factors, which will be the focus of future work.

Another interesting topic is to extend our method to some unsupervised or semisupervised learning, whose goal is to recover the subgroup structure within the functional data when the subgroup information is unknown or not completely observable. It is challenging because unwanted variations may be correlated with the subgroup information. For example, it is of great interest to conduct the clustering analysis in terms of brain atrophy variations among patients with Alzheimer's disease (Poulakis et al., 2018), and there is increasing evidence that the patients' cluster information has strong association with some unknown factors like marital status (Sommerlad et al., 2018). Thus, it would be interesting to extend our model to simultaneously investigate the latent subgroup structure, while accounting for unknown latent factors. We leave these extensions to future research.

Acknowledgement

Huang gratefully acknowledges financial support from the National Science Foundation and Zhu from the National Institutes of Health. The authors are grateful to the editor, associate editor and reviewers for their feedback that helped to improve the manuscript.

Supplementary material

Supplementary Material available at *Biometrika* online includes another example illustrating the heterogeneity in different imaging datasets, assumptions of theorems, proofs of the theoretical results, and additional simulation and real data analysis results.

References

Arenaza-Urquijo, E. M., Landeau, B., La Joie, R., Mevel, K., Mézenge, F., Perrotin, A., Desgranges, B., Bartrés-Faz, D., Eustache, F. & Chételat, G. (2013). Relationships between years of education and gray matter volume, metabolism and functional connectivity in healthy elders. *NeuroImage* **83**, 450–7.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

Buja, A. & Eyuboglu, N. (1992). Remarks on parallel analysis. *Mult. Behav. Res.* **27**, 509–40.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C. & Gur, R. E. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161**, 149–70.

Frisoni, G. B., Ganzola, R., Canu, E., Rüb, U., Pizzini, F. B., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C. & Thompson, P. M. (2008). Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. *Brain* **131**, 3266–76.

Guillaume, B., Wang, C., Poh, J., Shen, M. J., Ong, M. L., Tan, P. F., Karnani, N., Meaney, M. & Qiu, A. (2018). Improving mass-univariate analysis of neuroimaging data by modelling important unknown covariates: application to epigenome-wide association studies. *NeuroImage* **173**, 57–71.

Helmer, C., Damon, D., Letenneur, L., Fabrigoule, C., Barberger-Gateau, P., Lafont, S., Fuhrer, R., Antonucci, T., Commenges, D. & Orgogozo, J. (1999). Marital status and risk of Alzheimer's disease: a French population-based cohort study. *Neurology* **53**, 1953–8.

Johnson, W. E., Li, C. & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–27.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T. E., Mandl, R. C., Almasy, L., Booth, T., Brouwer, R. M., Curran, J. E. & de Zubicaray, G. I. (2014). Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. *NeuroImage* **95**, 136–50.

Kong, D., An, B., Zhang, J. & Zhu, H. (2019). L2RM: low-rank linear regression models for high-dimensional matrix responses. *J. Am. Statist. Assoc.* **115**, 403–24.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *J. Mult. Anal.* **84**, 299–318.

Lee, S., Sun, W., Wright, F. A. & Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104**, 303–16.

Leek, J. T. & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161.

Leek, J. T. & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc. Nat. Acad. Sci. U.S.A.* **105**, 18718–23.

Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Statist.* **7**, 523–42.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C., Morey, R. A. & Flashman, L. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage* **135**, 311–23.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W. & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am.* **15**, 869–77.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Statist.* **92**, 1004–16.

Owen, A. B. & Wang, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31**, 119–39.

Poulakis, K., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Wahlund, L.-O. et al. (2018). Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol. Aging* **65**, 98–108.

Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–70.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. & Nichols, T. E. (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage* **45**, 810–23.

Shi, J., Thompson, P. M., Gutman, B. & Wang, Y. (2013). Surface fluid registration of conformal representation: application to detect disease burden and genetic influence on hippocampus. *NeuroImage* **78**, 111–34.

Sommerlad, A., Ruegger, J., Singh-Manoux, A., Lewis, G. & Livingston, G. (2018). Marriage and risk of dementia: systematic review and meta-analysis of observational studies. *J. Neurol. Neurosurg. Psychiat.* **89**, 231–8.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J. & Landray, M. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779.

Sun, Y., Zhang, N. R. & Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *Ann. Appl. Statist.* **6**, 1664–88.

Sundström, A., Westerlund, O. & Kotyrlo, E. (2016). Marital status and risk of dementia: a nationwide population-based prospective study from Sweden. *BMJ Open* **6**, e008565.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K. & Consortium, W.-M. H. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage* **80**, 62–79.

Wang, J., Zhao, Q., Hastie, T. & Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45**, 1863–94.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016). Functional data analysis. *Ann. Rev. Statist.* **3**, 257–95.

Yu, Q., Risk, B. B., Zhang, K. & Marron, J. (2017). Jive integration of imaging and behavioral data. *NeuroImage* **152**, 38–49.

Zhang, J. & Chen, J. (2007). Statistical inference for functional data. *Ann. Statist.* **35**, 1052–79.

Zhu, H., Ibrahim, J. G., Tang, N., Rowe, D. B., Hao, X., Bansal, R. & Peterson, B. S. (2007). A statistical analysis of brain morphology using wild bootstrapping. *IEEE Trans. Med. Imag.* **26**, 954–66.

Zhu, H., Li, R. & Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40**, 2634–66.

[*Received on* 29 *April* 2020. *Editorial decision on* 4 *January* 2022]