Multivariate Functional Shape Data Analysis

1 INTRODUCTION

2 BACKGROUND AND METHODS

2.1 Background

Suppose that we observe an image dataset for n unrelated subjects. Without loss of generality, we focus on a compact set, denoted as $\mathcal{D} \subset \mathbb{R}^t$, which is general enough to cover curves ($t = 1$), contours ($t = 2$), and surfaces ($t = 3$). It is assumed that $\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{N_V}\}$ are N_V grid points (or vertices) on $\mathcal{D}$ from the template file. Specifically, for the i-th subject, we observe a $J \times 1$ vector of shape measurements corresponding to each grid point $\boldsymbol{d}$, denoted as $\boldsymbol{y}_i(\boldsymbol{d}) = (y_{i1}(\boldsymbol{d}), \ldots, y_{iJ}(\boldsymbol{d}))^T$, and a $p \times 1$ vector of covariates (e.g., age, gender, group information, and biological markers), denoted as $\boldsymbol{x}_i$ with its first component being 1.

2.2 Multivariate Functional Shape Data Analysis (MFSDA)

The MFSDA is defined as

$$y_{ij}(\boldsymbol{d}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j(\boldsymbol{d}) + \eta_{ij}(\boldsymbol{d}) + \epsilon_{ij}(\boldsymbol{d}), \tag{1}$$

where $\boldsymbol{\beta}_j(\boldsymbol{d})$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{\eta}_i(\boldsymbol{d}) = (\eta_{i1}(\boldsymbol{d}), \ldots, \eta_{iJ}(\boldsymbol{d}))^T$ characterizes both subject-specific and location-specific variability, and $\boldsymbol{\epsilon}_i(\boldsymbol{d}) = (\epsilon_{i1}(\boldsymbol{d}), \ldots, \epsilon_{iJ}(\boldsymbol{d}))^T$ are measurement errors. It is also assumed that $\boldsymbol{\eta}_i(\boldsymbol{d})$ and $\boldsymbol{\epsilon}_i(\boldsymbol{d})$ are mutually independent and identical

copies of SP$(\mathbf{0}, \Sigma_\eta)$ and SP$(\mathbf{0}, \Sigma_\epsilon)$, respectively, where SP$(\boldsymbol{\mu}, \Sigma)$ denotes a stochastic process vector with mean function $\boldsymbol{\mu}(\boldsymbol{d})$ and covariance function $\Sigma(\boldsymbol{d}, \boldsymbol{d}')$. Moreover, $\Sigma_\epsilon(\boldsymbol{d}, \boldsymbol{d}')$ takes the form of $\Omega_\epsilon(\boldsymbol{d})\mathbf{1}(\boldsymbol{d} = \boldsymbol{d}')$, where $\Omega_\epsilon(\boldsymbol{d})$ is a nonnegative function of $\boldsymbol{d}$ and $\mathbf{1}(\cdot)$ is an indicator function of an event. Compared with the standard linear regression model, MFSDA explicitly accounts for spatial smoothness, spatial correlation, and the low-dimensional representation of functional shape responses [Zhu et al., 2014, 2011].

2.3 Hypothesis Testing

Under model (1), we start with a hypothesis testing problem on $\boldsymbol{\beta}_j(\boldsymbol{d}), j = 1, \ldots, J$, to investigate whether there is statistically significant morphological difference caused by some covariate of interest or linear combination of covariates of interest:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta}(\boldsymbol{d}) = \mathbf{0} \quad \text{v.s.} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta}(\boldsymbol{d}) \neq \mathbf{0} \text{ for each } \boldsymbol{d}, \tag{2}$$

where $\boldsymbol{C} = \boldsymbol{I}_J \otimes \boldsymbol{a}^T$, $\boldsymbol{\beta}(\boldsymbol{d}) = \text{vec}([\boldsymbol{\beta}_1(\boldsymbol{d}), \ldots, \boldsymbol{\beta}_J(\boldsymbol{d})])$, and $\boldsymbol{a}$ is a $p \times 1$ vector. In particular, if the $k-$th covariate is of interest, $\boldsymbol{a}$ can be written as $\mathbf{1}_k$, a $p \times 1$ vector with the $k-$th element 1 and rest 0. $\otimes$ is the Kronecker product operator, and vec$(\cdot)$ is the vectorization of a matrix.

As an example, in our analysis of condylar shape data from 34 subjects (17 normal controls and 17 OA patients), we are interested in testing whether there is significant morphological difference between two groups (normal controls v.s. OA). We consider MFSDA (1) on the spatial x, y, and z coordinates of 1002 vertices on the condylar surface with $(y_{i1}(\boldsymbol{d}), y_{i2}(\boldsymbol{d}), y_{i3}(\boldsymbol{d}))^T =$ (x coordinate, y coordinate, z coordinate)^T, and $\boldsymbol{x}_i =$(intercept, group, age, gender, facial pain rate, the first principal component scores of all biological markers)^T. In addition, the coefficient matrix $\boldsymbol{C}$ in the hypothesis testing (2) can be written as $\boldsymbol{I}_3 \otimes \mathbf{1}_2^T$.

We introduce a local Wald-type test statistic $T_n(\boldsymbol{d})$ as follows:

$$T_n(\boldsymbol{d}) = \boldsymbol{r}(\boldsymbol{d})^T \left[\{\widehat{\Sigma}_\eta(\boldsymbol{d}, \boldsymbol{d})\}^{-1} \otimes \{[\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}]^{-1}\boldsymbol{a}^{\otimes 2}\} \right] \boldsymbol{r}(\boldsymbol{d}), \tag{3}$$

where $\boldsymbol{r}(\boldsymbol{d}) = \hat{\boldsymbol{\beta}}(\boldsymbol{d}) - \text{Bias}(\hat{\boldsymbol{\beta}}(\boldsymbol{d}))$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$. The $\hat{\boldsymbol{\beta}}(\boldsymbol{d})$ and $\widehat{\Sigma}_\eta$ are estimates of the corresponding parameters, $\text{Bias}(\hat{\boldsymbol{\beta}}(\boldsymbol{d}))$ is the bias term of $\hat{\boldsymbol{\beta}}(\boldsymbol{d})$. Moreover, under the null hypothesis H_0, the limiting distribution of $T_n(\boldsymbol{d})$ can be approximated by a weighted χ^2 distribution [Zhang and Chen, 2007].

To estimate all unknown parameters in model (1), we employ a weighted least squares (WLS) method based on the multivariate local polynomial kernel smoothing technique [Fan and Gijbels, 1996]. Let $K(\cdot)$ be a kernel function, and H be a bandwidth matrix with a simple diagonal form. We also denote that $K_{H,m}(\boldsymbol{d}) = |H|^{-1}K(H^{-1}(\boldsymbol{d}_m - \boldsymbol{d}))$ and $\boldsymbol{w}_H(\boldsymbol{d}_m - \boldsymbol{d}) = (1, (\boldsymbol{d}_m - \boldsymbol{d})^T H^{-1})^T$. For each j and a fixed bandwidth matrix H_β, the WLS estimator of $\boldsymbol{\beta}_j(\boldsymbol{d})$ is given by

$$\hat{\boldsymbol{\beta}}_j(\boldsymbol{d}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sum_{m=1}^{N_V} a_m(H_\beta, \boldsymbol{d})\boldsymbol{y}_{\cdot,j}(\boldsymbol{d}_m), \tag{4}$$

where $a_m(H_\beta, \boldsymbol{d}) = \boldsymbol{e}^T[\sum_{m=1}^{N_V} K_{H_\beta,m}(\boldsymbol{d})\{\boldsymbol{w}_{H_\beta}(\boldsymbol{d}_m - \boldsymbol{d})\}^{\otimes 2}]^{-1}K_{H_\beta,m}(\boldsymbol{d})\boldsymbol{w}_{H_\beta}(\boldsymbol{d}_m - \boldsymbol{d})$, $\boldsymbol{e} = (1, \boldsymbol{0}_t^T)^T$, and $\boldsymbol{y}_{\cdot,j}(\boldsymbol{d}) = (y_{1,j}(\boldsymbol{d}), \ldots, y_{n,j}(\boldsymbol{d}))^T$. Based on (4), for a fixed bandwidth matrix H_η, the WLS estimate of $\eta_{i,j}^{(g)}(\boldsymbol{d})$ is given by

$$\hat{\eta}_{i,j}^{(g)}(\boldsymbol{d}) = \sum_{m=1}^{N_V} a_m(H_\eta, \boldsymbol{d})[y_{ij}(\boldsymbol{d}_m) - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j(\boldsymbol{d}_m)]. \tag{5}$$

Finally, we can estimate $\Sigma_\eta^{(g)}(\boldsymbol{d}, \boldsymbol{d}')$ by using the sample covariance function of $\hat{\boldsymbol{\eta}}_i^{(g)}(\boldsymbol{d})$, denoted as $\widehat{\Sigma}_\eta^{(g)}(\boldsymbol{d}, \boldsymbol{d}')$.

To select the optimal bandwidth in $\hat{\boldsymbol{\beta}}_j(\boldsymbol{d})$ (or $\hat{\eta}_{i,j}(\boldsymbol{d})$), we use the generalized cross-validation score method [Zhu et al., 2012]. We standardize all covariates to have mean zero and standard deviation one; thus, we may choose a common bandwidth for all covariates. Moreover, following the arguments of Fan and Zhang [1999], a small bandwidth leads to a small value of $\text{Bias}(\hat{\boldsymbol{\beta}}(\boldsymbol{d}))$, which can be dropped from the test statistics hereafter.

The global Wald-type statistic, denoted as T_n, is an integral of $T_n(\boldsymbol{d})\mu(\boldsymbol{d})$ with respect to $\boldsymbol{d} \in \mathcal{D}$; that is, $T(g) = \int_{\boldsymbol{d}\in\mathcal{D}} T_n(\boldsymbol{d})\mu(\boldsymbol{d})dL(\boldsymbol{d})$, where $L(\boldsymbol{d})$ is the Lebesgue measure. Selecting

different $\mu(\boldsymbol{d})$ allows us to introduce the prior information of specific regions of interest (ROIs). If there is no such prior information, then a uniform prior can be used. In this case, except for a constant scalar, T_n can be approximated by

$$T_n = \frac{Q_X^a}{N_V}\text{tr}\Big([\sum_{m=1}^{N_V} \boldsymbol{Y}_w(\boldsymbol{d}_m)\{\widehat{\Sigma}_\eta(\boldsymbol{d}_m)\}^{-1}\boldsymbol{Y}_w^T(\boldsymbol{d}_m)] \otimes [\boldsymbol{P}_X^a \boldsymbol{P}_X^{a\,T}]\text{vec}(\boldsymbol{X}^T)^{\otimes 2}\Big), \tag{6}$$

where $\boldsymbol{Y}_w(\boldsymbol{d}) = \sum_{m=1}^{N_V} a_m(H_\beta, \boldsymbol{d})[\boldsymbol{y}_{1,\cdot}(\boldsymbol{d}_m), \ldots, \boldsymbol{y}_{n,\cdot}(\boldsymbol{d}_m)]^T$, $Q_x^a = [\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}]^{-1}$, and $\boldsymbol{P}_X^a = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}$. If the true region of interest, denoted by $\mathcal{D}^*$, is relatively large and its corresponding measurements are moderate, then the value of T_n should be relatively large. Thus, if the value of T_n is large, then morphological difference is far more likely to be found caused by covariates of interest.

We use the wild bootstrap method to approximate the null distribution of T_n under the assumption that the null hypothesis H_0 in (2) holds for all $\boldsymbol{d} \in \mathcal{D}$:

- Step 1. Fit model (1) under the null hypothesis H_0, which yields $\underline{\hat{\boldsymbol{\beta}}}^*(\boldsymbol{d})$, $\underline{\hat{\boldsymbol{\eta}}}_i^*(\boldsymbol{d})$ and $\underline{\hat{\boldsymbol{\epsilon}}}_i^*(\boldsymbol{d})$ for all i and $\boldsymbol{d}$.

- Step 2. Generate a random sample ν_i^b and $\upsilon_i^b(\boldsymbol{d}_m)$ from a $N(0,1)$ generator for $i = 1, \ldots, n$ and $m = 1, \ldots, N_V$. B bootstrap samples are constructed as

$$\boldsymbol{y}_i^{(b)}(\boldsymbol{d}_m) = \boldsymbol{x}_i^T\underline{\hat{\boldsymbol{\beta}}}^*(\boldsymbol{d}_m) + \nu_i^b\underline{\hat{\boldsymbol{\eta}}}_i^*(\boldsymbol{d}_m) + \upsilon_i^b(\boldsymbol{d}_m)\underline{\hat{\boldsymbol{\epsilon}}}_i^*(\boldsymbol{d}_m),\ b = 1, \ldots, B$$

  for all i and $\boldsymbol{d}_m \in \mathcal{D}$.

- Step 3. For the $b-$th bootstrap samples, $b = 1, \ldots, B$, calculate the global Wald-type statistic $T_n^{(b)}$ based on the formula in (6).

- Step 4. The $p-$value of T_n is calculated via using an approximation method [Huang et al., 2015]. In fact, T_n can be approximated by a χ^2-type random variable $\alpha_1\chi^2(\alpha_2) + \alpha_3$,

where α_1, α_2, and α_3 are respectively given by

$$\alpha_1 = \frac{\kappa_3(T)}{4\kappa_2(T)}, \quad \alpha_2 = \frac{8\kappa_2^3(T)}{\kappa_3^2(T)}, \text{ and } \alpha_3 = \kappa_1(T) - \frac{2\kappa_2^2(T)}{\kappa_3(T)}, \tag{7}$$

where $\kappa_k(T), k = 1, 2, 3$, are respectively the first three sample cumulants of $\{T_n^{(b)}\}_{b=1}^B$. Finally, the $p-$value of T_n can be approximated by using $P(\chi^2(\alpha_2) \geq [T_n - \alpha_3]/\alpha_1)$.

References

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

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27(5):1491–1518.

Huang, M., Nichols, T., Huang, C., Yang, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., and Zhu, H. T. (2015). Fvgwas: fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage*, 118:613–627.

Zhang, J. and Chen, J. (2007). Statistical inference for functional data. *The Annals of Statistics*, 35:1052–1079.

Zhu, H., Fan, J., and Kong, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109:977–990.

Zhu, H., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W., and Gilmore, J. H. (2011). Fadtts: functional analysis of diffusion tensor tract statistics. *NeuroImage*, 56:1412–1425.

Zhu, H. T., Li, R. Z., and Kong, L. L. (2012). Multivariate varying coefficient model for functional responses. *Annals of Statistics*, 40:2634–2666.