The Annals of Statistics
2016, Vol. 44, No. 1, 183–218
DOI: 10.1214/15-AOS1363
© Institute of Mathematical Statistics, 2016

FUNCTIONAL DATA ANALYSIS FOR DENSITY FUNCTIONS BY TRANSFORMATION TO A HILBERT SPACE

By Alexander Petersen and Hans-Georg Müller[1]

University of California, Davis

Functional data that are nonnegative and have a constrained integral can be considered as samples of one-dimensional density functions. Such data are ubiquitous. Due to the inherent constraints, densities do not live in a vector space and, therefore, commonly used Hilbert space based methods of functional data analysis are not applicable. To address this problem, we introduce a transformation approach, mapping probability densities to a Hilbert space of functions through a continuous and invertible map. Basic methods of functional data analysis, such as the construction of functional modes of variation, functional regression or classification, are then implemented by using representations of the densities in this linear space. Representations of the densities themselves are obtained by applying the inverse map from the linear functional space to the density space. Transformations of interest include log quantile density and log hazard transformations, among others. Rates of convergence are derived for the representations that are obtained for a general class of transformations under certain structural properties. If the subject-specific densities need to be estimated from data, these rates correspond to the optimal rates of convergence for density estimation. The proposed methods are illustrated through simulations and applications in brain imaging.

1. Introduction.

Data that consist of samples of one-dimensional distributions or densities are common. Examples giving rise to such data are income distributions for cities or states, distributions of the times when bids are submitted in online auctions, distributions of movements in longitudinal behavior tracking or distributions of voxel-to-voxel correlations in fMRI signals (see Figure 1). Densities may also appear in functional regression models as predictors or responses.

Received December 2014; revised July 2015.

[1] Supported in part by NSF Grants DMS-11-04426, DMS-12-28369 and DMS-14-07852.

AMS 2000 subject classifications. Primary 62G05; secondary 62G07, 62G20.

Key words and phrases. Basis representation, kernel estimation, log hazard, prediction, quantiles, samples of density functions, rate of convergence, Wasserstein metric.

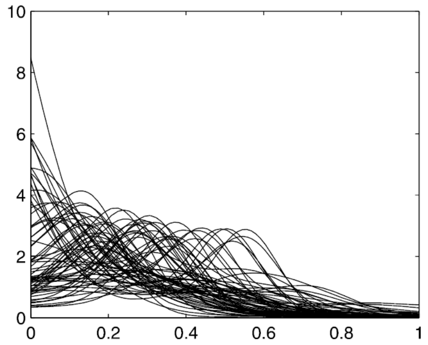


FIG. 1. *Densities based on kernel density estimates for time course correlations of BOLD signals obtained from brain fMRI between voxels in a region of interest. Densities are shown for $n = 68$ individuals diagnosed with Alzheimer's disease. For details on density estimation, see Section 2.3. Details regarding this data analysis, which illustrates the proposed methods, can be found in Section 6.2.*

The functional modeling of density functions is difficult due to the two constrains $\int f(x)\,dx = 1$ and $f \geq 0$. These characteristics imply that the functional space where densities live is convex but not linear, leading to problems for the application of common techniques of functional data analysis (FDA) such as functional principal components analysis (FPCA). This difficulty has been recognized before and an approach based on compositional data methods has been sketched in [17], applying theoretical results in [21], which define a Hilbert structure on the space of densities. Probably the first work on a functional approach for a sample of densities is [32], who utilized FPCA directly in density space to analyze samples of time-varying densities and focused on the trends of the functional principal components over time as well as the effects of the preprocessing step of estimating the densities from actual observations. Box–Cox transformations for a single nonrandom density function were considered in [48], who aimed at improving global bandwidth choice for kernel estimation of a single density function.

Density functions also arise in the context of warping, or registration, as time-warping functions correspond to distribution functions. In the context of functional data and shape analysis, such time-warping functions have been represented as square roots of the corresponding densities [42–44], and these square root densities reside in the Hilbert sphere, about which much is known. For instance, one can define the Fréchet mean on the sphere and also implement a nonlinear PCA method known as Principal Geodesic Analysis (PGA) [23]. We will compare this alternative methodology with our proposed approach in Section 6.

In this paper, we propose a novel and straightforward transformation approach with the explicit goal of using established methods for Hilbert space valued data once the densities have been transformed. The key idea is to map probability densities into a linear function space by using a suitably chosen continuous and invertible map ψ. Then FDA methodology, which might range anywhere from exploratory techniques to predictive modeling, can be implemented in this linear space. As an example of the former, functional modes of variation can be constructed by applying linear methods to the transformed densities, then mapping back into the density space by means of the inverse map. Functional regression or classification applications that involve densities as predictors or responses are examples of the latter.

We also present theoretical results about the convergence of these representations in density space under suitable structural properties of the transformations. These results draw from known results for estimation in FPCA and reflect the additional uncertainty introduced through both the forward and inverse transformations. One rarely observes data in the form of densities; rather, for each density, the data are in the form of a random sample generated by the underlying distribution. This fact will need to be taken into account for a realistic theoretical analysis, adding a layer of complexity. Specific examples of transformations that satisfy the requisite structural assumptions are the log quantile density and the log hazard transformations.

A related approach can be found in a recent preprint by [29], where the compositional approach of [17] was extended to define a version of FPCA on samples of densities. The authors represent densities by a centered log-ratio, which provides an isometric isomorphism between the space of densities and the Hilbert space L^2, and emphasize practical applications, but do not provide theoretical support or consider the effects of density estimation. Our methodology differs in that we consider a general class of transformations rather than one specific transformation. In particular, the transformation can be chosen independent of the metric used on the space of densities. This provides flexibility since, for many commonly-used metrics on the space of densities (see Section 2.2) corresponding isometric isomorphisms do not exist with the L^2 distance in the transformed space.

The paper is organized as follows: Pertinent results on density estimation and background on metrics in density space can be found in Section 2. Section 3 describes the basic techniques of FPCA, along with their shortfalls when dealing with density data. The main ideas for the proposed density transformation approach are in Section 4, including an analysis of specific transformations. Theory for this method is discussed in Section 5, with all proofs relegated to the Appendix. In Section 6.1, we provide simulations that illustrate the advantages of the transformation approach over the direct functional analysis of density functions, also including methods derived from properties of the Hilbert sphere. We also demonstrate how densities can serve as predictors in a functional regression analysis by using distributions of correlations of fMRI brain imaging signals to predict cognitive performance. More details about this application can be found in Section 6.2.

2. Preliminaries.

2.1. *Density modeling.* Assume that data consist of a sample of n (random) density functions $f_1, \ldots, f_n$, where the densities are supported on a common interval $[0, T]$ for some $T > 0$. Without loss of generality, we take $T = 1$. The assumption of compact support is for convenience, and does not usually present a problem in practice. Distributions with unbounded support can be handled analogously if a suitable integration measure is used. The main theoretical challenge for spaces of functions defined on an unbounded interval is that the uniform norm is no longer weaker than the L^2 norm, if the Lebesgue measure is used for the latter. This can be easily addressed by replacing the Lebesgue measure dx with a weighted version, for example, $e^{-x^2}\,dx$.

Denote the space of continuous and strictly positive densities on $[0, 1]$ by $\mathcal{G}$. The sample consists of i.i.d. realizations of an underlying stochastic process, that is, each density is independently distributed as $f \sim \mathfrak{F}$, where $\mathfrak{F}$ is an L^2 process [3] on $[0, 1]$ taking values in some space $\mathcal{F} \subset \mathcal{G}$. A basic assumption we make on the space $\mathcal{F}$ is:

(A1) For all $f \in \mathcal{F}$, f is continuously differentiable. Moreover, there is a constant $M > 1$ such that, for all $f \in \mathcal{F}$, $\|f\|_\infty$, $\|1/f\|_\infty$ and $\|f'\|_\infty$ are all bounded above by M.

Densities f can equivalently be represented as cumulative distribution functions (c.d.f.) F with domain $[0, 1]$, hazard functions $h = f/(1 - F)$ (possibly on a subdomain of $[0, 1]$ where $F(x) < 1$) and quantile functions $Q = F^{-1}$, with support $[0, 1]$. Occasionally of interest is the equivalent notion of the quantile-density function $q(t) = Q'(t) = \frac{d}{dt}F^{-1}(t) = [f(Q(t))]^{-1}$, from which we obtain $f(x) = [q(F(x))]^{-1}$, where we use the notation of [30]. This concept goes back to [37] and [46]. Another classical notion of interest is the density-quantile function $f(Q(t))$, which can be interpreted as a time-synchronized version of the density function [50]. All of these functions provide equivalent characterizations of distributions.

In many situations, the densities themselves will not be directly observed. Instead, for each i, we may observe an i.i.d. sample of data W_{il}, $l = 1, \ldots, N_i$, that are generated by the random density f_i. Thus, there are two random mechanisms at work that are assumed to be independent: the first generates the sample of densities and the second generates the samples of real-valued random data; one sample for each random density in the sample of densities. Hence, the probability space can be thought of as a product space $(\Omega_1 \times \Omega_2, \mathcal{A}, P)$, where $P = P_1 \otimes P_2$.

2.2. *Metrics in the space of density functions.* Many metrics and semimetrics on the space of density functions have been considered, including the L^2, L^1 [18], Hellinger and Kullback–Leibler metrics, to name a few. In previous applied and methodological work [8, 34, 50], it was found that a metric d_Q based on quantile functions $d_Q(f,g)^2 = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2\,dt$ is particularly promising from a practical point of view.

This quantile metric has connections to the optimal transport problem [47], and corresponds to the Wasserstein metric between two probability measures,

$$d_W(f,g)^2 = \inf_{X\sim f, Y\sim g} E(X-Y)^2, \tag{2.1}$$

where the expectation is with respect to the joint distribution of (X,Y). The equivalence $d_Q = d_W$ can be most easily seen by applying a covariance identity due to [28]; details can be found in the supplemental article [38]. We will develop our methodology for a general metric, which will be denoted by d in the following, and may stand for any of the above metrics in the space of densities.

2.3. *Density estimation.* A common occurrence in functional data analysis is that the functional data objects of interest are not completely observed. In the case of a sample of densities, the information about a specific density in the sample usually is available only through a random sample that is generated by this density. Hence, the densities themselves must first be estimated. Consider the estimation of a density $f \in \mathcal{F}$ from an i.i.d. sample (generated by f) of size N by an estimator $\check{f}$. Here, $N = N(n)$ will implicitly represent a sequence that depends on n, the size of the sample of random densities. In practice, any reasonable estimator can be used that produces density estimates that are bona fide densities and which can then be transformed into a linear space. For the theoretical results reported in Section 5, a density estimator $\check{f}$ must satisfy the following consistency properties in terms of the L^2 and uniform metrics (denoted as d_2 and d_∞, resp.):

(D1) For a sequence $b_N = o(1)$, the density estimator $\check{f}$, based on an i.i.d. sample of size N, satisfies $\check{f} \geq 0$, $\int_0^1 \check{f}(x)\,dx = 1$ and

$$\sup_{f\in\mathcal{F}} E(d_2(f,\check{f})^2) = O(b_N^2).$$

(D2) For a sequence $a_N = o(1)$ and some $R > 0$, the density estimator $\check{f}$, based on an i.i.d. sample of size N, satisfies

$$\sup_{f\in\mathcal{F}} P(d_\infty(f,\check{f}) > Ra_N) \to 0.$$

When this density estimation step is performed for densities on a compact interval, which is the case in our current framework, the standard kernel density estimator does not satisfy these assumptions, due to boundary effects. Much work has been devoted to rectify the boundary effects when estimating densities with compact support [15, 35], but the resulting estimators leave the density space and have not been shown to satisfy (D1) and (D2). Therefore, we introduce here a modified density estimator of kernel type that is guaranteed to satisfy (D1) and (D2).

Let κ be a kernel that corresponds to a continuous probability density function and $h < 1/2$ be the bandwidth. We define a new kernel density estimator to estimate the density $f \in \mathcal{F}$ on $[0,1]$ from a sample $W_1, \dots, W_N \overset{\text{i.i.d.}}{\sim} f$ by

$$\check{f}(x) = \sum_{l=1}^{N} \kappa\left(\frac{x - W_l}{h}\right) w(x,h) \Big/ \sum_{l=1}^{N} \int_0^1 \kappa\left(\frac{y - W_l}{h}\right) w(y,h)\, dy, \tag{2.2}$$

for $x \in [0,1]$ and 0 elsewhere. Here, the kernel κ is assumed to satisfy the following additional conditions:

(K1) The kernel κ is of bounded variation and is symmetric about 0.

(K2) The kernel κ satisfies $\int_0^1 \kappa(u)\, du > 0$, and $\int_{\mathbb{R}} |u|\kappa(u)\, du$, $\int_{\mathbb{R}} \kappa^2(u)\, du$ and $\int_{\mathbb{R}} |u|\kappa^2(u)\, du$ are finite.

The weight function

$$w(x,h) = \begin{cases} \left(\displaystyle\int_{-x/h}^{1} \kappa(u)\, du\right)^{-1}, & \text{for } x \in [0,h), \\ \left(\displaystyle\int_{-1}^{(1-x)/h} \kappa(u)\, du\right)^{-1}, & \text{for } x \in (1-h,1], \text{ and} \\ 1, & \text{otherwise}, \end{cases}$$

is designed to remove boundary bias.

The following result demonstrates that this modified kernel estimator indeed satisfies conditions (D1) and (D2). Furthermore, this result provides the rate in (D1) for this estimator as $b_N = N^{-1/3}$, which is known to be the optimal rate under our assumptions [45], where the class of densities $\mathcal{F}$ is assumed to be continuously differentiable, and it also shows that rates $a_N = N^{-c}$, for any $c \in (0, 1/6)$ are possible in (D2).

PROPOSITION 1. *If assumptions* (A1), (K1) *and* (K2) *hold, then the modified kernel density estimator (2.2) satisfies assumption* (D1) *whenever $h \to 0$ and $Nh \to \infty$ as $N \to \infty$ with $b_N^2 = h^2 + (Nh)^{-1}$. By taking $h = N^{-1/3}$ and $a_N = N^{-c}$ for any $c \in (0, 1/6)$,* (D2) *is also satisfied. In* (S1), *we may take $m(n) = n^r$ for any $r > 0$.*

Alternative density estimators could also be used. In particular, the beta kernel density estimator proposed in [14] is a promising prospect. The convergence of the expected squared L^2 metric was established in [14], while weak uniform consistency was proved in [10]. This density estimator is nonnegative, but requires additional normalization to guarantee that it resides in the density space.

3. Functional data analysis for the density process. For a generic density function process $f \sim \mathfrak{F}$, denote the mean function by $\mu(x) = E(f(x))$, the covariance function by $G(x,y) = \mathrm{Cov}(f(x), f(y))$, and the orthonormal eigenfunctions and eigenvalues of the linear covariance operator $(Af)(t) = \int G(s,t)f(s)\,ds$ by $\{\phi_k\}_{k=1}^{\infty}$ and $\{\lambda_k\}_{k=1}^{\infty}$, where the latter are positive and in decreasing order. If $f_1, \dots, f_n$ are i.i.d. distributed as f, then by the Karhunen–Loève expansion, for each i,

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

where $\xi_{ik} = \int_0^1 (f_i(x) - \mu(x))\phi_k(x)\,dx$ are the uncorrelated principal components with zero mean and variance λ_k. The Karhunen–Loève expansion constitutes the foundation for the commonly used FPCA technique [4, 6, 7, 16, 26, 27, 33].

The mean function μ of a density process $\mathfrak{F}$ is also a density function, as the space of densities is convex, and can be estimated by

$$\tilde{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{respectively} \quad \hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} \check{f}_i(x),$$

where the version $\tilde{\mu}$ corresponds to the case when the densities are fully observed and the version $\hat{\mu}$ corresponds to the case when they are estimated using suitable estimators such as (2.2); this distinction will be used throughout. However, in the common situation where one encounters horizontal variation in the densities, this mean is not a good measure of center. This is because the cross-sectional mean can only capture vertical variation. When horizontal variation is present, the L^2 metric does not induce an adequate geometry on the density space. A better method is *quantile synchronization* [50], a version of which has been introduced in [8] in the context of a genomics application. Essentially, this involves considering the cross-sectional mean function, $Q_\oplus(t) = E(Q(t))$, of the corresponding quantile process, Q. The synchronized mean density is then given by $f_\oplus = (Q_\oplus^{-1})'$.

The quantile synchronized mean can be interpreted as a Fréchet mean with respect to the Wasserstein metric $d = d_W$, where for a metric d on $\mathcal{F}$ the Fréchet mean of the process $\mathfrak{F}$ is defined by

$$f_\oplus = \arg\inf_{g \in \mathcal{F}} E(d(f,g)^2), \tag{3.1}$$

and the Fréchet variance is $E(d(f, f_{\oplus})^2)$. Hence, for the choice $d = d_W$, the Fréchet mean coincides with the quantile synchronized mean. Further discussion of this Wasserstein–Fréchet mean and its estimation is provided in the supplemental article [38]. Noting that the cross-sectional mean corresponds to the Fréchet mean for the choice $d = d_2$, the Fréchet mean provides a natural measure of center, adapting to the chosen metric or geometry.

Modes of variation [13] have proved particularly useful in applications to interpret and visualize the Karhunen–Loève representation and FPCA [31, 39]. They focus on the contribution of each eigenfunction ϕ_k to the stochastic behavior of the process. The kth mode of variation is a set of functions indexed by a parameter $\alpha \in \mathbb{R}$ that is given by

$$g_k(x, \alpha) = \mu(x) + \alpha\sqrt{\lambda_k}\phi_k(x). \tag{3.2}$$

In order to construct estimates of these modes, and generally to perform FPCA, the following estimates of the covariance function G of $\mathfrak{F}$ are needed:

$$\widetilde{G}(x, y) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) f_i(y) - \tilde{\mu}(x)\tilde{\mu}(y) \quad \text{respectively}$$

$$\widehat{G}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \check{f}_i(x) \check{f}_i(y) - \hat{\mu}(x)\hat{\mu}(y).$$

The eigenfunctions of the corresponding covariance operators, $\tilde{\phi}_k$ or $\hat{\phi}_k$, then serve as estimates of ϕ_k. Similarly, the eigenvalues λ_k are estimated by the empirical eigenvalues ($\tilde{\lambda}_k$ or $\hat{\lambda}_k$).

The empirical modes of variation are obtained by substituting estimates for the unknown quantities in the modes of variation (3.2),

$$\tilde{g}_k(x, \alpha) = \tilde{\mu}(x) + \alpha\sqrt{\tilde{\lambda}_k}\tilde{\phi}_k(x) \quad \text{respectively} \quad \hat{g}_k(x, \alpha) = \hat{\mu}(x) + \alpha\sqrt{\hat{\lambda}_k}\hat{\phi}_k(x).$$

These modes are useful for visualizing the FPCA in a Hilbert space. In a nonlinear space such as the space of densities, they turn out to be much less useful. Consider the eigenfunctions ϕ_k. In [32], it was observed that estimates of these eigenfunctions for samples of densities satisfy $\int_0^1 \hat{\phi}_k(x)\,dx = 0$ for all k. Indeed, this is true of the population eigenfunctions as well. To see this, consider the following argument. Let $\mathbf{1}(x) \equiv 1$ so that $\langle f - \mu, \mathbf{1}\rangle = 0$. Take φ to be the projection of ϕ_1 onto $\{\mathbf{1}\}^{\perp}$. It is clear that $\|\varphi\|_2 \le 1$ and $\mathrm{Var}(\langle f - \mu, \phi_1\rangle) = \mathrm{Var}(\langle f - \mu, \varphi\rangle)$. However, by definition, $\mathrm{Var}(\langle f - \mu, \phi_1\rangle) = \max_{\|\phi\|_2 = 1} \mathrm{Var}(\langle f - \mu, \phi\rangle)$. Hence, in order to avoid a contradiction, we must have $\|\varphi\|_2 = 1$, so that $\langle \phi_1, \mathbf{1}\rangle = 0$. The proof for all of the eigenfunctions follows by induction.

At first, this seems like a desirable characteristic of the eigenfunctions since it enforces $\int g_k(x, \alpha)\,dx = 1$ for any k and α. However, for $|\alpha|$ large enough, the resulting modes of variation leave the density space since $\langle \phi_k, 1\rangle = 0$ implies at least one sign change for all eigenfunctions. This also has the unfortunate consequence that the modes of variation intersect at a fixed point which, as we will see in Section 6, is an undesirable feature for describing variation of samples of densities.

In practical applications, it is customary to adopt a finite-dimensional approximation of the random functions by a truncated Karhunen–Loève representation, including the first K expansion terms,

$$f_i(x, K) = \mu(x) + \sum_{k=1}^{K} \xi_{ik}\phi_k(x). \tag{3.3}$$

Then the functional principal components (FPC) $\xi_{ik}, k = 1, \ldots, K$, are used to represent each sample function. For fully observed densities, estimates of the FPCs are obtained through their interpretation as inner products,

$$\tilde{\xi}_{ik} = \int_0^1 (f_i(x) - \tilde{\mu}(x))\tilde{\phi}_k(x)\, dx.$$

The truncated processes in (3.3) are then estimated by simple plug-in. Since the truncated finite-dimensional representations as derived from the finite-dimensional Karhunen–Loève expansion are designed for functions in a linear space, they are good approximations in the L^2 sense, but (i) may lack the defining characteristics of a density and (ii) may not be good approximations in a nonlinear space.

Thus, while it is possible to directly apply FPCA to a sample of densities, this approach provides an extrinsic analysis as the ensuing modes of variation and finite-dimensional representations leave the density space. One possible remedy would be to project these quantities back onto the space of densities, say by taking the positive part and renormalizing. In the applications presented in Section 6, we compare this ad hoc procedure with the proposed transformation approach.

4. Transformation approach.

The proposed transformation approach is to map the densities into a new space $L^2(\mathcal{T})$ via a functional transformation ψ, where $\mathcal{T} \subset \mathbb{R}$ is a compact interval. Then we work with the resulting L^2 process $X := \psi(f)$. By performing FPCA in the linear space $L^2(\mathcal{T})$ and then mapping back to density space, this transformation approach can be viewed as an intrinsic analysis, as opposed to ordinary FPCA. With ν and H denoting the mean and covariance functions, respectively, of the process X, $\{\rho_k\}_{k=1}^{\infty}$ denoting the orthonormal eigenfunctions of the covariance operator with kernel H with corresponding eigenvalues $\{\tau_k\}_{k=1}^{\infty}$, the Karhunen–Loève expansion for each of the transformed processes $X_i = \psi(f_i)$ is

$$X_i(t) = \nu(t) + \sum_{k=1}^{\infty} \eta_{ik}\rho_k(t), \qquad t \in \mathcal{T},$$

with principal components $\eta_{ik} = \int_{\mathcal{T}}(X_i(t) - \nu(t))\rho_k(t)\,dt$.

Our goal is to find suitable transformations $\psi : \mathcal{G} \to L^2(\mathcal{T})$ from density space to a linear functional space. To be useful in practice and to enable derivation of consistency properties, the maps ψ and ψ^{-1} must satisfy certain continuity requirements, which will be given at the end of this section. We begin with two specific examples of relevant transformations. For clarity, for functions in the native density space $\mathcal{G}$ we denote the argument by x, while for functions in the transformed space $L^2(\mathcal{T})$ the argument is t.

The log hazard transformation. Since hazard functions diverge at the right endpoint of the distribution, which is 1, we consider quotient spaces induced by identifying densities which are equal on a subdomain $\mathcal{T} = [0, 1_\delta]$, where $1_\delta = 1 - \delta$ for some $0 < \delta < 1$. With a slight abuse of notation, we denote this quotient space as $\mathcal{G}$ as well. The log hazard transformation $\psi_H : \mathcal{G} \to L^2(\mathcal{T})$ is

$$\psi_H(f)(t) = \log(h(t)) = \log\left\{\frac{f(t)}{1 - F(t)}\right\}, \qquad t \in \mathcal{T}.$$

Since the hazard function is positive but otherwise not constrained on $\mathcal{T}$, it is easy to see that ψ indeed maps density functions to $L^2(\mathcal{T})$. The inverse map can be defined for any continuous function X as

$$\psi_H^{-1}(X)(x) = \exp\left\{X(x) - \int_0^x e^{X(s)}\,ds\right\}, \qquad x \in [0, 1_\delta].$$

Note that for this case one has a strict inverse only modulo the quotient space. However, in order to use metrics such as d_W, we must choose a representative. A straightforward way to do this is to assign the remaining mass uniformly, that is,

$$\psi_H^{-1}(X)(x) = \delta^{-1}\exp\left\{-\int_0^{1_\delta} e^{X(s)}\,ds\right\}, \qquad x \in (1_\delta, 1].$$

The log quantile density transformation. For $\mathcal{T} = [0, 1]$, the log quantile density (LQD) transformation $\psi_Q : \mathcal{G} \to L^2(\mathcal{T})$ is given by

$$\psi_Q(f)(t) = \log(q(t)) = -\log\{f(Q(t))\}, \qquad t \in \mathcal{T}.$$

It is then natural to define the inverse of a continuous function X on $\mathcal{T}$ as the density given by $\exp\{-X(F(x))\}$, where $Q(t) = F^{-1}(t) = \int_0^t e^{X(s)}\,ds$. Since the value $F^{-1}(1)$ is not fixed, the support of the densities is not fixed within the transformed space, and as the inverse transformation should map back into the space of densities with support on $[0, 1]$, we make a slight adjustment when defining the inverse by

$$\psi_Q^{-1}(X)(x) = \theta_X \exp\{-X(F(x))\}, \qquad F^{-1}(t) = \theta_X^{-1}\int_0^t e^{X(s)}\,ds,$$

where $\theta_X = \int_0^1 e^{X(s)}\,ds$. Since $F^{-1}(1) = 1$ whenever $X \in \psi_Q(\mathcal{G})$, this definition coincides with the natural definition mentioned above on $\psi_Q(\mathcal{G})$.

To avoid the problems that afflict the linear-based modes of variation as described in Section 3, in the transformation approach we construct modes of variation in the transformed space for processes $X = \psi(f)$ and then map these back into the density space, defining transformation modes of variation

$$g_k(x, \alpha, \psi) = \psi^{-1}(\nu + \alpha\sqrt{\tau_k}\rho_k)(x). \tag{4.1}$$

Estimation of these modes is done by first estimating the mean function ν and covariance function H of the process X. Letting $\widehat{X}_i = \psi(\check{f}_i)$, the empirical estimators are

$$\tilde{\nu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \quad \text{respectively} \quad \hat{\nu}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i(t); \tag{4.2}$$

$$\begin{aligned} \widetilde{H}(s,t) &= \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t) - \tilde{\nu}(s)\tilde{\nu}(t) \quad \text{respectively} \\ \widehat{H}(s,t) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i(s)\widehat{X}_i(t) - \hat{\nu}(s)\hat{\nu}(t). \end{aligned} \tag{4.3}$$

Estimated eigenvalues and eigenfunctions ($\tilde{\tau}_k$ and $\tilde{\rho}_k$, resp., $\hat{\tau}_k$ and $\hat{\rho}_k$) are then obtained from the mean and covariance estimates as before, yielding the transformation mode of variation estimators

$$\begin{aligned} \tilde{g}_k(x, \alpha, \psi) &= \psi^{-1}(\tilde{\nu} + \alpha\sqrt{\tilde{\tau}_k}\tilde{\rho}_k)(x) \quad \text{respectively} \\ \hat{g}_k(x, \alpha, \psi) &= \psi^{-1}(\hat{\nu} + \alpha\sqrt{\hat{\tau}_k}\hat{\rho}_k)(x). \end{aligned} \tag{4.4}$$

In contrast to the modes of variation resulting from ordinary FPCA in (3.2), the transformation modes are bona fide density functions for any value of α. Thus, for reasonably chosen transformations, the transformation modes can be expected to provide a more interpretable description of the variability contained in the sample of densities. Indeed, the data application in Section 6.2 shows that this is the case, using the log quantile density transformation as an example.

The truncated representations of the original densities in the sample are then given by

$$f_i(x, K, \psi) = \psi^{-1}\left(\nu + \sum_{k=1}^{K} \eta_{ik}\rho_k\right)(x). \tag{4.5}$$

Utilizing (4.2), (4.3) and the ensuing estimates of the eigenfunctions, the (transformation) principal components, for the case of fully observed densities, are obtained in a straightforward manner,

$$\tilde{\eta}_{ik} = \int_{\mathcal{T}} (X_i(t) - \tilde{\nu}(t))\tilde{\rho}_k(t)\,dt, \tag{4.6}$$

whence

$$\tilde{f}_i(x, K, \psi) = \psi^{-1}\left(\tilde{\nu} + \sum_{k=1}^{K} \tilde{\eta}_{ik}\tilde{\rho}_k\right)(x).$$

In practice, the truncation point K can be selected by choosing a cutoff for the fraction of variance explained. This raises the question of how to quantify *total variance*. For the chosen metric d, we propose to use the Fréchet variance

$$V_\infty := E(d(f, f_\oplus)^2), \tag{4.7}$$

which is estimated by its empirical version

$$\tilde{V}_\infty = \frac{1}{n}\sum_{i=1}^{n} d(f_i, \tilde{f}_\oplus)^2, \tag{4.8}$$

using an estimator $\tilde{f}_\oplus$ of the Fréchet mean. Truncating at K included components as in (3.3) or in (4.5) and denoting the truncated versions as $f_{i,K}$, the variance explained by the first K components is

$$V_K := V_\infty - E(d(f_1, f_{1,K})^2), \tag{4.9}$$

which is estimated by

$$\tilde{V}_K = \tilde{V}_\infty - \frac{1}{n}\sum_{i=1}^{n} d(f_i, \tilde{f}_{i,K})^2. \tag{4.10}$$

The ratio V_K/V_∞ is called the fraction of variance explained (FVE), and is estimated by $\tilde{V}_K/\tilde{V}_\infty$. If the truncation level is chosen so that a fraction p, $0 < p < 1$, of total variation is to be explained, the optimal choice of K is

$$K^* = \min\left\{K : \frac{V_K}{V_\infty} > p\right\}, \tag{4.11}$$

which is estimated by

$$\tilde{K}^* = \min\left\{K : \frac{\tilde{V}_K}{\tilde{V}_\infty} > p\right\}. \tag{4.12}$$

As will be demonstrated in the data illustrations, this more general notion of variance explained is a useful concept when dealing with densities or other functions that are not in a Hilbert space. Specifically, we will show that density representations in (4.5), obtained via transformation, yield higher FVE values than the ordinary representations in (3.3), thus giving more efficient representations of the sample of densities.

For the theoretical analysis of the transformation approach, certain structural assumptions on the transformations need to be satisfied. The required smoothness properties for maps ψ and ψ^{-1} are implied by the three conditions (T0)–(T3) below. Here, the L^2 and uniform metrics are denoted by d_2 and d_∞, respectively, and the uniform norm is denoted by $\|\cdot\|_\infty$.

(T0) Let f, $g \in \mathcal{G}$ with f differentiable and $\|f'\|_\infty < \infty$. Set

$$D_0 \geq \max(\|f\|_\infty, \|1/f\|_\infty, \|g\|_\infty, \|1/g\|_\infty, \|f'\|_\infty).$$

Then there exists C_0 depending only on D_0 such that

$$d_2(\psi(f), \psi(g)) \leq C_0 d_2(f, g), \qquad d_\infty(\psi(f), \psi(g)) \leq C_0 d_\infty(f, g).$$

(T1) Let $f \in \mathcal{G}$ be differentiable with $\|f'\|_\infty < \infty$ and let D_1 be a constant bounded below by $\max(\|f\|_\infty, \|1/f\|_\infty, \|f'\|_\infty)$. Then $\psi(f)$ is differentiable and there exists $C_1 > 0$ depending only on D_1 such that $\|\psi(f)\|_\infty \leq C_1$ and $\|\psi(f)'\|_\infty \leq C_1$.

(T2) Let d be the selected metric in density space, Y be continuous and X be differentiable on $\mathcal{T}$ with $\|X'\|_\infty < \infty$. There exist constants $C_2 = C_2(\|X\|_\infty, \|X'\|_\infty) > 0$ and $C_3 = C_3(d_\infty(X, Y)) > 0$ such that

$$d(\psi^{-1}(X), \psi^{-1}(Y)) \leq C_2 C_3 d_2(X, Y)$$

and, as functions, C_2 and C_3 are increasing in their respective arguments.

(T3) For a given metric d on the space of densities and $f_{1,K} = f_1(\cdot, K, \psi)$ [see (4.5)], $V_\infty - V_K \to 0$ and $E(d(f, f_{1,K})^4) = O(1)$ as $K \to \infty$.

Here, assumptions (T0) and (T2) relate to the continuity of ψ and ψ^{-1}, while (T1) means that bounds on densities in the space $\mathcal{G}$ are accompanied by corresponding bounds of the transformed processes X. Assumption (T3) is needed to ensure that the finitely truncated versions in the transformed space are consistent, as the truncation parameter increases.

To establish these properties for the log hazard and log quantile density transformations, denoting as before the mean function, covariance function, eigenfunctions and eigenvalues associated with the process X by (ν, H, ρ_k, τ_k), assumption (T1) implies that ν, H, ρ_k, ν' and ρ_k' are bounded for all k (see Lemma 2 in the Appendix for details). In turn, these bounds imply a nonrandom Lipschitz constant for the residual process $X - X_K = \sum_{k=K+1}^\infty \eta_k \phi_k$ as follows. Under (A1), the constant C_1 in (T1) can be chosen uniformly over $f \in \mathcal{F}$. As a consequence, we have $\|X\|_\infty < C_1$ almost surely so that $\|\nu\|_\infty < C_1$ and

$$|\eta_k| = \left|\int_{\mathcal{T}} (X(t) - \nu(t))\phi_k(t)\, dt\right| \leq 2C_1 \int_{\mathcal{T}} |\phi_k(t)|\, dt \leq 2C_1 |\mathcal{T}|^{1/2}, \tag{4.13}$$

almost surely. Additionally, $\|\nu'\|_\infty < C_1$ and $\|\rho_k'\|_\infty < \infty$ for all k by dominated convergence, so that

$$\|X_K'\|_\infty \le \|\nu'\|_\infty + \sum_{k=1}^{K} |\eta_k| \|\rho_k'\|_\infty \le C_1 \left(1 + 2|\mathcal{T}|^{1/2} \sum_{k=1}^{K} \|\rho_k'\|_\infty\right).$$

Since $\|X'\|_\infty < C_1$ almost surely, setting

$$L_K := 2C_1 \left(1 + |\mathcal{T}|^{1/2} \sum_{k=1}^{K} \|\rho_k'\|_\infty\right) \tag{4.14}$$

then yields the almost sure bound

$$|(X - X_K)(s) - (X - X_K)(t)| \le L_K |s - t|.$$

The following result demonstrates the continuity of the log hazard and log quantile density transformations for classes of processes X that have suitably fast declining eigenvalues and suitable smoothness of the finite approximations.

PROPOSITION 2. *Assumptions* (T0)–(T2) *are satisfied for both* ψ_H *and* ψ_Q *with either* $d = d_2$ *or* $d = d_W$. *Let* L_K *denote the Lipschitz constant given in (4.14). If:*

(i) $L_K \sum_{k=K+1}^{\infty} \tau_k = O(1)$ *as* $K \to \infty$ *and*

(ii) *there is a sequence* r_m, $m \in \mathbb{N}$, *such that* $E(\eta_{1k}^{2m}) \le r_m \tau_k^m$ *for large* k *and* $(\frac{r_{m+1}}{r_m})^{1/3} = o(m)$,

are satisfied, then assumption (T3) *is also satisfied for both* ψ_H *and* ψ_Q *with either* $d = d_2$ *or* $d = d_W$.

As example, consider the Gaussian case for transformed processes X [or, similarly, the truncated Gaussian case in light of (4.13)] with components $\eta_{1k} \sim N(0, \lambda_k)$. Then $E(\eta_{1k}^{2m}) = \tau_k^m (2m-1)!!$, whence $r_m = (2m-1)!!$ so that $(r_{m+1}/r_m)^{1/3} = o(m)$ in (ii) is trivially satisfied. If the eigenfunctions correspond to the trigonometric basis, then $\|\rho_k'\|_\infty = O(k)$, so that $L_K = O(K^2)$. Hence, any eigenvalue sequence satisfying $\tau_k = O(k^{-4})$ would satisfy (i) in this case.

5. Theoretical results. The transformation modes of variation as defined in (4.1), together with the FVE values and optimal truncation points in (4.11), constitute the main components of the proposed approach. In this section, we investigate the weak consistency of the estimators of these quantities, given in (4.4) and (4.12), respectively, for the case of a generic density metric d, as $n \to \infty$. While asymptotic properties of estimates in FPCA are well established [9, 33], the effects of density estimation and transformation need to be studied in order to validate the proposed transformation approach. When densities are estimated, a lower bound m on the sample sizes available for estimating each density is required, as stipulated in the following assumption:

(S1) Let $\check{f}$ be a density estimator that satisfies (D2), and suppose densities $f_i \in \mathcal{F}$ are estimated by $\check{f}_i$ from i.i.d. samples of size $N_i = N_i(n)$, $i = 1, \dots, n$, respectively. There exists a sequence of lower bounds $m(n) \leq \min_{1 \leq i \leq n} N_i$ such that $m(n) \to \infty$ as $n \to \infty$ and

$$n \sup_{f \in \mathcal{F}} P(d_\infty(f, \check{f}) > R a_m) \to 0,$$

where, for generic $f \in \mathcal{F}$, $\check{f}$ is the estimated density from a sample of size $N(n) \geq m(n)$.

Proposition 1 in Section 2.3 implies that, for the density estimator in (2.2), property (S1) is satisfied for sequences of the form $m(n) = n^r$ for arbitrary $r > 0$. For $r < 3/2$, this rate dominates the rate of convergence in Theorem 1 below, which thus cannot be improved under our assumptions. While the theory we provide is general in terms of the transformation and metric, of particular interest are the specific transformations discussed in Section 4 and the Wasserstein metric d_W. Proofs and auxiliary lemmas are in the Appendix.

To study the transformation modes of variation, auxiliary results involving convergence of the mean, covariance, eigenvalue and eigenfunction estimates in the transformed space are needed. These auxiliary results are given in Lemma 3 and Corollary 1 in the Appendix. A critical component in these rates is the spacing between eigenvalues

$$\delta_k = \min_{1 \leq j \leq k} (\tau_j - \tau_{j+1}). \tag{5.1}$$

These spacings become important as one aims to estimate an increasing number of transformation modes of variation simultaneously.

The following result provides the convergence of estimated transformation modes of variation in (4.4) to the true modes $g_k(\cdot, \alpha, \psi)$ in (4.1), uniformly over mode parameters $|\alpha| \leq \alpha_0$ for any constant $\alpha_0 > 0$. For the case of estimated densities, if (D1), (D2) and (S1) are satisfied, $m = m(n)$ denotes the increasing sequence of lower bounds in (S1), and b_m is the rate of convergence in (D1), indexed by the bounding sequence m.

THEOREM 1. *Fix K and $\alpha_0 > 0$. Under assumptions* (A1), (T1) *and* (T2), *and with $\tilde{g}_k, \hat{g}_k$ as in (4.4),*

$$\max_{1 \leq k \leq K} \sup_{|\alpha| \leq \alpha_0} d(g_k(\cdot, \alpha, \psi), \tilde{g}_k(\cdot, \alpha, \psi)) = O_p(n^{-1/2}).$$

Additionally, there exists a sequence $K(n) \to \infty$ *such that*

$$\max_{1\le k\le K(n)} \sup_{|\alpha|\le\alpha_0} d(g_k(\cdot,\alpha,\psi),\tilde{g}_k(\cdot,\alpha,\psi)) = o_p(1).$$

If assumptions (T0), (D1), (D2) *and* (S1) *are also satisfied and* K, α_0 *are fixed,*

$$\max_{1\le k\le K} \sup_{|\alpha|\le\alpha_0} d(g_k(\cdot,\alpha,\psi),\hat{g}_k(\cdot,\alpha,\psi)) = O_p(n^{-1/2}+b_m).$$

Moreover, there exists a sequence $K(n) \to \infty$ *such that*

$$\max_{1\le k\le K(n)} \sup_{|\alpha|\le\alpha_0} d(g_k(\cdot,\alpha,\psi),\hat{g}_k(\cdot,\alpha,\psi)) = o_p(1).$$

In addition to demonstrating the convergence of the estimated transformation modes of variation for both fully observed and estimated densities, this result also provides uniform convergence over increasing sequences of included components $K = K(n)$. Under assumptions on the rate of decay of the eigenvalues and the upper bounds for the eigenfunctions, one also can get rates for the case $K(n) \to \infty$. For example, suppose the densities are fully observed, $\tau_k = ce^{-\theta k}$ for $c, \theta > 0$ and $\sup_k \|\rho_k\|_\infty \le A$ (as would be the case for the trigonometric basis, but this could be easily replaced by a sequence A_k of increasing bounds). Additionally, suppose $C_2 = a_0 e^{a_1\|X\|_\infty}$ in (T2), as is the case for the log quantile density transformation with the metric d_W (see the proof of Proposition 2). Then, following the proof of Theorem 1, one finds that, for $K(n) = \lfloor \frac{1}{4\theta}\log n \rfloor$,

$$\max_{1\le k\le K(n)} \sup_{|\alpha|\le\alpha_0} d(g_k(\cdot,\alpha,\psi),\tilde{g}_k(\cdot,\alpha,\psi)) = O_p(n^{-1/4}).$$

For the truncated representations in (4.5), the truncation point K may be viewed as a tuning parameter. When adopting the fraction of variance explained criterion [see (4.7) and (4.9)] for the data-adaptive selection of K, a user will typically choose the fraction $p \in (0,1)$, for which the corresponding optimal value K^* is given in (4.11), with the data-based estimate in (4.12). This requires estimation of the Fréchet mean $f_\oplus$ (3.1), for which we assume the availability of an estimator $\tilde{f}_\oplus$ that satisfies $d(f_\oplus, \tilde{f}_\oplus) = O_p(\gamma_n)$ for the given metric d in density space and some sequence $\gamma_n \to 0$. For the choice $d = d_W$, $\gamma_n = n^{-1/2}$ is admissible [38].

This selection procedure for the truncation parameter is a generalization of the scree plot in multivariate analysis, where the usual fraction of variance concept that is based on the eigenvalue sequence is replaced here with the corresponding Fréchet variance. As more data become available, it is usually desirable to increase the fraction of variance explained in order to more accurately represent the true underlying functions. Therefore, it makes sense to choose a sequence $p_n \in (0,1)$, with $p_n \uparrow 1$. The following result provides consistent recovery of the fraction of variance explained values V_K/V_∞ as well as the optimal choice K^* for such sequences.

THEOREM 2. *Assume* (A1) *and* (T1)–(T3) *hold. Additionally, suppose an estimator* $\tilde{f}_\oplus$ *of* $f_\oplus$ *satisfies* $d(f_\oplus, \tilde{f}_\oplus) = O_p(\gamma_n)$ *for a sequence* $\gamma_n \to 0$. *Then there is a sequence* $p_n \uparrow 1$ *such that*

$$\max_{1 \le K \le K^*} \left| \frac{V_K}{V_\infty} - \frac{\tilde{V}_K}{\tilde{V}_\infty} \right| = o_p(1)$$

and, consequently,

$$P(K^* \neq \tilde{K}^*) \to 0.$$

Specific choices for the sequence p_n and their implications for the corresponding sequence $K^*(n)$ can be investigated under additional assumptions. For example, consider the case where $\tau_k = ce^{-\theta k}$, $\sup_k \|\rho_k\|_\infty \le A$, $V_\infty - V_K = be^{-\omega K}$, $C_2 = a_0 e^{a_1 \|X\|_\infty}$ in (T2) and $\gamma_n = n^{-1/2}$. Then, by following the proofs of Lemma 4 and Theorem 2, we find that if $r < [2(2a_1 C_1 \times |\mathcal{T}|^{1/2} A + \theta + \omega)]^{-1}$, the choice

$$p_n = 1 - \frac{b(1+e^\omega)}{2V_\infty} n^{-\omega r}$$

leads to a corresponding sequence of tuning parameters $K^*(n) = \lfloor r \log n \rfloor$. In particular, this means that

$$\max_{1 \le K \le K^*} \left| \frac{V_K}{V_\infty} - \frac{\tilde{V}_K}{\tilde{V}_\infty} \right| = O_p\left(\left(\frac{\log n}{n} \right)^{1/2} \right)$$

and the relative error $(\tilde{K}^* - K^*)/K^*$ converges at the rate $o_p(1/\log n)$ under these assumptions.

6. Illustrations.

6.1. *Simulation studies.* Simulation studies were conducted to compare the performance between ordinary FPCA applied to densities, the proposed transformation approach using the log quantile density transformation, ψ_Q, and methods derived for the Hilbert sphere [23, 42–44] for three simulation settings that are listed in Table 1. The first two settings represent vertical and horizontal variation, respectively, while the third setting is a combination of both. We considered the case where the densities are fully observed, as well as the more realistic case where only a random sample of data generated by a density is available for each density. In the latter case, densities were estimated from a sample of size 100 each, using the density estimator in

Table 1
Simulation designs for comparison of methods

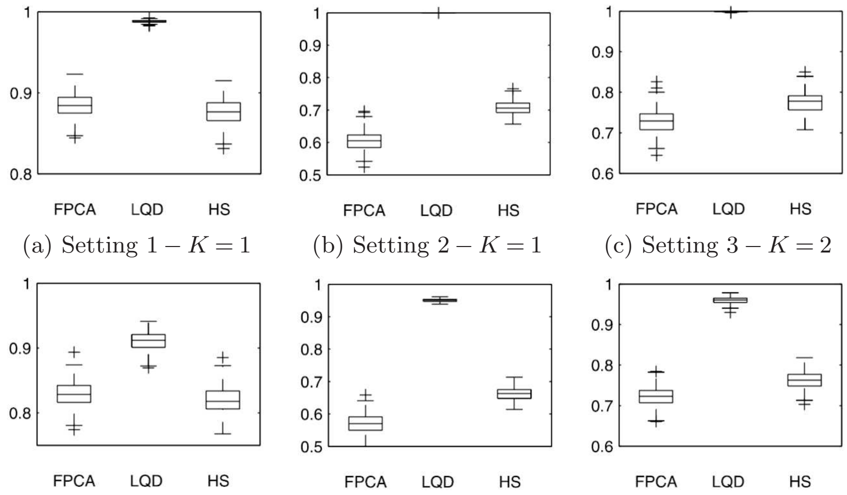
Setting	**Random component**	**Resulting density**
1	$\log(\sigma_i) \sim \mathcal{U}[-1.5, 1.5]$, $i = 1, \ldots, 50$	$\mathcal{N}(0, \sigma_i^2)$ truncated on $[-3, 3]$
2	$\mu_i \sim \mathcal{U}[-3, 3]$, $i = 1, \ldots, 50$	$\mathcal{N}(\mu_i, 1)$ truncated on $[-5, 5]$
3	$\log(\sigma_i) \sim \mathcal{U}[-1, 1]$, $\mu_i \sim \mathcal{U}[-2.5, 2.5]$, μ_i and σ_i independent, $i = 1, \ldots, 50$	$\mathcal{N}(\mu_i, \sigma_i^2)$ truncated on $[-5, 5]$

(2.2) with the kernel κ being the standard normal density and a bandwidth of $h = 0.2$.

In order to compare the different methods, we assessed the efficiency of the resulting representations. Efficiency was quantified by the fraction of variance explained (FVE), $\tilde{V}_K/\tilde{V}_\infty$, as given by the Fréchet variance [see (4.8) and (4.10)], so that higher FVE values reflect superior representations. As this quantity depends on the chosen metric d, we computed these values for both the L^2 and Wasserstein metrics. The FVE results for the two metrics were similar, so we only present the results using the L^2 metric here. Those corresponding to the Wasserstein metric d_W are given in the supplemental article [38]. As mentioned in Section 3, the truncated representations in (3.3) given by ordinary FPCA are not guaranteed to be bona fide densities. Hence, the representations were first projected onto the space of densities by taking the positive part and renormalizing, a method that has been systematically investigated by [24].

Boxplots for the FVE values (using the metric d_2) for the three simulation settings are shown in Figure 2, where the first row corresponds to fully observed densities and the second row to estimated densities. The number of components used to compute the fraction of variance explained was $K = 1$ for settings 1 and 2, and $K = 2$ for setting 3, reflecting the true dimensions of the random process generating the densities. Even in the first simulation setting, where the variation is strictly vertical, the transformation method outperformed both the standard FPCA and Hilbert sphere methods. The advantage of the transformation is most noticeable in settings 2 and 3 where horizontal variation is prominent.

As a qualitative comparison, we also computed the Fréchet means corresponding to three metrics: The L^2 metric (cross-sectional mean), Wasserstein metric and Fisher–Rao metric. This last metric corresponds to the geodesic metric on the Hilbert sphere between square-root densities. This fact was exploited in [42], where an estimation algorithm was introduced that we have implemented in our analyses. For details on the estimation of the Wasserstein–Fréchet mean, see the supplemental article [38]. To summarize these mean estimates across simulations, we again took the Fréchet mean (i.e., a Fréchet mean of Fréchet means), using the respective metric.



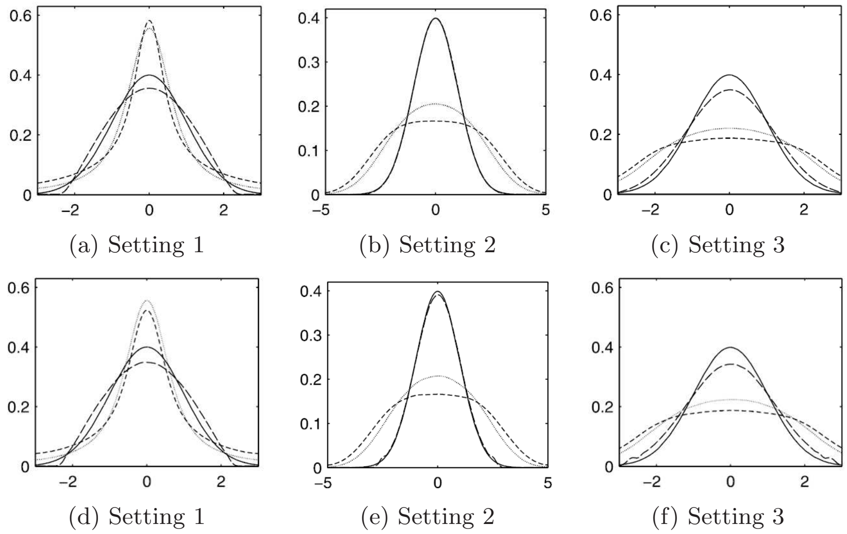
(a) Setting $1 - K = 1$ (b) Setting $2 - K = 1$ (c) Setting $3 - K = 2$

(d) Setting $1 - K = 1$ (e) Setting $2 - K = 1$ (f) Setting $3 - K = 2$

FIG. 2. *Boxplots of FVE (fraction of Fréchet variance explained, larger is better) values for 200 simulations, using the L^2 distance d_2. The first row corresponds to fully observed densities and the second corresponds to estimated densities. The columns correspond to settings 1, 2 and 3 from left to right (see Table 1). The methods are denoted by "FPCA" for ordinary FPCA on the densities, "LQD" for the transformation approach with ψ_Q and "HS" for the Hilbert sphere method.*

Note that a natural center for each simulation, if one knew the true random mechanism generating the densities, is the (truncated) standard normal density. Figure 3 plots the average mean estimates across all simulations (in the Fréchet sense) for the different settings along with the truncated standard normal density. One finds that in setting 2 for fully observed densities, the Wasserstein–Fréchet mean is visually indistinguishable from truncated normal density. Overall, it is clear that the Wasserstein–Fréchet mean yields a better concept for the "center" of the distribution of data curves than either the cross-sectional or Fisher–Rao–Fréchet means.

6.2. *Intra-hub connectivity and cognitive ability.* In recent years, the problem of identifying functional connectivity between brain voxels or regions has received a great deal of attention, especially for resting state fMRI [2, 22, 41]. Subjects are asked to relax while undergoing a fMRI brain scan, where blood-oxygen-level dependent (BOLD) signals are recorded and then processed to yield voxel-specific time courses of signal strength. Functional connectivity between voxels is customarily quantified in this area by the Pearson product-moment correlation [1, 5, 49] which, from a functional data analysis point of view, corresponds to a special case of dynamic correlation for random functions [19]. These correlations can be used for a variety of purposes. A



(a) Setting 1 (b) Setting 2 (c) Setting 3

(d) Setting 1 (e) Setting 2 (f) Setting 3

FIG. 3. *Average Fréchet means across 200 simulations. The first row corresponds to fully observed densities and the second corresponds to estimated densities. The columns correspond to settings 1, 2 and 3 from left to right (see Table 1). Truncated $\mathcal{N}(0,1)$—solid line; Cross-sectional—short-dashed line; Fisher–Rao—dotted line; Wasserstein—long-dashed line.*

traditional focus has been on characterizing voxel regions that have high correlations [11], which have been referred to as "hubs." For each such hub, a so-called seed voxel is identified as the voxel with the signal that has the highest correlation with the signals of nearby voxels.

As a novel way to characterize hubs, we analyzed the distribution of the correlations between the signal at the seed voxel of a hub and the signals of all other voxels within an $11 \times 11 \times 11$ cube of voxels that is centered at the seed voxel. For each subject, the target is the density within a specified hub that is then estimated from the observed correlations. The resulting sample of densities is then an i.i.d. sample across subjects. To demonstrate our methods, we select the Right inferior/superior Parietal Lobule hub (RPL) that is thought to be involved in higher mental processing [11].

The signals for each subject were recorded over the interval $[0, 470]$ (in seconds), with 236 measurements available at 2 second intervals. For the fMRI data recorded for $n = 68$ subjects that were diagnosed with Alzheimer's disease at UC Davis, we performed standard preprocessing that included the steps of slice-time correction, head motion correction and normalization to the Montreal Neurological Institute (MNI) fMRI template, in addition to linear detrending to account for signal drift, band-pass filtering to include only frequencies between 0.01 and 0.08 Hz and regressing out certain

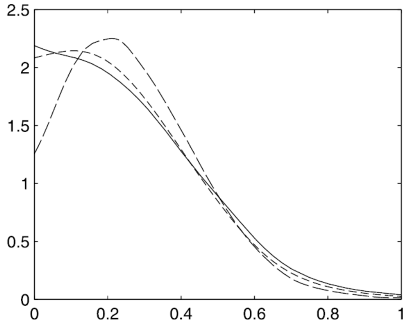
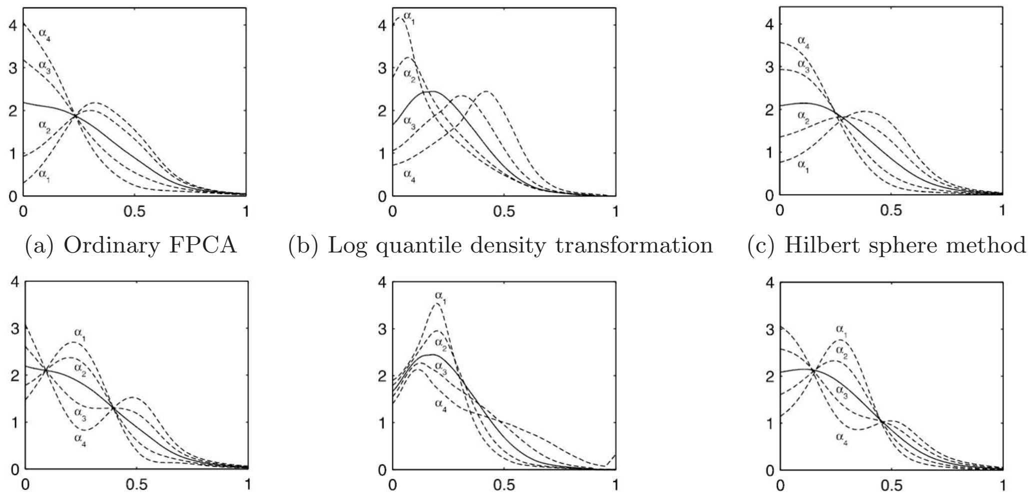


FIG. 4. *Comparison of means for distributions of seed voxel correlations for the RPL hub. Cross-sectional mean—solid line; Fisher–Rao–Fréchet mean—short-dashed line; Wasserstein–Fréchet mean—long-dashed line.*

time-dependent covariates (head motion parameters, white matter and CSF signal).

For the estimation of the densities of seed voxel correlations, the density estimator in (2.2) was utilized, with kernel κ chosen as the standard Gaussian density and a bandwidth of $h = 0.08$. As negative correlations are commonly ignored in connectivity analyses, the densities were estimated on $[0, 1]$. Figure 1 shows the estimated densities for all 68 subjects. A notable feature is the variation in the location of the mode, as well as the associated differences in the sharpness of the density at the mode. The Fréchet means that one obtains with different approaches are plotted in Figure 4. As in the simulations, the cross-sectional and Fisher–Rao–Fréchet means are very similar, and neither reflects the characteristics of the distributions in the sample. In contrast, the Wasserstein–Fréchet mean displays a sharper mode of the type that is seen in the sample of densities. Therefore, it is clearly more representative of the sample.

Next, we examined the first and second modes of variation, which are shown in Figure 5. The first mode of variation for each method reflects the horizontal shifts in the density modes, the location of which varies by subject. The modes for the Hilbert sphere method closely resemble those for ordinary FPCA and both FPCA and Hilbert sphere modes of variation do not adequately reflect the nature of the main variability in the data, which is the shift in the modes and associated shape changes. In contrast, the transformation modes of variation using the log quantile density transformation retain the sharp peaks seen in the sample and give a clear depiction of the horizontal variation. The second mode describes vertical variation. Here, the superiority of the transformation modes is even more apparent. The modes of ordinary FPCA and, to a lesser extent, those for the Hilbert sphere method, capture this form of variation awkwardly, with the extreme values of α moving toward bimodality—a feature that is not present in the



(a) Ordinary FPCA (b) Log quantile density transformation (c) Hilbert sphere method

(d) Ordinary FPCA (e) Log quantile density transformation (f) Hilbert sphere method

Fig. 5. *Modes of variation for distributions of seed voxel correlations. The first row corresponds to the first mode and the second row to the second mode of variation. The values of α used in the computation of the modes are quantiles ($\alpha_1 = 0.1$, $\alpha_2 = 0.25$, $\alpha_3 = 0.75$, $\alpha_4 = 0.9$) of the standardized estimates of the principal component (geodesic) scores for each method, and the solid line corresponds to $\alpha = 0$.*

data. In contrast, the log quantile density modes of variation capture the variation in the peaks adequately, representing all densities as unimodal density functions, where unimodality is clearly present throughout the sample of density estimates.

In terms of connectivity, the first transformation mode reflects mainly horizontal shifts in the densities of connectivity with associated shape changes that are less prominent, and can be characterized as moving from low to higher connectivity. The second transformation mode of variation provides a measure of the peakedness of the density, and thus to what extent connectivity is focused around a central value. The fraction of variance explained as shown in Figure 6 demonstrates that the transformation method provides not only more interpretable modes of variation, but also more efficient representations of the distributions than both ordinary FPCA and the Hilbert sphere methods. Thus, while the transformation modes of variation provide valuable insights into the variation of connectivity across subjects, this is not the case for the ordinary or Hilbert sphere modes of variation.

We also compared the utility of the densities and their transformed versions to predict a cognitive test score which assesses executive performance in the framework of a functional linear regression model. As the Hilbert sphere method does not give a linear representation, it cannot be used in this context. Denote the densities by f_i with functional principal compo-

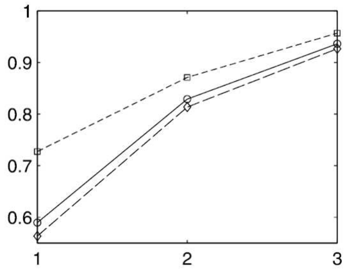


FIG. 6. *Fraction of variance explained for $K = 1, 2, 3$ components, using the metric d_2. Ordinary FPCA—solid line/circle marker; log quantile density transformation—short-dashed line/square marker; Hilbert Sphere method—long-dashed line/diamond marker.*

nents ξ_{ik}, the log quantile density functions by $X_i = \psi_Q(f_i)$ with functional principal components η_{ik} and the test scores by Y_i. Then the two models [12, 25] are

$$Y_i = B_{10} + \sum_{k=1}^{\infty} B_{1k}\xi_{ik} + \varepsilon_{1i} \quad \text{and}$$

$$Y_i = B_{20} + \sum_{k=1}^{\infty} B_{2k}\eta_{ik} + \varepsilon_{2i}, \qquad i = 1, \ldots, 65,$$

where three subjects who had missing test scores were removed. In practice, the sums are truncated in order to produce a model fit. These models were fit for different values of the truncation parameter K [see (3.3) and (4.5)] using the PACE package for MATLAB (code available at `http://anson.ucdavis.edu/~mueller/data/pace.html`) and 10-fold cross validation (averaged over 50 runs) was used to obtain the mean squared prediction error estimates give in Table 2.

In addition, the models were fitted using all data points to obtain an R^2 goodness-of-fit measurement for each truncation value K. The transformed densities were found to be better predictors of executive function than the ordinary densities for all values of K, both in terms of prediction error and

TABLE 2
Estimated mean squared prediction errors as obtained by 10-fold cross validation, averaged over 50 runs. Functional R^2 values for the fitted model using all data points are given in parentheses

K	**1**	**2**	**3**	**4**
FPCA	0.180 (0.0031)	0.185 (0.0135)	0.193 (0.0233)	0.201 (0.0244)
LQD	0.180 (0.0030)	0.176 (0.0715)	0.169 (0.1341)	0.173 (0.1431)

R^2 values. While the R^2 values were generally small, as only a relatively small fraction of the variation of the cognitive test score can generally be explained by connectivity, they were much larger for the model that used the transformation scores as predictors. These regression models relate transformation components of brain connectivity to cognitive outcomes, and thus shed light on the question of how patterns of intra-hub connectivity relate to cognitive function.

7. Discussion.

Due to the nonlinear nature of the space of density functions, ordinary FPCA is problematic for functional data that correspond to densities, both theoretically and practically, and the alternative transformation methods as proposed in this paper are more appropriate. The transformation based representations always satisfy the constraints of the density space and retain a linear interpretation in a suitably transformed space. The latter property is particularly useful for functional regression models with densities as predictors. Notions of mean and fraction of variance explained can be extended by the corresponding Fréchet quantities once a metric has been chosen. The Wasserstein metric is often highly suitable for the modeling of samples of densities.

While it is well known that for the L^2 metric d_2 the representations provided by ordinary FPCA are optimal in terms of maximizing the fraction of explained variance among all K-dimensional linear representations using orthonormal eigenfunctions, this is not the case for other metrics or if the representations are constrained to be in density space. In the transformation approach, the usual notion of explained variance needs to be replaced. We propose to do this by adopting the Fréchet variance, which in general will depend on the chosen transformation space and metric. As the data analysis indicates, even in the case of the L^2 metric, the log quantile density transformation performs better compared to FPCA or the Hilbert sphere approach in explaining most of the variation in a sample of densities by the first few components. The FVE plots, as demonstrated in Section 6, provide a convenient characterization of the quality of a transformation and can be used to compare multiple transformations or even to determine whether or not a transformation is better than no transformation.

In terms of interpreting the variation of functional density data, the transformation modes of variation emerge as clearly superior in comparison to the ordinary modes of variation, which do not keep the constraints to which density functions are subject. Overall, ordinary FPCA emerges as ill-suited to represent samples of density functions. When using such representations as an intermediate step, for example, if prediction of an outcome or classification with densities as predictors is of interest, it is likely that transformation methods are often preferable, as demonstrated in our data example.

Various transformations can be used that satisfy certain continuity conditions that imply consistency. In our experience, the log quantile density transformation emerges as the most promising of these. While we have only dealt with one-dimensional densities in this paper, extensions to densities with more complex support are possible. Since hazard and quantile functions are not immediately generalizable to multivariate densities, there is no obvious extension of the transformations based on these concepts to the multivariate case. However, for multivariate densities, a relatively straightforward approach is to apply the one-dimensional methodology to the conditional densities used by the Rosenblatt transformation [40] to represent higher-dimensional densities, although this approach would be computationally demanding and is subject to the curse of dimensionality and reduced rates of convergence as the dimension increases. However, it would be quite feasible for two- or three-dimensional densities. In general, the transformation approach is flexible, as it can be adopted for any transformation that satisfies some regularity conditions and maps densities to a Hilbert space.

APPENDIX: DETAILS ON THEORETICAL RESULTS

A.1. Proofs of propositions and theorems. This section contains proofs of Propositions 1 and 2 and Theorems 1 and 2. We also include some auxiliary lemmas. Additional proofs and a complete listing of all assumptions can be found in [38].

Proof of Proposition 1. Clearly, $\check{f} \geq 0$ and $\int_0^1 \check{f}(x)\,dx = 1$. Set

$$\mathring{f}(x) = \frac{1}{Nh}\sum_{l=1}^{N} \kappa\left(\frac{x - W_l}{h}\right) w(x,h),$$

so that $\check{f} = \mathring{f}/\int \mathring{f}$. Set $c_\kappa = (\int_0^1 \kappa(u)\,du)^{-1}$. For any $x \in [0,1]$ and $h < 1/2$, we have $1 \leq w(x,h) \leq c_\kappa$, so that

$$c_\kappa^{-1} \leq \inf_{y\in[0,1]} \int_{-yh^{-1}}^{(1-y)h^{-1}} \kappa(u)\,du \leq \int_0^1 \mathring{f}(x)\,dx \leq c_\kappa.$$

This implies

$$\left|1 - \left(\int_0^1 \mathring{f}(x)\,dx\right)^{-1}\right| \leq \min\{c_\kappa - 1, c_\kappa d_2(\mathring{f}, f), c_\kappa d_\infty(\mathring{f}, f)\},$$

which, together with assumption (A1), implies

$$d_2(\check{f}, f) \leq c_\kappa(M+1)\, d_2(\mathring{f}, f) \quad \text{and} \quad d_\infty(\check{f}, f) \leq c_\kappa(M+1) d_\infty(\mathring{f}, f).$$

Thus, we only need prove the remaining requirements in assumptions (D1) and (D2) for the estimator $\mathring{f}$.

The expected value is given by

$$E(\mathring{f}(x)) = h^{-1}\int_0^1 \kappa\left(\frac{x-y}{h}\right)w(x,h)f(y)\,dy$$
$$= f(x) + hw(x,h)\int_{-xh^{-1}}^{(1-x)h^{-1}} f'(x^*)u\kappa(u)\,dv,$$

for some x^* between x and $x+uh$. Thus, $E(\mathring{f}(x)) = f(x) + O(h)$, where the $O(h)$ term is uniform over $x \in [0,1]$ and $f \in \mathcal{F}$. Here, we have used the fact that $\sup_{f\in\mathcal{F}} \|f'\|_\infty < M$ and $\int_{\mathbb{R}} |u|\kappa(u)\,du < \infty$. Similarly,

$$\mathrm{Var}(\mathring{f}(x)) \le \frac{c_\kappa^2}{Nh}\left(f(x)\int_0^1 \kappa^2(u)\,du + h\int_0^1 u\kappa^2(u)f'(x^*)\,du\right),$$

for some x^* between x and $x+uh$, so that the variance is of the order $(Nh)^{-1}$ uniformly over $x \in [0,1]$ and $f \in \mathcal{F}$. This proves (D1) for $b_N^2 = h^2 + (Nh)^{-1}$.

To prove assumption (D2), we use the triangle inequality to see that

$$d_\infty(f,\mathring{f}) \le d_\infty(f, E(\mathring{f}(\cdot))) + d_\infty(\mathring{f}, E(\mathring{f}(\cdot))).$$

Using the DKW inequality [20], there are constants c_1, c_2 and a sequence $L_h = O(h)$ such that, for any $R > 0$,

$$P(d_\infty(f,\mathring{f}) > 2Ra_N) \le c_1\exp\{-c_2R^2a_N^2Nh^2\} + I\{L_h > Ra_N\},$$

where I is the indicator function. Notice that the bound is independent of $f \in \mathcal{F}$. By taking $h = N^{-1/3}$ and $a_N = N^{-c}$ for $c \in (0, 1/6)$, we have $L_h < Ra_N$ for large enough N, and thus, for such N,

$$\sup_{f\in\mathcal{F}} P(d_\infty(f,\mathring{f}) > 2Ra_N) \le c_1\exp\{-c_2R^2N^{1/3-2c}\} = o(1) \qquad \text{as } N \to \infty.$$

In assumption (S1), we may then take $m = n^r$ for any $r > 0$, since

$$n\sup_{f\in\mathcal{F}} P(d_\infty(f,\mathring{f}) > 2Ra_N) \le c_1 n\exp\{-c_2R^2n^{r/3-2rc}\} = o(1)$$
$$\text{as } n \to \infty. \qquad \square$$

Proof of Proposition 2. First, we deal with the log hazard transformation. Let f and g be two densities as specified in assumption (T0), with distribution functions F and G. Then

$$d_\infty(F,G) \le d_2(f,g) \le d_\infty(f,g).$$

Also, $1-F$ and $1-G$ are both bounded below by δD_0^{-1} on $[0, 1_\delta]$. Then, for $x \in [0, 1_\delta]$,

$$|\psi_H(f)(x) - \psi_H(g)(x)| \le \left|\log\left(\frac{f(x)}{g(x)}\right)\right| + \left|\log\left(\frac{1-F(x)}{1-G(x)}\right)\right|$$
$$\le D_0[|f(x) - g(x)| + \delta^{-1}|F(x) - G(x)|],$$

whence

$$\begin{aligned} d_\infty(\psi_H(f),\psi_H(g)) &\le D_0(1+\delta^{-1})d_\infty(f,g),\\ d_2(\psi_H(f),\psi_H(g))^2 &\le 2D_0^2\left[\int_0^{1_\delta}(f(x)-g(x))^2\,dx+\delta^{-2}d_2(f,g)^2\right]\\ &\le 2D_0^2(1+\delta^{-2})d_2(f,g)^2. \end{aligned}$$

These bounds provide the existence of C_0 in (T0). For (T1), observe that

$$\delta D_1^{-2} < \frac{f(x)}{1-F(x)} \le \delta^{-1}D_1^2,$$

so that

$$\begin{aligned} \|\psi_H(f)\|_\infty &= \sup_{x\in[0,1_\delta]}\left|\log\frac{f(x)}{1-F(x)}\right| \le 2\log D_1-\log\delta \quad \text{and}\\ \|\psi_H(f)'\|_\infty &= \sup_{x\in[0,1_\delta]}\left|\frac{f'(x)(1-F(x))+f(x)^2}{f(x)(1-F(x))}\right| \le 2\delta^{-1}D_1^4, \end{aligned}$$

which proves the existence of C_1.

Next, let X and Y be functions as in (T2) for $\mathcal{T}=[0,1_\delta]$ and set $f=\psi_H^{-1}(X)$ and $g=\psi_H^{-1}(Y)$. Let $\Lambda_X(x)=\int_0^x e^{X(s)}\,ds$ and $\Lambda_Y(x)=\int_0^x e^{Y(s)}\,ds$. Then

$$|\Lambda_X(x)-\Lambda_Y(x)| \le \int_0^x |e^{X(s)}-e^{Y(s)}|\,ds \le e^{\|X\|_\infty+d_\infty(X,Y)}d_2(X,Y),$$

whence

$$\begin{aligned} &d_2(\psi_H^{-1}(X),\psi_H^{-1}(Y))^2\\ &\qquad\le 2e^{2\|X\|_\infty}[d_2(\Lambda_X,\Lambda_Y)^2\,dx+e^{2d_\infty(X,Y)}d_2(X,Y)^2]\\ &\qquad\quad+\delta^{-1}(\Lambda_X(1_\delta)-\Lambda_Y(1_\delta))^2\\ &\qquad\le 2e^{2\|X\|_\infty}[(e^{2\|X\|_\infty}+\delta^{-1})+1]e^{2d_\infty(X,Y)}d_2(X,Y)^2. \end{aligned} \tag{A.1}$$

Taking $C_2=\sqrt{2}e^{\|X\|_\infty}[(e^{2\|X\|_\infty}+\delta^{-1})+1]^{1/2}$ and $C_3=e^{d_\infty(X,Y)}$, (T2) is established for $d=d_2$.

For $d=d_W$, the cdf's of f and g for $x\in[0,1_\delta]$ are given by $F(x)=1-e^{-\Lambda_X(x)}$ and $G(x)=1-e^{-\Lambda_Y(x)}$, respectively. For $x\in(1_\delta,1]$,

$$\begin{aligned} F(x)&=F(1_\delta)+\delta^{-1}(1-F(1_\delta))(x-1_\delta),\\ G(x)&=G(1_\delta)+\delta^{-1}(1-G(1_\delta))(x-1_\delta), \end{aligned}$$

so that $|F(x)-G(x)|\le|F(1_\delta)-G(1_\delta)|$ for such x. Hence, for all $x\in[0,1]$

$$|F(x)-G(x)| \le \sup_{x\in[0,1_\delta]}|\Lambda_X(x)-\Lambda_Y(x)| \le e^{\|X\|_\infty+d_\infty(X,Y)}d_2(X,Y).$$

Note that for $t \in [0,1]$ and $t \neq F(1_\delta)$,

$$(F^{-1})'(t) = [f(F^{-1}(t))]^{-1} \leq \exp\{e^{\|X\|_\infty}\} \max(\delta^{-1}, e^{\|X\|_\infty}) =: c_L,$$

so that F^{-1} is Lipschitz with constant c_L. Thus, letting $t \in [0,1]$ and $x = G^{-1}(t)$,

$$|F^{-1}(t) - G^{-1}(t)| = |F^{-1}(G(x)) - F^{-1}(F(x))| \leq c_L e^{\|X\|_\infty + d_\infty(X,Y)} d_2(X,Y),$$

whence

$$d_W(\psi_H^{-1}(X), \psi_H^{-1}(Y)) = d_2(F^{-1}, G^{-1}) \leq c_L e^{\|X\|_\infty} e^{d_\infty(X,Y)} d_2(X,Y). \tag{A.2}$$

Using (A.2), we establish (T2) for d_W by setting $C_2 = c_L e^{\|X\|_\infty}$ and $C_3 = e^{d_\infty(X,Y)}$.

To establish (T3), we let $X = \psi_H(f_1)$ and $X_K = \nu + \sum_{k=1}^K \eta_{1k}\rho_k$. Set $f_{1,K} = \psi_H^{-1}(X_K)$ and take C_1 as in (T1). Then, by assumption (A1) and equations (A.1) and (A.2),

$$E(d_2(f_1, f_{1,K})^2) \leq b_1 \sqrt{E(e^{4d_\infty(X,X_K)}) E(d_2(X, X_K)^4)} \quad \text{and}$$
$$E(d_W(f_1, f_{1,K})^2) \leq b_2 \sqrt{E(e^{4d_\infty(X,X_K)}) E(d_2(X, X_K)^4)},$$

where $b_1 = 2e^{2C_1}[(e^{2C_1} + \delta^{-1}) + 1]$ and $b_2 = \exp\{2(e^{C_1} + C_1)\} \max(\delta^{-2}, e^{2C_1})$. Note that $d_2(X, X_K)^2 = \sum_{k=K+1}^\infty \eta_{1k}^2 \leq \|X\|_2^2 \leq C_1^2|\mathcal{T}|$, so that

$$E(d_2(X, X_K)^4) \leq C_1^2|\mathcal{T}| E\left(\sum_{k=K+1}^\infty \eta_{1k}^2\right) = C_1^2|\mathcal{T}| \sum_{k=K+1}^\infty \tau_k \to 0.$$

So, we just need to show that $E(e^{4d_\infty(X,X_K)}) = O(1)$.

For the following, we need two lemmas that are listed below, and whose proofs are in the online supplement [38]. By applying assumptions (A1) and (T1), Lemma 2 implies the existence of the Lipschitz constant L_K for the residual process $X - X_K$ [see (4.14)]. By Lemma 1, we have

$$E(e^{4d_\infty(X,X_K)}) \leq E(\exp\{8|A|^{-1/2} d_2(X, X_K)\} + \exp\{8L_K^{1/3} d_2(X, X_K)^{2/3}\}).$$

Since $d_2(X, X_K) \leq \|X\|_2 < C_1|\mathcal{T}|^{1/2}$, the first expectation is bounded. For the second, we use Jensen's inequality to find

$$\begin{aligned} &E(\exp\{8L_K^{1/3} d_2(X, X_K)^{2/3}\}) \\ &\quad \leq 1 + \sum_{m=1}^\infty \frac{8^m [L_K^m E(d_2(X, X_K)^{2m})]^{1/3}}{m!}. \end{aligned} \tag{A.3}$$

For r.v.s. $Y_1, \ldots, Y_m$, $E(\prod_{i=1}^m Y_i) \leq \prod_{i=1}^m E(Y_i^m)^{1/m}$, so that

$$E(d_2(X, X_K)^{2m}) = \sum_{k_1=K+1}^{\infty} \cdots \sum_{k_m=K+1}^{\infty} E\left(\prod_{i=1}^m \eta_{1k_i}^2\right)$$

$$\leq \sum_{k_1=K+1}^{\infty} \cdots \sum_{k_m=K+1}^{\infty} \prod_{i=1}^m E(\eta_{1k_i}^{2m})^{1/m} = \left(\sum_{k=K+1}^{\infty} E(\eta_{1k}^{2m})^{1/m}\right)^m.$$

Next, by assumption, there exists B such that $L_K \sum_{k=K+1}^{\infty} \tau_k \leq B$ for large K. Then, by the assumption on the higher moments of η_{1k}^{2m}, for large K

$$L_K^m E(d_2(X, X_K)^{2m}) \leq \left(L_K \sum_{k=K+1}^{\infty} E(\eta_{1k}^{2m})^{1/m}\right)^m \leq \left(L_K \sum_{k=K+1}^{\infty} (r_m \tau_k^m)^{1/m}\right)^m$$

$$\leq r_m B^m.$$

Inserting this into (A.3), for large K

$$E(\exp\{8L_K^{1/3} d_2(X, X_K)^{2/3}\}) \leq 1 + \sum_{m=1}^{\infty} \frac{8^m B^{m/3} r_m^{1/3}}{m!}.$$

Using the assumption that $(\frac{r_{m+1}}{r_m})^{1/3} = o(m)$, the ratio test shows the sum converges. Since the sum is independent of K for K large, this establishes that $E(d_W(f_1, f_{1,K})^2) = o(1)$ and $E(d_2(f_1, f_{1,K})^2) = o(1)$. Using similar arguments, we can show that $E(d_W(f_1, f_{1,K})^4)$ and $E(d_2(f_1, f_{1,K})^4)$ are both $O(1)$, which completes the proof.

Next, we prove (T0)–(T3) for the log quantile density transformation. Let f and g be two densities as specified in assumption (T0) with cdf's F and G. For $t \in [0, 1]$,

$$\begin{aligned}
&|\psi_Q(f)(t) - \psi_Q(g)(t)| \\
&\quad = |\log f(F^{-1}(t)) - \log g(G^{-1}(t))| \\
&\quad \leq D_0(|f(F^{-1}(t)) - f(G^{-1}(t))| + |f(G^{-1}(t)) - g(G^{-1}(t))|) \\
&\quad \leq D_0^2 |F^{-1}(t) - G^{-1}(t)| + D_0 |f(G^{-1}(t)) - g(G^{-1}(t))|.
\end{aligned}$$

Since $F' = f$ is bounded below by D_0^{-1}, for any $t \in [0, 1]$ and $x = G^{-1}(t)$,

$$|F^{-1}(t) - G^{-1}(t)| = |F^{-1}(G(x)) - F^{-1}(F(x))| \leq D_0 |F(x) - G(x)|.$$

Recall that $d_\infty(F, G) \leq d_2(f, g) \leq d_\infty(f, g)$. Hence,

$$d_\infty(\psi_Q(f), \psi_Q(g)) \leq D_0(D_0^2 + 1) d_\infty(f, g),$$

$$d_2(\psi_Q(f), \psi_Q(g))^2 \leq 2D_0^2 \left[D_0^4 d_2(f, g)^2 + \int_0^1 (f(x) - g(x))^2 g(x)\, dx\right]$$

$$\leq 2D_0^3(D_0^3 + 1) d_2(f, g)^2,$$

whence C_0 in (T0). Next, we find that

$$\|\psi_Q(f)\|_\infty \leq \log D_1 \quad \text{and} \quad \|\psi_Q(f)'\|_\infty \leq D_1^3,$$

whence C_1 in (T1).

Now, let X and Y be as stated in (T2). Let F and G be the quantile functions corresponding to $f = \psi_Q^{-1}(X)$ and $g = \psi_Q^{-1}(Y)$, respectively. Then

$$\begin{aligned}|F^{-1}(t) - G^{-1}(t)| &\leq \theta_X^{-1}\left|\int_0^t (e^{X(s)} - e^{Y(s)})\,ds\right| + |\theta_X^{-1} - \theta_Y^{-1}|\int_0^t e^{Y(s)}\,ds \\ &\leq 2\theta_X^{-1}|\theta_X - \theta_Y|,\end{aligned}$$

where $\theta_X = \int_0^1 e^{X(s)}\,ds$ and $\theta_Y = \int_0^1 e^{Y(s)}\,ds$. It is clear that $\theta_X^{-1} \leq e^{\|X\|_\infty}$ and $|\theta_X - \theta_Y| \leq e^{\|X\|_\infty + d_\infty(X,Y)} d_2(X,Y)$, whence

$$|F^{-1}(t) - G^{-1}(t)| \leq 2e^{2\|X\|_\infty + d_\infty(X,Y)} d_2(X,Y).$$

This implies

$$d_W(\psi_Q^{-1}(X), \psi_Q^{-1}(Y)) \leq 2e^{4\|X\|_\infty} e^{2d_\infty(X,Y)} d_2(X,Y). \tag{A.4}$$

For $d = d_2$, using similar arguments as above, we find that

$$\begin{aligned}&d_2(\psi_Q^{-1}(X), \psi_Q^{-1}(Y)) \\ &\qquad \leq \sqrt{2}e^{6\|X\|_\infty}(4\|X'\|_\infty^2 + 3)^{1/2} e^{2d_\infty(X,Y)} d_2(X,Y).\end{aligned} \tag{A.5}$$

Equations (A.4) and (A.5) can then be used to find the constants C_2 and C_3 in (T2) for both $d = d_W$ and $d = d_2$, and also to prove (T3) in a similar manner to the log hazard transformation. □

The following auxiliary results, which are proved in the online supplement, are needed.

Lemma 1. *Let A be a closed and bounded interval of length $|A|$ and assume $X : A \to \mathbb{R}$ is continuous with Lipschitz constant L. Then*

$$\|X\|_\infty \leq 2\max(|A|^{-1/2}\|X\|_2, L^{1/3}\|X\|_2^{2/3}).$$

Lemma 2. *Let X be a stochastic process on a closed interval $\mathcal{T} \subset \mathbb{R}$ such that $\|X\|_\infty < C$ and $\|X'\|_\infty < C$ almost surely. Let ν and H be the mean and covariance functions associated with X, and ρ_k and τ_k, $k \geq 1$, be the eigenfunctions and eigenvalues of the integral operator with kernel H. Then $\|\nu\|_\infty < C$, $\|H\|_\infty < 4C^2$ and $\|\rho_k\|_\infty < 4C^2|\mathcal{T}|^{1/2}\tau_k^{-1}$ for all $k \geq 1$. Additionally, $\|\nu'\|_\infty < C$ and $\|\rho_k'\|_\infty < 4C^2|\mathcal{T}|^{1/2}\tau_k^{-1}$ for all $k \geq 1$.*

LEMMA 3. *Under assumptions* (A1) *and* (T1), *with* $\hat{\nu}, \tilde{\nu}, \widehat{H}, \widetilde{H}$ *as in (4.2) and (4.3),*

$$d_2(\nu, \tilde{\nu}) = O_p(n^{-1/2}), \qquad d_2(H, \widetilde{H}) = O_p(n^{-1/2}),$$
$$d_\infty(\nu, \tilde{\nu}) = O_p\biggl(\biggl(\frac{\log n}{n}\biggr)^{1/2}\biggr), \qquad d_\infty(H, \widetilde{H}) = O_p\biggl(\biggl(\frac{\log n}{n}\biggr)^{1/2}\biggr).$$

Under the additional assumptions (D1), (D2) *and* (S1), *we have*

$$d_2(\nu, \hat{\nu}) = O_p(n^{-1/2} + b_m), \qquad d_2(H, \widehat{H}) = O_p(n^{-1/2} + b_m),$$
$$d_\infty(\nu, \hat{\nu}) = O_p\biggl(\biggl(\frac{\log n}{n}\biggr)^{1/2} + a_m\biggr), \qquad d_\infty(H, \widehat{H}) = O_p\biggl(\biggl(\frac{\log n}{n}\biggr)^{1/2} + a_m\biggr).$$

LEMMA 4. *Assume* (A1), (T1) *and* (T2) *hold. Let* $A_k = \|\rho_k\|_\infty$, M *as in* (A1), δ_k *as in (5.1), and* C_1 *as in* (T1) *with* $D_1 = M$. *Let* $K^*(n) \to \infty$ *be any sequence which satisfies* $\tau_{K^*} n^{1/2} \to \infty$ *and*

$$\sum_{k=1}^{K^*} [(\log n)^{1/2} + \delta_k^{-1} + A_k + \tau_{K^*} \delta_k^{-1} A_k] = O(\tau_{K^*} n^{1/2}).$$

Let C_2 *be as in* (T2), $X_{i,K} = \nu + \sum_{k=1}^K \eta_{ik}\rho_k$, $\widetilde{X}_{i,K} = \tilde{\nu} + \sum_{k=1}^K \tilde{\eta}_{ik}\tilde{\rho}_k$, *and set*

$$S_{K^*} = \max_{1 \le K \le K^*} \max_{1 \le i \le n} C_2(\|X_{i,K}\|_\infty, \|X'_{i,K}\|_\infty).$$

Then

$$\max_{1 \le K \le K^*} \max_{1 \le i \le n} d(f_i(\cdot, K, \psi), \tilde{f}_i(\cdot, K, \psi)) = O_p\biggl(\frac{S_{K^*} \sum_{k=1}^{K^*} \delta_k^{-1}}{n^{1/2}}\biggr).$$

We now can also state the following corollary, the proof of which utilizes a lemma from [36].

COROLLARY 1. *Under assumption* (A1) *and* (T1), *letting* $A_k = \|\rho_k\|_\infty$, *with* δ_k *as in (5.1),*

$$|\tau_k - \tilde{\tau}_k| = O_p(n^{-1/2}),$$
$$d_2(\rho_k, \tilde{\rho}_k) = \delta_k^{-1} O_p(n^{-1/2}) \quad \textit{and}$$
$$d_\infty(\rho_k, \tilde{\rho}_k) = \tilde{\tau}_k^{-1} O_p\biggl(\frac{(\log n)^{1/2} + \delta_k^{-1} + A_k}{n^{1/2}}\biggr),$$

where all O_p terms are uniform over k. If the additional assumptions (D1), (D2) *and* (S1) *hold,*

$$\begin{aligned}|\tau_k - \hat{\tau}_k| &= O_p(n^{-1/2} + b_m),\\ d_2(\rho_k, \hat{\rho}_k) &= \delta_k^{-1} O_p(n^{-1/2} + b_m) \qquad \textit{and}\\ d_\infty(\rho_k, \hat{\rho}_k) &= \hat{\tau}_k^{-1} O_p\left(\frac{(\log n)^{1/2} + \delta_k^{-1} + A_k}{n^{1/2}} + a_m + b_m[\delta_k^{-1} + A_k]\right),\end{aligned}$$

where again all O_p terms are uniform over k.

Proof of Theorem 1. We will show the result for the fully observed case. The same arguments apply to the case where the densities are estimated.

First, suppose K is fixed. We may use the results of Lemma 2 due to (A1) and (T1) and define A_k as in Corollary 1. From

$$Y_{k,\alpha} = \nu + \alpha\sqrt{\tau_k}\rho_k \quad \text{and} \quad \widetilde{Y}_{k,\alpha} = \tilde{\nu} + \alpha\sqrt{\tilde{\tau}_k}\tilde{\rho}_k,$$

$g_k(\cdot, \alpha, \psi) = \psi^{-1}(Y_{k,\alpha})$ and similarly for $\tilde{g}_k$. Observe that, if $|\alpha| \le \alpha_0$,

$$\text{(A.6)} d_\infty(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}) \le d_\infty(\nu, \tilde{\nu}) + \alpha_0(\sqrt{\tilde{\tau}_1} d_\infty(\rho_k, \tilde{\rho}_k) + A_k|\sqrt{\tau_k} - \sqrt{\tilde{\tau}_k}|).$$

Next, $\max_{1\le k\le K} |\sqrt{\tau_k} - \sqrt{\tilde{\tau}_k}| = O_p(n^{-1/2})$ and $\max_{1\le k\le K} d_\infty(\rho_k, \tilde{\rho}_k) = O_p(1)$ by Corollary 1, so that $d_\infty(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}) = O_p(1)$, uniformly in k and $|\alpha| \le \alpha_0$. For $C_{2,k,\alpha} = C_2(\|Y_{k,\alpha}\|_\infty, \|Y'_{k,\alpha}\|_\infty)$ and $C_{3,k,\alpha} = C_3(d_\infty(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}))$ as in (T2),

$$\max_{1\le k\le K} \max_{|\alpha|\le\alpha_0} C_{2,k,\alpha} < \infty \quad \text{and} \quad \max_{1\le k\le K} \max_{|\alpha|\le\alpha_0} C_{3,k,\alpha} = O_p(1).$$

Furthermore,

$$d_2(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}) \le d_2(\nu, \tilde{\nu}) + \alpha_0(\sqrt{\tilde{\tau}_1} d_2(\rho_k, \tilde{\rho}_k) + |\sqrt{\tau_k} - \sqrt{\tilde{\tau}_k}|) = O_p(n^{-1/2}),$$

uniformly in k and $|\alpha| \le \alpha_0$, by Lemma 3. This means

$$\begin{aligned}\max_{1\le k\le K} \max_{|\alpha|\le\alpha_0} d(g_k(\cdot, \alpha, \psi), \tilde{g}_k(\cdot, \alpha, \psi)) &\le \max_{1\le k\le K} \max_{|\alpha|\le\alpha_0} C_{2,k,\alpha} C_{3,k,\alpha} d_2(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha})\\ &= O_p(n^{-1/2}).\end{aligned}$$

Next, we consider $K = K(n) \to \infty$. Define

$$S_K = \max_{|\alpha|\le\alpha_0} \max_{1\le k\le K} C_{2,k,\alpha}.$$

Let $B_K = \max_{1\le k\le K} A_k$ and take K to be a sequence which satisfies:

(i) $\tau_K n^{1/2} \to \infty$,

(ii) $(\log n)^{1/2} + \delta_K^{-1} + B_K = O(\tau_K n^{1/2})$, and
(iii) $S_K = o(\delta_K n^{1/2})$.

For $|\alpha| \leq \alpha_0$, we still have inequality (A.6). The term $d_\infty(\nu, \tilde{\nu})$ is $o_p(1)$ independently of K. From (i) and the above, it follows that $\max_{1\leq k\leq K} \tilde{\tau}_k^{-1} = O_p(\tau_K^{-1})$ and we find

$$\max_{1\leq k\leq K} |\sqrt{\tau_k} - \sqrt{\tilde{\tau}_k}| = O_p\left(\frac{1}{(\tau_K n)^{1/2}}\right).$$

Using Corollary 1 and (ii), this implies $\max_{1\leq k\leq K} d_\infty(\rho_k, \tilde{\rho}_k) = o_p(1)$, so that $d_\infty(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}) = O_p(1)$, uniformly over $k \leq K$ and $|\alpha| \leq \alpha_0$. Hence, $\max_{1\leq k\leq K} \max_{|\alpha|\leq\alpha_0} C_{3,k,\alpha} = O_p(1)$.

Similarly, we find that

$$d_2(Y_{k,\alpha}, \widetilde{Y}_{k,\alpha}) = O_p\left(\frac{1}{\delta_K n^{1/2}}\right),$$

uniformly over $k \leq K(n)$ and $|\alpha| \leq \alpha_0$. With (iii), this yields

$$\max_{|\alpha|\leq\alpha_0} \max_{1\leq k\leq K} d(g_k(\cdot, \alpha, \psi), \tilde{g}_k(\cdot, \alpha, \psi)) \leq O_p\left(\frac{S_K}{\delta_K n^{1/2}}\right) = o_p(1). \qquad \square$$

Proof of Theorem 2. We begin by placing the following restrictions on the sequence p_n:

(i) $p_n \uparrow 1$ and
(ii) for large n, $p_n \neq V_K V_\infty^{-1}$ for any K.

Furthermore, the corresponding sequence K^* must satisfy the assumption of Lemma 4. Set $\epsilon_K = \epsilon_K(n) = |V_K V_\infty^{-1} - p_n|$, $K = 1, \ldots, K^*$, where K^* is given in (4.11), and define $\pi_{K^*} = \min\{\epsilon_1, \ldots, \epsilon_{K^*}\}$. Letting S_{K^*} be defined as in Lemma 4 and $\beta_{K^*} = n^{-1/2}(S_{K^*} \sum_{k=1}^{K^*} \delta_k^{-1})$, we also require that

$$\left(\left(\frac{K^*}{n}\right)^{1/2} + \beta_{K^*} + \gamma_n\right)\pi_{K^*}^{-1} \to 0. \tag{A.7}$$

None of these restrictions are contradictory.

Next, let $f_{i,K} = f_i(\cdot, K, \psi)$ and define

$$\hat{V}_\infty = \frac{1}{n}\sum_{i=1}^n d(f_i, f_\oplus)^2 \quad \text{and} \quad \hat{V}_K = \hat{V}_\infty - \frac{1}{n}\sum_{i=1}^n d(f_i, f_{i,K})^2.$$

Observe that $\hat{V}_\infty - V_\infty = O_p(n^{-1/2})$ by the law of large numbers. Also, by (T3), for any $R > 0$,

$$\begin{aligned} P\Big(\max_{1\leq K\leq K^*} |(\hat{V}_\infty - \hat{V}_K) - (V_\infty - V_K)| > R\Big) &\leq \frac{K^*}{R^2 n} \max_{1\leq K\leq K^*} E(d(f_1, f_{1,K})^4) \\ &= O\left(\frac{K^*}{R^2 n}\right). \end{aligned}$$

Hence,

$$\max_{1\le K\le K^*}\left|\frac{\hat{V}_K}{\hat{V}_\infty}-\frac{V_K}{V_\infty}\right|=\max_{1\le K\le K^*}\left|\frac{\hat{V}_\infty-\hat{V}_K}{\hat{V}_\infty}-\frac{V_\infty-V_K}{V_\infty}\right|=O_p\left(\left(\frac{K^*}{n}\right)^{1/2}\right).$$

Define $\tilde{f}_{i,K}=\tilde{f}_i(\cdot,K,\psi)$. Then observe that

$$\begin{aligned}|(\hat{V}_\infty-\hat{V}_K)-(\tilde{V}_\infty-\tilde{V}_K)| &\le \frac{1}{n}\sum_{i=1}^n |d(f_i,f_{i,K})^2-d(f_i,\tilde{f}_{i,K})^2|\\ &\le \frac{1}{n}\sum_{i=1}^n d(f_{i,K},\tilde{f}_{i,K})(2d(f_i,f_{i,K})+d(f_{i,K},\tilde{f}_{i,K})).\end{aligned}$$

By using (T3), Lemma 4 and the assumptions on the sequence K^*, we find that

$$\max_{1\le K\le K^*}|(\hat{V}_\infty-\hat{V}_K)-(\tilde{V}_\infty-\tilde{V}_K)|=O_p(\beta_{K^*}).$$

By using similar arguments, we find that $\hat{V}_\infty-\tilde{V}_\infty=O_p(\gamma_n)$, which yields

$$\max_{1\le K\le K^*}\left|\frac{V_K}{V_\infty}-\frac{\tilde{V}_K}{\tilde{V}_\infty}\right|=O_p\left(\left(\frac{K^*}{n}\right)^{1/2}+\beta_{K^*}+\gamma_n\right). \tag{A.8}$$

To finish, observe that, since $p_n\ne V_K V_\infty^{-1}$ for any K when n is large, for such n

$$\{K^*\ne\tilde{K}^*\}=\left\{\max_{1\le K\le K^*}\left|\frac{V_K}{V_\infty}-\frac{\tilde{V}_K}{\tilde{V}_\infty}\right|>\pi_{K^*}\right\}.$$

Then, by (A.8), for any $\varepsilon>0$ there is $R>0$ such that

$$P\left(\max_{1\le K\le K^*}\left|\frac{V_K}{V_\infty}-\frac{\tilde{V}_K}{\tilde{V}_\infty}\right|>R\left(\left(\frac{K^*}{n}\right)^{1/2}+\beta_{K^*}+\gamma_n\right)\right)<\varepsilon$$

for all n. Then, by (A.7), for n large enough we have $P(K^*\ne\tilde{K}^*)<\varepsilon$. □

Acknowledgments. We wish to thank the Associate Editor and three referees for helpful remarks that led to an improved version of the paper.

SUPPLEMENTARY MATERIAL

The Wasserstein metric, Wasserstein–Fréchet mean, simulation results and additional proofs (DOI: 10.1214/15-AOS1363SUPP; .pdf). The supplementary material includes additional discussion on the Wasserstein distance and the rate of convergence of the Wasserstein–Fréchet mean is derived. Additional simulation results are presented for FVE values using the Wasserstein metric, similar to the boxplots in Figure 2, which correspond to FVE values using the L^2 metric. All assumptions are listed in one place. Lastly, additional proofs of auxiliary results are provided.

REFERENCES

[1] Achard, S., Salvador, R., Whitcher, B., Suckling, J. and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* **26** 63–72.

[2] Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T. and Calhoun, V. D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex* bhs352.

[3] Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes.* Academic Press, New York. MR0448463

[4] Bali, J. L., Boente, G., Tyler, D. E. and Wang, J.-L. (2011). Robust functional principal components: A projection-pursuit approach. *Ann. Statist.* **39** 2852–2882. MR3012394

[5] Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* **12** 512–523.

[6] Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343

[7] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311. MR0848110

[8] Bolstad, B. M., Irizarry, R. A., Åstrand, M. and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

[9] Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications. Lecture Notes in Statistics* **149**. Springer, New York. MR1783138

[10] Bouezmarni, T. and Rolin, J.-M. (2003). Consistency of the beta kernel density function estimator. *Canad. J. Statist.* **31** 89–98. MR1985506

[11] Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. R., Sperling, R. A. and Johnson, K. A. (2009). Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci.* **29** 1860–1873.

[12] Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

[13] Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

[14] Chen, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** 131–145. MR1718494

[15] Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 551–563. MR1394366

[16] Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. MR0650934

[17] Delicado, P. (2011). Dimensionality reduction when data are density functions. *Comput. Statist. Data Anal.* **55** 401–420. MR2736564

[18] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York. MR0780746

[19] Dubin, J. A. and Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *J. Amer. Statist. Assoc.* **100** 872–881. MR2201015

[20] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669. MR0083864

[21] Egozcue, J. J., Diaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Math. Sin. (Engl. Ser.)* **22** 1175–1182. MR2245249

[22] Ferreira, L. K. and Busatto, G. F. (2013). Resting-state functional connectivity in normal brain aging. *Neuroscience & Biobehavioral Reviews* **37** 384–400.

[23] Fletcher, P. T., Lu, C., Pizer, S. M. and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging* **23** 995–1005.

[24] Gajek, L. (1986). On improving density estimators which are not Bona fide functions. *Ann. Statist.* **14** 1612–1618. MR0868324

[25] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

[26] Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 109–126. MR2212577

[27] Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

[28] Höffding, W. (1940). Maszstabinvariante Korrelationstheorie. *Schr. Math. Inst. U. Inst. Angew. Math. Univ. Berlin* **5** 181–233. MR0004426

[29] Hron, K., Menafoglio, A., Templ, M., Hruzova, K. and Filzmoser, P. (2014). Simplicial principal component analysis for density functions in Bayes spaces. *MOX-report* **25** 2014.

[30] Jones, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Ann. Inst. Statist. Math.* **44** 721–727.

[31] Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.* **46** 140–145.

[32] Kneip, A. and Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.* **96** 519–542. MR1946423

[33] Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. MR2766854

[34] Mallows, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515. MR0298812

[35] Müller, H. G. and Stadtmüller, U. (1999). Multivariate boundary kernels and a continuous least squares principle. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 439–458. MR1680306

[36] Müller, H.-G. and Yao, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202

[37] Parzen, E. (1979). Nonparametric statistical modeling. *J. Amer. Statist. Assoc.* **74** 105–121.

[38] Petersen, A. and Müller, H.-G. (2015). Supplement to "Functional data analysis for density functions by transformation to a Hilbert space." DOI:10.1214/15-AOS1363SUPP.

[39] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

[40] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23** 470–472. MR0049525

[41] Sheline, Y. I. and Raichle, M. E. (2013). Resting state functional connectivity in preclinical Alzheimer's disease. *Biol. Psychiatry* **74** 340–347.

[42] Srivastava, A., Jermyn, I. and Joshi, S. (2007). Riemannian analysis of probability density functions with applications in vision. *Proceedings from IEEE Conference on Computer Vision and Pattern Recognition* **25** 1–8.

[43] Srivastava, A., Klassen, E., Joshi, S. H. and Jermyn, I. H. (2011a). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33** 1415–1428.

[44] Srivastava, A., Wu, W., Kurtek, S., Klassen, E. and Marron, J. S. (2011b). Registration of functional data using Fisher–Rao metric. Available at arXiv:1103.3817v2 [math.ST].

[45] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[46] Tukey, J. W. (1965). Which part of the sample contains the information? *Proc. Natl. Acad. Sci. USA* **53** 127–134. MR0172387

[47] Villani, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483

[48] Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation. *J. Amer. Statist. Assoc.* **86** 343–361. MR1137118

[49] Worsley, K. J., Chen, J.-I., Lerch, J. and Evans, A. C. (2005). Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360** 913–920.

[50] Zhang, Z. and Müller, H.-G. (2011). Functional density synchronization. *Comput. Statist. Data Anal.* **55** 2234–2249. MR2786984

Department of Statistics
University of California, Davis
Mathematical Sciences Building 4118
399 Crocker Lane
One Shields Avenue
Davis, California 95616
USA
E-mail: alxpetersen@gmail.com
hgmueller@ucdavis.edu

SUPPLEMENTARY MATERIAL TO "FUNCTIONAL DATA ANALYSIS FOR DENSITY FUNCTIONS BY TRANSFORMATION TO A HILBERT SPACE"

The Wasserstein metric, Wasserstein-Fréchet mean, simulation results and additional proofs

Alexander Petersen and Hans-Georg Müller

S.1. The Wasserstein Metric. The equivalence of the metrics

$$d_Q(f,g)^2 = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 \, dt \quad \text{and} \quad d_W(f,g)^2 = \inf_{X \sim f, Y \sim g} E(X - Y)^2$$

is well known. It can be easily seen by applying a covariance identity due to [28]. If $X \sim F$, $Y \sim G$ and $(X,Y) \sim H$, then this identity states that

$$\mathrm{Cov}(X,Y) = \int \int \{H(u,v) - F(u)G(v)\} \, du \, dv.$$

Expanding the expectation $E(X - Y)^2$, one finds that the distance is obtained by maximizing $E(XY)$, or, equivalently, by maximizing $\mathrm{Cov}(X,Y)$. For a random variable U that is uniformly distributed on $[0,1]$, take $X^* = F^{-1}(U)$ and $Y^* = G^{-1}(U)$. Then $X^* \sim F$, $Y^* \sim G$ and the distribution function of (X^*, Y^*) is given by $H^*(u,v) = \min(F(u), G(v))$. Clearly, for any joint distribution of $X \sim F$ and $Y \sim G$, we have $H \leq H^*$. By Hoeffding's inequality, this means $\mathrm{Cov}(X,Y) \leq \mathrm{Cov}(X^*, Y^*)$. Thus,

$$\begin{aligned} d_W(f,g)^2 = E[(X^* - Y^*)^2] &= E[(F^{-1}(U) - G^{-1}(U))^2] \\ &= \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 \, dt. \end{aligned}$$

Let Q be the quantile process corresponding to the density process $f \sim \mathfrak{F}$ and set $Q_\oplus(t) = E(Q(t))$. For $q_\oplus = Q'_\oplus$ and $F_\oplus = Q_\oplus^{-1}$, the *Wasserstein-Fréchet* mean is

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

Its estimation can thus be reduced to estimating the function $q_\oplus$. Due to the restrictions on the space $\mathcal{F}$ (see assumption (A1)), we can pass differentiation inside the expectation so that $E(Q'(t)) = q_\oplus(t)$. This suggests averaging the quantile densities of the sample to obtain an estimator for $q_\oplus$.

Starting with either the densities, f_i, or their estimates, $\check{f}_i$, $i = 1, \ldots, n$, we therefore use the corresponding quantile densities (q_i or $\check{q}_i$) to estimate $q_\oplus$ by

$$\tilde{q}_\oplus(t) = \frac{1}{n}\sum_{i=1}^{n} q_i(t), \quad \text{respectively,} \quad \hat{q}_\oplus(t) = \frac{1}{n}\sum_{i=1}^{n} \check{q}_i(t).$$

Computing the corresponding distribution functions, we thus estimate the Wasserstein-Fréchet mean by

$$\tilde{f}_\oplus(x) = \frac{1}{\tilde{q}_\oplus(\tilde{F}_\oplus(x))}, \quad \text{respectively,} \quad \hat{f}_\oplus(x) = \frac{1}{\hat{q}_\oplus(\hat{F}_\oplus(x))}.$$

As Theorem 2 requires a rate of convergence γ_n for the Wasserstein-Fréchet mean estimator based on fully observed densities, the following result shows that we make take $\gamma_n = n^{-1/2}$ in the case of fully observed densities.

PROPOSITION 3. *Under assumption (A1), the estimator $\tilde{f}_\oplus$ of $f_\oplus$ for the Wasserstein-Fréchet mean satisfies*

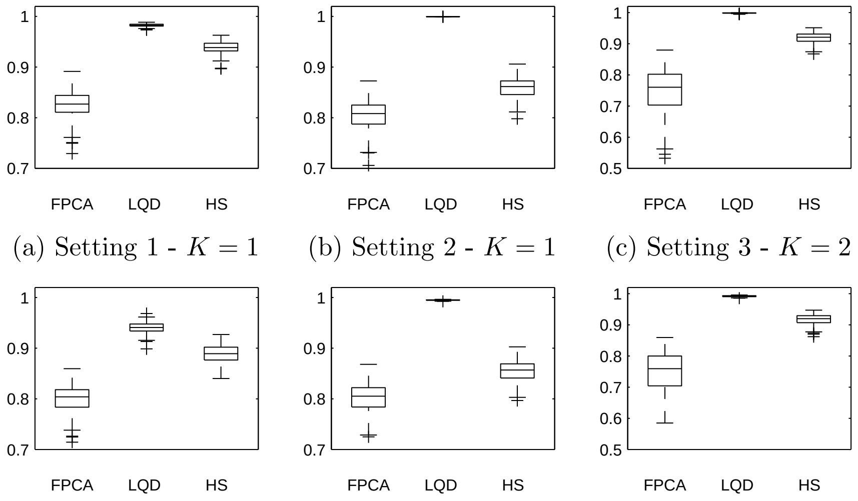
$$d_W(f_\oplus, \tilde{f}_\oplus) = O_p(n^{-1/2}).$$

PROOF. By Thm 3.9 in [9], $d_2(q_\oplus, \tilde{q}_\oplus) = O_p(n^{-1/2})$. As $|Q_\oplus(t) - \tilde{Q}_\oplus(t)| \leq d_2(q_\oplus, \tilde{q}_\oplus)$, we also have

$$d_W(f_\oplus, \tilde{f}_\oplus) = d_2(Q_\oplus, \tilde{Q}_\oplus) = O_p(n^{-1/2}).$$

□

S.2. Simulation Results for the Wasserstein Metric. Figure 7 shows the distribution of fraction of variance explained values in terms of the distance d_W for all simulation settings, similar to Figure 2 in the main text which shows the results for the ordinary L^2 distance. The use of the Wasserstein distance more clearly demonstrates the weakness of ordinary FPCA. The Hilbert sphere method performs relatively better in the context of metric d_W than the L^2 metric, but is still outperformed by the transformation method using the log quantile density transformation, ψ_Q.



(a) Setting 1 - $K = 1$ (b) Setting 2 - $K = 1$ (c) Setting 3 - $K = 2$

(d) Setting 1 - $K = 1$ (e) Setting 2 - $K = 1$ (f) Setting 3 - $K = 2$

Fig 7: Boxplots of fraction of variance explained for 200 simulations, using the Wasserstein metric, d_W. The first row corresponds to fully observed densities and the second corresponds to estimated densities. The columns correspond to settings 1, 2 and 3 from left to right (see Table 1). The methods are denoted by 'FPCA' for ordinary FPCA on the densities, 'LQD' for the transformation approach with ψ_Q and 'HS' for the Hilbert sphere method.

S.3. Listing of All Assumptions. The following is a systematic compilation of all assumptions, subsets of which are used for various results and some of which have been stated in the main text. Recall that d_2 and d_∞ denote the L^2 and uniform metrics, respectively, and $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the corresponding norms.

(A1) For all $f \in \mathcal{F}$, f is continuously differentiable. Moreover, there is a constant $M > 1$ such that, for all $f \in \mathcal{F}$, $\|f\|_\infty$, $\|1/f\|_\infty$ and $\|f'\|_\infty$ are all bounded above by M.

(D1) For a sequence $b_N = o(1)$, the density estimator $\check{f}$, based on an i.i.d. sample of size N, satisfies $\check{f} \geq 0$, $\int_0^1 \check{f}(x)\,dx = 1$ and

$$\sup_{f\in\mathcal{F}} E(d_2(f,\check{f})^2) = O(b_N^2).$$

(D2) For a sequence $a_N = o(1)$ and some $R > 0$, the density estimator $\check{f}$, based on an i.i.d. sample of size N, satisfies

$$\sup_{f\in\mathcal{F}} P(d_\infty(f,\check{f}) > Ra_N) \to 0.$$

(S1) Let $\check{f}$ be a density estimator that satisfies (D2), and suppose densities $f_i \in \mathcal{F}$ are estimated by $\check{f}_i$ from i.i.d. samples of size $N_i = N_i(n)$, $i = 1,\ldots,n$, respectively. There exists a sequence of lower bounds $m(n) \leq \min_{1\leq i\leq n} N_i$ such that $m(n) \to \infty$ as $n \to \infty$ and

$$n \sup_{f\in\mathcal{F}} P(d_\infty(f,\check{f}) > Ra_m) \to 0$$

where, for generic $f \in \mathcal{F}$, $\check{f}$ is the estimated density from a sample of size $N(n) \geq m(n)$.

(K1) The kernel κ is of bounded variation and is symmetric about 0.

(K2) The kernel κ satisfies $\int_0^1 \kappa(u)\,du > 0$, and $\int_{\mathbb{R}} |u|\kappa(u)\,du$, $\int_{\mathbb{R}} \kappa^2(u)\,du$ and $\int_{\mathbb{R}} |u|\kappa^2(u)\,du$ are finite.

(T0) Let f, $g \in \mathcal{G}$ with f differentiable and $\|f'\|_\infty < \infty$. Set

$$D_0 \geq \max\left(\|f\|_\infty, \|1/f\|_\infty, \|g\|_\infty, \|1/g\|_\infty, \|f'\|_\infty\right).$$

There exists C_0 depending only on D_0 such that

$$d_2(\psi(f),\psi(g)) \leq C_0\, d_2(f,g), \quad d_\infty(\psi(f),\psi(g)) \leq C_0\, d_\infty(f,g).$$

(T1) Let $f \in \mathcal{G}$ be differentiable with $\|f'\|_\infty < \infty$ and let D_1 be a constant bounded below by $\max(\|f\|_\infty, \|1/f\|_\infty, \|f'\|_\infty)$. Then $\psi(f)$ is differentiable and there exists $C_1 > 0$ depending only on D_1 such that $\|\psi(f)\|_\infty \leq C_1$ and $\|\psi(f)'\|_\infty \leq C_1$.

(T2) Let d be the selected metric in density space, Y be continuous and X be differentiable on $\mathcal{T}$ with $\|X'\|_\infty < \infty$. There exist constants $C_2 = C_2(\|X\|_\infty, \|X'\|_\infty) > 0$ and $C_3 = C_3(d_\infty(X,Y)) > 0$ such that

$$d(\psi^{-1}(X), \psi^{-1}(Y)) \leq C_2\, C_3\, d_2(X,Y)$$

and, as functions, C_2 and C_3 are increasing in their respective arguments.

(T3) For a given metric d on the space of densities and $f_{1,K} = f_1(\cdot, K, \psi)$ (see (4.5)), $V_\infty - V_K \to 0$ and $E(d(f, f_{1,K})^4) = O(1)$ as $K \to \infty$.

S.4. Additional Proofs.

Lemma 1. *Let A be a closed and bounded interval of length $|A|$ and assume $X : A \to \mathbb{R}$ is continuous with Lipschitz constant L. Then*

$$\|X\|_\infty \leq 2 \max\left(|A|^{-1/2}\|X\|_2,\ \ L^{1/3}\|X\|_2^{2/3}\right).$$

Proof of Lemma 1. Let t^* satisfy $|X(t^*)| = \|X\|_\infty$ and define $I = [t^* - \|X\|_\infty/(2L), t^* + \|X\|_\infty/(2L)] \cap A$. Then, for $t \in I$, $|X(t)| \geq \|X\|_\infty/2$. If $I = A$,

$$\|X\|_2^2 = \int_A X^2(s)\, ds \geq \frac{|A|\|X\|_\infty^2}{4},$$

so $\|X\|_\infty \leq 2|A|^{-1/2}\|X\|_2$. If $I \neq A$, suppose without loss of generality that $t^* + \|X\|_\infty/(2L) \in A$. Then

$$\|X\|_2^2 \geq \int_{t^*}^{t^*+\|X\|_\infty/(2L)} X^2(s)\, ds \geq \frac{\|X\|_\infty^2}{4} \cdot \frac{\|X\|_\infty}{2L} = \frac{\|X\|_\infty^3}{8L},$$

so $\|X\|_\infty \leq 2L^{1/3}\|X\|_2^{2/3}$. □

Lemma 2. *Let X be a stochastic process on a closed interval $\mathcal{T} \subset \mathbb{R}$ such that $\|X\|_\infty < C$ and $\|X'\|_\infty < C$ almost surely. Let ν and H be the mean and covariance functions associated with X, and ρ_k and τ_k, $k \geq 1$, be the eigenfunctions and eigenvalues of the integral operator with kernel H. Then $\|\nu\|_\infty < C$, $\|H\|_\infty < 4C^2$ and $\|\rho_k\|_\infty < 4C^2|\mathcal{T}|^{1/2}\tau_k^{-1}$ for all $k \geq 1$. Additionally, $\|\nu'\|_\infty < C$ and $\|\rho_k'\|_\infty < 4C^2|\mathcal{T}|^{1/2}\tau_k^{-1}$ for all $k \geq 1$.*

Proof. Since the bounds on X and X' are deterministic, $\|\nu\|_\infty$ and $\|H\|_\infty$ are both bounded by the given constants. The bound on $\|\rho_k\|_\infty$ follows since $\rho_k(t) = \tau_k^{-1}\int_{\mathcal{T}} H(s,t)\rho_k(s)\, ds$ and $\|\rho_k\|_2 = 1$. Dominated convergence implies that ν' exists and is bounded by C, and also implies the bound of $4C^2$ for the partial derivatives of H, which then leads to the bounds on ρ_k' for all k. □

LEMMA 3. *Under assumptions (A1) and (T1), with $\hat{\nu}, \tilde{\nu}, \widehat{H}, \widetilde{H}$ as in (4.2) and (4.3),*

$$\begin{aligned} & d_2(\nu, \tilde{\nu}) = O_p(n^{-1/2}), && d_2(H, \widetilde{H}) = O_p(n^{-1/2}), \\ & d_\infty(\nu, \tilde{\nu}) = O_p\left(\left(\frac{\log n}{n}\right)^{1/2}\right), && d_\infty(H, \widetilde{H}) = O_p\left(\left(\frac{\log n}{n}\right)^{1/2}\right). \end{aligned} \tag{S.1}$$

Under the additional assumptions (D1), (D2) and (S1), we have

$$\begin{aligned} & d_2(\nu, \hat{\nu}) = O_p(n^{-1/2} + b_m), && d_2(H, \widehat{H}) = O_p(n^{-1/2} + b_m), \\ & d_\infty(\nu, \hat{\nu}) = O_p\left(\left(\frac{\log n}{n}\right)^{1/2} + a_m\right), && d_\infty(H, \widehat{H}) = O_p\left(\left(\frac{\log n}{n}\right)^{1/2} + a_m\right). \end{aligned} \tag{S.2}$$

PROOF. Assumptions (A1) and (T1) imply $E\|X\|_2^2 < \infty$, so the first line in (S.1) follows from Theorems 3.9 and 4.2 in [9]. The second line in (S.1) follows from Corollaries 2.3(b) and 3.5(b) in [33]. We will show the argument for the mean estimate in (S.2), and the covariance follows similarly.

Let M be as given in assumption (A1) and set $D_1 = 2M$. Define

$$E_n = \bigcap_{i=1}^{n} \left\{ d_\infty(f_i, \check{f}_i) \le D_1^{-1} \right\}.$$

Then $P(E_n^c) \to 0$ by assumptions (D2) and (S1). Take C_1 as given in (T1) for D_1 as defined above. Also by (S1), there is $R > 0$ such that

$$P(\{d_\infty(\tilde{\nu}, \hat{\nu}) > R a_m\} \cap E_n) \le n \max_{1 \le i \le n} P(d_\infty(f_i, \check{f}_i) > C_1^{-1} R a_m) \to 0$$

as $n \to \infty$, so $d_\infty(\tilde{\nu}, \hat{\nu}) = O_p(a_m)$. Thus, by the triangle inequality, $d_\infty(\nu, \hat{\nu}) = O_p\left(\left(\frac{\log n}{n}\right)^{1/2} + a_m\right)$.

Next, letting $\widehat{X}_i = \psi(\check{f}_i)$,

$$\begin{aligned} P\left(\{d_2(\tilde{\nu},\hat{\nu}) > R\} \cap E_n\right) &\leq P\left(\left\{\sum_{i=1}^n d_2(X_i, \widehat{X}_i) > Rn\right\} \cap E_n\right) \\ &\leq P\left(\sum_{i=1}^n d_2(f_i, \check{f}_i) > C_1^{-1}Rn\right) \\ &\leq C_1 R^{-1} n^{-1} \sum_{i=1}^n \sqrt{E(d_2(f_i,\check{f}_i)^2)} = R^{-1}O(b_m), \end{aligned}$$

which shows that $d_2(\tilde{\nu},\hat{\nu}) = O_p(b_m)$, so the result holds by the triangle inequality. □

Corollary 1. *Under assumption (A1) and (T1), letting $A_k = \|\rho_k\|_\infty$, with δ_k as in (5.1),*

$$\begin{aligned} |\tau_k - \tilde{\tau}_k| &= O_p(n^{-1/2}), \\ d_2(\rho_k, \tilde{\rho}_k) &= \delta_k^{-1} O_p(n^{-1/2}), \text{ and} \\ d_\infty(\rho_k, \tilde{\rho}_k) &= \tilde{\tau}_k^{-1} O_p\left(\frac{(\log n)^{1/2} + \delta_k^{-1} + A_k}{n^{1/2}}\right), \end{aligned} \tag{S.3}$$

where all O_p terms are uniform over k. If the additional assumptions (D1), (D2) and (S1) hold,

$$\begin{aligned} |\tau_k - \hat{\tau}_k| &= O_p(n^{-1/2} + b_m), \\ d_2(\rho_k, \hat{\rho}_k) &= \delta_k^{-1} O_p(n^{-1/2} + b_m), \text{ and} \\ d_\infty(\rho_k, \hat{\rho}_k) &= \hat{\tau}_k^{-1} O_p\left(\frac{(\log n)^{1/2} + \delta_k^{-1} + A_k}{n^{1/2}} + a_m + b_m[\delta_k^{-1} + A_k]\right), \end{aligned} \tag{S.4}$$

where again all O_p terms are uniform over k.

Proof. First, observe that (A1) and (T1) together imply that X satisfies the assumptions of Lemma 2. The first two lines of both (S.3) and (S.4) follow by applying Lemmas 4.2 and 4.3 of [9] with the rates given in Lemma 3, above. For the uniform metric on the eigenfunctions, we follow the argument given in the proof of Lemma 1 in [36] to find that

$$d_\infty(\tau_k\rho_k, \tilde{\tau}_k\tilde{\rho}_k) \leq |\mathcal{T}|^{1/2}\left[d_\infty(H, \widetilde{H}) + \|H\|_\infty d_2(\rho_k, \tilde{\rho}_k)\right] = O_p\left(\frac{(\log n)^{1/2} + \delta_k^{-1}}{n^{1/2}}\right).$$

It follows that

$$\begin{aligned}|\rho_k(s) - \tilde{\rho}_k(s)| &\leq \tilde{\tau}_k^{-1}\left(|\tau_k\rho_k(s) - \tilde{\tau}_k\rho_k(s)| + |\rho_k(s)|\,|\tau_k - \tilde{\tau}_k|\right)\\ &= \tilde{\tau}_k^{-1} O_p\left(\frac{(\log n)^{1/2} + \delta_k^{-1} + A_k}{n^{1/2}}\right).\end{aligned}$$

Since this last expression is independent of s, this proves the third line of (S.3). The third line of (S.4) is proven in a similar manner. □

LEMMA 4. *Assume (A1), (T1) and (T2) hold. Let $A_k = \|\rho_k\|_\infty$, M as in (A1), δ_k as in (5.1), and C_1 as in (T1) with $D_1 = M$. Let $K^*(n) \to \infty$ be any sequence which satisfies $\tau_{K^*} n^{1/2} \to \infty$ and*

$$\sum_{k=1}^{K^*}\left[(\log n)^{1/2} + \delta_k^{-1} + A_k + \tau_{K^*}\delta_k^{-1}A_k\right] = O(\tau_{K^*}n^{1/2}).$$

Let C_2 be as in (T2), $X_{i,K} = \nu + \sum_{k=1}^K \eta_{ik}\rho_k$, $\widetilde{X}_{i,K} = \tilde{\nu} + \sum_{k=1}^K \tilde{\eta}_{ik}\tilde{\rho}_k$, and set

$$S_{K^*} = \max_{1\leq K\leq K^*}\max_{1\leq i\leq n} C_2(\|X_{i,K}\|_\infty, \|X'_{i,K}\|_\infty).$$

Then

$$\max_{1\leq K\leq K^*}\max_{1\leq i\leq n} d(f_i(\cdot, K, \psi), \tilde{f}_i(\cdot, K, \psi)) = O_p\left(\frac{S_{K^*}\sum_{k=1}^{K^*}\delta_k^{-1}}{n^{1/2}}\right).$$

PROOF. First, observe that $f_i(\cdot, K, \psi) = \psi^{-1}(X_{i,K})$ and $\tilde{f}_i(\cdot, K, \psi) = \psi^{-1}(\widetilde{X}_{i,K})$. Recall that $|\eta_{ik}| \leq 2C_1|\mathcal{T}|^{1/2}$ for all i and k (see (4.13). Then, by (A1) and Corollary 1,

$$|\eta_{ik} - \tilde{\eta}_{ik}| \leq d_2(X_i, \nu)d_2(\rho_k, \tilde{\rho}_k) + d_2(\nu, \tilde{\nu}) = \delta_k^{-1}O_p(n^{-1/2}),$$

where the O_p term is uniform over i and k. Next, using Lemma 3 and Corollary 1, along with the requirement that $\tau_{K^*}n^{1/2} \to \infty$, for $K \leq K^*$

$$\begin{aligned}d_\infty(X_{i,K}, \widetilde{X}_{i,K}) &\leq d_\infty(\nu, \tilde{\nu}) + \sum_{k=1}^K d_\infty(\eta_{ik}\rho_k, \tilde{\eta}_{ik}\tilde{\rho}_k)\\ &\leq d_\infty(\nu, \tilde{\nu}) + \sum_{k=1}^K |\eta_{ik}|d_\infty(\rho_k, \tilde{\rho}_k) + \sum_{k=1}^K \|\rho_k\|_\infty|\eta_{ik} - \tilde{\eta}_{ik}|\\ &= O_p\left(\frac{\sum_{k=1}^K\left[(\log n)^{1/2} + \delta_k^{-1} + A_k + \tau_K\delta_k^{-1}A_k\right]}{\tau_K n^{1/2}}\right).\end{aligned}$$

Since the O_p term does not depend on i or K, by the first assumption in the statement of the Lemma, we have

$$\max_{1\le K\le K^*}\max_{1\le i\le n} d_\infty(X_{i,K},\widetilde{X}_{i,K}) = O_p(1).$$

For $C_{3,K,i} = C_3(d_\infty(X_{i,K},\widetilde{X}_{i,K}))$ as in (T2),

$$\max_{1\le K\le K^*}\max_{1\le i\le n} C_{3,K,i} = O_p(1),$$

whence

$$\begin{aligned}
d_2(X_{i,K},\widetilde{X}_{i,K}) &\le d_2(\nu,\tilde{\nu}) + \sum_{k=1}^{K} d_2(\eta_{ik}\rho_k,\tilde{\eta}_{ik}\tilde{\rho}_k)\\
&\le d_2(\nu,\tilde{\nu}) + \sum_{k=1}^{K}|\eta_{ik}|d_2(\rho_k,\tilde{\rho}_k) + \sum_{k=1}^{K}|\eta_{ik}-\tilde{\eta}_{ik}|\\
&= O_p\left(n^{-1/2}\sum_{k=1}^{K}\delta_k^{-1}\right).
\end{aligned}$$

Again, this O_p term does not depend on i or K, so

$$\max_{1\le K\le K^*}\max_{1\le i\le n} d_2(X_{i,K},\widetilde{X}_{i,K}) = O_p\left(n^{-1/2}\sum_{k=1}^{K^*}\delta_k^{-1}\right),$$

leading to

$$\begin{aligned}
\max_{1\le K\le K^*}\max_{1\le i\le n} d(f_i(\cdot,K,\psi),\tilde{f}_i(\cdot,K,\psi)) &\le S_{K^*}\max_{1\le K\le K^*}\max_{1\le i\le n} C_{3,K,i}\, d_2(X_{i,K},\widetilde{X}_{i,K})\\
&= O_p\left(\frac{S_{K^*}\sum_{k=1}^{K^*}\delta_k^{-1}}{n^{1/2}}\right).
\end{aligned}$$

□

Biometrika (2018), **xx**, x, *pp*. 1–12
Printed in Great Britain

Wasserstein Covariance for Multiple Random Densities

BY ALEXANDER PETERSEN
Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106, U.S.A.
petersen@pstat.ucsb.edu

AND HANS-GEORG MÜLLER
Department of Statistics, University of California, Davis, California 95616, U.S.A.
hgmueller@ucdavis.edu

SUMMARY

A common feature of methods for analyzing samples of probability density functions is that they respect the geometry inherent to the space of densities. Once a metric is specified for this space, the Fréchet mean is typically used to quantify and visualize the average density from the sample. For one-dimensional densities, the Wasserstein metric is popular due to its theoretical appeal and interpretive value as an optimal transport metric, leading to the Wasserstein–Fréchet mean or barycenter as the mean density. We extend the existing methodology for samples of densities in two key directions. First, motivated by applications in neuroimaging, we consider dependent density data, where a p-vector of univariate random densities is observed for each sampling unit. Second, introduce Wasserstein covariance and propose intuitively appealing estimators for both fixed and diverging p, where the latter corresponds to continuously-indexed densities. We also give theory demonstrating consistency and asymptotic normality, while accounting for errors introduced in the unavoidable preparatory density estimation step. The utility of the Wasserstein covariance matrix is demonstrated through applications to functional connectivity in the brain using functional magnetic resonance imaging data and to the secular evolution of mortality for various countries.

Some key words: Barycenter; Fréchet Mean; Fréchet Variance; Functional Connectivity; Mortality; Random Density.

1. INTRODUCTION

The analysis of samples of density functions or distributions is an important and challenging problem for modern statistical practice (Delicado, 2011). Examples include distributions of age at death or mortality for different countries, warping functions, or distributions of voxel-to-voxel correlations of functional magnetic resonance imaging signals for a sample of subjects. While functional principal component analysis using cross-sectional averaging can be directly applied for samples of density functions (Kneip & Utikal, 2001), more recently techniques have been developed that incorporate the geometric constraints inherent to the space of density functions. A popular metric for data where each data atom corresponds to a randomly sampled distribution or density is the Wasserstein metric, both for its theoretical appeal and its convincing empirical performance in various applications (Bolstad et al., 2003; Zhang & Müller, 2011; Bigot et al., 2016, 2017; Panaretos & Zemel, 2016).

We consider samples of density data where multiple random densities per subject are observed, that is, repeated realizations of a stochastic process defined on $\mathcal{D}^p$, where $\mathcal{D}$ is a space of one-dimensional density functions. For $p = 1$, a variety of methods have been proposed, focusing on concepts of mean, modes of variation and dimensionality reduction (Delicado, 2011; Petersen & Müller, 2016; Hron et al., 2016). However, generalizations of these methods for $p > 1$, where one has a vector of densities, have not yet been developed, even though such data arise in various applications. An essential first step is the extension of the mean and variance concepts for samples of densities to a vector of means and covariance matrix. We demonstrate here the usefulness of these concepts for applications. The type of data we consider is different from data where one observes, for each subject, a random sample of identically distributed p-vectors, for which the joint distribution is of interest. Instead, we focus on the joint modeling of p univariate densities.

Using the Wasserstein geometry of optimal transport on the space of densities, we extend the concepts of Fréchet mean and variance to a measure of Wasserstein covariance between component densities for a p-variate density-valued process. As one does not observe the densities themselves but rather samples of univariate data that they generate, a preliminary density estimation step is necessary and is taken into account in our analysis. Our theoretical arguments show that the population Wasserstein covariance can be estimated consistently, with a limiting Gaussian distribution under sufficient conditions for the estimation error associated with the preliminary estimation step to be negligible. Motivated by one of the applications, we also consider continuously varying densities, where one observes a discretized version of a continuously evolving density-valued process, similar to densely observed repeated functional data (Park & Staicu, 2015; Chen et al., 2017). For this situation we target a Wasserstein covariance surface and develop theory for its consistent estimation while still accounting for errors in density estimation.

The utility of the proposed methodology is demonstrated through the analysis of functional magnetic resonance imaging and mortality data. In the application to brain imaging, we investigate differences in intra-regional functional connectivity between Alzheimer's patients, cognitively normal subjects, and a third group of individuals diagnosed with mild cognitive impairment, a sign of increased risk for developing dementia. Here, intra-regional functional connectivity is quantified by the distribution of voxel-to-voxel correlations within a neighborhood, called a functional hub, corresponding to a small region in the brain. Multiple densities per subject are obtained for a number of hubs simultaneously, leading to a random vector of densities obtained for each subject in the sample. In the mortality application, we consider distributions of age at death over a range of calendar years, and compare the resulting Wasserstein covariance surfaces between a group of Eastern European countries and a second group of other nations.

2. WASSERSTEIN COVARIANCE

2·1. *The Wasserstein Metric and Geometry*

Let $\mathcal{D}$ be a class of one-dimensional densities such that $\int_{\mathbb{R}} u^2 f(u)\mathrm{d}u < \infty$ for all $f \in \mathcal{D}$. For $f,\, g \in \mathcal{D}$, suppose $Y \sim f$ and consider the collection of nondecreasing maps $T^* : \mathbb{R} \to \mathbb{R}$, known as transports, such that $T^*(Y) \sim g$. The optimal transport problem that leads to the Wasserstein metric is to find the transport that minimizes

$$\int_{\mathbb{R}} \{T^*(u) - u\}^2 \, f(u)\mathrm{d}u, \tag{1}$$

and the solution is known to be $T = G^{-1} \circ F$, where F and G are the distribution functions of f and g, respectively (Ambrosio et al., 2008). The resulting squared Wasserstein distance is

$$d_W^2(f, g) = \int_{\mathbb{R}} \{T(u) - u\}^2 \, f(u)\mathrm{d}u. \tag{2}$$

This metric arises from a local inner product (Ambrosio et al., 2008). If $g_j \in \mathcal{D}$, $j = 1, 2$, have distribution functions G_j, the optimal transports $T_j = G_j^{-1} \circ F$ for given F reside in the tangent space of f, and for each f an inner product between T_1 and T_2 can be defined by

$$\langle T_1, T_2 \rangle_f = \int_{\mathbb{R}} \{T_1(u) - u\} \{T_2(u) - u\} \; f(u)\mathrm{d}u, \tag{3}$$

so that $d_W(f, g_j)^2 = \langle T_j, T_j \rangle_f = \langle T_j^{-1}, T_j^{-1} \rangle_{g_j}$.

2·2. *Wasserstein Mean, Variance and Covariance*

For a random density $\mathfrak{F}$ in $\mathcal{D}$, its Fréchet mean and Fréchet variance (Fréchet, 1948) are natural tools for relating the distributional properties of $\mathfrak{F}$ to the Wasserstein geometry of $\mathcal{D}$ in terms of first and second order behaviour. When the space $\mathcal{D}$ is endowed with the metric d_W, the Wasserstein–Fréchet, or simply Wasserstein, mean and variance of $\mathfrak{F}$ are

$$f_\oplus = \operatorname*{argmin}_{f \in \mathcal{D}} E\left\{d_W(\mathfrak{F}, f)^2\right\}, \quad \mathrm{var}_\oplus(\mathfrak{F}) = E\left\{d_W(\mathfrak{F}, f_\oplus)^2\right\}. \tag{4}$$

For a single density process $\mathfrak{F}$, the theoretical and practical properties of the Wasserstein mean have been thoroughly investigated (Bolstad et al., 2003; Zhang & Müller, 2011; Panaretos & Zemel, 2016; Bigot et al., 2017), and recently the Wasserstein variance was adopted to quantify variability explained when performing dimension reduction for densities (Petersen & Müller, 2016). To quantify the dependence between two random densities, we propose here the extension of these concepts to a Wasserstein covariance measure.

For a bivariate density process $(\mathfrak{F}_1, \mathfrak{F}_2)$, where the $\mathfrak{F}_j$ are random elements of $\mathcal{D}$, $j = 1, 2$, denote by $F_{\oplus,j}$ the distribution function of the Wasserstein mean of $\mathfrak{F}_j$ (4) and the corresponding density by $f_{\oplus,j}$. The random optimal transport from $f_{\oplus,j}$ to $\mathfrak{F}_j$ is $T_j = F_j^{-1} \circ F_{\oplus,j}$, where F_j is the distribution function of $\mathfrak{F}_j$, so that the Wasserstein variances are

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

This suggests defining a Wasserstein covariance measure as an expected inner product between T_1 and T_2. Since in general $f_{\oplus,1} \neq f_{\oplus,2}$, these transports reside in different tangent spaces. Adopting a common technique for manifold-valued data (Yuan et al., 2012), we push T_1 to a new transport map $\tilde{T}_1$ in the tangent space of $f_{\oplus,2}$ by a parallel transport map. Intuitively, as T_1 is transported to $\tilde{T}_1$ along the geodesic connecting $f_{\oplus,1}$ to $f_{\oplus,2}$, its angle with the geodesic is preserved. In the Wasserstein geometry, setting $T_{\oplus,12} = F_{\oplus,1}^{-1} \circ F_{\oplus,2}$, one obtains $\tilde{T}_1 = T_1 \circ T_{\oplus,12} - T_{\oplus,12} + \mathrm{id}$, where id is the identity map. Due to symmetry of this operation, if $\widetilde{T}_2 = T_2 \circ T_{\oplus,12}^{-1} - T_{\oplus,12}^{-1} + \mathrm{id}$, then $\langle \widetilde{T}_1, T_2 \rangle_{f_{\oplus,2}} = \langle T_1, \widetilde{T}_2 \rangle_{f_{\oplus,1}}$. Thus, the Wasserstein geometry motivates

$$\mathrm{cov}_\oplus(\mathfrak{F}_1, \mathfrak{F}_2) = E\left(\langle \widetilde{T}_1, T_2 \rangle_{f_{\oplus,2}}\right) = E\left(\langle T_1, \widetilde{T}_2 \rangle_{f_{\oplus,1}}\right) = \mathrm{cov}_\oplus(\mathfrak{F}_2, \mathfrak{F}_1) \tag{5}$$

as the Wasserstein covariance between the two random densities.

2·3. *Expression in Terms of Quantile Functions*

The well known fact that for one-dimensional densities, the Wasserstein geometry is closely related to quantile functions (Villani, 2003; Panaretos & Zemel, 2016; Petersen & Müller, 2016) leads to a second characterization of Wasserstein covariance, which is useful for practice. The change of variable $t = F(u)$ applied to (3) gives the alternative expression

$$\langle T_1, T_2\rangle_f = \int_0^1 \left\{F^{-1}(t) - G_1^{-1}(t)\right\}\left\{F^{-1}(t) - G_2^{-1}(t)\right\} \mathrm{d}t, \tag{6}$$

so that $d_W^2(f, g) = \int_0^1 \{F^{-1}(t) - G^{-1}(t)\}^2 \mathrm{d}t$. We can then express the quantities in (4) and (5) in terms of the random quantile functions F_j^{-1}. The Wasserstein means $f_{\oplus,j}$ are characterized by their quantile functions $F_{\oplus,j}^{-1}(t) = E\{F_j^{-1}(t)\}$, $0 \leq t \leq 1$, and for $j, k = 1, 2$, the Wasserstein variances and covariances are

$$\mathrm{cov}_\oplus(\mathfrak{F}_j, \mathfrak{F}_k) = E\left[\int_0^1 \left\{F_j^{-1}(t) - F_{\oplus,j}^{-1}(t)\right\}\left\{F_k^{-1}(t) - F_{\oplus,k}^{-1}(t)\right\} \mathrm{d}t\right]. \tag{7}$$

Expressions (6), (7) reveal a connection to functional data analysis. Viewing (F_1^{-1}, F_2^{-1}) as bivariate functional data, key objects are the mean functions $F_{\oplus,j}^{-1}(t)$ and covariance surfaces

$$\mathcal{C}_{jk}(s, t) = \mathrm{cov}\left\{F_j^{-1}(s), F_k^{-1}(t)\right\},\ 1 \leq j, k \leq 2,\ 0 \leq s, t \leq 1,$$

that characterize the first- and second-order behavior of the processes (Li & Hsing, 2010). Writing $\mathcal{M}_{jk}$ for the integral operator with kernel $\mathcal{C}_{jk}$, Fubini's theorem implies that

$$\mathrm{cov}_\oplus(\mathfrak{F}_j, \mathfrak{F}_k) = \int_0^1 \mathcal{C}_{jk}(t, t)\mathrm{d}t = \mathrm{Tr}(\mathcal{M}_{jk}),$$

where $\mathrm{Tr}(\cdot)$ is the operator trace. Accordingly, the Wasserstein variance of each component distribution can be interpreted as a summary of the variability in the quantile process, and Wasserstein covariance as a summary of covariability.

Using quantile functions has two major advantages when multiple densities are observed per subject. First, the derived notions of mean, variance, and covariance have geometric interpretations in the manifold induced by the Wasserstein metric in the space of distributions, as in (4) and (5). Second, quantile functions always have the same support $[0, 1]$ regardless of the distributional supports of the $\mathfrak{F}_j$, so that the Wasserstein covariance remains well-defined even when the latter differ. In contrast, attempts to define similar covariance summaries based on cross-covariance operators of density or compositional representations are bound to fail when distributional supports differ. For example, if $\mathcal{G}_{jk}$ is the ordinary cross-covariance operator between densities $\mathfrak{F}_j$ and $\mathfrak{F}_k$, the operator trace is well-defined only when the supports coincide. Even when they are the same, taking $\mathrm{Tr}(\mathcal{G}_{jk})$ as a summary covariance measure has no intuitive geometric meaning.

2·4. *Wasserstein Covariance Matrices and Kernels*

Considering a p-variate density process $\mathfrak{F} = (\mathfrak{F}_1, \ldots, \mathfrak{F}_p)$ with component Wasserstein means and variances $f_{\oplus,j}$ and $\mathrm{var}_\oplus(\mathfrak{F}_j)$, $j = 1, \ldots, p$, we initially assume that p is fixed, as in the brain imaging example in Section 4·1. The Wasserstein covariance matrix for $\mathfrak{F}$ is the $p \times p$ matrix with elements

$$(\Sigma_\oplus)_{jk} = \mathrm{cov}_\oplus(\mathfrak{F}_j, \mathfrak{F}_k), \tag{8}$$

which is easily seen to be a valid covariance matrix.

Motivated by the mortality example in Section 4·2, suppose the components of $\mathfrak{F}$ are indexed by a continuous variable $y_j \in [0,1]$, which might represent time. To model the setting of repeatedly observed densities that are measured densely in time, we allow $p \to \infty$. Then $\mathfrak{F}$ can be thought of as a discretized version of an unobservable process $\mathcal{F}(y)$, $0 \leq y \leq 1$, with $\mathfrak{F}_j = \mathcal{F}(y_j)$. We target the Wasserstein mean surface $f_\oplus(\cdot; y)$ and covariance kernel

$$\Sigma_\oplus(y,z) = \mathrm{cov}_\oplus \left\{ \mathcal{F}(y), \mathcal{F}(z) \right\}, \quad 0 \leq y, z \leq 1. \tag{9}$$

3. Estimation of Wasserstein Covariance Objects

3·1. *Density Estimation*

While we defined the Wasserstein covariance for fully observed densities, in reality these densities are rarely if ever directly observed. Rather, the data actually available are collections of scalar random variables W_{ijr} ($i = 1, \ldots, n$; $j = 1, \ldots p$; $r = 1, \ldots, N_{ij}$), where n is the number of subjects i, p is the number of densities or distributions j per subject, and N_{ij} is the number of independent observations r distributed according to the j-th density that are available for the i-th subject and may vary across i, j. The observed data can be viewed as resulting from two independent random mechanisms, where the first random mechanism generates the independent vectors of densities $f_i = \{f_{i1}, \ldots, f_{ip}\}$, $i = 1, \ldots, n$, while the second generates the observations that are sampled from these distributions, $W_{ijr} \sim f_{ij}$. The W_{ijr} are all independent and for each fixed (i,j) are also identically distributed.

Obtaining density estimates $\hat{f}_{ij}$ from the observed data W_{ijr} and using these as proxies for the f_{ij}, for the asymptotic analysis we need to address the challenge that these estimates are noisy and deviate from the true density targets. Since the targets $\Sigma_\oplus$ can be expressed as integrated moments of the multivariate quantile process, an obvious route would be to estimate the empirical quantile functions and proceed by averaging. However, this has some practical drawbacks and a preferred approach is to first construct a sample of smooth density estimates $\hat{f}_{ij}$, then obtaining smooth distribution functions by integration, quantile functions as inverse functions and the target quantities from the estimated quantile functions.

The theoretical analysis of the Wasserstein covariance estimates in Section 3·2 below requires the following assumption, where a preliminary density estimator is generically denoted by $\hat{f}$.

Assumption 1. There is a compact interval I such that, for any $f \in \mathcal{D}$, its support D_f is a compact interval contained in I. If $W_1, \ldots, W_N$ is a random sample from f, the density estimate $\hat{f}$ based on this sample is a probability density function on D_f such that, for some decreasing sequence $b_N = o(1)$ as $N \to \infty$,

$$\sup_{f \in \mathcal{D}} E\left\{ d_W(f, \hat{f}) \right\} = O(b_N).$$

A density estimator that satisfies Assumption 1 is described in Petersen & Müller (2016). If the support D_f is known and the condition

$$\sup_{f \in \mathcal{D}} \sup_{u \in D_f} \max\left\{ f(u), 1/f(u), |f'(u)| \right\} < \infty$$

is satisfied, then one may take $b_N = N^{-1/3}$; see Proposition 1 in the Supplementary Material. With no assumed uniform lower bound on the random densities, Panaretos & Zemel (2016) proposed a density estimator for which $b_N = N^{-1/4}$.

Lastly, because np densities need to be simultaneously estimated, each from data of varying sample sizes N_{ij}, these need to be tied to the number of independent subjects n, as follows.

Assumption 2. There exists a sequence $N = N(n)$ such that $\min_{i,j} N_{ij} \geq N$ and $N \to \infty$ as $n \to \infty$.

Here N is a uniform lower bound on the sample sizes for the np densities to be estimated that must diverge with the number of subjects n to ensure uniform consistency of the densities.

3·2. *Wasserstein Covariance Estimation*

For fixed p, given densities $f_i = (f_{i1}, \ldots, f_{ip})$, $i = 1, \ldots, n$, that are independently and identically distributed according to $\mathfrak{F} = (\mathfrak{F}_1, \ldots, \mathfrak{F}_p)$, our main goal is to estimate the Wasserstein covariance matrix $\Sigma_\oplus$, with elements defined in (8). We compute density estimates $\hat{f}_{ij}$, which are then mapped 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 $\hat{\mathcal{C}}_{jk}(s,t) = n^{-1}\sum_{i=1}^n \hat{X}_{ij}^c(s)\hat{X}_{ik}^c(t)$. To target the Wasserstein covariance $\Sigma_\oplus$, the results of Section 2·3 suggest the estimator

$$\left(\hat{\Sigma}_\oplus\right)_{jk} = \int_0^1 \hat{\mathcal{C}}_{jk}(t,t)\mathrm{d}t. \tag{10}$$

Theorem 1 demonstrates the overall rate of convergence of the Wasserstein covariance estimator, establishing asymptotic normality when N diverges sufficiently fast. While our focus is on the Wasserstein covariance, the same rate of convergence is also obtained for the full covariance surface $\hat{\mathcal{C}}_{jk}(s,t)$ as an estimator of $\mathcal{C}_{jk}(s,t)$; see Theorem 3 in the Supplementary Material, where also auxiliary results and proofs can be found.

THEOREM 1. *Suppose Assumptions* 1 *and* 2 *hold, and that* $\mathfrak{F}_j \in \mathcal{D}$ *almost surely,* $j = 1, \ldots, p$. *Then* $\|\Sigma_\oplus - \hat{\Sigma}_\oplus\|_F = O_p(n^{-1/2} + b_N)$, *where* $\|\cdot\|_F$ *denotes the Frobenius norm. Additionally, if* $n^{1/2}b_N$ *converges to zero, then there exists a zero-mean* $p \times p$ *Gaussian matrix* $\mathfrak{C}$ *such that* $n^{1/2}\left(\hat{\Sigma}_\oplus - \Sigma_\oplus\right)$ *converges weakly to* $\mathfrak{C}$.

The covariance structure of $\mathfrak{C}$ is a four-dimensional array defined in the Supplementary Material. Under regularity conditions, the density estimator in Petersen & Müller (2016) satisfies $b_N = N^{-1/3}$ so that asymptotic Gaussianity is obtained for $N = n^q$, $q > 3/2$. Faster rates of convergence $b_N = N^{-\rho}$, with $1/3 < \rho < 1/2$, can be obtained under additional smoothness conditions and for suitable density estimators, and then weaker conditions on q will suffice.

In the continuously indexed case, the vectors of densites $f_i = (f_{i1}, \ldots, f_{ip})$ correspond to discretized versions of independent and identically distributed realizations $\tilde{f}_i(\cdot\,; y)$ of a latent dynamic density surface $\mathcal{F}(y)$, $0 \leq y \leq 1$, where f_{ij} are independently distributed for different i as $\mathcal{F}(y_{ij})$. For simplicity, we require

Assumption 3. The number of observation times $p = p(n)$ satisfies $np^{-1} = O(1)$, and these are equidistant and common for all subjects, i.e., $y_{ij} = (j-1)/(p-1)$, $1 \leq j \leq p$.

We then estimate $\Sigma_\oplus(y_j, y_k)$ by the sample Wasserstein covariance kernel estimator

$$\hat{\Sigma}_\oplus(y_j, y_k) = \int_0^1 \hat{C}_{jk}(t,t)\mathrm{d}t, \tag{11}$$

followed by linearly interpolating these discretized estimates to obtain $\hat{\Sigma}_\oplus(y,z)$ for any $0 \leq y, z \leq 1$. For this interpolation to be negligible, we require two additional assumptions.

Assumption 4. There exists a constant $L_1 > 0$ such that

$$|\Sigma_\oplus(y_1, z_1) - \Sigma_\oplus(y_2, z_2)| \leq L_1(|y_1 - y_2| + |z_1 - z_2|), \quad 0 \leq y_1, y_2, z_1, z_2 \leq 1.$$

Assumption 5. With $X_i(t\,;y)$ denoting the random quantile function corresponding to $\tilde{f}_i(\cdot\,;y)$ and $X_i^c(t\,;y) = X_i(t\,;y) - E\{X_i(t\,;y)\}$,

$$\int_{[0,1]^4} \mathrm{cov}\left\{X_1^c(s\,;y)X_1^c(s\,;z), X_1^c(t\,;y)X_1^c(t\,;z)\right\} \mathrm{d}s\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z < \infty.$$

THEOREM 2. *Suppose Assumptions* 1–5 *hold and that* $\mathcal{F}(y) \in \mathcal{D}$ *almost surely. Then*

$$\int_0^1 \int_0^1 \left\{\Sigma_\oplus(y,z) - \hat{\Sigma}_\oplus(y,z)\right\}^2 \mathrm{d}y\,\mathrm{d}z = O_p(n^{-1} + b_N^2).$$

4. APPLICATIONS

4·1. *Functional Connectivity in Brain Imaging*

The study of functional connectivity in the brain involves the identification of voxels or regions which exhibit similar behaviour, as quantified by neuroimaging techniques such as electroencephalography and functional magnetic resonance imaging. Of special interest are connections that are present when subjects are in the resting state (Allen et al., 2014). Studies of connections between spatially remote regions revealed the so-called default mode network in the resting brain, including nodes of high centrality that have been characterized as functional connectivity hubs (Buckner et al., 2009). The strength of connections between neighboring voxels, as opposed to those between remote regions, contains important information related to various biological factors (Gao et al., 2016) and neurological diseases (Zalesky et al., 2012). The strength of local connections in a particular brain region has been quantified in various ways (Tomasi & Volkow, 2010; Zang et al., 2004).

Petersen & Müller (2016) demonstrated the utility of a relatively simple approach to quantifying local connectivity within functional connectivity hubs that uses probability density functions formed by smoothing histograms of pairwise temporal correlations between the signals of each voxel within a hub and the signal at its central seed voxel. We implement this approach for the $p = 10$ hubs in Table 3 of Buckner et al. (2009). The resulting 10-dimensional vectors of densities that are obtained for each subject are then analyzed with the proposed Wasserstein covariance. The ten hubs of interest are located in the left and right parietal lobes, medial superior frontal lobe, medial prefrontal lobe, left and right middle frontal lobes, posterior cingulate/precuneus region, right supramarginal lobe and the left and right middle temporal lobes.

The densities were constructed for 171 cognitively normal subjects, 65 subjects with Alzheimer's disease, and a third group of 86 subjects with mild cognitive impairment, with details about data preprocessing as in Petersen et al. (2016). The densities were then used to compute estimated Wasserstein mean densities $\hat{f}_{\oplus,j}$ for each group separately, which are displayed for each of the 10 hubs in Figure 1. The various groups show remarkable similarities in terms of the average Wasserstein connectivity distributions across these ten functional hubs. In contrast, the Wasserstein covariance and correlation matrices in Figure 2 exhibit clear differences between the groups and provide valuable additional information. Correlations are overall stronger for the normal subjects and those with mild cognitive impairment and exhibit different patterns of dependency across hubs. Complete second order behaviour can be studied through the estimated quantile covariance function estimates $\hat{\mathcal{C}}_{jk}$, where for $p = 10$, there are 45 such functions for each group. We examine the off-diagonal elements for $j < k$ by plotting slices corresponding to the three quartiles $s = 0.25, 0.5, 75$; see the Supplementary Material.

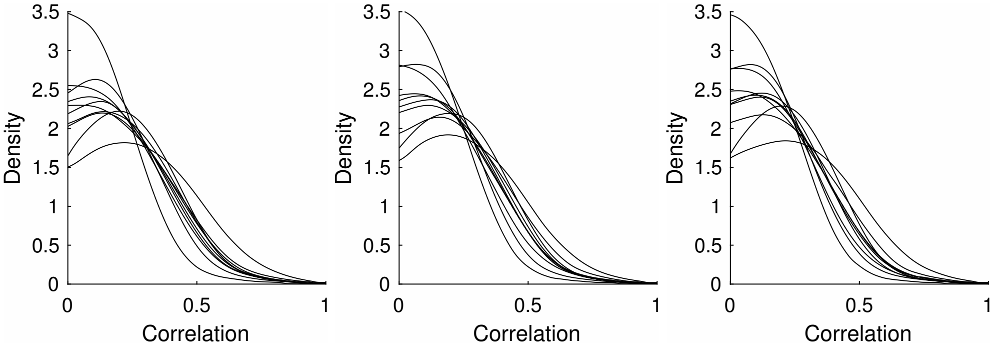


Fig. 1. Wasserstein means for ten functional connectivity hubs for normal (left), mild cognitive impairment (middle) and Alzheimer's (right) subjects, displayed as densities.

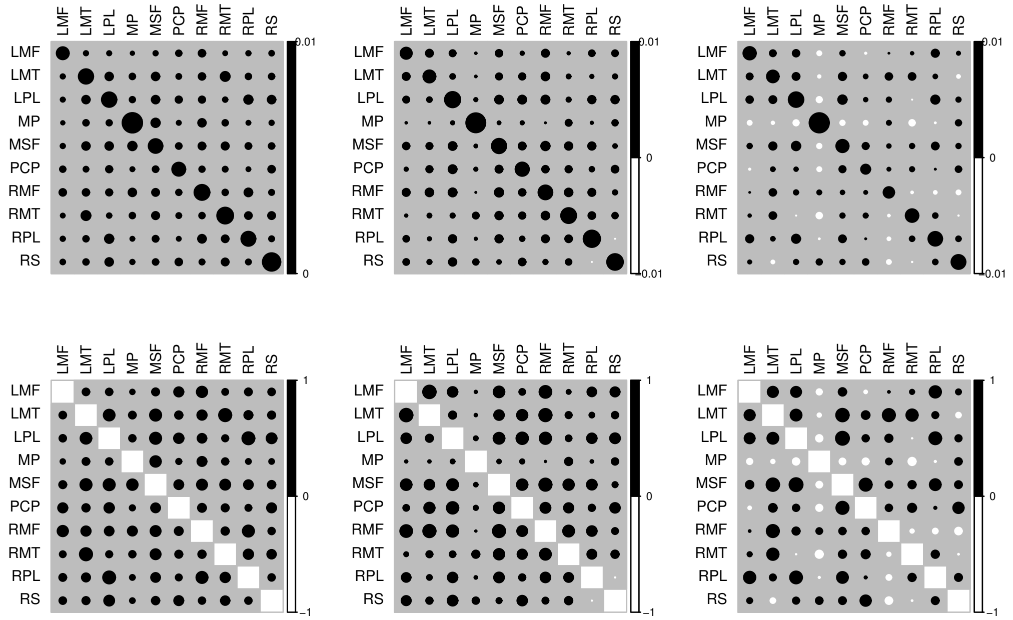


Fig. 2. Estimated Wasserstein covariance (top row) and correlation (bottom row) matrices for normal (left), mild cognitive impairment (middle) and Alzheimer's (right) subjects. Labels: LMF and RMF (left and right middle frontal), LPL and RPL (left and right parietal), LMT and RMT (left and right middle temporal), MSF (medial superior frontal), MP (medial prefrontal), PCP (posterior cingulate/precuneus) and RS (right supramarginal). Positive (negative) values are drawn in black (white) and larger circles correspond to larger absolute values.

We can also test for group differences in Wasserstein covariance using bootstrap samples obtained by centering all quantile functions with respect to their group means. To construct a bootstrap sample under the null hypothesis that all groups have the same Wasserstein covariance matrix, we simply center each multivariate quantile process at its corresponding group Wasser-

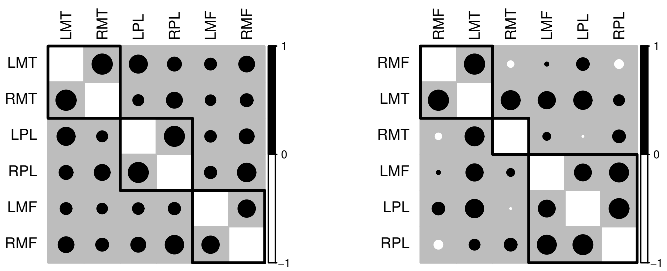


Fig. 3. Estimated Wasserstein correlation submatrices corresponding to lateral hub pairs for normal (left) and Alzheimer's (right) subjects, after reordering of hubs by hierarchical clustering using Ward's criterion. Rectangles indicate the groupings when three clusters are used. Labels correspond to those in Figure 2 and further explanations can be found in the caption of Figure 2.

stein mean. By pooling these centered processes, a bootstrap sample is obtained by sampling with replacement and then dividing into three groups of proper size. With estimates $\hat\Sigma^n_\oplus$, $\hat\Sigma^m_\oplus$ and $\hat\Sigma^a_\oplus$ for the groups of normal, mildly cognitively impaired, and Alzheimer's subjects, the test statistic for the global null hypothesis that all groups are equal is

$$S = \left\|\log\hat\Sigma^n_\oplus - \log\hat\Sigma^m_\oplus\right\|_F^2 + \left\|\log\hat\Sigma^n_\oplus - \log\hat\Sigma^a_\oplus\right\|_F^2 + \left\|\log\hat\Sigma^m_\oplus - \log\hat\Sigma^a_\oplus\right\|_F^2 .$$

Using 1000 bootstrap samples, the global p-value was found to be $p = 0.015$. An alternative test, obtained by replacing the estimated Wasserstein covariance matrices with the corresponding correlations in the above statistic, resulted in $p = 0.093$.

To further explore the differences between normal and Alzheimer's groups, the matrices were reduced to a subset of rows/columns corresponding to the lateral hub pairs in the middle frontal, middle temporal and parietal regions, then reordered using Ward's hierarchical clustering algorithm with three clusters, visualized in Figure 3. This demonstrates the presence of asymmetry in the Alzheimer's disease group, which is absent in the normal group. Derflinger et al. (2011) and others report findings of similar asymmetries in the brains of Alzheimer's patients.

One can ask how the Wasserstein covariance approach compares to established approaches in brain imaging. A commonly used measure of local connectivity is regional homogeneity (Zang et al., 2004), corresponding to Kendall's coefficient of concordance between the BOLD signals at the seed voxel and those of its immediate neighbors. This scalar measure of regional homogenity can be computed for each hub and each subject, resulting in a sample of ordinary multivariate data. The group covariance and correlation matrices are shown in the Supplementary Material. Some common patterns are seen, most notably an increased number of negative correlations for the Alzheimer's group. However, the differences are not as stark, and the regional homogeneity covariances do not reveal the asymmetry seen in the Wasserstein covariance analysis.

4·2. *Distribution of Age at Death for Period Cohorts*

To gain a better understanding of human longevity, the study of the temporal evolution of the distributions of age at death and their dependency structure over calendar time is of interest. The Human Mortality Database provides yearly mortality and population data for 38 countries at

<www.mortality.org>, which have been previously analyzed with various functional data analysis techniques (Hyndman & Shang, 2010; Chiou & Müller, 2009).

For a given country and calendar year, the probability distribution for mortality can be represented by its density. Consider a country for which life tables are available for the years $y_j \ (j = 1, \dots, p)$. For integer-valued ages $a = 0, \dots, 110$, the life table provides the size of the population m_a at least a years old, normalized so that $m_0 = 100000$. These life tables were converted to histograms of age-at-death, which we then smoothed, applying the hades package, available at <http://www.stat.ucdavis.edu/hades/>, with a smoothing bandwidth of $h = 2$. This led to estimated densities of age-at-death on the age domain $[20\text{years}, 110\text{years}]$.

To illustrate the proposed methods for continuously varying densities, we considered 32 countries and identify a subgroup of $n_E = 8$ countries located in Eastern Europe for comparison with the remaining $n_O = 24$ countries. Densities were estimated for every year between 1985 and 2005. Wasserstein means are depicted in the Supplementary Material and are found to be quite similar for the two groups, demonstrating increasing longevity over calendar time. In contrast, the estimated Wasserstein covariance and correlation surfaces of these two groups differ quite drastically, as seen in Figure 4.

Examining the diagonals of the Wasserstein covariance plots, the Eastern European countries are characterized by stagnant Wasserstein variability until 1993 when it increases sharply, followed by a steady increase between 1998 and 2005. For the non-Eastern European countries, the Wasserstein variability is maximal in 1994. Wasserstein correlations reveal that mortality dependencies are high and roughly constant over time for the non-Eastern European countries, while they are weak for Eastern European countries between years before 1990 and those after 1995.

These exploratory findings are not altogether surprising, given the economic and societal upheaval in Eastern Europe beginning around 1990, which is seen to be reflected in the Wasserstein covariance. In addition to exploratory analysis, we implemented a bootstrap test as described in the previous section. With estimated Wasserstein covariance surfaces $\hat{\Sigma}_{\oplus}^{E}$ and $\hat{\Sigma}_{\oplus}^{O}$, the test statistic was computed as the square root distance between the associated operators (Pigoli et al., 2014) and implemented by computing the Frobenius distance between the principal square roots of the discretized matrix estimates. Computing this statistic for 1000 bootstrap samples, the p-value for the difference between Wasserstein covariances of Eastern European and other countries was 0.287, while it was 0.007 using Wasserstein correlations, providing some evidence for differences in the distributions of the density processes.

5. DISCUSSION

For studying the covariance structure of vectors of random densities, the Wasserstein approach is preferred over possible alternatives such as the transformation approach (Petersen & Müller, 2016) or compositional methods based on the Aitchison geometry (Egozcue et al., 2006). This is because of its convincing practical behavior for the construction of barycenters (Bolstad et al., 2003; Zhang & Müller, 2011) and the theoretically appealing connections with optimal transport.

Specifically, the transformation approach where densities are mapped to the entire Hilbert space L^2 by means of a suitable transformation such as the log-quantile density transformation, could be applied to the components of the vectors of densities, which would lead to unconstrained multivariate functional data. For such data, any one of numerous available measures of functional covariance and correlation (Leurgans et al., 1993; Dubin & Müller, 2005; Eubank & Hsing, 2008; Yang et al., 2011) could then be harnessed. One would need to choose among many possible covariance measures and transformation maps, none of which is isometric to the Wasserstein distance. The resulting metric distortions make such an approach difficult to interpret.

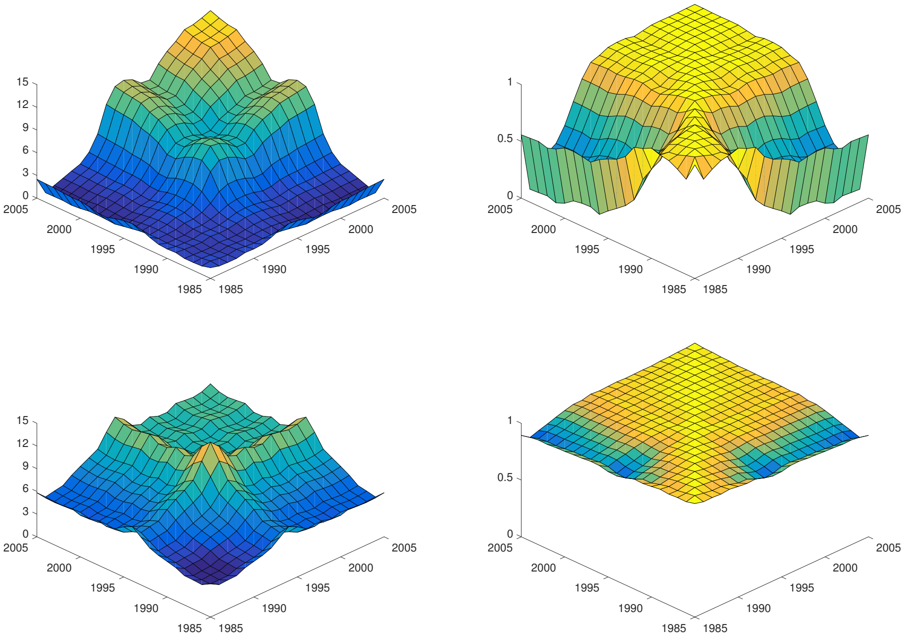


Fig. 4. Estimated Wasserstein covariance (left column) and correlation (right column) matrices for Eastern European (top row) and other (bottom row) countries.

In contrast, the proposed Wasserstein covariance has a canonical interpretation as an expected value of inner products of optimal transport maps. This means that the proposed Wasserstein covariance is similar to the notion of a regular covariance for an appropriate inner product, and can be interpreted as a measure of the degree of synchronization of the movement of probability mass from the Fréchet means to the random components of a bivariate density process. It thus emerges as a natural and compelling extension of the Wasserstein–Fréchet variance. The quantile function representation of the Wasserstein covariance in (6) facilitates the joint Wasserstein analysis of p one-dimensional distributions with Wasserstein covariance matrices and surfaces, enhancing the appeal of the proposed approach for practical applications.

While the geometric notion of covariance in (5) can be extended to the case where the sample densities have a multivariate domain, the implementation via quantile functions cannot be extended for this case, and therefore the asymptotic theory we provide remains limited to the case of one-dimensional densities. An extension to multivariate domains and analogous notions of covariance with respect to other metrics are open problems for future research.

References

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T. & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex* **24**, 663–676.

Ambrosio, L., Gigli, N. & Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Boston: Birkhäuser.

Bigot, J., Cazelles, E. & Papadakis, N. (2016). Regularization of barycenters in the Wasserstein space. *arXiv preprint arXiv:1606.01025* .

Bigot, J., Gouet, R., Klein, T. & López, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics* **53**, 1–26.

Bolstad, B. M., Irizarry, R., Åstrand, M. & Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. R., Sperling, R. A. & Johnson, K. A. (2009). Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *The Journal of Neuroscience* **29**, 1860–1873.

Chen, K., Delicado, P. & Müller, H.-G. (2017). Modeling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society, Series B (Theory and Methodology)* **79**, 177–196.

Chiou, J.-M. & Müller, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104**, 572–585.

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis* **55**, 401–420.

Derflinger, S., Sorg, C., Gaser, C., Myers, N., Arsic, M., Kurz, A., Zimmer, C., Wohlschläger, A. & Mühlau, M. (2011). Grey-matter atrophy in Alzheimer's disease is asymmetric but not lateralized. *Journal of Alzheimer's Disease* **25**, 347–357.

Dubin, J. A. & Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association* **100**, 872–881.

Egozcue, J. J., Diaz-Barrero, J. L. & Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica* **22**, 1175–1182.

Eubank, R. L. & Hsing, T. (2008). Canonical correlation for stochastic processes. *Stochastic Processes and their Applications* **118**, 1634–1661.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré* **10**, 215–310.

Gao, Q., Xu, F., Jiang, C., Chen, Z., Chen, H., Liao, H. & Zhao, L. (2016). Decreased functional connectivity density in pain-related brain regions of female migraine patients without aura. *Brain Research* **1632**, 73–81.

Hron, K., Menafoglio, A., Templ, M., Hruzová, K. & Filzmoser, P. (2016). Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis* **94**, 330–350.

Hyndman, R. J. & Shang, H. L. (2010). Rainbow plots, bagplots and boxplots for functional data. *Journal of Computational and Graphical Statistics* **19**, 29–45.

Kneip, A. & Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96**, 519–542.

Leurgans, S. E., Moyeed, R. A. & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society: Series B* **55**, 725–740.

Li, Y. & Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics* **38**, 3321–3351.

Panaretos, V. M. & Zemel, Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics* **44**, 771–812.

Park, S. Y. & Staicu, A.-M. (2015). Longitudinal functional data analysis. *Stat* **4**, 212–226.

Petersen, A. & Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics* **44**, 183–218.

Petersen, A., Zhao, J., Carmichael, O. & Müller, H.-G. (2016). Quantifying individual brain connectivity with functional principal component analysis for networks. *Brain Connectivity* **6**, 540–547.

Pigoli, D., Aston, J. A., Dryden, I. L. & Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika* **101**, 409–422.

Tomasi, D. & Volkow, N. D. (2010). Functional connectivity density mapping. *Proceedings of the National Academy of Sciences* **107**, 9885–9890.

Villani, C. (2003). *Topics in Optimal Transportation*. Providence, Rhode Island: American Mathematical Society.

Yang, W., Müller, H.-G. & Stadtmüller, U. (2011). Functional singular component analysis. *Journal of the Royal Statistical Society: Series B* **73**, 303–324.

Yuan, Y., Zhu, H., Lin, W. & Marron, J. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 697–719.

Zalesky, A., Fornito, A., Egan, G. F., Pantelis, C. & Bullmore, E. T. (2012). The relationship between regional and inter-regional functional connectivity deficits in Schizophrenia. *Human Brain Mapping* **33**, 2535–2549.

Zang, Y., Jiang, T., Lu, Y., He, Y. & Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *Neuroimage* **22**, 394–400.

Zhang, Z. & Müller, H.-G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis* **55**, 2234–2249.

Biometrika (2018), **xx**, x, *pp*. 1–6
Printed in Great Britain

Supplementary Material to "Wasserstein Covariance for Multiple Random Densities"

BY ALEXANDER PETERSEN

Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106, U.S.A.

petersen@pstat.ucsb.edu

AND HANS-GEORG MÜLLER

Department of Statistics, University of California, Davis, California 95616, U.S.A.

hgmueller@ucdavis.edu

SUMMARY

Section S.1 demonstrates a valid density estimator with the properties required by Assumption 1 in the paper. Section S.2 contains proofs of Theorems 1 and 2, while Section S.3 contains an additional theoretical result. Section S.4 provides additional visualizations for the connectivity and mortality analyses.

S.1. MODIFIED KERNEL DENSITY ESTIMATOR

A density estimator $\hat{f}$ that satisfies Assumption 1 is defined as follows. Assume one has an i.i.d. sample $W_1, \ldots, W_N \sim f$. Let K be a bounded pdf on $[-1, 1]$, symmetric about 0. We use an adapted version of the modified kernel density estimator (Petersen and Müller 2016)

$$\hat{f}(u) = \sum_{r=1}^{m} K\left(\frac{W_r - u}{h}\right) w(u,h) \Bigg/ \left\{\sum_{r=1}^{m} \int_{D_f} K\left(\frac{W_r - v}{h}\right) w(v,h)\mathrm{d}v\right\}, \qquad (1)$$

for $u \in D_f = [a_f, b_f]$ and 0 elsewhere. The weight function

$$w(u,h) = \begin{cases} \left\{\int_{(a_f-u)h^{-1}}^{1} K(v)\mathrm{d}v\right\}^{-1}, & a_f \leq u \leq a_f + h, \\ \left\{\int_{-1}^{(b_f-u)h^{-1}} K(v)\mathrm{d}v\right\}^{-1}, & b_f - h \leq u \leq b_f, \\ 1, & \text{otherwise}, \end{cases}$$

is designed to remove the boundary bias of the standard kernel density estimator, which is a major issue due to the compact support of he densities, while simultaneously maintaining the integrated squared error rate and rendering an estimate that is nonnegative and integrates to one. This leads to the following proposition.

PROPOSITION 1. *With $\mathcal{D}$ as in Assumption 1, assume further that*

$$\sup_{f \in \mathcal{D}} \sup_{u \in D_f} \max\left\{f(u), 1/f(u), |f'(u)|\right\} < M < \infty.$$

For generic $f \in \mathcal{D}$, given an i.i.d. sample $W_r \sim f$, $(r = 1, \ldots, N)$ and assuming D_f is known, then $\hat{f}$ given in (1) *satisfies Assumption 1 with $b_N = N^{-1/3}$, using a bandwidth $h = tN^{-1/3}$ for any $t > 0$.*

Proof. Proposition 1 of Petersen and Müller (2016) immediately implies that

$$\sup_{f \in \mathcal{D}} E\left\{d_2(\hat{f}, f)\right\} = O\left(N^{-1/3}\right)$$

under the given assumptions, where d_2 is the $L^2(I)$ metric and I is the interval in Assumption 1.

Choose $C > 0$ such that $I \subset [-C/2, C/2]$ and let F and $\hat{F}$ be the cdfs of f and $\hat{f}$. It is clear that $d_2(F, \hat{F}) \leq C d_2(f, \hat{f})$, and that $(F^{-1})'$ is uniformly bounded by M over $f \in \mathcal{D}$. By the mean value theorem, and using the change of variables $u = \hat{F}^{-1}(t)$,

$$\begin{aligned} d_W(f, \hat{f})^2 &= \int_0^1 \left\{F^{-1}(t) - \hat{F}^{-1}(t)\right\}^2 \mathrm{d}t \\ &\leq M^2 \left[\int_{D_f} \left\{F(u) - \hat{F}(u)\right\}^2 f(u)\mathrm{d}u \right.\\ &\qquad \left. + \int_{D_f} \left\{F(u) - \hat{F}(u)\right\}^2 \left\{\hat{f}(u) - f(u)\right\} \mathrm{d}u\right] \\ &\leq M^2 C(MC + 1) d_2(f, \hat{f})^2 \end{aligned}$$

by Cauchy-Schwarz, and the result follows. □

S.2. PROOFS OF THEOREMS 1 AND 2

Throughout the proofs, we use the notations $\langle f, g\rangle = \int_0^1 f(u)g(u)\mathrm{d}u$ for the standard inner product on $L^2[0,1]$, $\|f\| = (\langle f, f\rangle)^{1/2}$ for the corresponding L^2 norm and $\|\cdot\|_{[0,1]^2}$ for the standard $L^2([0,1]^2)$ norm. Let $X_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^\top$ be the vector of quantile functions for subject i, and set $\nu(t) = E\{X_1(t)\}$, $\tilde{\nu}(t) = n^{-1}\sum_{i=1}^n X_i(t)$, and $X_i^c(t) = X_i(t) - \nu(t)$. Letting $U_i(s,t) = X_i^c(s)X_i^c(t)^\top$ and $T_n(s,t) = \{\tilde{\nu}(s) - \nu(s)\}\{\tilde{\nu}(t) - \nu(t)\}^\top$, set

$$\tilde{\mathcal{C}}_{jk}(s,t) = \frac{1}{n}\sum_{i=1}^n \{U_i(s,t)\}_{jk} - \{T_n(s,t)\}_{jk}, \quad 0 \leq s, t \leq 1.$$

Proof of Theorem 1.

Define

$$\tilde{\Sigma}_\oplus = \int_0^1 \left\{\frac{1}{n}\sum_{i=1}^n U_i(t,t) - T_n(t,t)\right\} \mathrm{d}t$$

and, for $0 \leq s, t \leq 1$, define the four dimensional array $D(s,t)$ with elements

$$\{D(s,t)\}_{jklm} = \mathrm{cov}\left[\{U_i(s,s)\}_{jk}, \{U_i(t,t)\}_{lm}\right].$$

Since $E\left[\{U_i(t,t)\}_{jk}\right] = \mathcal{C}_{jk}(t,t)$, if $R_i = \int_0^1 U_i(t,t)\mathrm{d}t$, then $E(R_i) = \Sigma_\oplus$. With $\tau_n = \int_0^1 T_n(t,t)\mathrm{d}t$,

$$\tilde{\Sigma}_\oplus - \Sigma_\oplus = \left(\frac{1}{n}\sum_{i=1}^n R_i\right) - \Sigma_\oplus - \tau_n. \tag{2}$$

Let $\mathcal{S}_{jklm} = \mathrm{cov}\{(R_i)_{jk}, (R_i)_{lm}\} = \int_0^1\int_0^1 \{D(s,t)\}_{jklm}\mathrm{d}s\,\mathrm{d}t$, so that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n R_i - \Sigma_\oplus\right) \to \mathfrak{C} \quad \text{weakly},$$

where $\mathfrak{C}$ is a zero-mean $p \times p$ Gaussian matrix with covariance structure $\mathcal{S}$. Next, with $\|\cdot\|_F$ denoting the Frobenius norm, it is easy to show that $\|\tau_n\|_F \leq \sum_{j=1}^p \|\hat{\nu}_j - \nu_j\|^2$. Since each summand is $O_p(n^{-1})$, this proves that $\tau_n = o_p(n^{-1/2})$. Hence,

$$\sqrt{n}(\tilde{\Sigma}_\oplus - \Sigma_\oplus) \to \mathfrak{C} \quad \text{weakly}.$$

Next, due to Assumption 1, the functions X_{ij} and $\hat{X}_{ij}$ are uniformly bounded for all i and j, and this bound clearly carries over for the quantile averages $\tilde{\nu}_j = n^{-1}\sum_{i=1}^n X_{ij}$ and $\hat{\nu}_j = n^{-1}\sum_{i=1}^n \hat{X}_{ij}$. Hence, for some positive constant C,

$$\begin{aligned}
\left|\left(\hat{\Sigma}_\oplus\right)_{jk} - \left(\tilde{\Sigma}_\oplus\right)_{jk}\right| &= \left|\frac{1}{n}\sum_{i=1}^n \Big(\langle \hat{X}_{ij} - \hat{\nu}_j, \hat{X}_{ik} - \hat{\nu}_k\rangle - \langle X_{ij} - \tilde{\nu}_j, X_{ik} - \tilde{\nu}_k\rangle\Big)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \Big(\langle \hat{X}_{ij} - \hat{\nu}_j, \hat{X}_{ik} - X_{ik}\rangle + \langle \hat{X}_{ij} - X_{ij}, X_{ik} - \tilde{\nu}_k\rangle\Big)\right| \\
&\leq \frac{1}{n}\left(\sum_{i=1}^n \left|\langle \hat{X}_{ij} - \hat{\nu}_j, \hat{X}_{ik} - X_{ik}\rangle\right| + \sum_{i=1}^n \left|\langle \hat{X}_{ij} - X_{ij}, X_{ik} - \tilde{\nu}_k\rangle\right|\right) \\
&\leq \frac{C}{n}\left(\sum_{i=1}^n \|\hat{X}_{ik} - X_{ik}\| + \sum_{i=1}^n \|\hat{X}_{ij} - X_{ij}\|\right) \\
&= \frac{1}{n}\left\{\sum_{i=1}^n d_W(\hat{f}_{ik}, f_{ik}) + \sum_{i=1}^n d_W(\hat{f}_{ij}, f_{ij})\right\}.
\end{aligned}$$

The result then follows from Assumptions 1 and 2.

Proof of Theorem 2.

Adopting the same notations as in the proof of Theorem 1, let $R_{ijk} = (R_i)_{jk}$ and

$$\tilde{\Sigma}_\oplus(y_j, y_k) = n^{-1}\sum_{i=1}^n R_{ijk} - \tau_{njk},$$

where $\tau_{njk} = \int_0^1 T_{njk}(t,t)\mathrm{d}t$ can be shown to satisfy $p^{-2}\sum_{j,k=1}^{p}\tau_{njk}^2 = O_p(n^{-2})$. Using Assumptions 3 and 4, we immediately obtain

$$\begin{aligned}\left\|\Sigma_\oplus - \tilde{\Sigma}_\oplus\right\|_{[0,1]^2}^2 &= O\left[n^{-1} + \frac{1}{p^2}\sum_{j,k=1}^{p}\left\{\Sigma_\oplus(y_j,y_k) - \tilde{\Sigma}_\oplus(y_j,y_k)\right\}^2\right] \\ &= O\left[n^{-1} + \frac{1}{p^2}\sum_{j,k=1}^{p}\left\{\Sigma_\oplus(y_j,y_k) - n^{-1}\sum_{i=1}^{n}R_{ijk}\right\}^2\right].\end{aligned}$$

Observing $E(R_{ijk}) = \Sigma_\oplus(y_j,y_k)$ and $\mathrm{var}(R_{ijk}) = G(y_j,y_k)$, where

$$G(y,z) = \int_{[0,1]^2}\mathrm{cov}\left\{X_1^c(s;y)X_1^c(s;z), X_1^c(t;y)X_1^c(t;z)\right\}\mathrm{d}s\mathrm{d}t,$$

Assumption 5 implies

$$E\left[\frac{1}{p^2}\sum_{j,k=1}^{p}\left\{\Sigma_\oplus(y_j,y_k) - n^{-1}\sum_{i=1}^{n}R_{ijk}\right\}^2\right] = \frac{1}{np^2}\sum_{j,k=1}^{p}G(y_j,y_k) = O(n^{-1}),$$

so that $\left\|\Sigma_\oplus - \tilde{\Sigma}_\oplus\right\|_{[0,1]^2}^2 = O_p(n^{-1})$, as claimed.

Finally,

$$\max_{j,k=1,\ldots,p}\left\{\tilde{\Sigma}_\oplus(y_j,y_k) - \hat{\Sigma}_\oplus(y_j,y_k)\right\} = O_p(b_N)$$

by Assumptions 1 and 2, so that

$$\left\|\tilde{\Sigma}_\oplus - \hat{\Sigma}_\oplus\right\|_{[0,1]^2}^2 = O_p(b_N^2),$$

completing the proof.

S.3. THEOREM 3

THEOREM 3. *Under Assumptions 1 and 2, the covariance function estimates satisfy*

$$\int_0^1\int_0^1\left\{\hat{\mathcal{C}}_{jk}(s,t) - \mathcal{C}_{jk}(s,t)\right\}^2\mathrm{d}s\,\mathrm{d}t = O_p(n^{-1} + b_N^2).$$

Proof of Theorem 3.

Write $U_{ijk}(s,t) = \{U_i(s,t)\}_{jk}$. Since $E\{U_{ijk}(s,t)\} = \mathcal{C}_{jk}(s,t)$, Assumption 1 implies $E\left(\|X_{1j}\|^4\right) < \infty$, whence

$$E\left(\left\|\mathcal{C}_{jk} - \frac{1}{n}\sum_{i=1}^{n}U_{ijk}\right\|_{[0,1]^2}^2\right) = O(n^{-1}).$$

Furthemore, with $T_{njk}(s,t) = \{T_n(s,t)\}_{jk}$,

$$\left\|T_{njk}\right\|_{[0,1]^2}^2 = \|\tilde{\nu}_j - \nu_j\|^2\|\tilde{\nu}_k - \nu_k\|^2 = O_p(n^{-2}),$$

so that $\|\mathcal{C}_{jk} - \tilde{\mathcal{C}}_{jk}\|^2_{[0,1]^2} = O_p(n^{-1})$. Lastly, under Assumption 1, there exists a constant C such that

$$\left\|\tilde{\mathcal{C}}_{jk} - \hat{\mathcal{C}}_{jk}\right\|^2_{[0,1]^2} \leq \frac{2C}{n}\left\{\sum_{i=1}^{n} d_W(f_{ij}, \hat{f}_{ij})^2 + \sum_{i=1}^{n} d_W(f_{ik}, \hat{f}_{ik})^2\right\} = O_p(b_N^2)$$

by applying Assumption 2, which proves the result.

S.4. Additional Figures

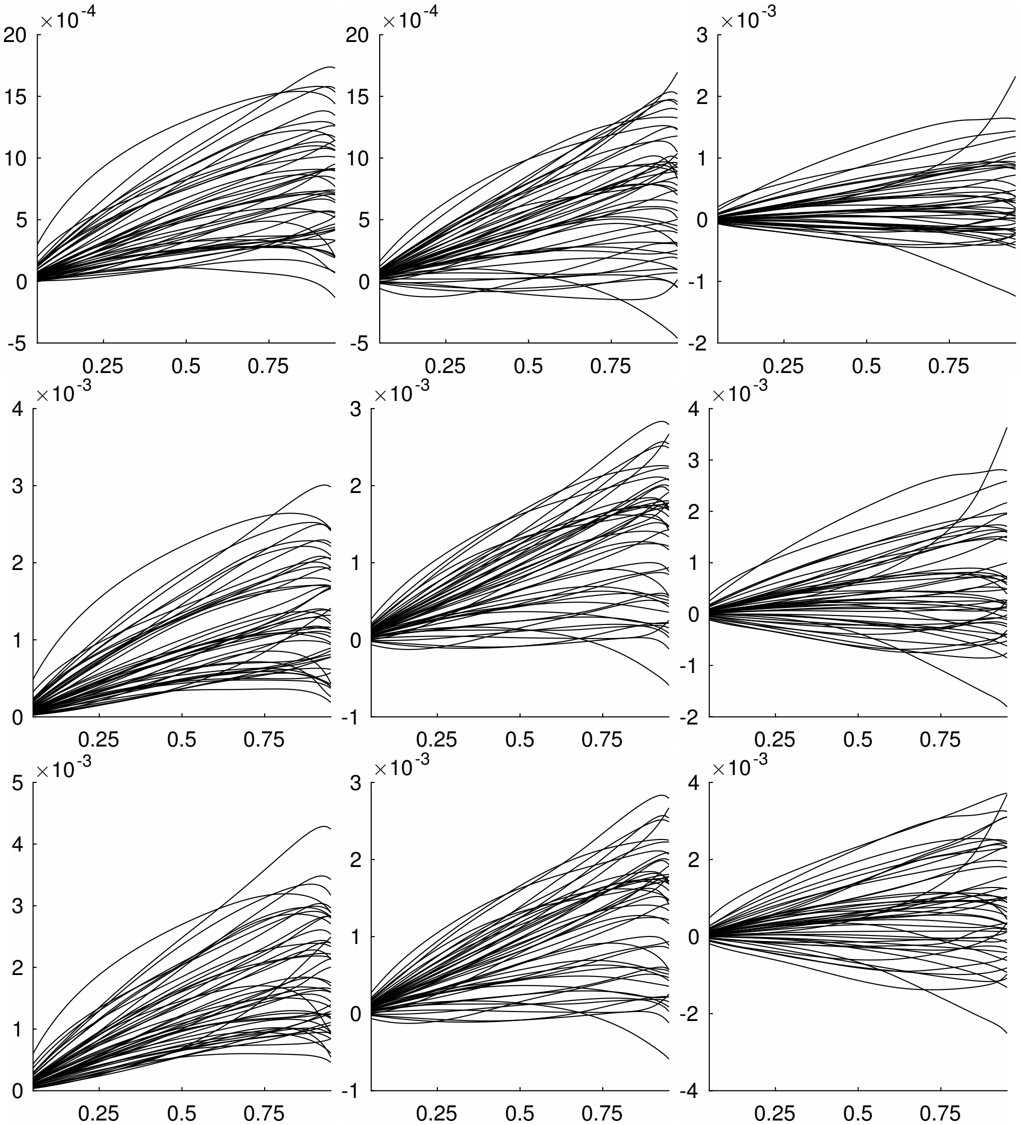


Fig. 5. Slices of covariance surface estimates $\hat{\mathcal{C}}_{jk}(s^*, t)$, $1 \leq j < k \leq 10$, for $s^* = 0.25, 0.5, 0.75$, corresponding to top, middle, and bottom rows, respectively. The left, middle, and right columns correspond to normal, mild cognitive impairment, and Alzheimer's subjects.

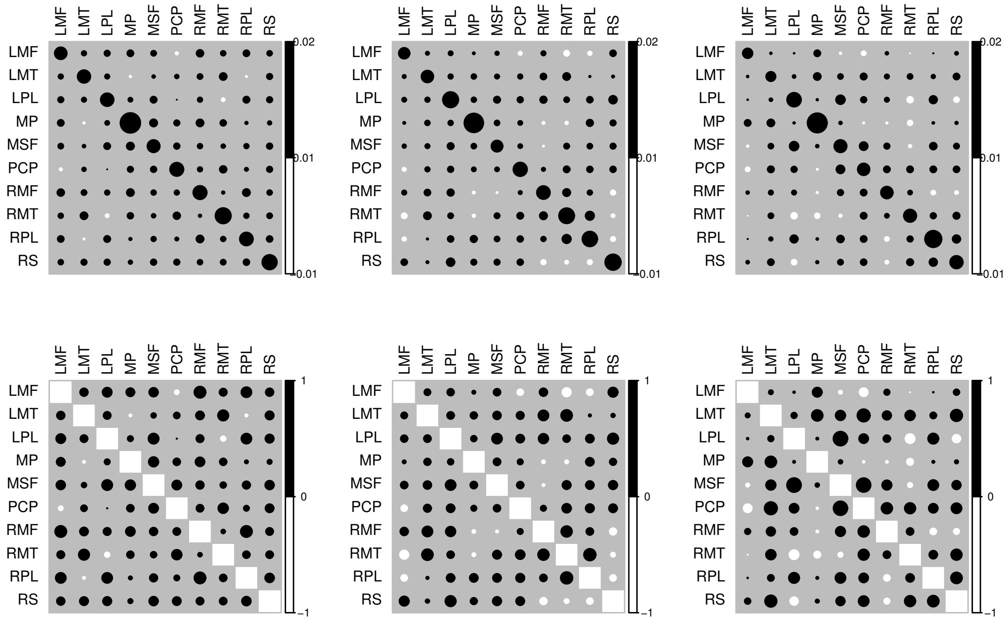


Fig. 6. Estimated covariance (top row) and correlation (bottom row) matrices of regional homogeneity scores for normal (left), mild cognitive impairment (middle) and Alzheimer's (right) subjects. Labels: LMF and RMF (left and right middle frontal), LPL and RPL (left and right parietal), LMT and RMT (left and right middle temporal), MSF (medial superior frontal), MP (medial prefrontal), PCP (posterior cingulate/precuneus) and RS (right supramarginal). Positive (negative) values are drawn in black (white) and larger circles correspond to larger absolute values.

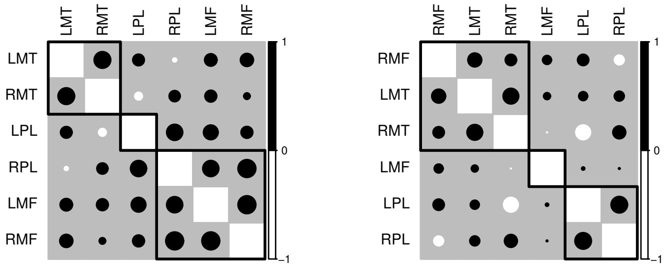


Fig. 7. Estimated regional homogeneity correlation submatrices corresponding to lateral hub pairs for normal (left) and Alzheimer's (right) subjects, after reordering of hubs by hierarchical clustering using Ward's criterion. Rectangles indicate the groupings when three clusters are used. Labels correspond to those in Figure 6 and further explanations can be found in the caption of Figure 6.

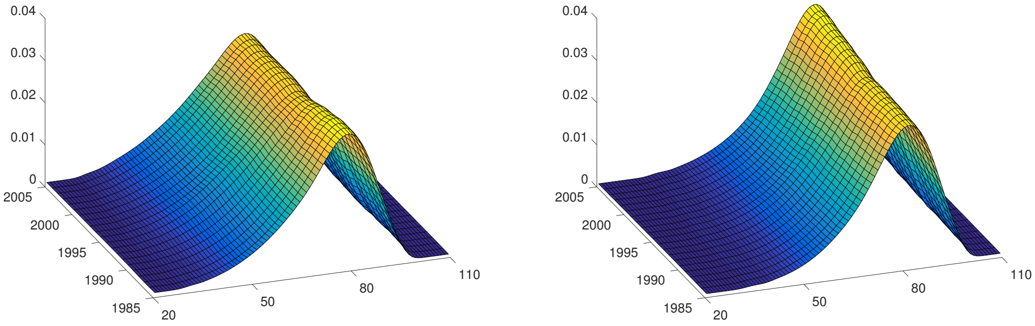


Fig. 8. Wasserstein mean density surfaces for mortality between 1980 and 2005, for Eastern European (left) and other (right) countries.