Functional Data Analysis for Probability Density Functions

Outline

- **Petersen, A., Muller, H.-G. (2018). Wasserstein Covariance for Multiple Random Densities.** *Biometrika*.
- **Petersen, A., Muller, H.-G. (2016). Functional Data Analysis For Density Functions By Transformation To A Hilbert Space.** *Annals of Statistics*.
- **R demo of "fdadensity" package.**

Introduction

- **For random variables defined in $\mathbb{R}$, we are able to define and compute its mean, variances, and covariances.**
- **Let $x_1, \ldots, x_n$ denotes n realization of random variable X in $\mathbb{R}$. We can compute the mean as $\sum_1^n x_i/n$.**
- **What if $x_1, \ldots, x_n$ is a series of probability density function?**
- **(a) How can we define the "mean" and "variance" of a collection of probability density functions?**
- **The authors of this paper and several other literatures established answer for question (a) via Wasserstein mean and variance.**

- **Additionally, let $y_1, \ldots, y_n$ denotes n realization of random variable Y in $\mathbb{R}$. We can compute the covariance between X and Y as $\sum_1^n x_i y_i / n - \bar{x}\bar{y}$.**
- **(b) What if instead of scalars in $\mathbb{R}$, we have two collections of probability density functions? How to define and compute the "covariance" between two collections of p.d.f?**
- **This paper proposes an answer for question (b).**

- **To qualify the question, we start by defining a metric on the space of probability density functions.**
- **A metric quantifies the distance between two elements in a set. Using this distance, the mean and variance can be defined based on the metrics.**
- **We start by introducing the Wasserstein Metric. A Wasserstein Metric measures the "distance" between two probability density functions, or probability measures.**

- **Consider two p.d.f, f and g. A Wasserstein distance between f and g is**

$$d_W^2(f,g) = \int_0^1 \left\{ F^{-1}(t) - G^{-1}(t) \right\}^2 \mathrm{d}t$$

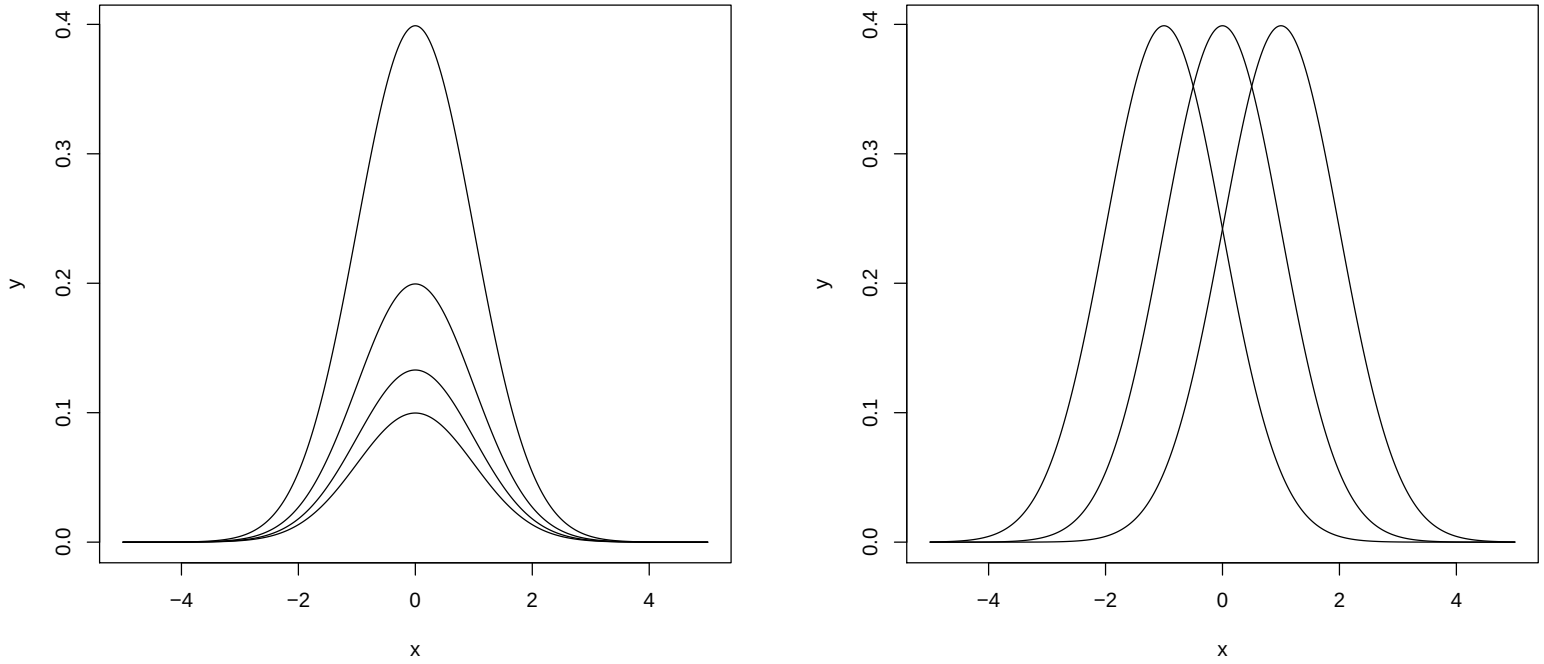
- **Here, $F(\cdot)$ denotes the cumulative distribution function of $f(\cdot)$, so $F^{-1}(\cdot)$ denotes the quantile function of $f(\cdot)$.**
- **A quantile function can be thought of as rotating the c.d.f. function by 90 degrees. It takes an input between 0 and 1 (support is always $[0,1]$) and output the percentile in the probability distribution.**

- **How can we define the "mean" and "variance" of a collection of probability density functions?**
- **Let $\mathfrak{F}$ be a collection of probability density functions.**
- **We use the Frechet mean and Frechet variance on the Wasserstein metric, which are defined as**

$$f_{\oplus} = \underset{f \in \mathcal{D}}{\operatorname{argmin}} E\left\{d_W(\mathfrak{F}, f)^2\right\}, \quad \operatorname{var}_{\oplus}(\mathfrak{F}) = E\left\{d_W\left(\mathfrak{F}, f_{\oplus}\right)^2\right\}$$

- **Basically, the Frechet-Wasserstein mean $f_{\oplus}$ is a new p.d.f. that has the smallest the average Wasserstein distances to the p.d.f's in $\mathfrak{F}$.**
- **The Frechet-Wasserstein variance is a scalar value in $\mathbb{R}$. It can be computed as the average Wasserstein distances between all the p.d.f's in $\mathfrak{F}$ to $f_{\oplus}$.**

- **How to find the Frechet-Wasserstein mean $f_{\oplus}$?**
- **Suppose we have a series of p.d.f. $f_1, \ldots, f_n$. Can we simply take the mean on L^2, i.e., summing them up and divided by n?**
- **This is a bad idea because it only captures vertical variation. It cannot capture horizontal variation in p.d.f.**



Horizontal variation versus vertical variation

- **For a p.d.f. f, let $Q(\cdot) = F^{-1}(\cdot)$ denote the quantile function. The *quantile density function* can be defined as taking the derivative of $Q(\cdot)$, i.e., $q(\cdot) = Q'(\cdot)$.**
- *Theorem* **:** *For the density process $f \sim \mathfrak{F}$, set $Q_{\oplus}(t) = E(Q(t))$. For $q_{\oplus} = Q'_{\oplus}$ and $F_{\oplus} = Q_{\oplus}^{-1}$, the Wasserstein-Frechet mean is*

$$f_{\oplus}(x) = \frac{1}{q_{\oplus}\left(F_{\oplus}(x)\right)}$$

- **In R, we may use a function called getWFmean() under package "fdadensity" to compute the Wasserstein-Frechet mean of a collection of p.d.f.**

- **As the Wasserstein-Frechet mean is $f_\oplus(x) = \frac{1}{q_\oplus\big(F_\oplus(x)\big)}$, it suggests the following approach for computing it.**
- **For a collection of densities, f_i, $i = 1, \ldots, n$, we use the corresponding quantile densities q_i to estimate $q_\oplus$ by $\tilde{q}_\oplus(t) = \frac{1}{n}\sum_{i=1}^n q_i(t)$.**
- **Then we take the integration of $q_\oplus$ to obtain $Q_\oplus$, and further take its inverse function to obtain $F_\oplus$.**
- **The estimated Wasserstein-Frechet mean is $\tilde{f}_\oplus(x) = \frac{1}{\tilde{q}_\oplus\big(\tilde{F}_\oplus(x)\big)}$.**
- **It can be shown theoretically that the Wasserstein between the estimator $\tilde{f}_\oplus$ and $f_\oplus$ is stochasticaly bounded, i.e., $d_W(\tilde{f}_\oplus, f_\oplus) = O_p(n^{-1/2})$.**

- **Once we obtain the $\tilde{f}_{\oplus}$, the Wasserstein variance can be computed as**

$$\begin{aligned}\mathrm{var}_{\oplus}(\mathfrak{F}) &= \frac{1}{n}\sum_{i=1}^{n} d_W\left(f_i, \tilde{f}_{\oplus}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\int\left\{F_i^{-1}(t) - F_{\oplus}^{-1}(t)\right\}^2 \mathrm{d}t\right]\end{aligned}$$

- **How about Wasserstein covariance?**
- **For two collection of p.d.f., $\mathfrak{F} = \{f_1, \ldots, f_n\}$ and $\mathfrak{G} = \{g_1, \ldots, g_n\}$. We may define Wasserstein covariance as**

$$\mathrm{cov}_{\oplus}(\mathfrak{F}, \mathfrak{G}) = E\left[\int \left\{F_i^{-1}(t) - F_{\oplus}^{-1}(t)\right\}\left\{G_i^{-1}(t) - G_{\oplus}^{-1}(t)\right\}\mathrm{d}t\right]$$

- **We now introduce some mathematical interpretation/connection of the Wasserstein metric.**
- **Essentially, the author try to support their definition of Wasserstein mean, variance and covariance by drawing connection to "optimal transport" on the space of densities.**
- **What is optimal transport, and its relation to Wasserstein metric?**

- **What is a transport?**
- **For probability densities f and g, suppose random variable $Y \sim f$. A transport is a map $T^* : \mathbb{R}$, such that $T^*(Y) \sim g$.**
- **The optimal transport from f to g is the transport that minimizes the following**

$$\int_{\mathbb{R}} \left\{T^*(u) - u\right\}^2 f(u)\mathrm{d}u$$

- **Imagine that we have a collection of particles (or sand), that are distributed according to the density f, that have to be moved so that they form a new distribution whose density is prescribed as g. The movement has to be chosen so as to minimize the average displacement. This movement corresponds to the optimal transport.**

- **The solution to the optimal transport problem is known to be $T = G^{-1} \circ F$ where F and G are the distribution functions of f and g, respectively. The resulting minimized objective function is exactly the Wasserstein distance**

$$\int \{T(u) - u\}^2 f(u) \mathrm{d}u = d_W^2(f, g)$$

- **To see this, apply the change of variable $t = F(u)$.**

- **Further, when there are two transport from f to g_1 and from f to g_2, let $T_1 = G_1^{-1} \circ F$ and $T_2 = G_2^{-1} \circ F$ be the corresponding optimal transport. we may define an inner product between T_1 and T_2 w.r.t. f as**

$$\langle T_1, T_2 \rangle_f = \int_{\mathbb{R}} \left\{ T_1(u) - u \right\} \left\{ T_2(u) - u \right\} f(u) \mathrm{d}u$$

- **It can be seen that $d_W \left(f, g_j \right)^2 = \langle T_j, T_j \rangle_f = \left\langle T_j^{-1}, T_j^{-1} \right\rangle_{g_j}$, $j = 1, 2$.**

- **We can interpret the Wasserstein variance using the defined inner product of optimal transport. We define the optimal transport from the Wasserstein mean $f_\oplus$ to a p.d.f. f_j in $\mathfrak{F}$ as $T = F_j^{-1} \circ F_\oplus$. The Wasserstein variance is thus**

$$\mathrm{var}_\oplus(\mathfrak{F}) = E\left\{d_W(\mathfrak{F}, f_\oplus)^2\right\} = E\left[\int_{\mathbb{R}} \left\{T(u) - u\right\}^2 f_\oplus(u)\mathrm{d}u\right] = E\left(\langle T, T\rangle_\oplus\right)$$

- **Here, the expectation is taken with respect to $\mathfrak{F}$, i.e., in practice we may simply take the average across all the f_j's in $\mathfrak{F}$.**

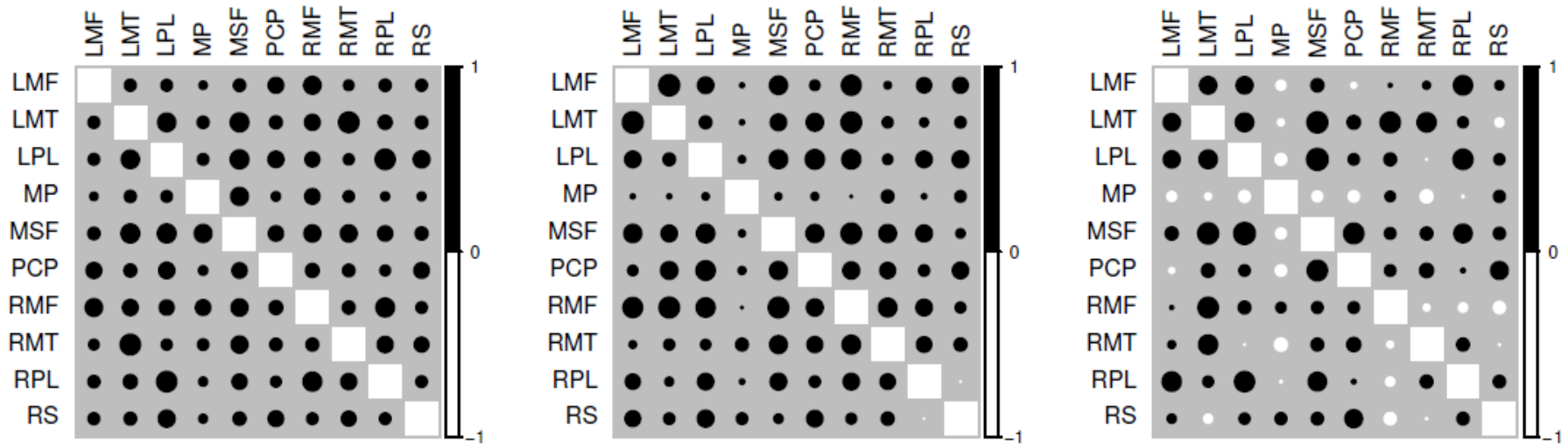
- **The Wasserstein covariance between two collections of p.d.f., $\mathfrak{F}_1$ and $\mathfrak{F}_2$, can be interpreted in terms of the inner product of $T_{j,1} = F_{j,1}^{-1} \circ F_{\oplus,1}$ and $T_{j,2} = F_{j,2}^{-1} \circ F_{\oplus,2}$ as follows**

$$\mathrm{cov}_{\oplus}\left(\mathfrak{F}_1, \mathfrak{F}_2\right) = E\left(\left\langle \widetilde{T}_{j,1}, T_{j,2} \right\rangle_{f_{\oplus,2}}\right) = E\left(\left\langle T_{j,1}, \widetilde{T}_{j,2} \right\rangle_{f_{\oplus,1}}\right)$$

where $\tilde{T}_1 = T_1 \circ T_{\oplus,12} - T_{\oplus,12} +$ identity map; and $T_{\oplus,12} = F_{\oplus,1}^{-1} \circ F_{\oplus,2}$.

- **Suppose we have p collections of p.d.f., $\mathfrak{F} = \left(\mathfrak{F}_1, \ldots, \mathfrak{F}_p\right)$, how to find the $p \times p$ Wasserstein covariance matrix of these p.d.f. collections?**
- **Recall that**
  $\operatorname{cov}_\oplus\left(\mathfrak{F}_j, \mathfrak{F}_k\right) = E\left[\int_0^1 \left\{F_j^{-1}(t) - F_{\oplus,j}^{-1}(t)\right\}\left\{F_k^{-1}(t) - F_{\oplus,k}^{-1}(t)\right\} \mathrm{d}t\right]$
- **Let $\hat{f}_{ij}$ be the density estimate, $i = 1, \ldots, n$, $j = 1, \ldots, p$. We map to their quantile function estimates $\hat{X}_{ij} = \hat{F}_{ij}^{-1}$. Write $\hat{X}_{ij}^c = \hat{X}_{ij} - n^{-1}\sum_{i=1}^n \hat{X}_{ij}$ and $\hat{\mathcal{C}}_{jk}(t,t) = n^{-1}\sum_{i=1}^n \hat{X}_{ij}^c(t)\hat{X}_{ik}^c(t)$.**
- **Denote Wasserstein covariance matrix as $\Sigma_\oplus$, it would then have the elements computed as $\left(\hat{\Sigma}_\oplus\right)_{jk} = \int_0^1 \hat{\mathcal{C}}_{jk}(t,t)\mathrm{d}t$,**
  $j, k = 1, \ldots, p$.
- **The author established theoretical results showing the consistency of the estimator $\hat{\Sigma}_\oplus$.**

- **We look at an application case of Alzheimer Disease.**
- **We have $p = 10$ brain regions. Each region has a collection of $n = 45$ density functions. Each density function comes from a smoothing histograms of pairwise temporal correlations between the signals of each voxel within a brain region and the signal at its central seed voxel. For simplicity, we may understood it as a density function of the neuro-connectivity for a certain brain region.**
- **Conduct such analysis on three subject groups: Normal, Mild Cognitive Impairment, and Alzheimer Disease.**



Correlation plot of ten brain regions for Normal (left), Mild Cognitive Impairment (Middle), and Alzheimer Disease (Right)

FPCA for density curves

- **We now look at another paper, which deals with FPCA for probability density curves.**
- **Suppose we have a collection of density curves $\mathfrak{F} = \{f_1, \ldots, f_n\}$. How do use FPCA to capture the variation within this collection of p.d.f.?**
- **The constraint here is $\int f_i(t)dt = 1$ and $f_i(t) > 0$.**

- **The usual approach that conducts the one-dimensional FPCA directly on these p.d.f curves does not work. As mentioned before, there are vertical/horizontal variation to be taken into account.**
- **More importantly, suppose a p.d.f. can be decomposed as**

$$f_i(x) = \mu(x) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(x)$$

- **In practice, we only use the first K FPC, i.e.,**

$$\tilde{f}_i(x) = \mu(x) + \sum_{k=1}^{K} \xi_{ik}\phi_k(x)$$

- **Obviously, $\tilde{f}_i(x)$ is smaller than $f_i(x)$ and does not satisfy the constraint of integral to 1.**

- **In FPCA, the mode of variation is defined as**

$$g_k(x, \alpha) = \mu(x) + \alpha\sqrt{\lambda_k}\phi_k(x)$$

  where $\lambda_k = Var(\xi_{ik})$ and α is a specified value. Essentially this is analogous to α s.d. away from the mean function.

- **Our goal is to construct a way of doing FPCA such that any mode of variation is a p.d.f function.**

- **The idea is to do a transformation on $f(t)$. After transformation, the constraints $\int f_i(t)dt = 1$ and $f_i(t) > 0$ should no longer exist, and regular FPCA can be applied.**
- **After doing the regular FPCA on the transformed $f(t)$, we may map the mode of variation in the the transformed $f(t)$ back to the density space where $f(t)$ lies.**

- **For a function $f(t)$, there are many candidates for transformation.**
- **Arithmatic $+ - \times /$**
- **log, exp**
- **integration, derivatives**
- **inverse function**

- **Consider the mapping $f(t)$ to the quantile function $Q(t) = F^{-1}(t)$.**
- **A quantile function can be thought of as rotating the c.d.f. function by 90 degrees. It is thus monotonically increasing.**
- **The support of $Q(t)$ is always $[0, 1]$.**
- **The advantage of this map is that the integral constraint $\int f_i(t)dt = 1$ is converted into the support $[0, 1]$, and the constraint $f_i(t) > 0$ no longer exists.**
- **However, $Q(t)$ suffers from the constraint of monotonically increasing.**

- **If we further take the derivative of $Q(t)$, resulting in the quantile density function $q(t)$. We now are able to convert the constraint of monotonically increasing into a constraint of $q(t) > 0$.**
- **To avoid the constraint of $q(t) > 0$, we take the log. Now all the constraints are gone.**
- **The author propose to use the log of the quantile density as the transformation. Any continuous and bounded function defined on $[0, 1]$, when mapped back from the unconstrained space of log quantile density to the space p.d.f., is automatically a p.d.f.**

- **One notable difference between such an approach and the regular FPCA lies in the selection of the nubmer of FPC, K.**
- **Traditionally, K is selected based on the fraction of variance explained (FVE). The variance is no longer the regular L^2 variance on the transformed space, but recommended to be the Wasserstein variance on the space of probability density functions.**
- **The total variance is**

$$\tilde{V}_\infty = \frac{1}{n}\sum_{i=1}^{n} d_W\left(f_i, \tilde{f}_\oplus\right)^2$$

- **The variance explained by the first K component is**

$$\tilde{V}_K = \tilde{V}_\infty - \frac{1}{n}\sum_{i=1}^{n} d_W\left(f_i, \tilde{f}_{i,K}\right)^2$$

where $\tilde{f}_{i,K} = \psi^{-1}\left(\tilde{\nu} + \sum_{k=1}^{K}\tilde{\eta}_{ik}\tilde{\rho}_k\right)(x)$. Here, $\tilde{\nu}$, $\tilde{\eta}_{ik}$, $\tilde{\rho}_k$ are the related FPCA quantities in the transformed space of log quantile densities, ψ^{-1} is the inverse map from the log quantile density to p.d.f.

- **We select the K as**

$$K^* = \min\left\{K : \frac{V_K}{V_\infty} > p\right\}$$

- **The major portion of the paper is devoted to theoretical work establishing certain consistency bound on the fraction of variance explained (FVE) and the mode of variation, i.e.,** $\left|\frac{V_K}{V_\infty} - \frac{\tilde{V}_K}{\tilde{V}_\infty}\right|$ **and** $d\left(g_k(\cdot, \alpha, \psi), \tilde{g}_k(\cdot, \alpha, \psi)\right)$**.**
- **Simulation studies demonstrate that for the same number of FPCs** K**, the proposed method achieves higher FVE compared to the naive approach of applying FPCA directly ignoring the constraint.**

- **The author develop a package for conducting FPCA on densities and I will now illustrate its use.**