Received: 10 September 2019 | Accepted: 21 May 2022 | Published on: 16 November 2022

DOI: 10.1111/rssb.12543

ORIGINAL ARTICLE

Empirical likelihood-based inference for functional means with application to wearable device data

Hsin-wen Chang[1] | **Ian W. McKeague**[2]

[1]Academia Sinica, Taipei, Taiwan

[2]Columbia University, New York, New York, USA

Correspondence
Hsin-wen Chang, Institute of Statistical Science, Academia Sinica, R5017, Environmental Changes Research Building, No.128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan.
Email: hwchang@stat.sinica.edu.tw

Abstract

This paper develops a nonparametric inference framework that is applicable to occupation time curves derived from wearable device data. These curves consider all activity levels within the range of device readings, which is preferable to the practice of classifying activity into discrete categories. Motivated by certain features of these curves, we introduce a powerful likelihood ratio approach to construct confidence bands and compare functional means. Notably, our approach allows discontinuities in the functional covariances while accommodating discretization of the observed trajectories. A simulation study shows that the proposed procedures outperform competing functional data procedures. We illustrate the proposed methods using wearable device data from an NHANES study.

KEYWORDS

accelerometry, bootstrap, functional data analysis, nonparametric likelihood ratio, occupation measures

1 | INTRODUCTION

The motivation for this paper comes from applications of physiological monitoring in which there is a need to compare groups of subjects in terms of health outcomes that are functional in nature. Inexpensive wearable sensors are now capable of generating massive amounts of data collected longitudinally (for weeks or months at a time), and they are playing an increasingly important role in epidemiological, public health and biomedical research; for example, in studies of congestive heart failure, pulmonary disease, diabetes, obesity and Alzheimer's disease. Various functional data methods have been proposed to analyse such data (see, e.g. Backenroth et al., 2020; Zhang, Li, et al., 2019, and references therein), and to our knowledge all of these methods implemented model-based and smoothing approaches. However, such approaches can result in loss of information from smoothing/dimension reduction (Dette et al., 2020), for example when faced with discontinuities in occupation time data (to be described later). Inspired by this problem, we propose a powerful nonparametric approach that does not require smoothing, for comparing functional means and constructing confidence bands. A distinct feature of our approach is in allowing discontinuities in the functional means *and* covariances while accounting for dense discretization of the observed trajectories, in contrast to the existing literature (see, e.g. Cao et al., 2012; Choi & Reimherr, 2018; Degras, 2011; Zhang, Cheng, et al., 2019) concerning dense functional data analysis.

Our application focuses on wearable device measurements used to assess physical activity, which is of great interest in human physiology and pathophysiology research (Wright et al., 2017). Physical activity is often quantified by time spent in activities of various intensities (e.g. sedentary, light, moderate, vigorous) during the study period (see, e.g. Staudenmayer et al., 2012). The amount and intensity of physical activity is typically determined from accelerometer readings (Migueles et al., 2017), for example, 'counts' from ActiGraph devices, as shown in Figure 1. Thresholds for the readings are often used to specify the various activity categories. For example, the time spent in sedentary behaviour could be represented by the accumulated amount of time below 100 counts/min (Matthews et al., 2008). Such thresholds, however, are arbitrary in the absence of separate validation studies.

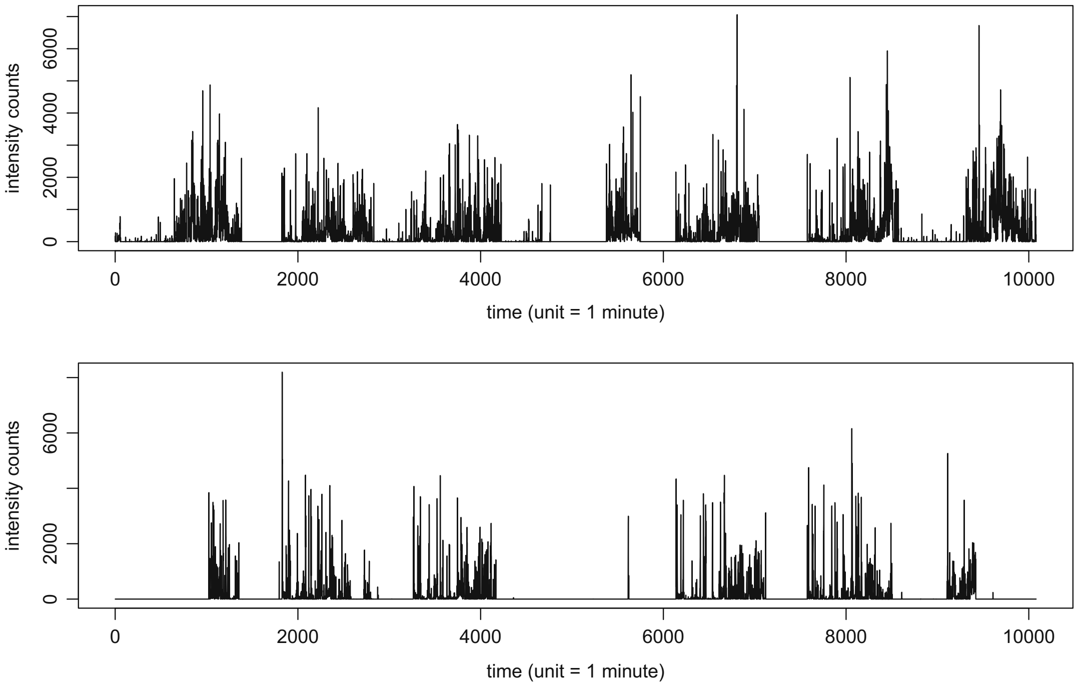

FIGURE 1 ActiGraph raw activity data during 1 week for two subjects in the 2005–2006 National Health and Nutrition Examination Survey (NHANES).

To deal with this issue, instead of just measuring time spent in discrete activity categories, we consider activity over a continuum of levels. This is done in terms of an *occupation time*: the total amount of time spent above an activity level as a function of that level over the range of sensor readings. This coincides with occupation time measures in the theory of stochastic processes (Samorodnitsky, 2016). Such measures can readily be used to obtain the time spent in activities of various intensities that is of interest in the physical activity literature (see, e.g. Matthews et al., 2008; Staudenmayer et al., 2012). To avoid confusing occupation time data with the original accelerometer readings as shown in Figure 1, we will refer to the latter as 'raw activity curves' or 'raw activity data'.

Occupation time curves constructed from the raw activity data will be viewed as non-increasing piecewise-constant functions with discontinuities on a (fixed) common grid having the resolution of the measurement unit of the device. These features will be exploited for more efficient inference than what can be provided by existing functional data methods. In particular, in the absence of substantial measurement errors, it is not necessary to smooth these curves before the proposed statistical analysis (see Section 2.1 for more details). This 'non-smoothing' approach is commonly implemented when the observed curves are available on a common grid of points (Górecki & Smaga, 2018; Zhang, Cheng, et al., 2019). It can allow discontinuities in the functional moments to be captured accurately; see Figure 2 for a simulated data example in which the band based on non-smoothing is accurate whereas the one based on smoothing is misleading in the neighbourhood of the discontinuity. On the other hand, an attractive feature of the occupation time curves is that they are automatically aligned on the grid of activity levels, as opposed to the raw activity curves being indexed by the follow-up time. The latter case needs curve alignment/registration methods (Wrobel et al., 2019) to deal with the difference between the chronological follow-up time and the internal time on which major features are aligned across the curves; see Figure 1 for an example of this difference, where the subject in the top panel has peak activity later in the day than the subject below.

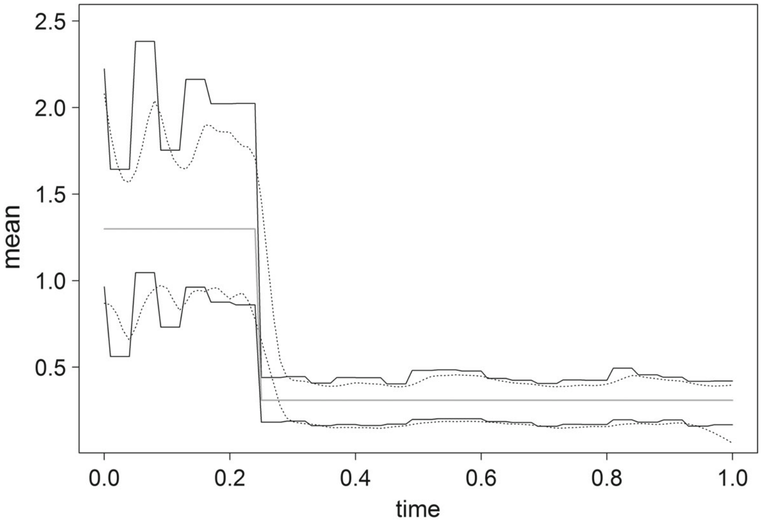

FIGURE 2 95% simultaneous confidence bands based on smoothing (dotted black) and non-smoothing (solid black) for a functional mean (solid dark grey) with one discontinuity at 0.25.

To justify our non-smoothing approach, we provide a novel theoretical framework that accommodates discretization of the observed trajectories without the need for smoothing. Viewing the discretized data as step functions obviates the need for further regularization. Moreover, the conditions we need are weaker than existing conditions imposed when using smoothing to estimate functional means under a fixed dense design (Cao et al., 2012; Degras, 2011). This allows us to deal with a wider range of functional data (beyond occupation time), such as curves with non-smooth latent mean and covariance functions (see Section 3.1 for an illustration). Cai and Yuan (2011) studied discretely observed functional data using spline interpolation (rather than piecewise-constant interpolation) to reconstruct the data between fixed common design points, and showed that the minimax convergence rate (for the mean) does not improve from further smoothing. Other non-smoothing methods for dense functional data include *linear* interpolation for two-sample tests (Yuan et al., 2020) and direct discretization for ANOVA tests (Zhang, Cheng, et al., 2019). However, these methods require the assumption of smooth covariance functions, which is not needed in our approach.

Under this discretization framework, we develop an empirical likelihood (EL) approach to nonparametric inference for functional means. We choose to utilize EL due to its proven optimality properties (see Remark 1 after Theorem 2). Moreover, EL confidence bands respect range and monotonicity constraints (see Section 2.3). Thus EL is well suited to the task of analysing occupation time curves, which are bounded and non-increasing, as mentioned previously. Specifically, we introduce an EL-based simultaneous confidence band for the functional mean, and an EL-based functional ANOVA test. Most of the literature on functional data analysis that is applicable in our fixed dense design setting focuses on less-than-optimal Wald-type procedures that are not range-respecting nor monotonicity-preserving, for example, confidence bands for the mean based on local linear smoothing (Degras, 2011, 2017), polynomial splines (Cao et al., 2012) and geometry (Choi & Reimherr, 2018), two-sample problems (see, e.g. Fan & Lin, 1998; Yuan et al., 2020; Zhang, 2013, and references therein) and Wald-type ANOVA tests (see, e.g. Cuevas et al., 2004; Górecki & Smaga, 2015, 2018; Zhang, Cheng, et al., 2019, and references therein). The use of EL in functional data analysis is a fairly recent development. Sang et al. (2019) constructed EL-based confidence intervals for dynamical correlation, and Wang et al. (2018) proposed EL-based tests in a concurrent linear model for functional data. However, both of these papers used local linear smoothing and assumed continuity of at least the second derivatives of the target functions. In contrast, we do not require smoothing, and we allow discontinuities in the target functions.

Our approach is developed for functional data in the space of functions of bounded variation (equipped with sup-norm), in contrast to $L^2[0,1]$ or $C[0,1]$ used in most functional data analysis literature (Cuevas, 2014). This is general enough to handle occupation time curves that are bounded and monotonic, but not necessarily continuous, and is well matched with our sup-norm formulation used for the EL-based confidence band and ANOVA test. The need to choose the function space (and metric) in a way that is compatible with the key sample-path properties of the functional data has been emphasized in a recent paper by Dette et al. (2020), who promote the use of $C[0,1]$ equipped with the sup-norm for analysing continuous functional data instead of the $L^2[0,1]$ framework. Note that the Lipschitz continuity and moment assumptions made in Dette et al. (2020) are stronger than our assumptions (see Section 2.2). An example of non-monotonic and discontinuous functional data of bounded variation is the extent covered by Arctic sea ice as a function of time (Witze, 2019), although not pursued in the sequel. Sea ice extent is calculated in terms of the number of ice-covered pixels in a satellite image (NASA, 2016), pixels being 'ice-covered' if the percentage of sea ice is no less than some threshold (commonly set at 15%).

Sea ice extent changes when there is a crossing of the threshold in any pixel, and thus it is a step function over time. Furthermore, this function is of bounded variation because sea ice reacts slowly to changes in temperature over short time scales.

The paper is organized as follows. In Section 2.1, we first present our application and show how to construct occupation time curves from raw activity data. Certain features of these curves motivate the general theory: the empirical estimator of the functional mean is given in Section 2.2, the EL-based confidence band in Section 2.3, and EL-based ANOVA test in Section 2.4. Section 3 presents simulation results showing that the proposed procedures outperform competing Wald-type procedures. In Section 4 we analyse the NHANES data. A discussion is given in Section 5. Proofs are presented in the Supplement. R code for implementing all the methods and for reproducing the results in Sections 3 and 4 can be obtained from https://github.com/news11/paper_fdEL.

2 | INFERENCE FOR FUNCTIONAL DATA OF BOUNDED VARIATION

We introduce the application to occupation time data in Section 2.1. In the subsequent sections we develop an inferential framework that can be applied to general functional data of bounded variation.

2.1 | Occupation time

Occupation time is defined as the amount of time a stochastic process spends in a given set (e.g. above some level a). Let $X = \{X(t), t \in [0, \tau]\}$ be a measurable stochastic process representing the fully observed trajectory of a raw activity curve (on a given subject). The (occupation) time that X spends above the level a is given by $L(a) = \mathrm{Leb}(\{t \in [0, \tau]: X(t) > a\})$, where Leb denotes Lebesgue measure on the real line. Here the index a varies over the range of activity levels of interest. As a function of a, $L(a)$ is bounded between 0 and τ, monotonically decreasing, and right-continuous, by definition. A flat in $X(t)$ can result in a left-discontinuity in $L(a)$, and hence its moments; in our data example, the number of flats per individual can be as high as 148 min/week. The boundedness and monotonicity can be seen from Figure 3, which illustrates the occupation time curves of the two subjects from Figure 1 with $\tau = 168$ h (10,080 min). The curves are smooth apart from hidden jump discontinuities, instead of having noisy fluctuations commonly seen in functional data.

Regarding observability, note that $L(a)$ is merely a transformation of the original data $X(t)$ via a known map, and thus if $X(t)$ is directly observable, so is $L(a)$. Instead of fully observed trajectories, both $L(a)$ and $X(t)$ will be observed on grids, with the grid of $L(a)$ having the resolution of the measurement unit of the wearable device. Such domain discretization will be incorporated into our theory, to be described in Section 2.2. Note that the discretization of the domain of $X(t)$ becomes a discretization in the range of $L(a)$, which is typically ignored in functional data analysis.

In the rest of Section 2, we introduce inference methods in terms of a more general stochastic process $T(a)$ that can be applied to these occupation time curves. Our framework does not smooth the observed discretized processes, due to the following reasons. First, smoothing is frequently implemented as a way to pool information from neighbourhood indices on the grid where the functional data are observed (Wang et al., 2016), but we find a way to pool information via

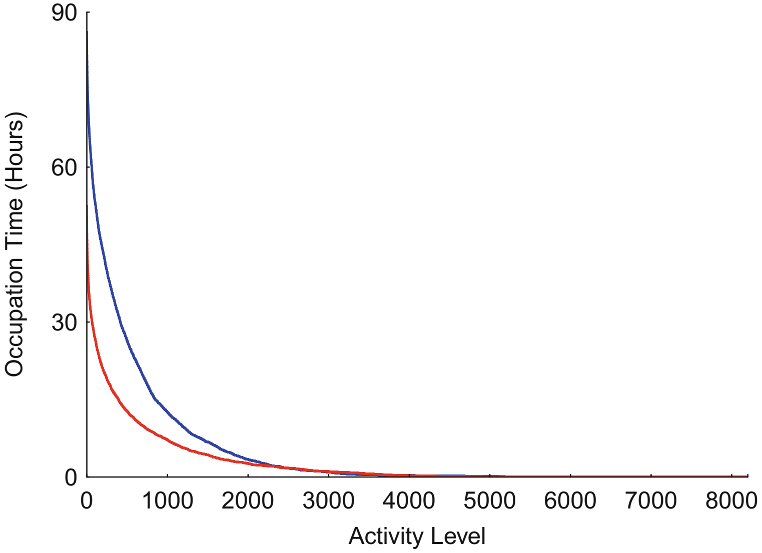

FIGURE 3 Occupation time (in hours) during 1 week for the first (blue) and second (red) subjects in Figure 1. Each occupation time curve is discretized using the function f_n defined in (1), although the jumps in the step function are not apparent due to line thickness and high density of the grid (> 8000 points). [Colour figure can be viewed at wileyonlinelibrary.com]

a step function approach instead of smoothing. The benefit of this step function approach is a more flexible framework and weaker assumptions compared to existing approaches in the literature, in particular in allowing discontinuities in the functional means and covariances under a dense design. Second, smoothing frequently serves the purpose of dampening fluctuations of the observations across the indices, in particular measurement errors. However, as the observed occupation time curves are smooth already (apart from barely visible jump discontinuities; see Figure 3) and an explicit modelling of measurement error is not necessarily implemented in the literature of wearable sensors (see, e.g. Huang et al., 2019; Song et al., 2019), we do not find smoothing necessary in our application to occupation time. Not implementing smoothing does not mean measurement errors cannot be taken into account in our framework, but just that they are incorporated implicitly into the covariance function of the limiting distributions of the empirical mean function and EL statistics. For more details about the effect of measurement error, see Supplement Section 13.3.

2.2 | Empirical mean for discretized observations

In this section we investigate the properties of a discretized version of the empirical mean function. We view the data through the lens of a discretization mechanism defined as follows. Let $\{T_1(a), \dots, T_n(a), a \in [\alpha_1, \alpha_2]\}$ be n i.i.d. realizations of a measurable stochastic process $T(a)$ having right-continuous sample paths of bounded variation. Let $\mu(a) \equiv ET(a)$ be the mean of $T(a)$ and $\hat{\mu}(a)$ the corresponding sample mean. Instead of fully observed trajectories (of the stochastic process $T = \{T(a)\}$), we can only observe $T(a)$ on $\mathbf{G}_n$, a (not necessarily equispaced) grid of points in $[\alpha_1, \alpha_2]$ (including the endpoints). Denote this discretized observation as $f_n(T)$, and $f_n(\mu)$ and $f_n(\hat{\mu})$ the corresponding mean and sample mean, respectively. Here $f_n(g)$ is the discretization of a function $g : [\alpha_1, \alpha_2] \to \mathbb{R}$ defined by

$$f_n(g)(a) = \begin{cases} g(a), & a \in \mathbf{G}_n, \\ g(b_a), & a \in [\alpha_1, \alpha_2] \backslash \mathbf{G}_n, \end{cases} \tag{1}$$

and b_a is the closest point on the grid to the right of a. This discretization $f_n(g)$ transforms the complete trajectory of g into a step function. The mesh of $\mathbf{G}_n$ (the maximal distance between adjacent grid points) is assumed to converge to zero as $n \to \infty$. Without requiring a rate of convergence for the mesh, we establish a parametric $\sqrt{n}$-convergence rate of $f_n(\hat{\mu})$ when centred on $f_n(\mu)$. To approximate the mean function outside $\mathbf{G}_n$, we need an infill condition in which the mesh of $\mathbf{G}_n$ shrinks more quickly than a certain negative power of n, as is typically assumed in dense functional data analysis settings (Wang et al., 2018).

Furthermore, we need a condition involving an extended version of the right-hand Dini derivatives (see, e.g. Hagood & Thomson, 2006):

$$D^+(\mu, \beta)(a) = \limsup_{h \to 0+} \frac{\mu(a+h) - \mu(a)}{h^\beta}, \text{ and } D_+(\mu, \beta)(a) = \liminf_{h \to 0+} \frac{\mu(a+h) - \mu(a)}{h^\beta}$$

for $\beta > 0$. Here we term them right-hand β-Dini derivatives, although these numbers have been studied under different names in the mathematics literature, such as right upper/lower Lipschitz numbers (Besicovitch, 1929) and β-right local derivatives (Ben Adda & Cresson, 2001). The special case of $\beta = 1$ leads to the usual right-hand Dini derivatives. These β-Dini derivatives always exist, if we allow values in the extended real line. They are reminiscent of β-Hölder continuity, but our requirement of boundedness of the right-hand β-Dini derivatives in the following theorem is much weaker than β-Hölder continuity. This is because we just focus on pointwise convergence (unlike β-Hölder continuity being a global feature) and right-hand limits.

The following theorem describes the asymptotic behaviour of the estimated mean function based on the aforementioned discretization. Here and in the sequel, the convergence in distribution (denoted by $\xrightarrow{d}$) of a sequence of random elements in a metric space $\mathbb{D}$ means convergence of the expectation of every bounded, continuous real-valued function applied to each element of the sequence (see, e.g. van der Vaart, 2000, p. 258). In the following theorem we use $\mathbb{D} = \ell^\infty([\alpha_1, \alpha_2])$, the space of all bounded real-valued functions on $[\alpha_1, \alpha_2]$, endowed with the supremum norm. The proof of the theorem is in Supplement Section 1.

Theorem 1. *Suppose the sample paths of $T(a)$ are right-continuous, of bounded variation, $ET^2(a)$ is bounded over $a \in [\alpha_1, \alpha_2]$, and $\mu(\cdot)$ has at most finitely many jump discontinuities. If the mesh of $\mathbf{G}_n$ is $o(1)$, then for all sufficiently small $\delta > 0$, there exists $I_\delta \subset [\alpha_1, \alpha_2]$ having Lebesgue measure δ such that*

$$\sqrt{n}\{f_n(\hat{\mu}) - f_n(\mu)\}(a) \xrightarrow{d} U(a)$$

in $\ell^\infty([\alpha_1, \alpha_2] \backslash I_\delta)$ as $n \to \infty$, where $U(a)$ is a zero-mean Gaussian process with $\mathrm{Cov}\{U(a), U(b)\} = \mathrm{Cov}\{T(a), T(b)\}$. *Suppose, in addition, that $D^+(\mu, \beta)(a)$* and *$D_+(\mu, \beta)(a)$ are bounded over $a \in [\alpha_1, \alpha_2]$ for some $\beta > 0$. If the mesh of $\mathbf{G}_n$* is *$o(n^{-1/(2\beta)})$, then*

$$\sqrt{n}\{f_n(\hat{\mu}) - \mu\}(a) \xrightarrow{d} U(a)$$

in $\ell^\infty([\alpha_1, \alpha_2] \backslash I_\delta)$ as $n \to \infty$.

Remark 1. Using the first part of Theorem 1, or more generally Theorem S.1 in Supplement Section 11, we can construct a confidence band for $\mu(a)$ over $a \in \mathbf{G}_n$, irrespective of how quickly its mesh shrinks as $n \to \infty$. Thus, there is no distinction between moderately dense and dense functional data here (Wang et al., 2018). Such a distinction only matters in the second part of the theorem, which can be used to construct a simultaneous confidence band for $\mu(\cdot)$ for *essentially all* $a \in [\alpha_1, \alpha_2]$ (see Supplement Section 2.1), provided the mesh of $\mathbf{G}_n$ tends to zero faster than $n^{-1/(2\beta)}$.

Remark 2. Since both $|D^+(\mu, \beta)(a)|$ and $|D_+(\mu, \beta)(a)|$ are non-decreasing in β, if they are bounded over $a \in [\alpha_1, \alpha_2]$, for some $\beta = \beta_1 > 0$, then the boundedness also holds for all $0 < \beta \leq \beta_1$. See Supplement Section 2.2 for some example functions in which such a boundedness holds.

Remark 3. The first part of the theorem applies to occupation time because the bounded variation assumption of the result is satisfied due to the monotonicity of $L(a)$. The moment condition is satisfied because $L(a)$ is bounded. Furthermore, the right-continuity holds by the strict inequality in the definition of $L(a)$. The second part of the theorem applies to occupation time for $\beta = 1$ because the instantaneous change in $EL(a)$ from the right is bounded between 0 and τ, by definition of $L(a)$.

Remark 4. The condition that $ET^2(a)$ is bounded over $a \in [\alpha_1, \alpha_2]$ can be reduced to assuming $EV^2(\alpha_2)$ and $ET^2(\alpha_1)$ are finite, where $V(a)$ is the total variation of $T(\cdot)$ over $[\alpha_1, a]$. This is because $|T(a)| \leq |T(\alpha_1)| + |T(a) - T(\alpha_1)| \leq |T(\alpha_1)| + V(a)$, by the assumption that each sample path $a \mapsto T(a)$ has bounded variation.

Remark 5. Here we compare our framework with the approach of using smoothing to accommodate fixed dense trajectory discretization in inference for the functional mean (Cao et al., 2012; Degras, 2011). Besides having similar moment assumptions on the data, the smoothness conditions we use, namely right-continuity, bounded variation, finite number of discontinuities, and boundedness of the right-hand β-Dini derivatives, are weaker than the assumptions on the data made in the papers mentioned above: Degras (2011) utilizes bounded second-order derivative and β_1-Hölder continuity for some $\beta_1 > 0$, whereas Cao et al. (2012) uses Lipschitz continuity of the qth order derivatives for some $q \in \{0\} \cup \mathbb{N}$ and β_2-Hölder continuity for some $\beta_2 \in (0, 1]$. Having fewer assumptions means we can deal with a wider range of data (beyond occupation time), such as processes with non-smooth latent moments (see Section 3.1 for an illustration). Furthermore, we do not require smoothing as Degras (2011) and Cao et al. (2012) do: the resulting rate at which the mesh of $\mathbf{G}_n$ tends to zero is characterized by mesh × $\log(\text{mesh}^{-1}) = o(n^{-1/4})$ and mesh $= o(n^{-1/\{2(q+1)\}})$ in Degras (2011) and Cao et al. (2012), respectively. To see that our rate is competitive, our $\beta = 1$ case is a weaker version of the least stringent $q = 0$ assumption in Cao et al. (2012), but both lead to mesh $= o(n^{-1/2})$. Last but not least, our framework provides insights on the price for constructing the confidence band at points outside $\mathbf{G}_n$, namely the additional need to control the rate at which the mesh of $\mathbf{G}_n$ tends to zero.

The distribution of the limiting process $U(a)$ needs to be estimated because it is not distribution-free. This can be done using the nonparametric bootstrap $f_n(U_n^*)(a)$ based on sampling n curves with replacement from the data $\{f_n(T_1)(a), \ldots, f_n(T_n)(a), a \in [\alpha_1, \alpha_2]\}$, where $U_n^*(a) = \sqrt{n}\{\hat{\mu}^*(a) - \hat{\mu}(a)\}$, $\hat{\mu}^*(a) = \sum_{i=1}^n W_{ni} T_i(a)/n$, and W_{ni} is the number of times that $f_n(T_i)(a)$ is redrawn from the data. We examine an alternative method of calibration based on simulating an estimated $U(a)$ in the Supplement Section 13.2, and we find similar results in one particular example.

Bootstrap consistency of $f_n(U_n^*)(a)$ is established as follows (see Supplement Section 3.1 for the proof). Interestingly, in contrast to the different conditions needed in the two parts of Theorem 1, this bootstrap consistency result holds irrespective of how quickly the mesh shrinks.

Corollary 1. *Under the conditions of the first part of Theorem 1, for all sufficiently small $\delta > 0$, there exists $I_\delta \subset [\alpha_1, \alpha_2]$ having Lebesgue measure δ such that $f_n(U_n^*)$ converges weakly to $U(a)$ in $\ell^\infty([\alpha_1, \alpha_2]\backslash I_\delta)$ as $n \to \infty$, given the data sequence $\{f_j(T_i), i = 1, \dots, j, j = 1, 2, \dots\}$, in probability.*

From this result, according to Theorem 1 and Supplement Section 2.1, we can construct an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\mu(\cdot)$ as $f_n(\hat{\mu})(a) \pm n^{-1/2}c^*_{NS,\alpha}$ for essentially all $a \in [\alpha_1, \alpha_2]$, where $c^*_{NS,\alpha}$ denotes the upper α-quantile of the $\sup_{a\in[\alpha_1,\alpha_2]}|f_n(U_n^*(a))|$ values obtained from B bootstrap samples; we use $B = 1000$ for implementation (see Supplement Section 3.2 for details). We refer to this as the Wald-type *NS band*, where NS stands for 'non-standardized', in contrast to existing bands in the literature (Cao et al., 2012; Choi & Reimherr, 2018; Degras, 2011) that use standardized estimators in forming the simultaneous confidence bands. Unfortunately, this band did not perform well in our simulation study (see Section 3.1). Besides the fact that it does not have the optimality EL enjoys, this band does not fully exploit the boundedness feature of the occupation time. An alternative approach is developed in Section 2.3.

2.3 | Empirical likelihood confidence band

In this section, we develop the proposed simultaneous confidence band for the mean $\mu(\cdot)$. Our approach is based on inverting a localized form of the EL statistic at each value of a. For simplicity of exposition, we define the observed EL ratio by discretizing the fully observed trajectories of the EL ratio process in the following, instead of defining the discretized version of each component that determines the EL ratio. But note that those components are available to us only in their discretized forms. Such exposition will be used in the next section, too.

For a given $a \in [\alpha_1, \alpha_2]$, the local EL ratio for $\mu(a)$ is $f_n(\mathcal{R}(\tilde{\mu}))(a)$ at a given value $\tilde{\mu}(a)$, where

$$\mathcal{R}(\tilde{\mu})(a) = \frac{\sup\{L(F_a) : m(a) = \tilde{\mu}(a), F_a \in \Gamma_a\}}{\sup\{L(F_a) : F_a \in \Gamma_a\}}, \tag{2}$$

$F_a(\cdot)$ is a candidate for the cumulative distribution function of $T(a)$, $m(a)$ is the mean of $F_a(\cdot)$, Γ_a is the set of distributions supported by the data $\{T_i(a), i = 1, \dots, n\}$, $L(F_a) = \prod_{i=1}^n p_i(a)$ is the nonparametric likelihood, $p_i(a) = F_a\{T_i(a)\} - F_a\{T_i(a)-\}$, and we follow the convention $\sup\emptyset = 0$. By the same reasoning as in EL for a univariate mean (Owen, 2001, p. 70), any feasible solution for $\tilde{\mu}(a)$ in computing $f_n(\mathcal{R}(\tilde{\mu}))(a)$ lies in the interval $[\min_{i=1,\dots,n} f_n(T_i)(a), \max_{i=1,\dots,n} f_n(T_i)(a)]$. This is what we mean by EL respecting the range restrictions inherent in the data, as mentioned in Section 1.

We now state our first key result, giving the asymptotic distribution of the EL statistic $-2\log f_n(\mathcal{R}(\mu))(a)$ viewed as a process indexed by a.

Theorem 2. *Suppose the conditions of the first part of Theorem 1 hold and in addition, $\inf_{a\in[\alpha_1,\alpha_2]} \sigma^2(a) > 0$ and $\sigma^2(\cdot)$ has at most finitely many jump discontinuities, where $\sigma^2(a) = \mathrm{Var}\{T(a)\}$. Then for all sufficiently small $\delta > 0$, there exists $I_\delta \subset [\alpha_1, \alpha_2]$ having Lebesgue* *measure δ such that $-2\log f_n(\mathcal{R}(\mu))(a) \xrightarrow{d} U^2(a)/\sigma^2(a)$ in $\ell^\infty([\alpha_1,\alpha_2]\backslash I_\delta)$ as $n \to \infty$, where the process $U(a)$ is defined in Section 2.2.*

Remark 1. The optimality of EL is obtained in terms of (i) the large deviation principle (Kitamura, 2007; Kitamura et al., 2012), and (ii) a second-order local maximinity property which also characterizes an ordinary parametric likelihood ratio (Bravo, 2003; Mukerjee, 1994). Here we explain the intuition as to why EL performs optimally in the large deviation sense. The reason is due to the fact that EL minimizes the Kullback–Leibler divergence between the empirical measure and the probability measure satisfying the (pointwise) null hypothesis. Since the large deviation principle for empirical measures (i.e. Sanov's Theorem) shows the probability that the empirical measure falls into any given set of probability measures is governed by the minimum value of the Kullback–Leibler divergence number, it can be expected that using the minimum (empirical) Kullback–Leibler divergence as a statistical criterion leads to an optimal procedure, in contrast to procedures based on minimizing any other objective function that contrasts the aforementioned two probability measures.

Remark 2. The proof (given in Supplement Section 4) is based on a uniform approximation of the EL statistic by $f_n(\hat{\Psi})^2(a)$, where $\hat{\Psi}(a) = \sqrt{n}\{\hat{\mu}(a) - \mu(a)\}/\sigma(a)$. Note, however, that the asymptotic equivalence of $f_n(\hat{\Psi})^2(a)$ to the EL statistic does not imply optimality of $f_n(\hat{\Psi})^2(a)$, because it is known that Pitman efficiency based on first-order linear approximations is not informative enough in distinguishing among the performance of procedures satisfying the same moment condition. Thus, higher-order asymptotics and large deviation theory have been used to show that EL enjoys optimality while other procedures (even the first-order asymptotically equivalent ones) do not, as mentioned in the previous remark.

Remark 3. The condition $\inf_{a\in[\alpha_1,\alpha_2]}\sigma^2(a) > 0$ is similar to the condition of a positive definite information or covariance matrix in the Wilks type theorem (Owen, 2001). To deal with data that violate this condition, we adapt a two-step approach that has been proposed in Nair (1984), as discussed in Supplement Section 5.1.

For calibration, we use a similar nonparametric bootstrap method as in Section 2.2, based on sampling n curves with replacement from the data $\{f_n(T_1)(a), \dots, f_n(T_n)(a), a \in [\alpha_1,\alpha_2]\}$. Since $M_n = \sup_{a\in[\alpha_1,\alpha_2]}\{-2\log f_n(\mathcal{R}(\mu))(a)\}$ is asymptotically equivalent to $\sup_{a\in[\alpha_1,\alpha_2]} f_n(\hat{\Psi})^2(a)$ by the above Remark 2, it suffices to bootstrap $f_n(\hat{\Psi})(a)$ by $f_n(\hat{\Psi}^*)(a)$, where $\hat{\Psi}^*(a) = U_n^*(a)/\hat{S}(a)$, $U_n^*(a)$ is defined in Section 2.2 and $\hat{S}(a) = \left[\sum_{i=1}^n\{T_i(a) - \hat{\mu}(a)\}^2/n\right]^{1/2}$ is the sample version of $\sigma(a)$. The resulting bootstrap for M_n is $M_n^* = \sup_{a\in[\alpha_1,\alpha_2]} f_n(\hat{\Psi}^*)^2(a)$. The relevant bootstrap consistency is established as follows (see Supplement Section 6.1 for the proof).

Corollary 2. *Under the conditions of Theorem 2, for all sufficiently small $\delta > 0$, there exists $I_\delta \subset [\alpha_1,\alpha_2]$ having Lebesgue measure δ such that $f_n(\hat{\Psi}^*)^2(a)$ converges weakly to $U^2(a)/\sigma^2(a)$ in $\ell^\infty([\alpha_1,\alpha_2]\backslash I_\delta)$ as $n \to \infty$, given the data sequence $\{f_j(T_i), i = 1, \dots, j, j = 1, 2, \dots\}$, in probability.*

Under the conditions of the second part of Theorem 1, Corollary 2 provides an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\mu(\cdot)$ for essentially all $a \in [\alpha_1,\alpha_2]$:

$$\{(a, \tilde{\mu}(a)) : -2\log f_n(\mathcal{R}(\tilde{\mu}))(a) \le c^*_{EL,\alpha}, \tilde{\mu} \in \mathcal{D}_n\},$$

where $c^*_{EL,\alpha}$ denotes the upper α-quantile of the M^*_n values obtained from $B = 1000$ (as in Section 2.2) bootstrap samples (see Supplement Section 6.2 for details), and $\mathcal{D}_n$ is the class of functions of the form in (1) (see Supplement Section 7.1 for details). We refer to this as the *EL band*. It can be shown that if the observed processes are monotone in a, as is the case for $L(a)$, then the lower and upper boundaries of the EL band will respect this monotonicity. See Supplement Section 7.2 for a proof and an illustration of this monotonicity in Figure S.1.

The NS band we introduced in Section 2.2 also respects such monotonicity, by the fact that $f_n(\hat{\mu})(a)$ is monotone in a and $n^{-1/2}c^*_{NS,\alpha}$ is constant in a. However, as mentioned in Section 2.2, it is not an optimal band and does not respect the range restrictions imposed by the data.

By the first-order asymptotic equivalence of $f_n(\hat{\Psi})^2(a)$ to $-2\log f_n(\mathcal{R}(\mu))(a)$ in Remark 2 after Theorem 2, another asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\mu(\cdot)$ is $f_n(\hat{\mu})(a) \pm n^{-1/2}c^*_{EP,\alpha}f_n(\hat{S})(a)$ for essentially all $a \in [\alpha_1, \alpha_2]$, where $c^*_{EP,\alpha}$ denotes the upper α-quantile of the $\sup_{a\in[\alpha_1,\alpha_2]}|f_n(\hat{\Psi}^*)(a)|$ values obtained from $B = 1000$ (as in Section 2.2) bootstrap samples (see Supplement Section 6.2 for details). This is another Wald-type band, but with equal precision across different values of a, in the sense that its pointwise width is proportional to its pointwise estimated standard deviation (Nair, 1984). We refer to this as the *EP band*. This band does not respect monotonicity of the observed processes in the sense described in the previous two paragraphs, nor the range restrictions imposed by the data. Furthermore, it is not optimal as explained in Remark 2 after Theorem 2.

2.4 | Empirical likelihood-based ANOVA test

Now we consider the setting of k independent samples, where we use the same notation as before except with a further subscript j indicating the jth sample, $j = 1, \ldots, k$. We want to test $H_0 : \mu_1(\cdot) = \ldots = \mu_k(\cdot) \equiv \mu_0(\cdot)$ versus the omnibus alternative H_1. Assume the proportion of data in the jth sample $n_j/n \to \gamma_j > 0$ for some fixed γ_j as $n \to \infty$, where n is the total sample size. The local EL ratio at a given activity level a is defined by $f_n(\mathcal{R}_k)(a)$, where

$$\mathcal{R}_k(a) = \frac{\sup\left\{\prod_{j=1}^k L(F_{aj}) : m_1(a) = \ldots = m_k(a), F_{aj} \in \Gamma_{aj}, j = 1, \ldots, k\right\}}{\sup\left\{\prod_{j=1}^k L(F_{aj}) : F_{aj} \in \Gamma_{aj}, j = 1, \ldots, k\right\}}, \tag{3}$$

where $L(F_{aj})$ is the nonparametric likelihood for the jth sample at level a.

To test H_0 versus H_1, we propose the following maximally selected EL statistic:

$$K_n = \sup_{a\in[\alpha_1,\alpha_2]} \left\{-2\log f_n(\mathcal{R}_k)(a)\right\}.$$

This maximal-deviation-type statistic is sensitive to any local difference among the functional means, and provides a consistent test against the omnibus hypothesis H_1. An alternative approach would be to use the integral-type statistic $\int_{\alpha_1}^{\alpha_2}\{-2\log f_n(\mathcal{R})(a)\}\, da$, which can detect differences dispersed over the range of indices. In our simulation studies, however, we found strong evidence that K_n consistently outperforms the above integral-type statistic, so in the sequel we restrict attention to K_n.

The following result gives the approximation of K_n (see Supplement Section 8 for the proof), expressed in terms of the Gaussian process $\Psi_j(a) = U_j(a)/\sigma_j(a)$ arising in the asymptotic distribution of $\hat{\Psi}_j(a)$.

Theorem 3. *Suppose the conditions of Theorem 2 hold for each group indexed by $j = 1, \ldots, k$. Then, under H_0,* as $n \to \infty$,

$$K_n = \sup_{a\in[\alpha_1,\alpha_2]} f_n(\widehat{\text{SSB}})(a) + o_p(1) \quad \textit{and} \quad \widehat{\text{SSB}}(a) \xrightarrow{d} \text{SSB}\ (a)$$

in $\ell^\infty([\alpha_1, \alpha_2])$, where

$$\widehat{\text{SSB}}(a) = \sum_{j=1}^{k} w_j(a)\left\{\frac{\hat{\Psi}_j(a)}{\sqrt{w_j(a)}} - \check{\Psi}(a)\right\}^2, \quad \text{SSB}\ (a) = \sum_{j=1}^{k} w_j(a)\left\{\frac{\Psi_j(a)}{\sqrt{w_j(a)}} - \overline{\Psi}(a)\right\}^2,$$

$\check{\Psi}(a) = \sum_{j=1}^{k} \sqrt{w_j(a)}\hat{\Psi}_j(a)$, $\overline{\Psi}(a) = \sum_{j=1}^{k} \sqrt{w_j(a)}\Psi_j(a)$, *and the varying weights $w_j(a) \propto \gamma_j/\sigma_j^2(a)$ are normalized to sum to 1 across the groups.*

Remark 1. Note that SSB(a) is a weighted sum of squares between blocks, with each block contrasting a weighted version of $\Psi_j(a)$ with the overall weighted average, in parallel with standard ANOVA. A similar structure emerges in k-sample EL-based tests for differences in survival functions (Chang & McKeague, 2019, section 3.4).

Remark 2. As in Remark 3 following Theorem 2, the procedure needs to be modified to deal with data that violate the nonzero variance condition. To this end, we adapt Uno et al. (2015)'s selection approach, as described in Supplement Section 5.2.

Remark 3. The approximation of K_n in Theorem 3 leads to a *Wald-type test* of H_0 versus H_1 as $K_{n,Wald} = \sup_{a\in[\alpha_1,\alpha_2]} f_n(\widehat{\text{SSB}})(a)$. Note, however, that the first-order asymptotic equivalence of $K_{n,Wald}$ to K_n does not imply optimality of $K_{n,Wald}$, as we pointed out in Remark 2 after Theorem 2. This will be seen in our simulation study in Section 3.2 as well.

For calibration, again we use a similar nonparametric bootstrap method as in Sections 2.2 and 2.3, based on sampling n_j curves with replacement from the jth sample for $j = 1, \ldots, k$. Since K_n is asymptotically equivalent to $\sup_{a\in[\alpha_1,\alpha_2]} f_n(\widehat{\text{SSB}})(a)$, it suffices to bootstrap $f_n(\widehat{\text{SSB}})(a)$ by $f_n(\widehat{\text{SSB}}^*)(a)$, where

$$\widehat{\text{SSB}}^*(a) = \sum_{j=1}^{k} \hat{w}_j(a)\left\{\frac{\hat{\Psi}_j^*(a)}{\sqrt{\hat{w}_j(a)}} - \check{\Psi}^*(a)\right\}^2,$$

$\hat{w}_j(a) \propto \hat{\gamma}_j/\hat{S}_j^2(a)$ are normalized to sum to 1 across the groups, $\hat{\gamma}_j = n_j/n$, $\hat{\Psi}_j^*(a)$ is defined in Section 2.3 for $j = 1$ and $\check{\Psi}^*(a) = \sum_{j=1}^{k}\sqrt{\hat{w}_j(a)}\hat{\Psi}_j^*(a)$. The resulting bootstrap for K_n is $K_n^* = \sup_{a\in[\alpha_1,\alpha_2]} f_n(\widehat{\text{SSB}}^*)(a)$. To calibrate the test, we compare the upper α-quantile of the K_n^* values obtained from $B = 1000$ bootstrap samples (as in Sections 2.2 and 2.3) with our test statistic K_n (see Supplement Section 10 for details). The Wald-type test $K_{n,Wald}$ can be calibrated the same way due to its asymptotic equivalence to K_n (see Supplement Section 10 for details).

3 | SIMULATION STUDY

In this section, we compare the performance of the proposed simultaneous confidence band with five other types of Wald-type simultaneous confidence bands for the mean of functional data: the NS band in Section 2.2, the EP band in Section 2.3, the band by Degras (2011, 2017) (MFD) with cross-validated bandwidth selection implemented in the R package **SCBmeanfd** (Degras, 2016), the band by Cao et al. (2012) (Cao1) with the initially smoothed covariance estimates projected onto the space of non-negative definite matrices (Hall et al., 2008), and the hyper-ellipsoid Scheffé-style band (Geo) of Choi and Reimherr (2018) implemented in the R package **fregion** (Choi, 2017). The MFD band is based on local linear smoothing and normal quantiles, which turns out to have poor performance in our simulation settings, so we provide an alternative band using bootstrap calibration and our step-function approach (MFDbs). The Cao1 band is based on spline smoothing and the recommended number of interior knots in the original paper; we provide an alternative (Cao2) band by using the number of grid points on which the functional trajectory is observed as the number of knots. Since the Geo band is based on any readily available functional mean estimators the user would like, we use our discretized mean estimates and incorporate our step-function approach into constructing the band. Note that besides MFDbs, the first two types of bands are also based on bootstrap, whereas the Cao1, Cao2 and Geo bands do not have readily available bootstrapped versions.

We also compare the proposed EL-based ANOVA test (in terms of accuracy and power) with four Wald-type functional ANOVA tests: the Wald-type test in the previous subsection, the Fmax test by Zhang, Cheng, et al. (2019) based on a maximally selected F-statistic, the GPF test by Zhang and Liang (2014) based on an integrated F-statistic and a test based on random projections with a Wald-type permutation statistic (Cuesta-Albertos & Febrero-Bande, 2010; Pauly et al., 2015) (TRP). The latter three tests are recommended in the literature based on extensive simulation studies (Górecki & Smaga, 2015, 2018; Zhang, Cheng, et al., 2019), and they are implemented in the R package **fdANOVA** (Górecki & Smaga, 2018). In implementing the confidence bands and tests, we use the default settings given in the aforementioned three packages, except the number of bootstrap or permutation replications is taken as 1000 in all procedures to make them comparable.

3.1 | Performance of simultaneous confidence bands

We consider two simulation examples in this subsection: general functional data (not occupation time) with non-smooth means and covariances, and occupation time curves. For the first, we generate $T(a) = \max(J(a), 0)$, $a \in [0,1]$, where $J(a)$ is a zero-mean Gaussian process having a non-smooth $\mathrm{Cov}\{J(a), J(b)\} = (0.6 + \nu_T)I\{a < 0.25, a = b\} + 0.6I\{a \geq 0.25, a = b\} + 1.5I\{a, b < 0.25, a \neq b\} + 0.5I\{a \text{ or } b \geq 0.25, a \neq b\}$ for some $\nu_T > 1$ (See Supplement Section 13.1 for further discussion). We expect the range of the EL band to be non-negative but not necessarily for the other bands. We use regular grids $\mathbf{G}_n$ of 26 and 51 points for $n = 100$ and 200, respectively. The coverage is evaluated on a regular grid of 101 points. Since the marginal variance of $T(a)$ is bounded away from 0, there is no need to modify any of the bands in the way described in Supplement Section 5.1.

For the second example, we generated occupation time data $L(a)$ from $X(t) = \lfloor \{10^{10}\pi(t)/\varepsilon\}^{1/4} \rfloor I\{\pi(t) > \xi\}$ for $t \in \mathbf{H}_n$, a regular grid on $[0, 1)$, where $\pi(\cdot)$ is a random permutation of $\mathbf{H}_n$, $\varepsilon \sim$ Log-normal$(0, \nu_L^2) + 10^{-6}$, $\nu_L > 0$ is the log-normal scale parameter, $\xi \sim$ Uniform$(0, 1)$, and $\varepsilon, \pi(\cdot), \xi$ are independent. $\mathbf{H}_n$ is taken to have 1000 and 2000 points for

$n = 100$ and 200, respectively. Here $X(t)$ takes only non-negative integer values, reflecting the same property found in activity count data from the wearable (ActiGraph) devices. The true mean occupation time $E\{L(a)\}$ is readily calculated, allowing us to assess the coverage of confidence bands; see Supplement Section 12 for the calculation, and Figure S.2 for a simulated sample path of X along with $E\{L(a)\}$ for $\nu_L = 2$. The grid $\mathbf{G}_n$ is taken to be every fourth and second non-negative integer for $n = 100$ and 200, respectively. The coverage is evaluated at each non-negative integer. All the bands except for NS utilize the two-step approach described in Supplement Section 5.1 to handle zero-variance situations. Comparing the mean functions in the two examples, there is a larger jump (0.9898) when $\nu_T = 10$ compared to jumps of sizes 0.0019 and 0.0016 in the second example when $\nu_L = 1.5$ and 2.

Empirical coverage rates, average widths, range-violation and monotonicity preservation of the various bands are given in Table 1, where we define the width of a band as the average width over the range of activity levels. The empirical coverage rates of our EL band and the NS band are closer to the nominal 95% level compared with other bands, but NS is much wider than the EL band (up to 2.5 times wider). Although first-order asymptotically equivalent to the EL band, the EP band still has worse coverage, as explained in Remarks 1 and 2 after Theorem 2. Existing bands based on smoothing, namely MFD, Cao1 and Cao2 bands, tend to have undercoverage due to the phenomenon illustrated in Figure 2, with severe undercoverage in the first example. Existing bands not based on smoothing tend to either undercover or overcover: MFDbs undercovers in the first example and overcovers in the second, whereas Geo has the opposite behaviour. Note that the range-respecting and monotonicity properties of EL are reflected in the results. None of the other bands have these properties, although our results only reflect this for NS, MFD, MFDbs, Cao1 and Geo. We conclude that the proposed EL confidence bands have the best performance in terms of the properties mentioned above.

3.2 | Performance of ANOVA tests

In this section, the empirical level and power of the proposed EL test is compared with the four Wald-type functional ANOVA tests mentioned earlier. We restrict attention to $k = 3$ groups and the application to occupation time. We study how the performance of the tests is affected by unequal sample sizes and unequal variance functions (heteroscedasticity) among the groups. For each group j, we generated the raw activity data $X_j(t) = \lfloor 300 \max(\Omega_j(t), 0) \rfloor$, where $\Omega_j(t)$ is an Ornstein–Uhlenbeck process, and the floor function and the positive part are to reflect non-negative integer values of activity count data from wearable (ActiGraph) devices. Here t belongs to a regular grid on $[0, 1)$ with 1000 points. The resulting functional data of interest are $L_j(a)\Sigma_j$, where Σ_j is an independent beta random variable to allow flexible control of $\mathrm{Var}\left\{L_j(a)\Sigma_j\right\}$ through different parameters from those that control $E\left\{L_j(a)\Sigma_j\right\}$ (see Supplement Section 13.4 for details). The grid $\mathbf{G}_n$ of $L_j(a)$ is taken to be every second non-negative integer. Each group has a distinct set of Ornstein–Uhlenbeck and beta parameters. These parameters are chosen to produce identical functional means (scenario A, upper left panel of Figure 4), crossing functional means (scenario B, upper middle panel of Figure 4) or ordered functional means (scenario C, upper right panel of Figure 4). Here the underlying functional mean is obtained by averaging 50,000 replicates in each group (closed-form expressions are not available). For each scenario, the deviations (of the group functional means) from the grand mean, along with the variance functions, are plotted beneath their respective mean functions in Figure 4. In all scenarios, the third group

TABLE 1 Simulation study for 95% simultaneous confidence bands: Empirical coverage (percentage), average width (in parenthesis), range-violation rate (percentage, in square brackets) and average number of confidence band boundaries that satisfy monotonicity (rounded to two decimal places, in curly brackets); 1000 Monte Carlo replications, 1000 bootstrap samples, jump parameter $\nu_T = 10,\ 15$, log-normal scale parameter $\nu_L = 1.5,\ 2$, $n = 100, 200$

	Case 1				**Case 2**			
	$\boldsymbol{\nu_T = 10}$		$\boldsymbol{\nu_T = 15}$		$\boldsymbol{\nu_L = 1.5}$		$\boldsymbol{\nu_L = 2}$	
Tests	$\boldsymbol{n = 100}$	$\boldsymbol{n = 200}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 200}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 200}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 200}$
EL	94.5 (0.47) [0]	94.8 (0.35) [0]	94.2 (0.53) [0]	94.9 (0.40) [0]	93.9 (0.10) [0] {2}	94.4 (0.08) [0] {2}	94.3 (0.10) [0] {2}	94.8 (0.09) [0] {2}
EP	88.7 (0.47) [0]	91.1 (0.35) [0]	91.5 (0.52) [0]	91.2 (0.39) [0]	91.7 (0.09) [0] {2}	93.7 (0.08) [0] {2}	92.2 (0.09) [0] {2}	93.8 (0.08) [0] {2}
NS	95.1 (1.03) [100]	96.1 (0.78) [100]	95.0 (1.25) [100]	96.5 (0.95) [100]	93.9 (0.14) [100] {2}	95.2 (0.10) [100] {2}	93.6 (0.14) [100] {2}	95.1 (0.10) [100] {2}
MFD	0 (0.40) [0]	0 (0.29) [0]	0 (0.45) [0]	0 (0.33) [0]	90.6 (0.09) [0.4] {0.19}	93.7 (0.08) [0] {0.12}	90.8 (0.09) [2.2] {1.24}	93.4 (0.08) [0] {0.84}
MFDbs	89.4 (0.47) [0]	91.2 (0.35) [0]	91.3 (0.53) [0]	91.6 (0.39) [0]	98.6 (0.09) [0] {2}	99.3 (0.08) [0] {2}	99.0 (0.09) [0.1] {2}	99.5 (0.08) [0] {2}
Cao1	0 (0.21) [17.6]	0 (0.15) [0]	0 (0.21) [51.6]	0 (0.15) [0]	89.1 (0.09) [0] {2}	91.1 (0.08) [0] {2}	88.4 (0.09) [0] {2}	91.9 (0.08) [0] {2}
Cao2	0 (0.21) [0]	0 (0.15) [0]	0 (0.22) [0]	0 (0.15) [0]	89.7 (0.09) [0] {2}	91.3 (0.08) [0] {2}	89.0 (0.09) [0] {2}	92.0 (0.08) [0] {2}
Geo	99.9 (0.82) [73.9]	100 (0.73) [12.2]	99.8 (0.94) [92.2]	100 (0.83) [41.4]	74.9 (0.16) [100] {0}	78.2 (0.12) [100] {0}	74.4 (0.16) [100] {0}	77.8 (0.12) [100] {0}

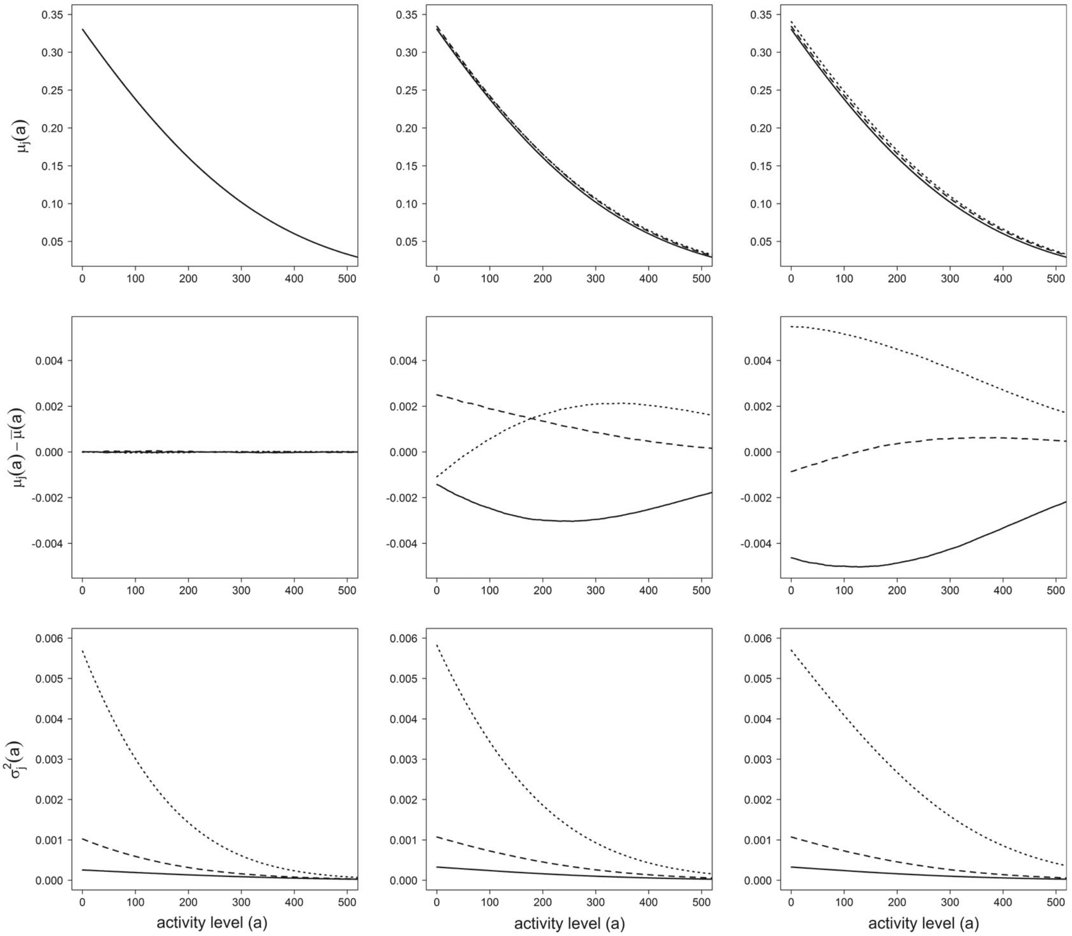

FIGURE 4 Simulation study for the functional ANOVA tests: the mean (top row), deviations (of the group means) from the grand mean (denoted as $\overline{\mu}(a)$) (middle row), and variance functions (bottom row) in the first (solid), second (dashed), and third (dotted) groups. Scenario A (top left panel): identical means (i.e. under H_0). Scenario B (top middle panel): crossing pattern. Scenario C (top right panel): ordered pattern.

has the largest variance function, followed by the second and first groups. All the tests except for TRP utilize the selection approach described in Supplement Section 5.2 to handle zero-variance situations.

In classical ANOVA testing, it is well known (Horsnell, 1953) that size and power are adversely affected by heteroscedasticity and lack-of-balance in the sample sizes. The above simulation model can address the question of whether our proposed EL-based approach mitigates this problem in the functional data setting.

The empirical rejection rates of the tests are given in Table 2, for sample size combinations $(n_1, n_2, n_3) = (70, 100, 130), (130, 100, 70)$. In scenario A (under H_0), the empirical levels of the EL test are close to the nominal 5% level. TRP becomes conservative when $(n_1, n_2, n_3) = (70, 100, 130)$, whereas the other tests are highly anti-conservative when $(n_1, n_2, n_3) = (130, 100, 70)$, reflecting the adverse performance pointed out by Horsnell (1953). In scenarios B and C (under H_1), the EL test has improved power in all cases, with an increase in power over existing tests ranging from

TABLE 2 Empirical rejection rates (percentages) for functional ANOVA tests under various scenarios (depicted in Figure 4) and sample sizes, based on 1000 Monte Carlo replications, 1000 bootstrap or permutation samples, and a nominal level of 5%

Scenario	(n_1, n_2, n_3)	EL test	Wald	Fmax	GPF	TRP
A	(70, 100, 130)	5.7	7.6	4.8	4.5	3.2
	(130, 100, 70)	6.3	9.1	12.8	15.1	5.5
B	(70, 100, 130)	73.3	72.7	39.9	36.1	65.6
	(130, 100, 70)	68.7	67.5	65.4	63.9	58.4
C	(70, 100, 130)	82.4	81.8	22.8	24.4	67.7
	(130, 100, 70)	84.2	83.1	58.8	60.1	71.0

3.3 to 59.6%. Although first-order asymptotically equivalent to the EL test, the Wald-type test in Section 2.4 is anti-conservative under H_0 and has less power under H_1; this less-than-optimal performance is discussed in Remark 3 about Theorem 3. In summary, our results show that the EL test outperforms the four Wald-type functional ANOVA tests.

4 | APPLICATION

We return to the occupation time application based on data from the 2005–2006 U.S. National Health and Nutrition Examination Survey (United States National Center for Health Statistics, 2005–2006). Each raw activity curve was measured in 1-min epochs using a wearable ActiGraph device for seven consecutive days (normalized to $[0, \tau] = [0, 1]$); we only keep measurements that NHANES flagged as both 'reliable' and 'in calibration'. We restrict attention to subjects aged 65-and-older and consider the following four subgroups: veterans aged 75-and-older ($n_1 = 160$), non-veterans aged 75-and-older ($n_2 = 279$), veterans aged 65–74 ($n_3 = 139$) and non-veterans aged 65–74 ($n_4 = 348$); systematically selected occupation time from each group are displayed in the left panel of Figure 5. We consider the maximal range of activity levels that has been categorized as sedentary in existing physical activity literature (Gorman et al., 2014), namely < 500 counts/min. The occupation time curves are restricted to the interval $[\alpha_1, \alpha_2] = [0, 499]$ and discretized at each intensity count (i.e. integers). There are five and two subjects with missing data in the second and fourth groups, respectively, and those subjects have 5%–62% missing readings that are handled using the imputation method described in Supplement Section 14. An alternative graphical comparison on the basis of sample means for the raw activity data is provided in Supplement Section 13.5.

The results of the functional ANOVA tests are summarized in Table 3. All the tests suggest that there are some significantly different mean occupation time curves among the four groups, as the p-values are all < 0.001 (see the row 'all groups' in Table 3). Next we investigate whether this result is driven by veteran status or age. Specifically, we conduct four pairwise comparisons among the groups, namely 1:2, 3:4, 1:3 and 2:4. The former two comparisons refer to the effect of veteran status, and the latter two refer to the effect of age, controlling for the other factor. Under a Bonferroni adjustment, we use $\alpha = 0.05/4 = 0.0125$. Regarding the effect of veteran status, in the older age group (i.e. the comparison 1:2), EL, Wald and Fmax indicate a significant difference (with EL being the most significant), whereas GPF and TRP fail to detect a difference. One possible explanation is that there are large local differences among the groups (e.g. in the tail of

TABLE 3 *p*-values from various functional ANOVA test statistics for comparing the mean occupation time of the four groups in the 2005–2006 NHANES study: Veterans aged 75-and-older (group 1), non-veterans aged 75-and-older (group 2), veterans aged 65–74 (group 3) and non-veterans aged 65–74 (group 4)

Comparison	EL test	Wald	GPF	Fmax	TRP
All groups	<0.001	<0.001	<0.001	<0.001	<0.001
Group 1 versus 2	0.007	0.010	0.040	0.012	0.030
Group 3 versus 4	0.333	0.331	0.436	0.375	0.375
Group 1 versus 3	<0.001	<0.001	<0.001	<0.001	<0.001
Group 2 versus 4	<0.001	<0.001	<0.001	<0.001	<0.001

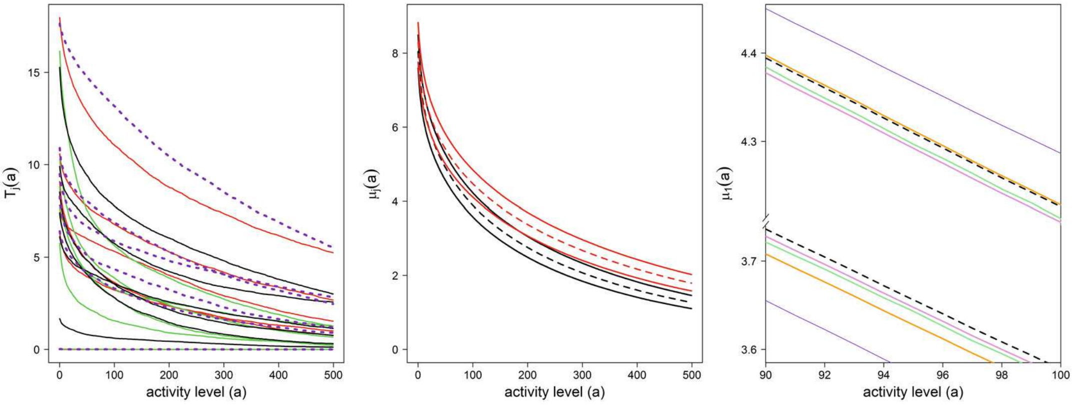

FIGURE 5 Left panel: occupation time (hours per day) from systematically selected (the smallest and largest observations, and every 20th percentile at $a = 0$) veterans aged 75-and-older (black), non-veterans aged 75-and-older (green), veterans aged 65–74 (red), and non-veterans aged 65–74 (dotted purple), based on the NHANES physical activity data. Middle panel: comparison of EL simultaneous confidence bands for the mean occupation time (estimates in dashed line) of veterans aged 75-and-older (black) and veterans aged 65–74 (red). Right panel: EL (dashed black), EP (pink), NS (purple), MFDbs (orange) and Geo (light green) 95% simultaneous confidence bands for the mean occupation time (dashed black line in the middle panel) of veterans aged-75-and-older, zooming-in on activity levels in the range 90–100 counts/min. [Colour figure can be viewed at wileyonlinelibrary.com]

the middle panel of Figure 5) that are better detected by statistics of maximal-deviation type (EL, Wald, Fmax) than integral type (GPF) or projection type (TRP) that can miss particular areas. In the younger age group (i.e. the comparison 3:4), none of the tests detect a significant difference between veterans and non-veterans. As for the effect of age, all the tests give significant results regardless of veteran status (see the comparisons 1:3 and 2:4). Such patterns are reflected in the confidence bands, for example, of older versus younger veterans (see the middle panel of Figure 5), where the younger group can be seen to spend more time at higher activity levels.

Turning to a comparison of the confidence bands based on EL, EP, NS, MFDbs and Geo, for veterans aged-75-and-older, see the right panel of Figure 5 (which zooms-in on activity levels in the range 90–100 counts/min). The comparison is restricted to bands that are not based on smoothing (see Section 3.1). EP and EL are close to one another, whereas NS is much wider, which is consistent with the simulation results in Table 1. On the other hand, MFDbs is wider than EL, and Geo has similar width to EL, in contrast to the simulation results in Table 1. Comparing the EL band with these two bands over the entire range of activity levels, MFDbs is consistently wider than EL. Geo is wider than EL in the extremes of the activity levels. All the bands respect range and monotonicity constraints of the occupation time data.

As mentioned in the Introduction, sedentary behaviour has typically been defined as < 100 counts/min (Matthews et al., 2008). Alternative cut-points have been suggested, however, namely 50, 200, 260 and 500 (Gorman et al., 2014). For veterans aged 75-and-older, we consider the cut-points 50, 100 and 500. A 95% confidence interval for the mean hours per day of sedentary behaviour using 50 as the cut-point can be obtained by subtracting from 24 the black confidence band in the middle panel of Figure 5 at $a = 49$, similarly for the other cut-points. This results in the EL confidence intervals 18.7–19.4, 19.8–20.4 and 22.5–22.9 for the mean hours per day of sedentary behaviour, for cut-points 50, 100 and 500, respectively. We see that there is a 3-h difference between the upper limit of the first confidence interval and the lower limit of the third confidence interval. This illustrates how dramatically the quantification of sedentary behaviour can change when the cut-points are changed, hence our preference for analysing the full occupation time curve.

5 | DISCUSSION

In this paper we developed a general nonparametric framework for the analysis of functional means that allows discontinuities in the functional means and covariances under a dense non-random design. We applied this framework to occupation time data derived from wearable devices. Indexed by activity level ranging continuously over the range of device readings, the occupation time curves are automatically aligned and contain more information than time spent in discrete activity categories (cf., Matthews et al., 2008; Staudenmayer et al., 2012). Taking advantage of optimality properties and the ability to handle the boundedness and monotonicity of these occupation time curves, our EL approach is used to construct a simultaneous confidence band and an ANOVA test for the functional means. We have shown via simulations that the new test adapts well to heteroscedasticity and imbalance in the sample sizes, and the proposed confidence band has more accurate coverage while being narrower than alternative approaches. In particular, when there is a discontinuity in the functional means and covariances, the EL band is shown to maintain accuracy, whereas alternative bands based on smoothing can have severe undercoverage. We applied the proposed procedures to wearable device data from the 2005–2006 NHANES study, obtaining narrower confidence bands than existing ones that are not based on smoothing; we also obtain more significant results for the ANOVA test.

By definition, the occupation time curve $L(a)$ given in Section 2.1 is proportional to a (random) survival function. In contrast to viewing $L(a)$ as an observed random measure, and then using functional data analysis of distributions (see, e.g. Bigot et al., 2018; Delicado, 2011), for us $L(a)$ is a directly observed (in the discretized form) quantity of interest. This aspect also distinguishes our problem from situations where the random measures are not directly observable (Bigot et al., 2018; Petersen & Müller, 2016). More specifically, methods of functional data analysis for densities do not apply to our setting because the functions need to be bounded below and have a constrained integral (see, e.g. Han et al., 2020; Petersen & Müller, 2016, and references therein), whereas $L(a)$ is bounded both above and below and does not have a constrained integral.

A future direction is to consider the occupation time $L(a)$ with a multi-dimensional index a. This scenario arises when the data contain additional physiological information such as heart rate or blood pressure, where it would be necessary to treat $X(\cdot)$ as a vector-valued stochastic process. Another future direction is to develop inference for the local times corresponding to occupation time measures. In order to estimate the mean local time, namely the derivative of $E\{L(a)\}$, we could use a kernel estimator based on the sample mean of $E\{L(a)\}$ (cf., hazard function estimation based on the Nelson—Aalen estimator; Andersen et al., 1993, section IV.2.1). This way, local differences in occupation time could be explored in a more detailed way. Still another interesting direction for future work would be goodness-of-fit testing for parametric models of mean occupation time. These could be used to provide parsimonious descriptions of mean occupation time for comparing groups of subjects. One possibility is a Weibull-type model (cf., the survival function of a Weibull distribution); our nonparametric estimates of the mean occupation time displayed in Figure 5 based on the NHANES data, however, suggest marked departures from a Weibull model, so more flexible types of parametric models would need to be developed.

ACKNOWLEDGEMENTS

The research of Hsin-wen Chang was partially supported by National Science and Technology Council of Taiwan under Grant 110-2118-M-001-013-MY2. The research of Ian McKeague was partially supported by NIH Grant AG062401. The authors thank the referees, Yifei Sun and Shih-Hao Huang for many helpful comments, and Yu-Ju Wang, Yan-Yu Chen and Shu-Hsiang Wang for computational support.

ORCID

Hsin-wen Chang https://orcid.org/0000-0003-4566-7047

REFERENCES

Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993) *Statistical models based on counting processes*. New York: Springer.

Backenroth, D., Shinohara, R.T., Schrack, J.A. & Goldsmith, J. (2020) Nonnegative decomposition of functional count data. *Biometrics*, 76, 1273–1284.

Ben Adda, F. & Cresson, J. (2001) About non-differentiable functions. *Journal of Mathematical Analysis and Applications*, 263, 721–737.

Besicovitch, A. (1929) On Lipschitz numbers. *Mathematische Zeitschrift*, 30, 514–519.

Bigot, J., Gouet, R., Klein, T. & López, A. (2018) Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics*, 12, 2253–2289. Available from: https://doi.org/10.1214/18-EJS1400

Bravo, F. (2003) Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika*, 90, 881–890. Available from: https://doi.org/10.1093/biomet/90.4.881

Cai, T.T. & Yuan, M. (2011) Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *The Annals of Statistics*, 39, 2330–2355. Available from: https://doi.org/10.1214/11-AOS898

Cao, G., Yang, L. & Todem, D. (2012) Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, 24, 359–377.

Chang, H.-w. & McKeague, I.W. (2019) Nonparametric testing for multiple survival functions with non-inferiority margins. *The Annals of Statistics*, 47, 205–232.

Choi, H. (2017) fregion: confidence regions and bands for functional data. R package version 0.0934. Available from: https://github.com/hpchoi/fregion

Choi, H. & Reimherr, M. (2018) A geometric approach to confidence regions and bands for functional parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 239–260.

Cuesta-Albertos, J. & Febrero-Bande, M. (2010) A simple multiway ANOVA for functional data. *Test*, 19, 537–557.

Cuevas, A. (2014) A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1–23.

Cuevas, A., Febrero, M. & Fraiman, R. (2004) An ANOVA test for functional data. *Computational Statistics & Data Analysis*, 47, 111–122.

Degras, D.A. (2011) Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, 21, 1735–1765.

Degras, D. (2016) SCBmeanfd: simultaneous confidence bands for the mean of functional data. R package version 1.2.2. Available from: https://CRAN.R-project.org/package=SCBmeanfd

Degras, D.A. (2017) Simultaneous confidence bands for the mean of functional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9, e1397.

Delicado, P. (2011) Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55, 401–420. Available from: https://doi.org/10.1016/j.csda.2010.05.008

Dette, H., Kokot, K. & Aue, A. (2020) Functional data analysis in the Banach space of continuous functions. *Annals of Statistics*, 48, 1168–1192.

Fan, J. & Lin, S.-K. (1998) Test of significance when data are curves. *Journal of the American Statistical Association*, 93, 1007–1021. Available from: https://doi.org/10.1080/01621459.1998.10473763

Górecki, T. & Smaga, L. (2015) A comparison of tests for the one-way ANOVA problem for functional data. *Computational Statistics*, 30, 987–1010.

Górecki, T. & Smaga, L. (2018) fdANOVA: an R software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*. Available from: https://doi.org/10.1007/s00180-018-0842-7

Gorman, E., Hanson, H., Yang, P., Khan, K., Liu-Ambrose, T. & Ashe, M. (2014) Accelerometry analysis of physical activity and sedentary behavior in older adults: a systematic review and data analysis. *European Review of Aging and Physical Activity: Official Journal of the European Group for Research into Elderly and Physical Activity*, 11, 35–49.

Hagood, J.W. & Thomson, B.S. (2006) Recovering a function from a Dini derivative. *The American Mathematical Monthly*, 113, 34–46.

Hall, P., Müller, H.-G. & Yao, F. (2008) Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 703–723. Available from: https://doi.org/10.1111/j.1467-9868.2008.00656.x

Han, K., Müller, H.-G. & Park, B.U. (2020) Additive functional regression for densities as responses. *Journal of the American Statistical Association*, 115, 997–1010. Available from: https://doi.org/10.1080/01621459.2019.1604365

Horsnell, G. (1953) The effect of unequal group variances on the *F*-test for the homogeneity of group means. *Biometrika*, 40, 128–136.

Huang, L., Bai, J., Ivanescu, A., Harris, T., Maurer, M., Green, P. et al. (2019) Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *Journal of the American Statistical Association*, 114, 553–564.

Kitamura, Y. (2007) *Empirical likelihood methods in econometrics: theory and practice, vol. 3 of Econometric Society Monographs*. Cambridge MA: Cambridge University Press, pp. 174–237.

Kitamura, Y., Santos, A. & Shaikh, A.M. (2012) On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 80, 413–423.

Matthews, C.E., Chen, K.Y., Freedson, P.S., Buchowski, M.S., Beech, B.M., Pate, R.R. et al. (2008) Amount of time spent in sedentary behaviors in the United States, 2003–2004. *American Journal of Epidemiology*, 167, 875–881.

Migueles, J.H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M. et al. (2017) Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations. *Sports Medicine*, 47, 1821–1845.

Mukerjee, R. (1994) Comparison of tests in their original forms. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)*, 56, 118–127.

Nair, V.N. (1984) Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26, 265–275.

NASA. (2016) NASA Earth Observatory: monitoring sea ice. Available from: https://earthobservatory.nasa.gov/features/SeaIce/page2.php

Owen, A.B. (2001) *Empirical likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

Pauly, M., Brunner, E. & Konietschke, F. (2015) Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 461–473.

Petersen, A. & Müller, H.-G. (2016) Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44, 183–218. Available from: https://doi.org/10.1214/15-AOS1363

Samorodnitsky, G. (2016) *Stochastic processes and long range dependence*. Berlin: Springer.

Sang, P., Wang, L. & Cao, J. (2019) Weighted empirical likelihood inference for dynamical correlations. *Computational Statistics & Data Analysis*, 131, 194–206.

Song, J., Swartz, M.D., Gabriel, K.P. & Basen-Engquist, K. (2019) A semiparametric model for wearable sensor-based physical activity monitoring data with informative device wear. *Biostatistics*, 20, 287–298.

Staudenmayer, J., Zhu, W. & Catellier, D.J. (2012) Statistical considerations in the analysis of accelerometry-based activity monitor data. *Medicine and Science in Sports and Exercise*, 44, S61–S67.

United States National Center for Health Statistics. (2005–2006) National health and nutrition examination survey data. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2005

Uno, H., Tian, L., Claggett, B. & Wei, L.J. (2015) A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves. *Statistics in Medicine*, 34, 3680–3695.

van der Vaart, A.W. (2000) *Asymptotic statistics*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge, MA: Cambridge University Press.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257–295.

Wang, H., Zhong, P.-S., Cui, Y. & Li, Y. (2018) Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 343–364.

Witze, A. (2019) Dramatic sea-ice melt caps tough arctic summer. *Nature*, 573, 320–321.

Wright, S.P., Brown, T.S.H., Collier, S.R. & Sandberg, K. (2017) How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 312, R358–R367.

Wrobel, J., Zipunnikov, V., Schrack, J. & Goldsmith, J. (2019) Registration for exponential family functional data. *Biometrics*, 75, 48–57.

Yuan, A., Fang, H.-B., Li, H., Wu, C.O. & Tan, M.T. (2020) Hypothesis testing for multiple mean and correlation curves with functional data. *Statistica Sinica*, 30, 1095–1116.

Zhang, J.-T. (2013) *Analysis of variance for functional data*. Boca Raton, FL: Chapman & Hall/CRC.

Zhang, J.-T. & Liang, X. (2014) One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics*, 41, 51–71.

Zhang, J.-T., Cheng, M.-Y., Wu, H.-T. & Zhou, B. (2019) A new test for functional one-way ANOVA with applications to ischemic heart screening. *Computational Statistics & Data Analysis*, 132, 3–17. Special Issue on Biostatistics.

Zhang, Y., Li, H., Keadle, S., Matthews, C. & Carroll, R. (2019) A review of statistical analyses on physical activity data collected from accelerometers. *Statistics in Biosciences*, 11, 465–476.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Chang, H. & McKeague, I.W. (2022) Empirical likelihood-based inference for functional means with application to wearable device data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5), 1947–1968. Available from: https://doi.org/10.1111/rssb.12543