

# HFPRM: Hierarchical Functional Principal Regression Model for Diffusion Tensor Image Bundle Statistics

Jingwen Zhang<sup>1</sup>, Chao Huang<sup>1</sup>, Joseph G. Ibrahim<sup>1</sup>, Shaili Jha<sup>2</sup>,  
Rebecca C. Knickmeyer<sup>3</sup>, John H. Gilmore<sup>3</sup>, Martin Styner<sup>3</sup>,  
and Hongtu Zhu<sup>1,4</sup>(✉)

<sup>1</sup> Department of Biostatistics, University of North Carolina at Chapel Hill,  
Chapel Hill, USA

<sup>2</sup> Curriculum in Neurobiology, University of North Carolina at Chapel Hill,  
Chapel Hill, USA

<sup>3</sup> Department of Psychiatry, University of North Carolina at Chapel Hill,  
Chapel Hill, USA

<sup>4</sup> Department of Biostatistics, University of Texas MD Anderson Cancer Center,  
Houston, USA

HZhu5@mdanderson.org

**Abstract.** Diffusion-weighted magnetic resonance imaging (MRI) provides a unique approach to understand the geometric structure of brain fiber bundles and to delineate the diffusion properties across subjects and time. It can be used to identify structural connectivity abnormalities and helps to diagnose brain-related disorders. The aim of this paper is to develop a novel, robust, and efficient dimensional reduction and regression framework, called hierarchical functional principal regression model (HFPRM), to effectively correlate high-dimensional fiber bundle statistics with a set of predictors of interest, such as age, diagnosis status, and genetic markers. The three key novelties of HFPRM include the simultaneous analysis of a large number of fiber bundles, the disentanglement of global and individual latent factors that characterizes between-tract correlation patterns, and a bi-level analysis on the predictor effects. Simulations are conducted to evaluate the finite sample performance of HFPRM. We have also applied HFPRM to a genome-wide association study to explore important genetic variants in neonatal white matter development.

---

Knickmeyer was partially supported by the National Institutes of Health grant MH083045.

Gilmore was partially supported by the National Institutes of Health grants MH064065, MH070890, and HD053000.

Styner was partially supported by the National Institutes of Health grant EB005149-01.

Zhu was partially supported by the National Institutes of Health grant MH086633, the National Science Foundation grants SES-1357666 and DMS-1407655, as well as a senior investigator grant from the Cancer Prevention Research Institute of Texas.

**Keywords:** Fiber bundle statistics · Varying coefficient model · Functional principal component analysis · Factor analysis · Imaging genetics

## 1 Introduction

Scientifically, investigation in the connectional organization of human brain and its variation across subjects is a critical step to understand the pathology of many neuro-related disorders. Diffusion-weighted MRI offers a non-invasive approach to study the tissue structure of white matter fiber bundles *in vivo*, including both the geometric shape and the diffusion properties [2, 6, 9, 12, 17, 24, 27]. Delineating diffusion statistics along fiber bundles may help identify structural connectivity abnormalities across different spatial-temporal scales. It could eventually inspire new approaches for disease preventions, diagnoses and clinical treatments.

Group analysis of fiber bundle statistics poses remarkable computational and mathematical challenges to existing statistical methods. The first challenge is to efficiently and simultaneously study multiple fiber bundles with heterogeneous geometric structures and variation patterns. The second challenge is to correlate fiber bundle statistics with a large number of covariates, such as millions of genetic markers. This challenge is motivated by the demand to carry out a genome-wide association study on fiber bundle statistics. Voxel-wise methods [21] and single tract analysis [8, 26, 28] suffer from performing massive multiple comparison adjustments, which would severely reduce detection power. The third challenge is to properly handle the potential correlation among multiple tracts and to disentangle tract-specific information from global information shared by a large portion of fiber bundles.

The aim of this paper is to develop a hierarchical functional principal regression model (HFPRM) framework to address the three challenges discussed above. HFPRM consists of three statistical models, including a varying coefficient model (VCM), a latent factor analysis (LFA) procedure, and a multivariate regression model (MRM). The path diagram of HFPRM is presented in Fig. 1. The VCM not only captures the functional structure of fiber bundle statistics for each single tract, but also maps the heterogeneous geometric structure of multiple fiber bundles onto a common coordinate system. The LFA is applied to characterize potential inter-tract correlation across multiple bundles. It allows us to explicitly identify both tract-specific and global latent signals. The integration of VCM and LFA dramatically reduces the dimension of fiber bundle statistics. Finally, using MRM, we are able to examine the effect of selected predictors on both global level and individual level.

In Sect. 2, we introduce the general framework of HFPRM and propose a two stage estimation procedure to study both global effect and individual tract effect. In Sects. 3 and 4, we use numerical simulations and a real data example to examine the finite sample performance of HFPRM. Section 5 concludes with some remarks.

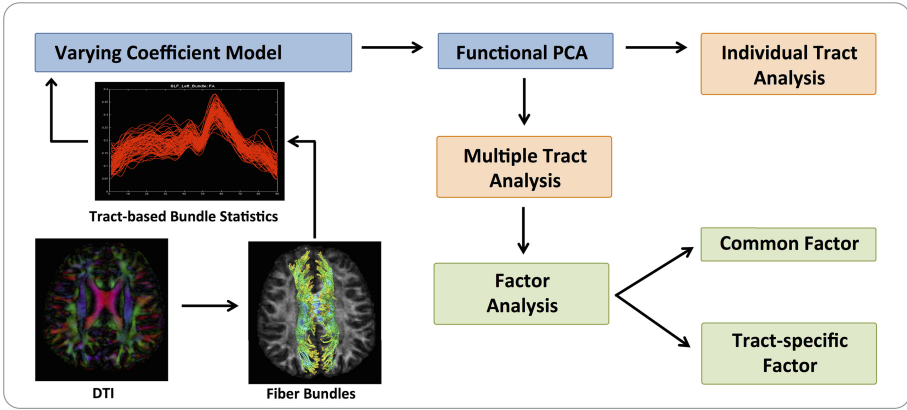


Fig. 1. A schematic overview of HFPRM

## 2 Methods

### 2.1 Data Structure

Suppose that we obtain a data set with clinical, genetic variables as well as DTI statistics along  $M$  fiber bundles from  $n$  subjects. For the  $m$ -th fiber bundle,  $m = 1, \dots, M$ , we use  $s_m \in [0, S_m]$  to denote the arc length of any point relative to a fixed end point, where  $S_m$  is the longest arc length on the tract. For the  $i$ -th subject where  $i = 1, \dots, n$ ,  $y_{i,m}(s_m)$  denotes a specific diffusion statistics observed at arc-length  $s_m$  along the  $m$ -th tract, and  $\mathbf{x}_i$  is a  $q \times 1$  vector of covariates.

### 2.2 HFPRM

HFRPM is proposed to study the association between diffusion properties (e.g., FA, MD or RD) along  $M$  fiber bundles with a set of covariates, such as age, gender, and genetic markers. It consists of three key components, a varying coefficient model (VCM), a latent factor analysis (LFA) procedure, and a multivariate regression model (MRM).

The VCM describes the functional association between  $\{y_{i,m}(s_m) : s_m \in [0, S_m]\}$  and  $\mathbf{x}_i$  for a single tract. It admits the following form,

$$y_{i,m}(s_m) = \mu_m(s_m) + \eta_{i,m}(s_m) + e_{i,m}(s_m), \tag{1}$$

where  $\mu_m(s_m)$  is the function of population mean,  $\eta_{i,m}(s_m)$  is an individual function characterizing subject-specific spatial variations along the  $m$ -th tract, and  $e_{i,m}(s_m)$  is the measurement error. Let  $SP(0, \Sigma)$  represent a stochastic process with mean zero and covariance operator  $\Sigma(s_m, s'_m)$ . It is assumed that  $\eta_{i,m}(s_m)$  and  $e_{i,m}(s_m)$  are mutually independent and identical copies of stochastic processes  $SP(0, \Sigma_{\eta_m})$  and  $SP(0, \Sigma_{e_m})$  respectively, in which  $\Sigma_{e_m}(s_m, s'_m) = \sigma_{e_m}^2(s_m)\mathbf{1}(s_m = s'_m)$  and  $\mathbf{1}(\cdot)$  is an indicator function.

The major challenge to simultaneously study  $M$  fiber bundles is the heterogeneity in their geometric structures. It is necessary to find a common coordinate system for  $\{\eta_{i,m}(s_m)\}_{m=1}^M$ . Specifically, we use functional principal component analysis (fPCA) to extract the key features in  $\eta_{i,m}(s_m)$ . Based on Mercer’s theorem,  $\Sigma_{\eta_m}(s_m, s'_m)$  admits a spectral decomposition as follows:

$$\Sigma_{\eta_m}(s_m, s'_m) = \sum_{d=1}^{+\infty} \lambda_{md} \phi_{md}(s_m) \phi_{md}(s'_m), \tag{2}$$

where  $\{\lambda_{md} \geq 0\}$  are eigenvalues in descending order with  $\sum_{d=1}^{+\infty} \lambda_{md} < \infty$  and  $\{\phi_{md}(s_m)\}$  are the corresponding orthonormal eigenfunctions. Using Karhunen-Loeve expansion [13, 16],  $\eta_{im}(s_m)$  can be expressed as

$$\eta_{i,m}(s_m) = \sum_{d=1}^{+\infty} z_{i,md} \phi_{md}(s_m) \quad \text{with} \quad z_{i,md} = \int_0^{S_m} \eta_{i,m}(s_m) \phi_{md}(s_m) ds_m. \tag{3}$$

Individual function  $\eta_{i,m}(s_m)$  can then be equivalently represented by a set of functional principal component (fPC) scores  $\{z_{i,md} : d = 1, \dots, \infty\}$ . In practice, a relatively small number of fPC scores would account for the majority of variation in  $\eta_{i,m}(s)$ . Therefore, we can approximate  $\eta_{i,m}(s_m)$  by a finite vector  $\mathbf{z}_{i,m} = (z_{i,m1}, \dots, z_{i,mD})^T$  of dimension  $D$ . For notational simplicity, it is assumed that  $D$  is the same across all  $M$  bundles. Now we use  $\mathbf{z}_{i,m}$  to integrate information across  $M$  bundles and denote  $\mathbf{z}_i$  as a  $p \times 1$  long vector that concatenates all  $\mathbf{z}_{i,m}$ s together, where  $p = DM$ .

A LFA is then proposed to account for potential inter-tract correlation across multiple bundles. Specifically,  $\mathbf{z}_i$  is assumed to have the following latent factor structure,

$$\mathbf{z}_i = \mathbf{A} \mathbf{f}_i + \mathbf{u}_i, \tag{4}$$

where  $\mathbf{A}$  is a  $p \times L$  loading matrix and  $\mathbf{f}_i$  and  $\mathbf{u}_i$ , respectively, represent global and individual latent factors. When there exist homogeneous signal patterns across multiple fiber bundles,  $L$  is expected to be much smaller than  $p$ . Global factor  $\mathbf{f}_i$  thus allows us to study the shared pattern in a low dimensional space. And tract-specific pattern can also be captured by each component in  $\mathbf{u}_i = (\mathbf{u}_{i,1}, \dots, \mathbf{u}_{i,M})^T$ .

Finally, a MLM is introduced to correlate the global and individual latent factors with covariate  $\mathbf{x}_i$ ,

$$\mathbf{f}_i = \mathbf{B}_f^T \mathbf{x}_i + \boldsymbol{\epsilon}_{f,i} \quad \text{and} \quad \mathbf{u}_{i,m} = \mathbf{B}_{u_m}^T \mathbf{x}_i + \boldsymbol{\epsilon}_{u_m,i}, \quad \text{for } m = 1, \dots, M, \tag{5}$$

where  $\mathbf{B}_f$  and  $\mathbf{B}_{u_m}$  are, respectively,  $q \times L$  and  $q \times D$  coefficient matrices and  $\boldsymbol{\epsilon}_{f,i}$  and  $\boldsymbol{\epsilon}_{u_m,i}$  are residual terms. Using (5), we are able to perform a hierarchical analysis on both global level and individual level.

### 2.3 Estimation and Inference Procedure

In practice, diffusion statistics are observed on discrete grid points along each tract. For the  $m$ -th tract, assume  $y_{i,m}(s_m)$  is observed on sample point set  $\mathcal{S}_m = \{s_{m,1}, \dots, s_{m,k}, \dots, s_{m,K_m}\} \subset [0, S_m]$ , we use the following two-stage procedure to estimate fPC scores  $\mathbf{Z} = \{\mathbf{z}_i\}_{1 \leq i \leq n}$ , global factors  $\mathbf{F} = \{\mathbf{f}_i\}_{1 \leq i \leq n}$  and individual factors  $\mathbf{U} = \{\mathbf{u}_i\}_{1 \leq i \leq n}$ .

- Stage I: For each tract,  $\mu_m(s_m)$  and  $\eta_{i,m}(s_m)$  are estimated from (1) and functional principal component analysis is applied to calculate  $\hat{\phi}_{md}(s_m)$  and  $\hat{\mathbf{z}}_i$ ,
- Stage II: Perform factor analysis on  $\hat{\mathbf{z}}_i$  to extract global factor  $\mathbf{f}_i$  and individual factor  $\mathbf{u}_i$ . Regression and hypothesis testing can then be applied on  $\hat{\mathbf{f}}_i$  and  $\hat{\mathbf{u}}_i$  respectively.

Details of the two stages are given below.

In Stage I, to estimate the mean curve from model (1), we apply the local linear kernel smoothing technique.  $\mu_m(s_m)$  is first approximated by the following Taylor expansion,

$$\mu_m(s_{m,k}) \approx \mu_m(s_m) + d\mu_m(s_m)(s_{m,k} - s_m). \tag{6}$$

Let  $K(s)$  be a predetermined smoothing kernel and denote  $K_h(s) = \frac{1}{h}K(\frac{s}{h})$  as the rescaled function with bandwidth  $h$ ,  $\hat{\mu}_m(s_m)$  and  $d\hat{\mu}_m(s_m)$  can be estimated as the minimizers of the following weighted least square function,

$$\sum_{i=1}^n \sum_{k=1}^{K_m} [y_{i,m}(s_{m,k}) - \mu_m(s_m) - d\mu_m(s_m)(s_{m,k} - s_m)]^2 K_h(s_{m,k} - s_m), \tag{7}$$

and solution  $\hat{\mu}_m(s_m)$  is smooth curve with local linearity. More complicated polynomial structure can be applied using higher order expansion if necessary.

Similarly, we expand individual function  $\eta_{i,m}(s_m)$  for subject  $i$  as follows,

$$\eta_{i,m}(s_{m,k}) \approx \eta_{i,m}(s_m) + d\eta_{i,m}(s_m)(s_{m,k} - s_m). \tag{8}$$

The corresponding weighted least square function is given by,

$$\sum_{k=1}^{K_m} [y_{i,m}(s_{m,k}) - \hat{\mu}_m(s_{m,k}) - \eta_{i,m}(s_m) - d\eta_{i,m}(s_m)(s_{m,k} - s_m)]^2 K_h(s_{m,k} - s_m). \tag{9}$$

When smoothed individual functions are obtained as  $\{\hat{\eta}_{i,m}(s_m)\}_{i=1}^n$ , we can calculate the empirical covariance function  $\hat{\Sigma}_{\eta_m}(s_m, s'_m) = \frac{1}{n} \sum_{i=1}^n \hat{\eta}_{i,m}(s_m) \hat{\eta}_{i,m}(s'_m)$ . And eigenbases  $\{\hat{\phi}_{md}(s_m)\}$  can be estimated from spectral decomposition,

$$\hat{\Sigma}_{\eta_m}(s_m, s'_m) = \sum_d \hat{\lambda}_{md} \hat{\phi}_{md}(s_m) \hat{\phi}_{md}(s'_m). \tag{10}$$

Then individual random effect  $\hat{\eta}_{i,m}(s_m)$  is projected onto basis functions  $\{\hat{\phi}_{md}(s_m)\}$  to get functional PC scores,

$$\hat{z}_{i,md} = \sum_{k=1}^{K_m} \hat{\eta}_{i,m}(s_{k,m}) \hat{\phi}_{md}(s_{k,m}). \quad (11)$$

There are several strategies to determine the number of fPCs to be extracted. For example, the analog of some model selection techniques have been generalized for this purpose, such as Akaike information criterion (AIC), Bayesian information criterion (BIC) [25] and cross-validation (CV) [20]. Alternatively, the percentage of explained variation has been widely used to give an appropriate cut-off in practice. Here, we choose  $D$  as the minimum number of fPCs that incorporates at least  $V\%$  of total variation in each tract. When the optimal  $D = D_m$  is different across tracts, the largest  $D_m$  will be used for all tracts.

In Stage II, a PCA-based factor analysis is performed. Let  $\hat{\xi}_1, \dots, \hat{\xi}_L$  be the first  $L$  eigenvectors of sample covariance matrix  $\hat{\Sigma}_{\mathbf{z}} = \frac{1}{n} \hat{\mathbf{Z}}^T \hat{\mathbf{Z}}$ . The loading matrix, the global factors and the individual factors are estimated as,

$$\hat{\mathbf{A}} = \sqrt{p}(\hat{\xi}_1, \dots, \hat{\xi}_L), \quad \hat{\mathbf{F}} = \frac{1}{p} \hat{\mathbf{Z}} \hat{\mathbf{A}}, \quad \text{and} \quad \hat{\mathbf{U}} = \hat{\mathbf{Z}} - \hat{\mathbf{F}} \hat{\mathbf{A}}^T \quad (12)$$

Finally, the MLM (5) is used to estimate regression coefficients. Standard test statistics, such as wald and score statistics, can be applied subsequently for inference purpose.

### 3 Simulations

In this section, numerical simulations are conducted to evaluate the proposed method. Particularly, we examine the performance of HFPRM to detect covariate effect in hypothesis testing.

#### 3.1 Setup

11 fiber tracts with FA measure shown in Table 1 were selected from diffusion tensor tractography in UNC Early Human Brain Development Studies [7]. Functional responses were simulated from a vary coefficient model with fixed covariate effects,

$$y_{i,m}(s_m) = \mu_m(s_m) + \boldsymbol{\beta}_m(s_m)^T \mathbf{x}_i + \eta_{i,m}(s_m) + e_{i,m}(s_m), \quad (13)$$

where  $i = 1, \dots, n$  and  $m = 1, \dots, 11$ ,  $\boldsymbol{\beta}_m(s_m)$  was a  $q \times 1$  vector of coefficient functions along the  $m$ -th tract, covariates  $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})^T$  were generated from  $N(0, 1)$  for continuous variables or from multinomial distribution with equal probabilities for categorical variables,  $\eta_{i,m}(s_m)$  followed gaussian process  $GP\{0, \Sigma_{\eta_m}\}$  and  $e_{i,m}(s_m)$  followed  $GP\{0, \Sigma_{e_m}\}$ . Compared to model (1), the above equation directly specified the covariates as fixed effect. Sample size  $n$

was set to be 100 and true parameters  $(\beta(s_m), \Sigma_{\eta_m}, \Sigma_{e_m})$  were estimated from real data using FADTTS [28].

To examine our method, the following two scenarios on  $\beta(s_m)^T x_i$  were simulated. In case I, the aim is to study shared effect of multiple tracts. Gender (G) and gestational age at birth (Gage) were included as covariates for all the 11 tracts,

$$y_{i,m}(s_m) = \mu_m(s_m) + c\beta_{m,1}(s_m)\text{Gage}_i + \beta_{m,2}(s_m)\text{G}_i + \eta_{i,m}(s_m) + e_{i,m}(s_m), \quad \forall m,$$

in which we assumed  $c = 0, 0.2, 0.4, 0.6$  and Gage effect was tested.

In case II, we want to examine a tract-specific effect. Birth weight (BW) was added as covariate to one particular tract, right uncinat fasciculus ( $m = 11$ ), in addition to case I,

$$\begin{aligned} y_{i,m}(s_m) &= \mu_m(s_m) + \beta_{m,1}(s_m)\text{Gage}_i + \beta_{m,2}(s_m)\text{G}_i + \eta_{i,m}(s_m) + e_{i,m}(s_m), \quad m \leq 10, \\ y_{i,11}(s_m) &= \mu_m(s_m) + \beta_{11,1}(s_m)\text{Gage}_i + \beta_{11,2}(s_m)\text{G}_i + c\beta_{11,3}(s_m)\text{BW}_i \\ &\quad + \eta_{i,11}(s_m) + e_{i,11}(s_m), \end{aligned}$$

where effect size  $c$  was set to take values 0, 0.5, 1, 1.5 and the effect of BW was tested.

We applied HFPRM to the simulated dataset. The varying coefficient model (1) was first fitted to estimate individual functions. Functional principal components were then extracted such that at least 85% of total variation is included for each tract. In factor analysis, the first elbow point in the scree plot was taken as a cut-off to determine the number of global factors. In testing step, type I error and statistical power were calculated at significance level  $\alpha = 0.05$  based on 1000 simulation replications. FADTTS was also applied on each single tract and the results were compared.

### 3.2 Results

In case I, the first five functional principal components were extracted for each tract and the first factor was identified as global factor. The rejection rates for global factor analysis and FADTTS on testing Gage effect are presented by Fig. 2(a). The global factor analysis is substantially more powerful than the single tract analysis when detecting commonly shared effect. Such results are expected since common effect tends to be accumulated in the global factor.

In case II, the first five functional principal components were extracted for each tract and the first two factors were identified as global factors. Figure 2(b) shows the rejection rates for global factor analysis, individual factor analysis and FADTTS on testing BW effect. As can be seen, individual factor analysis in HFPRM achieves comparable power to single tract analysis for detecting tract-specific effect.

## 4 Early Human Brain Development Study

To investigate how genetic factors influence brain structure in prenatal and early postnatal stage, we conducted a genome-wide association study on the fiber

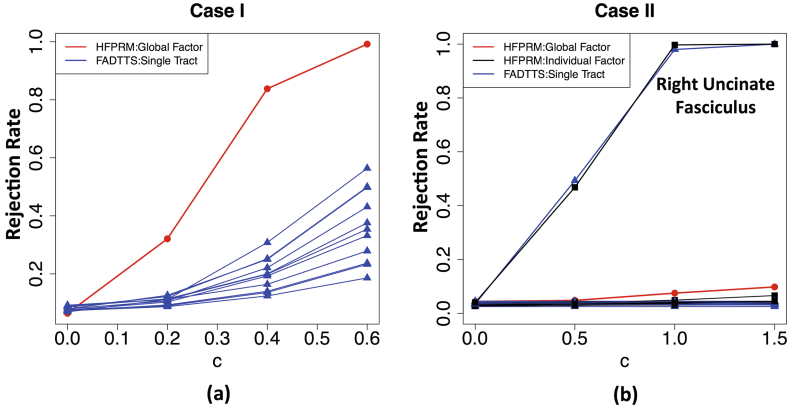


Fig. 2. Simulation result

bundle statistics in a unique cohort of infants. A total number of 662 neonatal twin subjects were taken from the UNC Early Brain Development Studies [7].

### 4.1 Data Acquisition and Preprocessing

MRI scans were acquired either on a 3T Siemens Allegra head-only scanner ( $N = 566$ ) or on a 3T Siemens TIM Trio 3T scanner ( $N = 96$ ). For the Allegra model, 339 diffusion weighted images were acquired by a single shot EPI DTI sequence with the following parameters:  $TR/TE = 5200/73$  ms, voxel resolution =  $2 \times 2 \times 2$  mm<sup>3</sup>, 6 non-collinear directions with  $b = 1000$  s/mm<sup>2</sup> and 1 baseline image with  $b = 0$ . To improve the signal-to-noise ratio, five scans were repeated and averaged. For the remaining subjects scanned on Allegra, DWI was acquired with the following parameters:  $TR/TE = 7680/82$  ms, voxel resolution =  $2 \times 2 \times 2$  mm<sup>3</sup>, 42 non-collinear directions with diffusion gradients of  $b = 1000$  s/mm<sup>2</sup> in addition to 7 baseline images. For the Trio model, DWIs were acquired using a similar protocol to that of the 42 direction Allegra model with  $TR/TE = 7200/83$  ms. Quality control was applied on raw DWIs using DTIPrep [18], and FSL [11, 22] was performed for skull stripping and brain masking. We used a weighted least squares method [8] to estimate diffusion tensors and followed the UNC-Utah NA-MIC framework [23] to create a study-specific atlas. Subsequently, a total number of 44 fiber tracts listed in Table 1 were reconstructed in the atlas space using a streamline algorithm [5]. For each subject, four scalar diffusion properties, FA, MD, AD and RD, were then calculated at each location along each tract using neighboring diffusion tensors.

Genotyping of single nucleotide polymorphisms (SNPs) was conducted on Affymetrix Axiom genome-wide LAT Array. Samples with call rates less than 95%, outliers for homozygosity, ancestry outliers and unexpected relatedness were excluded from the study. We also removed genetic markers with Hardy-Weinberg equilibrium p-value less than  $10^{-8}$ , call rate less than 95% and



**Table 1.** List of fiber tracts in simulation and real data experiment

Bundle group	Tract segments
Arcuate fasciculus	Left fronto-parietal, right fronto-parietal, left fronto-temporal*, right fronto-temporal*, right temporo-parietal
Corpus callosum	Motor body*, occipital splenium, parietal body*, premotor body, rostrum*, genu*, temporal tapetum*
Cingulum	Left premotor, left cingulate gyrus, right cingulate gyrus, right hippocampal, right prefrontal cortex
Corticothalamic	left motor, right motor, left premotor, right premotor, left parietal, right parietal, left prefrontal, right prefrontal
CorticoFugal	Left motor, right motor, left parietal, right parietal, left prefrontal cortex,
Others	Left fornix, right fornix, left inferior fronto-occipital fasciculi, right inferior fronto-occipital fasciculi, left inferior longitudinal fasciculi*, right inferior longitudinal fasciculi*, left medial lemniscus, right medial lemniscus, left optic, right optic, left superior longitudinal fasciculus, right superior longitudinal fasciculus, left uncinate fasciculus*, right uncinate fasciculus*

\*Selected tracts for simulation study

Mendelian error rate larger than 10%. Population stratification was assessed using PCA [19]. Imputation was performed with MaCH-Admix [15] using 1000G reference panel [3]. To evaluate the quality of imputed SNPs, we computed the mean  $R^2$  under varying minor allele frequency (MAF) categories and selected  $R^2$  cutoffs as described in [14]. SNPs with MAF less than 0.01 were excluded from imputed dataset. Eventually, 472 twin subjects (32 MZ pairs, 75 DZ pairs and 259 singletons or unpaired twin subjects) and 8,538,562 genetic markers were retained for further analysis.

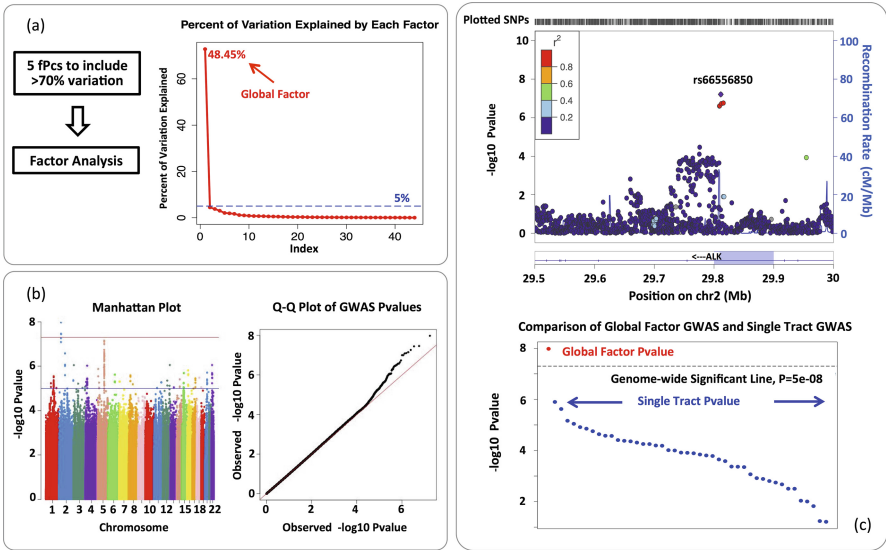
## 4.2 Data Analysis

In this experiment, we chose to focus on the fractional anisotropy (FA) measure. FA quantifies the extent of local directional water diffusion and partially reflects the degree of bundle maturation in premature brains [4]. To eliminate the heterogeneity in variance among different tracts,  $y_{i,m}(s_m)$  was rescaled by the total standard deviation along the tract. For the twin study, ACE model was fitted in (5) to account for correlation within twin pairs. Seven variables were added as covariates, including gestational age at birth, gender, DTI direction, scanner type and the first three genetic principal component to adjust for population stratification.

## 4.3 Results

In functional PCA, the first 5 functional principal components were extracted for each tract to include at least 70% of variation. Figure 3(a) shows the scree

plot in factor analysis and the elbow point is located at factor 2. Therefore, the first factor is identified as the global factor. We then performed GWAS on the global factor. The result is visualized by Fig. 3(b). In the Manhattan plot, we observed a significant region in anaplastic lymphoma kinase (ALK) gene on chromosome 2. The ALK gene is a neuronal orphan receptor tyrosine kinase that plays an important role in the nervous system development [1], and is highly expressed in the neonatal brain [10]. As a comparison, we also performed association analysis for top hit rs66556850 on each single tract. The result is presented by Fig. 3(c). A number of tracts have relatively small pvalue yet not small enough to be detected by a single tract GWAS. It indicates that the global factor analysis is more powerful to detect commonly shared genetic effect than single tract analysis.



**Fig. 3.** Real data analysis result: (a) Functional PCA and factor analysis. (b) Visualization of GWAS result of the global factor. (c) A comparison between global factor analysis and single tract analysis on marker rs66556850, the  $-\log_{10}p$  value in the association test is plotted. The majority of pvalues in single tract analysis are around  $10^{-2} \sim 10^{-6}$ .

## 5 Conclusion

We have developed a hierarchical functional principal regression model (HFPRM) to efficiently conduct joint analysis on diffusion statistics from multiple neurofiber bundles. A varying coefficient model is introduced and functional PCA is applied to capture major tract variation. Factor analysis is then adopted to extract key features at both global level and individual level. Finally, standard estimation and testing procedures can be applied to study global effect and

tract-specific effect. Simulation results demonstrated that HFPRM is powerful to detect common effect shared by multiple tracts. HFPRM has also been successfully applied to a genome-wide association study on neonatal twins. We are able to identify some important genetic variants related to early childhood brain development that were ignored by single tract analysis.

## References

1. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/gene/238>
2. Bach, M., Laun, F.B., Leemans, A., Tax, C.M., Biessels, G.J., Stieltjes, B., Maier-Hein, K.H.: Methodological considerations on tract-based spatial statistics (TBSS). *Neuroimage* **100**, 358–369 (2014)
3. Genomes Project Consortium, et al.: An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65 (2012)
4. Dubois, J., Hertz-Pannier, L., Dehaene-Lambertz, G., Cointepas, Y., Le Bihan, D.: Assessment of the early organization and maturation of infants' cerebral white matter fiber bundles: a feasibility study using quantitative diffusion tensor imaging and tractography. *Neuroimage* **30**(4), 1121–1132 (2006)
5. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* **30**(9), 1323–1341 (2012)
6. Garyfallidis, E., Ocegueda, O., Wassermann, D., Descoteaux, M.: Robust and efficient linear registration of white-matter fascicles in the space of streamlines. *NeuroImage* **117**, 124–140 (2015)
7. Gilmore, J.H., Schmitt, J.E., Knickmeyer, R.C., Smith, J.K., Lin, W., Styner, M., Gerig, G., Neale, M.C.: Genetic and environmental contributions to neonatal brain structure: a twin study. *Hum. Brain Mapp.* **31**(8), 1174–1182 (2010)
8. Goodlett, C.B., Fletcher, P.T., Gilmore, J.H., Gerig, G.: Group analysis of DTI fiber tract statistics with application to neurodevelopment. *NeuroImage* **45**, S133–S142 (2009)
9. Guevara, P., Poupon, C., Rivière, D., Cointepas, Y., Descoteaux, M., Thirion, B., Mangin, J.: Robust clustering of massive tractography datasets. *NeuroImage* **54**(3), 1975–1993 (2011)
10. Iwahara, T., Fujimoto, J., Wen, D., Cupples, R., Bucay, N., Arakawa, T., Mori, S., Ratzkin, B., Yamamoto, T.: Molecular characterization of ALK, a receptor tyrosine kinase expressed specifically in the nervous system. *Oncogene* **14**(4), 439–449 (1997)
11. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: FSL. *Neuroimage* **62**(2), 782–790 (2012)
12. Jin, Y., Shi, Y., Zhan, L., Gutman, B.A., de Zubizaray, G.I., McMahon, K.L., Wright, M.J., Toga, A.W., Thompson, P.M.: Automatic clustering of white matter fibers in brain diffusion mri with an application to genetics. *NeuroImage* **100**, 75–90 (2014)
13. Karhunen, K.: Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fennicae Ser. A* **1**, 34 (1946)

14. Liu, E.Y., Buyske, S., Aragaki, A.K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D.C., Haessler, J., Hindorff, L.A., et al.: Genotype imputation of metabochips using a study-specific reference panel of 4,000 haplotypes in african americans from the women's health initiative. *Genet. Epidemiol.* **36**(2), 107–117 (2012)
15. Liu, E.Y., Li, M., Wang, W., Li, Y.: Mach-admix: genotype imputation for admixed populations. *Genet. Epidemiol.* **37**(1), 25–37 (2013)
16. Loève, M.: Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique* **84**, 159–162 (1946)
17. O'Donnell, L.J., Westin, C.F., Golby, A.J.: Tract-based morphometry for white matter group analysis. *NeuroImage* **45**, 832–844 (2009)
18. Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., Johnson, H.J., Styner, M.A.: DTIPrep: quality control of diffusion-weighted images. *Front. Neuroinform.* **8**, 4 (2014)
19. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**(8), 904–909 (2006)
20. Rice, J.A., Silverman, B.W.: Estimating the mean and covariance structure non-parametrically when the data are curves. *J. R. Stat. Soc. Ser. B (Methodol.)* **53**, 233–243 (1991). JSTOR
21. Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., et al.: Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* **31**(4), 1487–1505 (2006)
22. Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al.: Advances in functional and structural mr image analysis and implementation as FSL. *Neuroimage* **23**, S208–S219 (2004)
23. Verde, A.R., Budin, F., Berger, J.B., Gupta, A., Farzinfar, M., Kaiser, A., Ahn, M., Johnson, H.J., Matsui, J., Hazlett, H.C., et al.: UNC-Utah NA-MIC framework for DTI fiber tract analysis. *Front. Neuroinform.* **7**, 51 (2014)
24. Wedeen, V.J., Rosene, D.L., Wang, R., Dai, G., Mortazavi, F., Hagmann, P., Kaas, J.H., Tseng, W.Y.I.: The geometric structure of the brain fiber pathways. *Science* **335**(6076), 1628–1634 (2012)
25. Yao, F., Müller, H.G., Wang, J.L.: Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* **100**(470), 577–590 (2005)
26. Yuan, Y., Gilmore, J.H., Geng, X., Martin, S., Chen, K., Wang, J.I., Zhu, H.: FMEM: functional mixed effects modeling for the analysis of longitudinal white matter tract data. *NeuroImage* **84**, 753–764 (2014)
27. Yushkevich, P.A., Zhang, H., Simon, T.J., Gee, J.C.: Structure-specific statistical mapping of white matter tracts. *NeuroImage* **41**, 448–461 (2008)
28. Zhu, H., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W., Gilmore, J.H.: FADTTS: functional analysis of diffusion tensor tract statistics. *NeuroImage* **56**(3), 1412–1425 (2011)