Covariate-Adjusted Hybrid Principal Components Analysis

Aaron Wolfe Scheffler[1](✉), Abigail Dickinson[2], Charlotte DiStefano[2], Shafali Jeste[2], and Damla Şentürk[3]

[1] Department of Epidemiology and Biostatistics, University of California, San Francisco, USA
aaron.scheffler@ucsf.edu

[2] Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, USA

[3] Department of Biostatistics, University of California, Los Angeles, USA

Abstract. Electroencephalography (EEG) studies produce region-referenced functional data in the form of EEG signals recorded across electrodes on the scalp. The high-dimensional data capture underlying neural dynamics and it is of clinical interest to model differences in neurodevelopmental trajectories between diagnostic groups, for example typically developing (TD) children and children with autism spectrum disorder (ASD). In such cases, valid group-level inference requires characterization of the complex EEG dependency structure as well as covariate-dependent heteroscedasticity, such as changes in variation over developmental age. In our motivating study, resting state EEG is collected on both TD and ASD children aged two to twelve years old. The peak alpha frequency (PAF), defined as the location of a prominent peak in the alpha frequency band of the spectral density, is an important biomarker linked to neurodevelopment and is known to shift from lower to higher frequencies as children age. To retain the most amount of information from the data, we model patterns of alpha spectral variation, rather than just the peak location, regionally across the scalp and chronologically across development for both the TD and ASD diagnostic groups. We propose a covariate-adjusted hybrid principal components analysis (CA-HPCA) for region-referenced functional EEG data, which utilizes both vector and functional principal components analysis while simultaneously adjusting for covariate-dependent heteroscedasticity. CA-HPCA assumes the covariance process is weakly separable conditional on observed covariates, allowing for covariate-adjustments to be made on the marginal covariances rather than the full covariance leading to stable and computationally efficient estimation. A mixed effects framework is proposed to estimate the model components coupled with a bootstrap test for group-level inference. The proposed methodology provides novel insights into neurodevelopmental differences between TD and ASD children.

Keywords: Electroencephalography · Autism spectrum disorder · Functional data analysis · Marginal covariances · Functional principal components analysis · Covariate-adjustments

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 391–404, 2020.
https://doi.org/10.1007/978-3-030-50153-2_30

1 Introduction

Despite the numerous developmental delays observed in children with autism spectrum disorder (ASD) compared to their typically developing peers (TD), the neural mechanisms underpinning these delays are not well characterized. To address this gap, our motivating study collected resting-state electroencephalograms (EEG) on TD and ASD children aged two to twelve years old, making it possible to contrast neural processes between the two diagnostic groups over a wide developmental range. EEG and magnetoencephalography (MEG) characterize cortical and intracortical brain activity, respectively, via the measurement of electrical potentials and their corresponding oscillatory dynamics (i.e. spectral characteristics). Recent studies in cognitive development using both EEG and MEG highlight the peak alpha frequency (PAF), defined as the location of a single prominent peak in the spectral density within the alpha frequency band (6–14 Hz), as a potential biomarker associated with autism diagnosis [7–9]. Specifically, the location of the PAF tends to shift from lower to higher frequencies as TD children age but this chronological shift is notably delayed or absent in ASD children [7,8,12,16]. This trend is observed in our motivating data from a temporal electrode (T8) where the PAF, identifiable as 'humps' in age-specific slices of the group-specific bivariate mean spectral density (across age and frequency), increases in frequency with age for TD children but not for ASD children (Fig. 1(a)).

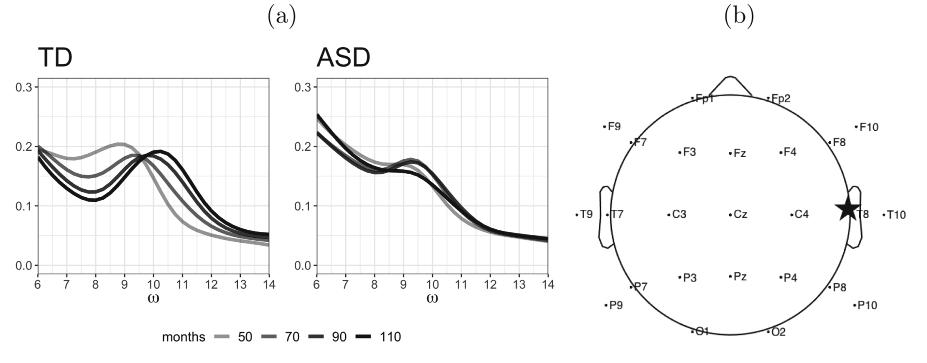

Fig. 1. (a) The group-specific bivariate mean alpha band spectral density (across age and frequency (6–14 Hz)) at ages 50, 70, 90 and 110 months from the T8 electrode. (b) A diagram of the 25 electrode montage used in our motivating data with the T8 electrode marked by a star.

Although the PAF holds promise as a biomarker for neural development in TD and ASD children, emphasis on the identification of a single peak produces considerable drawbacks. Estimation of a subject-electrode specific PAF can be error prone due to the presence of noise and multiple local maxima [6] and measurement of PAF inherently reduces the information from the alpha spectral band to a single scalar summary resulting in a loss of information. To avoid these limitations, we follow Scheffler et al. [15] and consider the entire spectral density across the alpha band as a functional measurement of neural activity. We focus on modeling and contrasting patterns of alpha spectral variation regionally across the scalp and chronologically across development for both the ASD and TD diagnostic groups. While Scheffler et al. [14] proposed a hybrid principal components analysis (HPCA) decomposition that models variation in region-referenced functional EEG data, it does not allow for the covariance structure to change across development as needed in our application. Previous research clearly shows that alpha spectral dynamics differ as a function of age between TD and ASD children and to assume a constant covariance structure across development risks missing important findings. To avoid this misspecification, we propose a covariate-adjusted hybrid principal components analysis (CA-HPCA) that models variation in high-dimensional functional data while simultaneously allowing the patterns of variation to change as a function of subject-specific covariates. CA-HPCA assumes the covariance process is weakly separable conditional on observed covariates, allowing for covariate-adjustments to be made on the marginal covariances rather than the full covariance leading to stable and computationally efficient estimation.

In the simplified context of one-dimensional functional data, existing methods allow for covariate-adjustments to the functional covariance in two ways: (1) both the eigenvalues and eigenfunctions of the functional covariance are allowed to change as a function of observed covariates or (2) the eigenfunctions are assumed to be constant across the covariate dimension but their corresponding eigenvalues (hence principal scores) are covariate-dependent. In the former class, Cardot [2] proposed a non-parametric covariate-adjusted functional principal components analysis (FPCA) in the context of dense functional data and Jiang and Wang [10] extended covariate-adjusted FPCA to noisy or sparse settings by estimating subject-specific scores using conditional expectation. In both cases, covariance estimation is performed non-parametrically by simultaneous smoothing across the covariate and functional domains via kernel methods. By fixing eigenfunctions across the covariate domain, Chio et al. [5] introduced a semi-parametric functional regression model that estimates covariate-dependent principal scores using a single-index model and Backenroth et al. [1] developed a heteroscedastic FPCA for repeatedly measured curves that models eigenvalues as an exponential function of covariate and subject-dependent effects.

Our proposed covariate-adjusted hybrid principal components analysis combines existing one-dimensional methods for covariate-dependent functional heteroscedasticity with recent advances in multi-dimensional FPCA to allow covariate-adjustments in the context of high-dimensional functional data. We briefly explore the methodological contributions of our proposed model and the resulting computational gains. A central theme in FPCA decompositions for multi-dimensional functional data is the use of simplifying assumptions regarding the covariance structure to ease estimation. A flexible approach in modeling two-dimensional functional data is to assume weak separability of the covariance process [4,11] in which the marginal covariances along each dimension are targeted and the full covariance is projected onto a tensor basis formed from the corresponding marginal eigenfunctions. Thus, estimation is reduced from that of the total covariance in four-dimensions to the marginal covariances in two-dimensions for which efficient two-dimensional smoothers exist. Scheffler et al. [14] extended weak separability to region-referenced functional EEG data (similar to our motivating study) by allowing a discrete regional dimension via HPCA. We leverage the simplifying assumptions and computational efficiency of HPCA through the proposed CA-HPCA which introduces covariate-dependence to the marginal covariances rather than the total covariance and allows the marginal eigenvalues and eigenfunctions to change across the covariate domain.

CA-HPCA provides a flexible modeling framework but introduces potential compute burden through the addition of a covariate dimension to estimation of the marginal covariances which for a scalar covariate requires smoothing across three dimensions. Previous methods such as Cardot [2] and Jiang and Wang [10] utilized kernel methods to estimate covariate-dependent marginal covariances but these approaches are computationally intensive and scale poorly with the introduction of additional covariates. To address this challenge, we extend the fast functional covariance smoothing proposed by Cederbaum et al. [3] to allow for covariate-adjustments by including an additional basis along the covariate dimension. Thus, CA-HPCA generalizes covariate-adjustments to high-dimensional functional covariances and substantially reduces the resulting computational burden by applying adjustments to the marginal covariances and introducing covariate-dependence to cutting-edge fast covariance smoothers. A mixed effects framework is proposed to estimate the model components and is paired with parametric bootstrap resampling to perform inference across the covariate domain. The remaining sections are organized as follows. Section 2 introduces the proposed CA-HPCA and Sect. 3 describes the corresponding estimation procedure. Application of the proposed method to our resting state EEG data follows in Sect. 4. Section 5 concludes with a brief summary and discussion.

2 Covariate-Adjusted Hybrid Principal Components Analysis (CA-HPCA)

Let $Y_{di}(a_i, r, \omega)$ be a random function observed in the presence of some continuous non-functional covariate $a_i \in \mathcal{A}$, for subject i, $i = 1, \dots, n_d$, from group d, $d = 1, \dots, D$, in region r, $r = 1, \dots, R$, and at frequency ω, $\omega \in \varOmega$. We decompose $Y_{di}(a_i, r, \omega)$ additively such that the expectation and covariance of the process depend on the covariate a_i,

$$Y_{di}(a_i, r, \omega) = \eta_d(a_i, r, \omega) + Z_{di}(a_i, r, \omega) + \epsilon_{di}(a_i, r, \omega),$$

where $\eta_d(a_i, r, \omega) = E\{Y_{di}(a_i, r, \omega)|a_i\}$ denotes the group-region mean function, $Z_{di}(a_i, r, \omega)$ denotes a mean zero region-referenced stochastic process with total variance $\boldsymbol{\Sigma}_{d,T}(a_i; r, \omega; r', \omega') = \mathrm{cov}\{Z_{di}(a_i, r, \omega), Z_{di}(a_i, r', \omega')|a_i\}$, and

$\epsilon_{di}(a_i, r, \omega)$ denotes measurement error with mean zero and variance σ_d^2 that is independent across the regional and functional domains. We assume the group-region means $\eta_d(a, r, \omega)$ are smooth in both the functional domain Ω and the non-functional domain $\mathcal{A}$ though we place no restrictions across the regional domain $\mathbb{R}^R$.

In the proposed CA-HPCA model, we assume that the total covariance $\boldsymbol{\Sigma}_{d,T}(a; r, \omega; r', \omega')$ is weakly separable for each $a \in \mathcal{A}$. Weak separability, a concept recently proposed by Lynch and Chen [11] and adapted by Scheffler et al. [14] to region-referenced functional EEG data, implies that a covariance can be approximated by a weighted sum of separable covariance components and that the direction of variation (i.e. eigenvectors/eigenfunctions) along one dimensions of the EEG data is the same across fixed slices of the other dimension. Note that weak separability is more flexible than strong separability (i.e. separability) commonly utilized in spatiotemporal modeling which requires the total covariance, not just the directions of variation, is the same up to a constant for fixed slices of the other dimensions. Unlike previously applications of weak separability, we assume that the total covariance is weakly separable conditional on observed covariates and the marginal covariance functions vary smoothly along the covariate domain. Let the covariate-adjusted regional and functional marginal covariances be defined as

$$\{\boldsymbol{\Sigma}_{d,\mathcal{R}}(a_i)\}_{r,r'} = \int_{\Omega} \text{cov}\{Z_{di}(a_i, r, \omega), Z_{di}(a_i, r', \omega)\} d\omega = \sum_{k=1}^{R} \tau_{dk,\mathcal{R}}(a_i) \mathrm{v}_{dk}(a_i, r) \mathrm{v}_{dk}(a_i, r'),$$

$$\boldsymbol{\Sigma}_{d,\Omega}(a_i, \omega, \omega') = \sum_{r=1}^{R} \text{cov}\{Z_{di}(a_i, r, \omega), Z_{di}(a_i, r, \omega')\} = \sum_{\ell=1}^{\infty} \tau_{d\ell,\Omega}(a_i) \phi_{d\ell}(a_i, \omega) \phi_{d\ell}(a_i, \omega'),$$

where $\mathrm{v}_{dk}(a, r)$ are covariate-adjusted marginal eigenvectors, $\phi_{d\ell}(a, \omega)$ are covariate-adjusted marginal eigenfunctions, and $\tau_{dk,\mathcal{R}}(a)$ and $\tau_{d\ell,\Omega}(a)$ are their respective covariate-adjusted marginal eigenvalues. Utilizing the covariate-dependent eigenvectors and eigenfunctions, the covariate-adjusted hybrid principal components decomposition (**CA-HPCA**) of $Y_{di}(a_i, r, \omega)$ is given as,

$$\begin{aligned} Y_{di}(a_i, r, \omega) &= \eta_d(a_i, r, \omega) + Z_{di}(a_i, r, \omega) + \epsilon_{di}(a_i, r, \omega) \\ &= \eta_d(a_i, r, \omega) + \sum_{k=1}^{R} \sum_{\ell=1}^{\infty} \xi_{di,k\ell}(a_i) \mathrm{v}_{dk}(a_i, r) \phi_{d\ell}(a_i, \omega) + \epsilon_{di}(a_i, r, \omega), \end{aligned}$$

where $\xi_{di,k\ell}(a_i)$ are subject-specific scores defined through the projection $\langle Z_{di}(a_i, r, \omega),\ \mathrm{v}_{dk}(a_i, r)\phi_{d\ell}(a_i, \omega)\rangle = \sum_{r=1}^{R} \int Z_{di}(a_i, r, \omega)\ \mathrm{v}_{dk}(a_i, r)\phi_{d\ell}(a_i, \omega) d\omega$.

Weak separability of the total covariance at each covariate value implies that the scores $\xi_{di,k\ell}(a_i)$ are uncorrelated leading to the decomposition of the total covariance $\boldsymbol{\Sigma}_{d,T}(a;r,\omega;r',\omega')$ as follows,

$$\begin{aligned}\boldsymbol{\Sigma}_{d,T}(a;r,\omega;r',\omega') &= \operatorname{cov}\{Z_{di}(a,r,\omega),Z_{di}(a,r',\omega')|a\}+\sigma_d^2\delta(a;r,\omega;r',\omega')\\ &= \sum_{k=1}^{R}\sum_{\ell=1}^{\infty}\tau_{d,k\ell}(a)\mathrm{v}_{dk}(a,r)\mathrm{v}_{dk}(a,r')\phi_{d\ell}(a,\omega)\phi_{d\ell}(a,\omega')+\sigma_d^2\delta(a;r,\omega;r',\omega'),\end{aligned}$$

where $\tau_{d,k\ell}(a)=\operatorname{var}\{\xi_{di,k\ell}(a)\}$. Note, the above model assumes that both the marginal directions of variation and their associated tensor weights are allowed to vary across the covariate domain. In practice, the CA-HPCA decomposition is truncated to include only K_d and L_d covariate-adjusted marginal eigencomponents for the regional and functional domains, respectively, with the number of components initially selected on the marginal fraction of variance explained (FVE). One guideline is to include the minimum number of covariate-adjusted marginal eigencomponents in the CA-HPCA expansion that explain at least 90% of variation in their respective covariate-adjusted marginal covariances. The final number of components can be fixed after the subject-specific scores and their associated variance components are estimated in Sect. 3 which allow enumeration of the overall FVE in the observed data not just the marginal covariances.

3 Estimation of Model Components and Inference

The following section outlines the CA-HPCA estimation procedure, provides detailed descriptions of each step, and outlines how to perform inference via parametric bootstrap.

(1) Estimation of group-region mean functions: We estimate the group-region mean function $\eta_d(a,r,\omega)$ for each region via smoothing performed by projection onto a tensor basis formed by penalized marginal B-splines in the regional and functional domains. Smoothing parameter selection and variance components are estimated using restricted maximum likelihood (REML) methods.

(2) Estimation of covariate-adjusted marginal covariances and measurement error variance: We estimate the covariate-adjusted marginal covariances by assuming each two-dimensional marginal covariance varies smoothly over the covariate dimension. For the functional marginal covariance, $\boldsymbol{\Sigma}_{d,\Omega}(a,\omega,\omega')$, we extend the fast bivariate covariance smoother of Cederbaum et al. [3] to include a third covariate dimension $a\in\mathcal{A}$. The resulting trivariate smoother maintains the computational efficiency of Cederbaum et al. [3] while simultaneously allowing the marginal functional covariance to vary smoothly along the covariate dimension. As an added bonus, we also obtain an initial estimate of the measurement error variance $\hat{\sigma}^2_{d,\Omega}$.

Algorithm: *CA-HPCA Estimation Procedure*

```
1. Estimation of group-region mean functions
   (a) Calculate $\hat{\eta}_d(a_i, r, \omega)$ by applying a bivariate penalized spline smoother
       to all observed data $\{a_i, \omega, Y_{di}(a_i, r, \omega) : \ i = 1, \ldots, n_d; a_i \in \mathcal{A}; \omega \in \Omega\}$.
   (b) Mean center each observation, $\widehat{Y}^c_{di}(a_i, r, \omega) = Y_{di}(a_i, r, \omega) - \hat{\eta}_d(a_i, r, \omega)$.
2. Estimation of covariate-adjusted marginal covariances and measurement
   error variance
   (a) Calculate $\widehat{\boldsymbol{\Sigma}}_{d,\Omega}(a, \omega, \omega')$ and $\hat{\sigma}^2_{d,\Omega}$ by applying trivariate penalized spline
       smoothers to the products,
       $\{a_i, \omega, \omega', \widehat{Y}^c_{di}(a_i, r, \omega)\widehat{Y}^c_{di}(a_i, r, \omega') :$b $i = 1, \ldots, n_d; a_i \in \mathcal{A}; \omega, \omega' \in \Omega\}$.
   (b) Calculate $\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}(a)$ by smoothing each $(r, r')$ entry across $\mathcal{A}$. For $r \neq r'$,
       estimate $\{\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}(a)\}_{(r,r')}$ by applying a univariate kernel smoother to
       $\{a_i, r, r', \widehat{Y}^c_{di}(a_i, r, \omega)\widehat{Y}^c_{di}(a_i, r', \omega) : \ i = 1, \ldots, n_d; a_i \in \mathcal{A}\}$. For $r = r'$,
       estimate $\{\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}(a)\}_{(r,r)}$ by applying a univariate penalized smoother to
       $\{a_i, r, r, \widehat{Y}^c_{di}(a_i, r, \omega)\widehat{Y}^c_{di}(a_i, r, \omega) - \hat{\sigma}^2_{d,\Omega} : i = 1, \ldots, n_d; a_i \in \mathcal{A}\}$.
3. Estimation of covariate-adjusted marginal eigencomponents
   (a) For each unique value of $a$ observed, employ FPCA on $\widehat{\boldsymbol{\Sigma}}_{d,\Omega}(a, \omega, \omega')$ to
       estimate the eigenvalue, eigenfunction pairs,
       $\{\tau_{d\ell,\Omega}(a), \phi_{d\ell}(a, \omega) : \ell = 1, \ldots, L_d\}$.
   (b) For each unique value of $a$ observed, employ PCA on $\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}(a)$ and to
       estimate the eigenvalue, eigenvector pairs $\{\tau_{dk,\mathcal{R}}(a), \mathrm{v}_{dk}(a, r) :$
       $k = 1, \ldots K_d\}$.
4. Estimation of covariate-adjusted variance components and subject-specific
   scores via linear mixed effects models
   (a) Calculate $\hat{\tau}_{dg}(a_i) = \mathrm{cov}\{\hat{\zeta}_{dig}(a_i)\}$ and $\hat{\sigma}^2_d$ by fitting the proposed linear
       mixed effects model.
   (b) Select $G'_d$ such that $FVE_{dG'} > .8$ for $d = 1, \ldots, D$ and form predictions
       $\widehat{Y}_{di}(a_i, r, \omega)$.
   (c) Calculate $\hat{\zeta}_{dig}\{a_i$ as the BLUP $\hat{\zeta}_{dig}(a_i) = E\left\{\zeta_{dig}(a_i) | \boldsymbol{Y}_{di}\right\}$.
5. Inference via parametric bootstrap.
```

For fixed slices of the covariate domain, the regional marginal covariance $\{\boldsymbol{\Sigma}_{d,\mathcal{R}}(a)\}_{r,r'}$ is discrete and thus not amenable to trivariate smoothers as the functional marginal covariance above. Therefore, we estimate the raw regional marginal covariance at each covariate-value, remove the measurement variance from the diagonals as in Scheffler et al. [14] and then apply a Nadarya-Watson kernel smoother to the resulting matrices entry-by-entry along the covariate domain. Our kernel smoother is the kernel weighted-average, $\{\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}(a_o)\}_{(r,r')} = \sum_{i=1}^{n_d} \sum_{\omega \in \Omega} K_\lambda(a_o, a_i)\widehat{Y}^c_{di}(a_i, r, \omega)\widehat{Y}^c_{di}(a_i, r', \omega) / |\Omega| \sum_{i=1}^{n_d} K_\lambda(a_o, a_i)$, where $K_\lambda(\cdot\,, \cdot\,)$ is some kernel with smoothness parameter λ and $|\Omega|$ is the number of observed functional grid points. The smoothing parameter λ is selected to minimize the LOSOCV(λ) statistic across all channel pairs (r, r'), $\text{LOSOCV}(\lambda) = \sum_{r=1}^{R} \sum_{r'<r} \text{LOSOCV}(\lambda, r, r')$, $\text{LOSOCV}(\lambda, r, r') = (1/|\Omega| n_d) \sum_{i=1}^{n_d} \sum_{\omega \in \Omega} \left[\widehat{Y}_{di}^{c}(a_i, r, \omega)\, \widehat{Y}_{di}^{c}(a_i, r', \omega) - \{\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}^{(-i)}(a_i)\}_{(r,r')}\right]^2$, where $\{\widehat{\boldsymbol{\Sigma}}_{d,\mathcal{R}}^{(-i)}(a_i)\}$ is the smoothed marginal covariance matrix with the ith subject left out. Thus, we introduce two novel covariate-adjusted smoothers that allow for calculation of the covariate-adjusted marginal covariances which are then used for subsequent covariate-adjusted eigendecompositions.

(3) Estimation of covariate-adjusted marginal eigencomponents: To estimate the covariate-adjusted marginal eigencomponents we perform eigendecompositions at each fixed covariate-value as described in Scheffler et al. [14] retaining a common number of K_d and L_d covariate-adjusted eigencomponents.

(4) Estimation of covariate-adjusted variance components and subject-specific scores via linear mixed effects models: We make use of the estimated functional fixed effects and marginal eigencomponents to propose a linear mixed effects framework for modeling covariate-adjusted region-referenced functional data. Under the assumption of joint normality of the covariate-adjusted subject-specific scores and measurement error, the proposed mixed effects framework induces regularization and stability in modeling the data by enforcing a low-rank structure on the covariate-adjusted variance components $\tau_{dg}(a)$. The resulting variance components can be used to select the number of eigencomponents to include in the CA-HPCA decomposition by quantifying the proportion of variance explained and for hypothesis testing and point-wise confidence bands via parametric bootstrap. We present the linear mixed effects modeling framework below.

To make the notation more compact, we replace the double index $k\ell$ in CA-HPCA truncated at K_d and L_d with a single index $g = (k-1) + K_d(\ell - 1) + 1$,

$$Y_{di}(a_i, r, \omega) = \eta_d(a_i, r, \omega) + \sum_{g=1}^{G_d} \xi_{di,g}(a_i) \varphi_{dg}(a_i, r, \omega) + \epsilon_{di}(a_i, r, \omega),$$

where $\varphi_{dg}(a_i, r, \omega) = \mathrm{v}_{dk}(a_i, r)\phi_{d\ell}(a_i, \omega)$, $\zeta_{dig}(a_i) = \langle Z_{di}(a_i, r, \omega), \varphi_{dg}(a_i, r, \omega)\rangle$, $\tau_{dg} = \text{cov}\{\zeta_{dig}(a_i)\}$, and $G_d = K_d L_d$. Let $\boldsymbol{Y}_{di}(a_i)$ represent the vectorized form of $Y_{di}(a_i, r, \omega)$ for subject i, $i = 1, \ldots, n_d$, observed along with covariate value a_i. Note, an argument for the covariate domain a is included to stress that a subject's covariance is covariate-dependent. Analogous vectorized forms for the covariate-adjusted functional fixed effects, $\eta_d(a_i, r, \omega)$, the region-referenced stochastic process $Z_{di}(a_i, r, \omega)$, multidimensional orthonormal basis $\varphi_{dg}(a_i, r, \omega)$, and the measurement error $\epsilon_{di}(a_i, r, \omega)$ are denoted by $\boldsymbol{\eta}_{di}(a_i)$, $\boldsymbol{Z}_{di}(a_i)$, $\boldsymbol{\varphi}_{dg}(a_i)$, and $\boldsymbol{\epsilon}_{di}$, respectively. Under the assumption that $\boldsymbol{\zeta}_{di}(a_i) = \{\zeta_{di1}(a_i), \ldots, \zeta_{diG_d}(a_i)\}$ and $\boldsymbol{\epsilon}_{di}$ are jointly Gaussian and $\text{cov}\{\boldsymbol{\zeta}_{di}(a_i), \boldsymbol{\epsilon}_{di}\} = \boldsymbol{0}$ at a fixed value of a_i, the proposed linear mixed effects model is given as

$$\begin{aligned}\boldsymbol{Y}_{di}(a_i) &= \boldsymbol{\eta}_{di}(a_i) + \boldsymbol{Z}_{di}(a_i) + \boldsymbol{\epsilon}_{di}\\ &= \boldsymbol{\eta}_{di}(a_i) + \sum_{g=1}^{G_d} \zeta_{dig}(a_i)\boldsymbol{\varphi}_{dg}(a_i) + \boldsymbol{\epsilon}_{di}, \quad \text{for} \quad i = 1, \ldots, n_d. \qquad (1)\end{aligned}$$

The model can be fit separately in each group, $d = 1, \ldots, D$ and the regional and functional dependencies in $\boldsymbol{Y}_{di}(a_i)$ are induced through the subject-specific random effects $\zeta_{dig}(a_i)$ in (1). Given the assumption that the total covariance is weakly separable for fixed values of a, $\text{cov}\{\zeta_{dig}(a_i), \zeta_{dig\prime}(a_i)\} = 0$ for $g \neq g\prime$ and thus the covariance of the subject-specific scores possess a diagonal diagonal structure, $\text{cov}\{\boldsymbol{\zeta}_{di}(a_i)\} = \boldsymbol{T}_d(a_i) = \text{diag}\{\boldsymbol{\tau}_d(a_i)\}$, where $\boldsymbol{\tau}_d(a_i) = \{\tau_{d1}(a_i), \ldots, \tau_{dG}(a_i)\}$. We further assume that $\boldsymbol{T}_d(a)$ evolves smoothly along the covariate domain and target the smooth variance components through their corresponding precision components. Given previous estimates for $\boldsymbol{\eta}_{di}(a_i)$ and $\boldsymbol{\varphi}_{dg}(a_i)$, estimates of the variance components and subject-specific scores are obtained using REML methods [17].

The assumption that the variance components evolve smoothly over the covariate domain resolves several challenges that emerge when modeling the covariate-adjusted subject-specific scores. First, the estimation procedure is able to borrow strength across the covariate-domain when modeling variation, a necessity when specific covariate values may only be observed once as in our motivating data. Second, we are able to project the precision components onto a smooth low-rank basis which induces regularization and control over the speed at which $\tau_g(a)$ is allowed to vary. Alternatively, a projection based approach is less computationally burdensome with estimates of the subject-specific scores obtained directly by numerical integration, $\hat{\zeta}_{dig}(a_i) = \langle Z_{di}(a_i, r, \omega), \hat{\varphi}_{dg}(a_i, r, \omega)\rangle$ and their corresponding variance components calculated empirically, $\hat{\tau}_{dg} = \text{cov}\{\hat{\zeta_{dig}}(a)\}$, but the resulting estimates are unstable due to the limited number of observations at each point along the covariate domain. Therefore, despite the added compute time, our proposed linear mixed effects framework is better suited for providing covariate-adjustments to the region-referenced functional process in a controlled and principled manner.

The estimated variance components are used to choose the number of eigencomponents included in the CA-HPCA decomposition where G'_d denotes a set of eigencomponents such that the total fraction of variance explained ($FVE_{dG'_d}$) is greater than 0.8 in each group $d = 1, \ldots, D$. We recommend starting with a larger number $G_d = K_d L_d$ of eigencomponents in the mixed effects modeling used for the estimation of $\{\tau_{dg}(a_i) : g = 1, \ldots, G_d\}$ and then reducing or adding components as appropriate to fix the final value of G'_d. In order to estimate the group-specific fraction of total variance explained via the G'_d eigencomponents, we consider the quantity, $FVE_{dG'_d} = \int\{\sum_{i=1}^{n_d}\sum_{g=1}^{G'_d}\hat{\tau}_{dg}(a_i)\}da/\int[\sum_{i=1}^{n_d}\{||Y_{di}(a_i, r, \omega) - \hat{\eta}_d(a_i, r, \omega)|| - R\int\hat{\sigma}_d^2 da\}]da$, where $||f(a_i, r, \omega)||^2 = \sum_{r=1}^{R}\int f(a_i, r, \omega)^2 d\omega$. Once G'_d is selected, the subject-specific scores can be obtained using their best linear unbiased predictor (BLUP) as in Scheffler et al. [14].

(5) Inference via parametric bootstrap: Inference in the form of hypothesis testing and point-wise confidence intervals can be performed via a parametric bootstrap based on the estimated CA-HPCA model components. To test the null hypothesis that all groups have equal means in the region r for a fixed covariate value $a \in \mathcal{A}$, i.e. $H_0 : \eta_d(a, r, \omega) = \eta(a, r, \omega)$ for $d = 1, \ldots D$, we propose a parametric bootstrap procedure based on the CA-HPCA decomposition. For $b = 1, \ldots, B$, the proposed parametric bootstrap generates outcomes based on the estimated model components under the null hypothesis as $Y^b_{di}(a_i, r, \omega) = \hat{\eta}(a_i, r, \omega) + Z^b_{di}(a_i, r, \omega) + \epsilon^b_{di}(a_i, r, \omega) = \hat{\eta}(a_i, r, \omega) + \sum_{g=1}^{G'_d} \zeta^b_{dig}(a_i)\hat{\varphi}_{dg}(a_i, r, \omega) + \epsilon^b_{di}(a_i, r, \omega)$ in region r and as $Y^b_{di}(a_i, r, \omega) = \hat{\eta}_d(a_i, r, \omega) + Z^b_{di}(a_i, r, \omega) + \epsilon^b_{di}(a_i, r, \omega) = \hat{\eta}_d(a_i, r, \omega) + \sum_{g=1}^{G'_d} \zeta^b_{dig}\hat{\varphi}_{dg}(a_i, r, \omega) + \epsilon^b_{di}(a_i, r, \omega)$ in the other regions, where subject-specific scores and measurement error are sampled from $\zeta^b_{dig}(a_i) \sim \mathcal{N}\{0, \hat{\tau}_{dg}(a_i)\}$ and $\epsilon^b_{di}(a_i, r, \omega) \sim \mathcal{N}(0, \hat{\sigma}^2_d)$. The proposed test statistic $T_r(a) = [\sum_{d=1}^{D} \int \{\hat{\eta}_d(a, \omega, s) - \hat{\eta}(a, r, \omega)\}^2 d\omega]^{1/2}$ is based on the norm of the sum of the departures of the estimated group-region shifts $\hat{\eta}_d(a, r, \omega)$ from the estimate of the common shift across groups, $\hat{\eta}(a, r, \omega)$. The common region shift estimate $\hat{\eta}(a, r, \omega)$, under the null, is set to the point-wise average of the group-region shift estimates, $\hat{\eta}_d(a, r, \omega)$, $d = 1, \ldots, D$. We utilize the proposed parametric bootstrap to estimate the distribution of the test statistic $T_r(a)$ which can be used to evaluate the null hypothesis along the covariate domain.

To generate point-wise confidence intervals for estimates of $\hat{\eta}_d(a, r, \omega)$, we repeat the above parametric bootstrap procedure but instead generate outcomes from the model $Y^b_{di}(a_i, r, \omega) = \hat{\eta}_d(a_i, r, \omega) + \sum_{g=1}^{G'_d} \zeta^b_{dig}\hat{\varphi}_{dg}(a_i, r, \omega) + \epsilon^b_{di}(a_i, r, \omega)$. At each iteration of the bootstrap, estimate $\hat{\eta}^b_d(a, r, \omega)$ from the simulated data and then form point-wise confidence intervals based on percentiles of the estimated group-region mean functions as a function of a, r and ω across iterations, $\{\hat{\eta}^b_{dg}(a, r, \omega) : b = 1, \ldots, B\}$.

4 Application to the Task-Free Paradigm Data

Data structure: In our motivating data application, EEG signals were sampled at 500 Hz for two minutes from a 128-channel HydroCel Geodesic Sensor Net on 58 ASD and 39 TD children aged 25 to 146 months old (diagnostic groups were age matched). EEG recordings were collected during an 'eyes-open' paradigm in which bubbles were displayed on a screen in a sound-attenuated room to subjects at rest [7]. We describe the dataset in our previous work and present an abbreviated description here, though the reader may reference Scheffler et al. [15] for technical details related to pre-processing and data acquisition. EEG data for each subject is interpolated down to a standard 10–20 system 25 electrode montage (R = 25) using spherical interpolation as detailed in Perrin et al. [13], producing 25 electrodes with continuous EEG signal (Fig. 1 (b)). Alpha spectral density (Ω = (6 Hz, 14 Hz)) estimates for each electrode were obtained and the resulting electrode-specific alpha spectral estimates form an instance of region-referenced functional data.

Data analysis results: We present the results from our application of CA-HPCA to the EEG data collected under the 'eyes-open' paradigm. While the main focus of our analysis is to characterize differences in alpha spectral dynamics between TD and ASD children over the course of development via inference on the group-region mean functions, we begin by briefly discussing the eigencomponents produced by the decomposition. The leading four; four and four; six covariate-adjusted regional and functional marginal eigencomponents are collectively found to explain 1.006 and 0.895 of the total FVE ($FVE_{dG'_d}$) in the TD and ASD groups, respectively. In the functional dimension along the covariate domain, the first least leading covariate-adjusted marginal eigenfunctions $\phi_{d1}(a, \omega)$ (Fig. 2(a), top row) display maximal variation at approximately 6 and 10 Hz (in opposing directions), where the location of maximal variation shifts in TD children from higher to lower frequencies as age increases but remains relatively constant in ASD children. The second leading covariate-adjusted marginal eigenfunctions $\phi_{d2}(a, \omega)$ (Fig. 2(a), bottom row) show maximal variation at 6 Hz, 8.5 Hz, 10.5 Hz and 6 Hz, 7.5 Hz in the TD and ASD groups, respectively, where again peak variation moves from higher to lower frequencies as age increases in the TD group but not the ASD group which instead displays shifts in the magnitude of maximal variation across development. The first two leading covariate-adjusted marginal eigenfunctions together explain at least 65% of the variation in the covariate-adjusted functional marginal covariances. In the regional dimension along the covariate domain, the first leading covariate-adjusted marginal eigenvectors $\mathrm{v}_{d1}(a, r)$ (Fig. 2(b), top row) display maximal variation in the central; right temporal; left posterior and central; middle posterior electrodes at younger ages with a shift to right posterior and frontal; right temporal electrodes at older ages in the TD and ASD groups, respectively. The second leading covariate-adjusted marginal eigenvector $\mathrm{v}_{d2}(a, r)$ (Fig. 2(b), bottom row) shows maximal variation in the frontal and right frontal; right temporal electrodes at younger ages which moves to frontal; right posterior (opposing directions) and central; left posterior (opposing directions) at older ages in the TD and ASD groups, respectively. The first two covariate-adjusted marginal eigenvectors together explain at least 70% of the variation in the covariate-adjusted regional marginal covariances.

To test for differences between TD and ASD groups in the alpha spectrum over development, we utilize the parametric bootstrap procedure described in Sect. 3 under the null hypothesis that the TD and ASD group-region mean functions are equal for every electrode r at each age $a = 25, \ldots, 145$ months which takes the form $H_0 : \eta_d(r, \omega, a) = \eta(r, \omega, a)$, $d = 1, 2$. Figure 3(a) displays the results of the hypothesis tests for all electrodes and ages with p-values transformed to the $-\log_{10}$ scale to better stratify results where values greater than $-\log_{10}(0.05) = 1.30$ denote significance at level $\alpha = .05$. Nearly all electrodes show significant differences between diagnostic groups in the alpha spectrum at

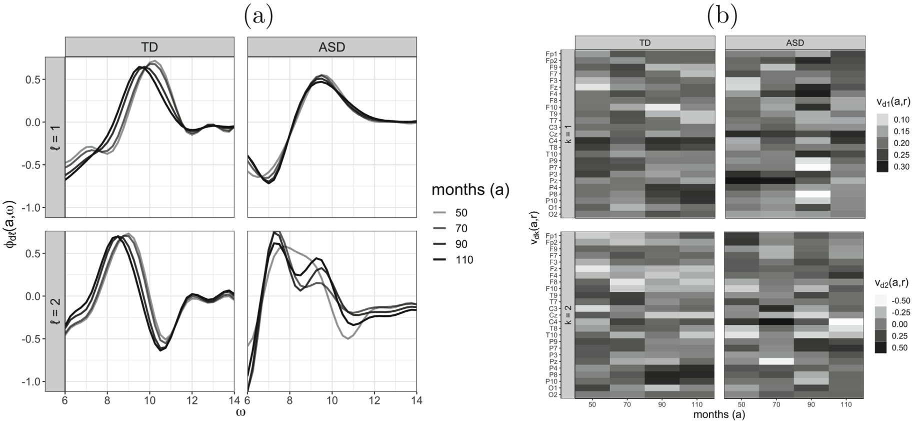

Fig. 2. (a) Estimated first and second leading covariate-adjusted eigenfunctions $\phi_{d1}(a,\omega)$ and $\phi_{d2}(a,\omega)$ at $a = 50, 70, 90, 110$ months. (b) Estimated first and second leading covariate-adjusted eigenvectors $\mathrm{v}_{d1}(a,r)$ and $\mathrm{v}_{d2}(a,r)$ at $a = 50, 70, 90, 110$ months.

some point over development (with the exception of the P3 and P7 electrodes) with the strongest group differences occurring at younger ages (30–50 months) and older ages (100–130 months) in the frontal, central, temporal, and posterior regions.

The greatest differences in the group-region mean functions across development are observed in the T8 and T10 electrodes displayed in Fig. 3(b) along with their 95% point-wise confidence intervals generated as described in Sect. 3. At both electrodes, the TD group displays a well-defined peak in the alpha spectrum that shifts from 8 Hz–10 Hz moving from 50–110 months, whereas the ASD group generally has less clearly-defined peaks that tend to center around 9 Hz throughout development. Differences in the estimated group-region mean functions mirror the results found from the parametric bootstrap procedure with separation of the point-wise confidence intervals occurring at 50; 90; 110 months and 110 months for the T8 and T10 electrodes, respectively. When aggregated, the observations and inferences obtained from the CA-HPCA model components provide evidence for differences in both the mean structure and patterns of covariation between the two diagnostic groups that shift and change over development highlighting the need to provide covariate-adjustments in modeling the high-dimensional EEG data.

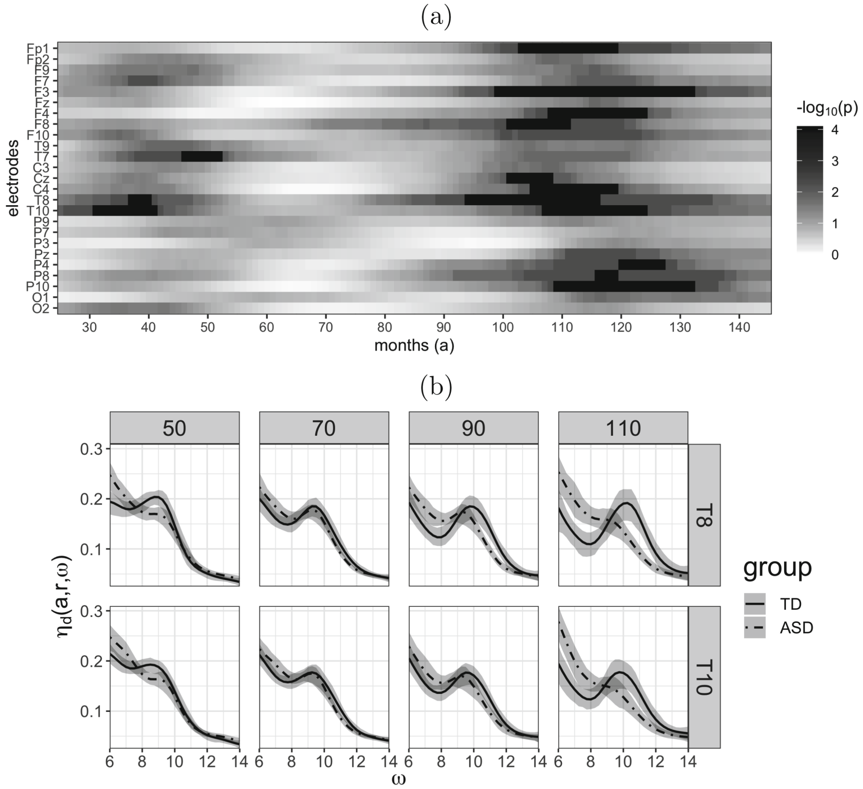

Fig. 3. (a) The $-\log_{10}$ transformed p-values from the hypothesis test for each electrode from the parametric bootstrap test for $a = 25, \ldots, 145$ months. (b) The estimated group-region mean functions $\eta_d(a, r, \omega)$ at ages $a = 50, 70, 90, 110$ months from the T8 and T10 electrodes. Grey shading denotes 95% point-wise confidence intervals for each estimate.

5 Discussion

We proposed a covariate-adjusted hybrid principal components analysis (CA-HPCA) which decomposes region-referenced functional data and accounts for covariate-dependent heteroscedasticity by assuming the high-dimensional covariance structure is weakly separable conditional on observed covariates. The proposed estimation procedure develops computationally efficient fast-covariance smoothers that incorporate covariate-dependence when estimating marginal covariances as well as a mixed effects framework which admits inference along the covariate-domain via bootstrap sampling. The CA-HPCA decomposition was developed to model EEG data over a broad developmental range but may be applied to other settings where high-dimensional data is expected to exhibit differential covariation as a function heterogenous covariates.

References

1. Backenroth, D., Goldsmith, J., Harran, M.D., Cortes, J.C., Krakauer, J.W., Kitago, T.: Modeling motor learning using heteroscedastic functional principal components analysis. J. Am. Stat. Assoc. **113**(523), 1003–1015 (2018)
2. Cardot, H.: Conditional functional principal components analysis. Scand. J. Stat. **34**(2), 317–335 (2007)
3. Cederbaum, J., Scheipl, F., Greven, S.: Fast symmetric additive covariance smoothing. Comput. Stat. Data Anal. **120**(C), 25–41 (2018)
4. Chen, K., Delicado, P., Müller, H.G.: Modelling function-valued stochastic processes, with applications to fertility dynamics. J. Roy. Stat. Soc.: Ser. B (Methodol.) **79**(1), 177–196 (2016)
5. Chiou, J.M., Müller, H.G., Wang, J.L.: Functional quasi-likelihood regression models with smooth random effects. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **65**(2), 405–423 (2003)
6. Corcoran, A.W., Alday, P.M., Schlesewsky, M., Bornkessel-Schlesewsky, I.: Toward a reliable, automated method of individual alpha frequency (IAF) quantification. Psychophysiology **55**(7), e13064 (2018)
7. Dickinson, A., DiStefano, C., Senturk, D., Jeste, S.S.: Peak alpha frequency is a neural marker of cognitive function across the autism spectrum. Eur. J. Neurosci. **47**(6), 643–651 (2018)
8. Edgar, J.C., et al.: Abnormal maturation of the resting-state peak alpha frequency in children with autism spectrum disorder. Hum. Brain Mapp. **40**(11), 3288–3298 (2019)
9. Edgar, J.C., et al.: Resting-state alpha in autism spectrum disorder and alpha associations with thalamic volume. J. Autism Dev. Disord. **45**(3), 795–804 (2015)
10. Jiang, C.R., Wang, J.L.: Covariate adjusted functional principal components analysis for longitudinal data. Ann. Statist. **38**(2), 1194–1226 (2010)
11. Lynch, B., Chen, K.: A test of weak separability for multi-way functional data, with application to brain connectivity studies. Biometrika **105**(4), 815–831 (2018)
12. Miskovic, V., et al.: Developmental changes in spontaneous electrocortical activity and network organization from early to late childhood. NeuroImage **118**(Supplement C), 237–247 (2015)
13. Perrin, F., Pernier, J., Bertrand, O., Echallier, J.: Spherical splines for scalp potential and current density mapping. Electroencephalogr. Clin. Neurophysiol. **72**(2), 184–187 (1989)
14. Scheffler, A.W., et al.: Hybrid principal components analysis for region-referenced longitudinal functional EEG data. Biostatistics **21**(1), 139–157 (2018)
15. Scheffler, A.W., et al.: Covariate-adjusted region-referenced generalized functional linear model for EEG data. Stat. Med. **38**(30), 5587–5602 (2019)
16. Valdas-Hernandez, P., et al.: White matter architecture rather than cortical surface area correlates with the EEG alpha rhythm. NeuroImage **49**(3), 2328–2339 (2010)
17. Wood, S.: Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, London (2017)