Received: 23 January 2022 | Accepted: 3 February 2023

DOI: 10.1111/biom.13846

BIOMETRIC PRACTICE

Biometrics WILEY

A nonparametric test of group distributional differences for hierarchically clustered functional data

Alexander S. Long[1] | **Brian J. Reich**[1] | **Ana-Maria Staicu**[1] | **John Meitzen**[2]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

[2]Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA

Correspondence
Brian J. Reich, Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA.
Email: bjreich@ncsu.edu

Abstract

Biological sex and gender are critical variables in biomedical research, but are complicated by the presence of sex-specific natural hormone cycles, such as the estrous cycle in female rodents, typically divided into phases. A common feature of these cycles are fluctuating hormone levels that induce sex differences in many behaviors controlled by the electrophysiology of neurons, such as neuronal membrane potential in response to electrical stimulus, typically summarized using a priori defined metrics. In this paper, we propose a method to test for differences in the electrophysiological properties across estrous cycle phase without first defining a metric of interest. We do this by modeling membrane potential data in the frequency domain as realizations of a bivariate process, also depending on the electrical stimulus, by adopting existing methods for longitudinal functional data. We are then able to extract the main features of the bivariate signals through a set of basis function coefficients. We use these coefficients for testing, adapting methods for multivariate data to account for an induced hierarchical structure that is a product of the experimental design. We illustrate the performance of the proposed approach in simulations and then apply the method to experimental data.

KEYWORDS

bivariate functional data, functional data analysis, hierarchically clustered data, multivariate testing

1 | INTRODUCTION

Biological sex and gender are critical variables for biomedical research, especially for addressing underserved aspects of women's health (Arnegard et al., 2020; Galea et al., 2020; Tannenbaum et al., 2019). Complicating this consideration is the presence of sex-specific natural hormone cycles in both females and males, such as the menstrual cycle in female humans and the estrous cycle in female rodents, which can influence experimental outcomes (Mamlouk et al., 2020; Proaño et al., 2018). These cycles can be divided into phases featuring different hormone concentrations. Hormone-level fluctuations induce sex differences in many behaviors, including those related to motivation and disorders such as depression and addiction. These behaviors are controlled by the electrophysiology of specific neurons that communicate with each other between designated brain regions via electrical impulses called action potentials. Thus, it is of high research interest to determine if the properties of neurons change throughout the estrous cycle.

The most prominent and widely employed experimental procedure is the whole-cell patch clamp (WHPC), which can analyze how the neuron membrane potential changes

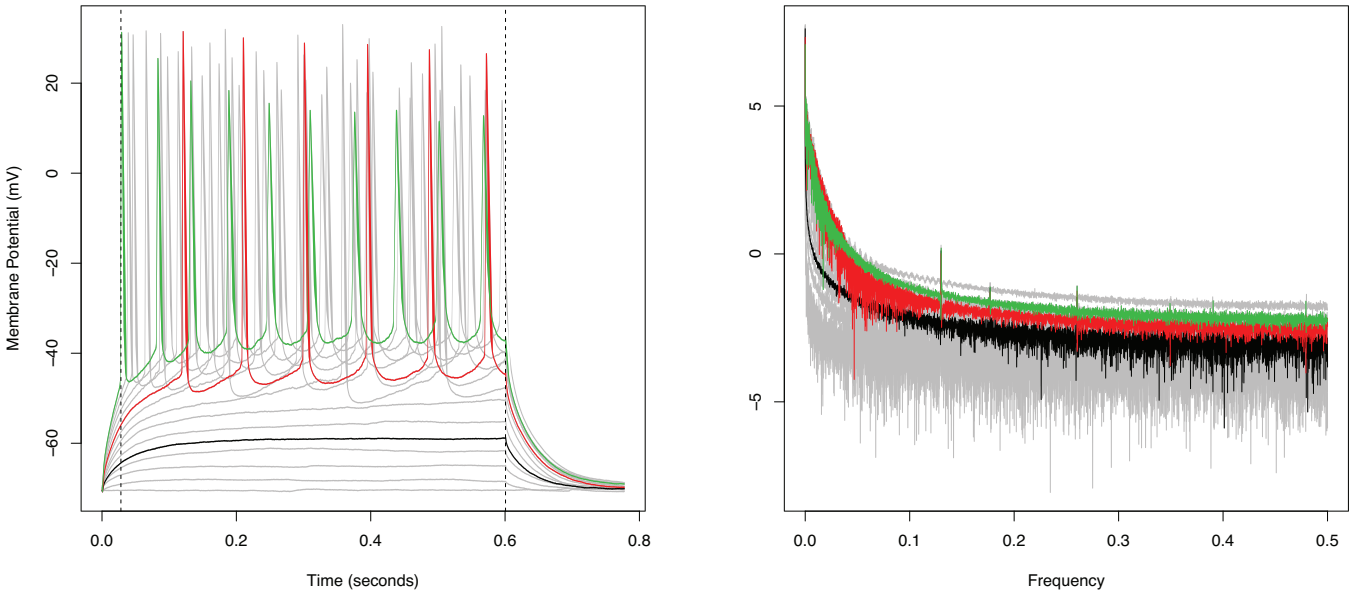

FIGURE 1 Left: The observed membrane potential curves are shown for all currents (0 to +0.14 nA) injected for one replicate of one medium spiny neuron from a rat in the diestrus phase of the estrous cycle. Curves corresponding to currents of +0.04, +0.09, and +0.14 nA are shown in black, red, and green, respectively. Right: The log-periodogram of the membrane potential curves shown in the left panel. The colored curves correspond to the membrane potential curves shown on the left in the same color. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

with various current voltage injected during a fixed period of time. These electrophysiological properties can be studied in vitro by measuring the membrane potential of a neuron in response to artificial stimulus like an electrical current. Proaño et al. (2018) and Cao et al. (2018) are two examples of these experiments. An example of the observed membrane potential of a neuron from such an experiment is shown in the left most plot in Figure 1; the membrane potential is depicted in response to a constant amount of current being applied to the neuron starting at 0 s and ending at 0.6 s for an increasing amount of current. Application of higher current results in the membrane potential increasing, and eventually generating action potentials seen in the plot as spikes.

Existing approaches to analyze such data rely heavily on summaries of the rich data produced by WHPC experiments. One example of this is the heavy emphasis on a priori-defined experimental metrics; while they have strong neurological justification, they limit analysis to only those assessed metrics. For example, the observed membrane potential curves can be summarized using features such as action potential frequency and amplitude. A one-way analysis of variance (ANOVA) or Kruskal–Wallis test can be used to test for group differences with an adjustment for multiple comparisons if necessary. Alternatively, principal components analysis (PCA) of these experimental metrics has been used in the analysis of similar membrane potential data (Druckmann et al., 2012; Hernáth et al., 2019). Developing methods that better account for the complex dependence and structure of these data may enhance discovery beyond what is possible using these neurologically relevant metrics.

In this paper, we describe a method to test for differences in the distribution of the membrane potential behavior in response to stimulus between phases of the estrous cycle in rats that does not require defining a priori metrics of interest. We accomplish this by working with the periodogram of the observed membrane potential and viewing it as the realization of a random function observed at a finite grid of timepoints. There has been a considerable amount of development of methods for testing for distributional differences between groups in independent functional data. Testing for equality of the mean function has been proposed by Cuesta-Albertos and Febrero-Bande (2010), Horváth et al. (2013), Zhang and Liang (2014), and Zhang et al. (2019). Testing for equality of the covariance function has been proposed by Fremdt et al. (2013) and Paparoditis and Sapatinas (2016). More generally, Pomann et al. (2016) and Wynne and Duncan (2020) tested for differences in distribution.

A limitation of these testing procedures is that they require independent functional data. In the motivating application, multiple neurons are observed from each rat, and further multiple observations are made on each neuron. This experimental design naturally induces a hierarchical structure on the data, and thus, an assumption of independence is not reasonable. Such clustered data are commonly modeled using functional mixed effects

models. For example, Di et al. (2009) developed multilevel functional PCA (MFPCA); Li et al. (2015) and Xu et al. (2018) proposed extensions to MFPCA for the analysis of three-level hierarchies. Testing procedures for clustered functional data have been proposed as well. Abramovich and Angelini (2006) and Antoniadis and Sapatinas (2007) consider testing for mean differences in a mixed model framework. Staicu et al. (2015) proposed an L^2-norm-based testing procedure for the group mean differences in clustered data. Xu et al. (2018) introduced a testing procedure for hierarchically clustered functional data, however only considered tests of the form of a smooth mean function.

An additional complication of this motivating dataset is the importance of the applied electrical stimulus. As Figure 1 shows, the membrane potential curves can vary significantly depending on the applied current. We could incorporate the effect of the stimulus using function-on-scalar regression (Ramsay & Silverman, 2005). Statistical inferences for such models have been studied without (Fan & Zhang, 2000) and with a hierarchical structure (Zhu et al., 2012) as in the motivating dataset. Such models require assumptions about the relationship between the stimulus and the observed membrane potential that may not be supported by the biological processes responsible for these data. Alternatively, we estimate this relationship using a set of empirical basis functions to characterize this relationship. Specifically, we consider the membrane potential to be the realization of a latent process that depends on the stimulus and we will use the eigenfunctions of an appropriate covariance matrix to describe the variation of the membrane potential associated with the stimulus. The membrane potentials are observed densely in time, but for comparatively few unique levels of current. We note the similarity to longitudinal functional data; rather than functional observations being made at several times for each subject, we make functional observations at several levels of applied current. Thus, we utilize existing methods for longitudinal functional data (Chen et al., 2017; Chen & Müller, 2012; Park & Staicu, 2015).

We present the membrane potential variation using the same data-driven basis with coefficients that depend on the applied current. The basis coefficients are recovered as the inner product between the response function and the estimated basis functions. As a result, they will preserve the dependence of the response profiles. We utilize methods from longitudinal functional data analysis to estimate the basis system. A multivariate testing procedure is then applied to the coefficients of the basis function expansion. To account for the known hierarchical structure in the data, we approximate the null distribution of the test statistic by bootstrapping over independent observational units.

2 | DATA DESCRIPTION

The dataset that motivates this work is from an experiment employing WHPC technique to assess the electrophysiological properties of medium spiny neurons in the acute brain slice preparation of the nucleus accumbens core of adult female rats (see Proaño et al., 2018 for details). The overall goal of this experiment was to test the hypothesis that these properties change across phases of the estrous cycle. We focus on data generated when the recorded neurons were injected with excitatory current for 0.6 s and the membrane potential was measured for the duration. The time series of the measured membrane potentials while current was being injected was observed. The amount of current injected started at 0 nA, to provide a reading of the baseline resting membrane potential of the neuron, and the current was then increased until there was an observed decrease in the number of action potentials as measured by the scientist performing the experiment.

The experiment included 26 rats, observed across three phases of the estrous cycle: diestrus (11), proestrus (8), and estrus (7). From each rat, one to four neurons are collected, and additionally, there are multiple replications per neuron. Thus, the data have a natural nested hierarchical structure: (i) estrous cycle phase, (ii) rat, (iii) neurons, and (iv) replicates. For a single replicate within a neuron, the increase in current, typically +0.01 nA, continues until an observed decrease in action potential frequency, indicating the limit of the physiological range of the neuron's response properties. All neurons had at least seven different, nonzero currents injected with over half of the neurons receiving at least 18 different currents.

An example of the data collected from a single replication of a neuron from a rat in the diestrus phase is provided in Figure 1. The left plot shows the membrane potential response to all levels of current injected into the neuron. During current injection, there is an increase in the membrane potential until it plateaus, typical of medium spiny neurons but not all neuron types. After 0.6 s, the current injection stops and the membrane potential returns to the resting membrane potential. If sufficient current is applied, causing a large enough depolarization in membrane potential, action potentials can be generated, seen as rapid spikes in the membrane potential. Once an initial action potential is generated for fixed level of current, in this neuron type, they typically repeat at an approximately constant frequency while current is being applied.

Due to data having both smooth and spiky features across varying currents, it is reasonable to represent the data in the frequency domain. We use a Fourier transform to decompose the current-specific curves into their constituent frequencies and estimate the spectral density for each curve. Before taking the Fourier transform, all the

current-specific curves measured on a fixed neuron and replicate are truncated to focus on a time interval that has scientific interpretation (see vertical lines in Figure 1). By restricting to this region, the process is appropriately stationary in which case the Fourier transformation retains all the information in the original data. To assess for sensitivity of the results to the selection of this truncation point, multiple points were considered and had minimal impact on the subsequent results.

On the right panel in Figure 1, we show the periodogram of each current-specific membrane potential curve for a single replicate from a single neuron, on a log scale. When the current injected to the neuron is such that no action potentials are generated, the log-periodogram has a spike at a frequency of 0 with very small values at all other frequencies. With increasing current causing a larger depolarizing change in the membrane potential, and eventually causing the generation of action potentials, the value of the log-periodogram at higher frequencies increased.

As seen in the periodogram plot in Figure 1, the fixed-current spectral profiles look like a noisy realization of smooth monotone decreasing signals. To account for the different current levels, we view the log-periodogram to be a realization of a bivariate function depending on both the frequency and the current applied to the neuron. In Figure 2, the log-periodogram is shown as a bivariate function for three neurons in the diestrus and estrus phases. It appears that the log-periodograms from the diestrus group have higher values at lower currents and low frequencies than those from the estrus group. If there are differences in the electrophysiological properties of the neuron across the phases of the estrous cycle, we expect those to be exhibited by differences in the bivariate log-periodogram.

3 | STATISTICAL FRAMEWORK

3.1 | Model framework

Consider the following hierarchical data: for each group $g = 1, \dots, G$, we observe measurements on a number of units $r = 1, \dots, n_g$, and for each subunit $i = 1, \dots, n_{gr}$ within a unit, the observed data are $[\{(t_{gri,k}, u_{gri,\ell}), Y_{gri,k\ell}\}_{(k,\ell)}]_i$, where $k = 1, \dots, K_{gri}$ and $\ell = 1, \dots, L_{gri}$. We assume that $Y_{gri,k\ell}$ is an evaluation of a bivariate function observed with noise at $(t_{gri,k}, u_{gri,\ell})$; in other words, let $X_{gri}(\cdot,\cdot) : \mathcal{T} \times \mathcal{U} \to \mathbb{R}$, such that $Y_{gri,k\ell} = X_{gri}(t_{gri,k}, u_{gri,\ell}) + \epsilon_{gri,kl}$, where $\epsilon_{gri,kl}$ is measurement error. We assume that K_{gri} is large and the grid of points $\{t_{gri,k} : k\}$ is fine in $\mathcal{T}$ and consider the case when L_{gri} is small for each i, but $\{u_{gri,\ell} : \ell, i, r, g\}$ is dense in $\mathcal{U}$. In our data application, g denotes estrous cycle phase, r denotes rat, i denotes replicate within rat, and $Y_{gri,k\ell}$ is the log-periodogram at frequency t_k and current u_ℓ; we do not explicitly account for the neuron level to simplify notation. Regarding current, we directly use a normalized current based on the maximum current for each observation and take $\mathcal{U} = [0, 1]$. By an abuse of notation, we assume that $X_{gri}(\cdot,\cdot) \overset{d}{=} X_g(\cdot,\cdot)$ for all r, i, where the notation $\overset{d}{=}$ denotes that the random quantities have the same distribution. This is justified because all neurons belong to the same region of the brain and information such as relative location of neurons is lost due to the collection methods. Our objective is to develop a testing procedure to formally assess

$$H_0 : X_1(\cdot,\cdot) \overset{d}{=} \dots \overset{d}{=} X_G(\cdot,\cdot), \tag{1}$$

versus the alternative that $X_g(\cdot,\cdot) \overset{d}{\neq} X_{g'}(\cdot,\cdot)$ for some $g \neq g' = 1, \dots, G$.

Testing the equality of a group of curves is not new; for example, Staicu et al. (2014), Pomann et al. (2016), and Zhang et al. (2019) study this problem for independent and/or univariate curves. In our situation, the bivariate structure of the curves, with mixed dense/sparse sampling design, along with the complex hierarchical dependence increase the challenge.

We propose to model $X_{gri}(t,u) = \mu(t,u) + V_{gri}(t,u)$, where $\mu(t,u)$ is the overall mean function and $V_{gri}(t,u)$ is the subunit deviation. Let $\{\phi_p(\cdot)\}_{p\geq 1}$ be an orthonormal basis in $L^2(\mathcal{T})$ and represent the deviations as $V_{gri}(t,u) = \sum_{p=1}^{\infty} \xi_{gri,p}(u)\phi_p(t)$ where the $\xi_{gri,p}(u) = \int_{\mathcal{T}} V_{gri}(t,u)\phi_p(t)dt$ are the corresponding basis coefficients that have mean zero and are uncorrelated across g, r, and p. As in Chen and Müller (2012), Park and Staicu (2015), and Chen et al. (2017), we then propose a similar decomposition of the $\xi_{gri,p}(u)$'s. That is, $\xi_{gri,p}(u) = \sum_{q=1}^{\infty} \zeta_{gri,pq}\psi_{pq}(u)$ where $\{\psi_{pq}(\cdot)\}_{q\geq 1}$ is an orthonormal basis in $L^2(\mathcal{U})$ and the $\zeta_{gri,pq}$'s are the corresponding basis coefficients that are mean zero. Thus, combining all components, we obtain $V_{gri}(t,u) = \sum_{p=1}^{\infty}\sum_{q=1}^{\infty} \zeta_{gri,pq}\phi_p(t)\psi_{pq}(u)$.

There are typically two possible ways to select the basis system for the above representation. One option is to use a prespecified set of basis functions. We pursue a different option: we select $\{\phi_p(\cdot)\}_{p\geq 1}$ to be the eigenbasis of the marginal covariance function $\Sigma_{\mathcal{T}}(t,t') = \int_{\mathcal{U}} \Sigma(t,t',u,u)f(u)du$, where $f(\cdot)$ is the sampling density of u; similar to Park and Staicu (2015). We also select $\{\phi_{p,q}(\cdot)\}_{q\geq 1}$ to be the eigenbasis of the covariance of the coefficients of the initial decomposition, $\Sigma_{\mathcal{U},p}(u,u') = Cov\{\xi_{gri,p}(u), \xi_{gri,p}(u')\}$. This representation allows us to explain the variation in the bivariate functional data by sets of eigenfunctions for each argument, t and u, separately. Furthermore, this framework allows us to extract

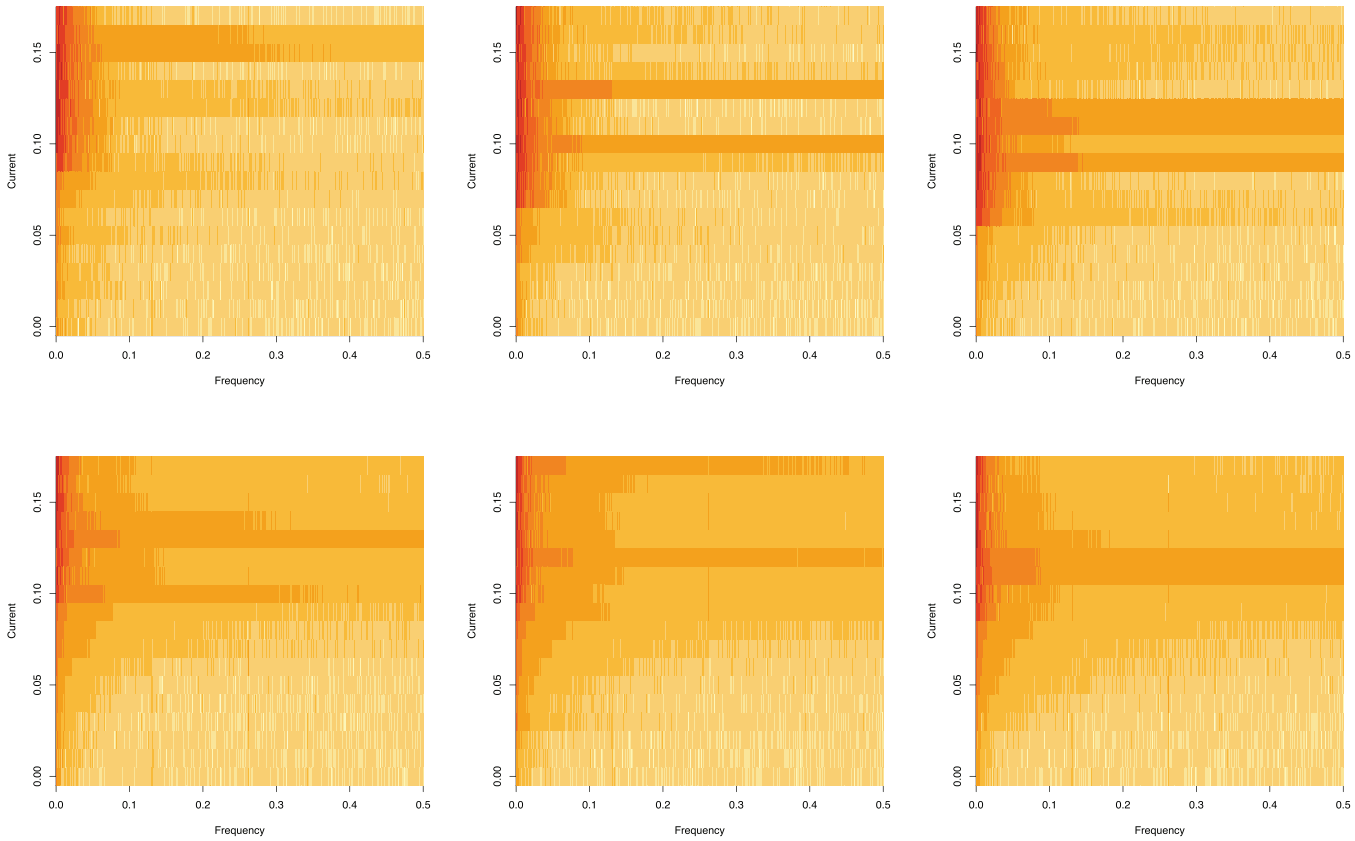

FIGURE 2 Log-periodogram across current and frequency for three neurons from a single rat in the diestrus (top row) and estrus phase (bottom). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

the main features of the bivariate signals through the set of basis function coefficients. This approach has recently been considered by Scheffler et al. (2018).

In practice, we truncate the infinite basis functions; let P and $Q_1, \dots, Q_P$ denote the truncation for the bases $\{\phi_p(\cdot)\}$ and $\{\psi_{p,q}(\cdot)\}$, respectively. It follows that the vector $\boldsymbol{\zeta}_{gri} = (\boldsymbol{\zeta}_{gri,1}^T, \dots, \boldsymbol{\zeta}_{gri,P}^T)^T$, where $\boldsymbol{\zeta}_{gri,p} = (\zeta_{gri,p1}, \dots, \zeta_{gri,pQ_p})^T$, represents a feature extraction of the bivariate signal, $X_{gri}(\cdot,\cdot)$. We thus reduce the testing the null hypothesis (1) to the hypothesis that the distribution of the $\boldsymbol{\zeta}_{gri}$ is not varying across the groups. That is, assume $\boldsymbol{\zeta}_{gri} \sim f_g$ where f_g is any probability distribution with sample space $\mathbb{R}^{\sum_{p=1}^{P} Q_p}$ that depends on the group, g; the null hypothesis of interest is reduced to

$$H_0 : f_1 = \cdots = f_G. \tag{2}$$

In this regard, we consider testing procedures from the multivariate statistics literature; to account for the hierarchical dependence in the data, we propose a bootstrap-based null distribution approximation.

In the next section, we discuss estimation of the model components, including selection of the number of basis functions and estimation of the basis function coefficients. In Section 4, we describe the testing procedure.

3.2 | Estimation

The roadmap of the estimation procedure is: first, estimate the marginal mean function. Using the centered data, we then estimate the marginal covariance function $\Sigma_{\mathcal{T}}(t,t')$ and its eigencomponents. The coefficients of this initial decomposition are then used to estimate the marginal covariance functions, $\Sigma_{\mathcal{U},p}(u,u')$, and their eigencomponents. We utilize existing methods for the estimation of all model components; additional details of these methods are provided in the Supporting Information.

We estimate the marginal mean function $\mu(t,u)$ by using the bivariate sandwich smoother (Xiao et al., 2013) and a working independence assumption. In the numerical investigation, we use the sandwich smoother constructed using cubic B-spline basis functions for t and u and select the tuning parameters by generalized cross validation (GCV).

Let $\widetilde{Y}_{gri,k\ell} = Y_{gri,k\ell} - \widehat{\mu}(t_k, u_\ell)$ be the demeaned data. We use the demeaned data to first estimate the marginal

covariance function $\Sigma_{\mathcal{T}}(t,t')$. To estimate this covariance function, and subsequently, the eigenfunctions of this covariance, we use the fast covariance estimation, or FACE estimator, (Xiao et al., 2016) that is a smoothing of the traditional sample covariance,

$$S(t,t') = \sum_{g=1}^{G}\sum_{r=1}^{n_g}\sum_{i=1}^{n_{gr}}\sum_{\ell=1}^{L_{gri}} \widetilde{Y}_{gri}(t,u_\ell)\widetilde{Y}_{gri}(t',u_\ell) \Bigg/ \left(\sum_{g=1}^{G}\sum_{r=1}^{n_g}\sum_{i=1}^{n_{gr}} L_{gri}\right).$$

This estimator is a special case of the sandwich smoother used to estimate the mean function. As with the sandwich smoother, this method depends on a smoothing parameter that can be selected using GCV. The final estimator is adjusted to be symmetric and positive definite by zeroing the negative eigenvalues. Let $\{\widehat{\phi}_p(\cdot),\widehat{\lambda}_p\}_{p\geq 1}$ be the pairs of estimated eigenfunctions and eigenvalues obtained by spectral decomposition of the smoothed estimate of $\Sigma_{\mathcal{T}}(\cdot,\cdot)$. The truncation parameter, P, can be determined based on a prespecified percentage of variance explained (PVE) using the estimated eigenvalues (Di et al., 2009).

Let $\widehat{\xi}_{gri,p}(u_\ell) = \int \widetilde{Y}_{gri}(t,u_\ell)\widehat{\phi}_p(t)dt$ be the estimated coefficient of the pth eigenfunction for the ℓth current applied to the ith neuron in the gth group and rth rat; $\widehat{\xi}_{gri,p}(u_\ell)$ can be approximated well via numerical integration because $\{t_{gri,k} : k\}$ is dense in $\mathcal{T}$. We use these estimated coefficients separately for each p to estimate $\Sigma_{\mathcal{U},p}(u,u')$ and its eigencomponents. As the data need not be observed on a regular grid in $\mathcal{U}$ as in $\mathcal{T}$, we estimate $\Sigma_{\mathcal{U},p}(u,u')$ and its eigencomponents using methods for sparse functional data (Yao et al., 2005) instead of using the same method for estimating $\Sigma_{\mathcal{T}}(t,t')$. This approach also obtains a smoothed estimate of $\Sigma_{\mathcal{U},p}(u,u')$ by smoothing the raw covariances, here calculated as $S_{p,gri}(u_\ell,u_{\ell'}) = \widehat{\xi}_{gri,p}(u_\ell)\widehat{\xi}_{gri,p}(u_{\ell'})$. Let $\{\widehat{\psi}_{pq}(\cdot),\widehat{\gamma}_{pq}\}_{q\geq 1}$ be the pairs of estimated eigenfunctions and eigenvalues obtained by spectral decomposition of the smoothed estimate of $\Sigma_{\mathcal{U},p}(\cdot,\cdot)$. As when choosing P, the truncation parameters, Q_p, can be determined using PVE. Upon estimation of the eigenfunctions of $\Sigma_{\mathcal{U},p}(\cdot,\cdot)$, the scores, $\zeta_{gri,pq}$, can be estimated using a mixed model framework as described in Yao et al. (2005).

4 | TESTING PROCEDURE

In this section, we describe the testing procedure. Recall that the null hypothesis of interest (1) is simplified to the null hypothesis that the vector of basis function coefficients has the same distributions across groups; see null hypothesis (2). When considering the alternative hypothesis, while we emphasized modeling the marginal mean and covariance functions (Section 3.2), we make no restrictions on how the distributions may differ between groups.

To test this hypothesis, k-sample multivariate testing procedures can be used. Examples of such testing procedures are Bathke et al. (2008) and Heller et al. (2013). We use the Heller–Heller–Gorfine (HHG) test (Heller et al., 2013) because of its minimal assumptions and sensitivity to many forms of deviations from the null hypothesis; we do note that many other multivariate tests can be used similarly depending on the objectives of the analysis. This test statistic is based on all pairwise norm differences of the data. We describe this test as if the vector of basis coefficients were known; in practice, we replace the basis coefficients by their estimates obtained as described in Section 3.2. Consider a fixed pair of observations, indexed by (1) g, r, and i and (2) g', r', and i', with $i \neq i'$; denote the norm difference between the estimated coefficients for these two observations, $R_0 = \|\boldsymbol{\zeta}_{gri} - \boldsymbol{\zeta}_{g'r'i'}\|$. This value R_0 depends on the indices g, g', r, r', i, i', but we suppress this dependence until the end for notational simplicity. Using R_0, we can create and summarize a $2 \times G$ contingency table using the remaining data as follows. For $g^* = 1, \ldots, G$, let

$$A_{1g^*} = \sum_{g''=1}^{G}\sum_{r''=1}^{n_g}\sum_{i''=1}^{n_{gr}} I(\|\boldsymbol{\zeta}_{gri} - \boldsymbol{\zeta}_{g''r''i''}\| > R_0)I(g'' = g^*) \text{ and}$$

$$A_{2g^*} = \sum_{g''=1}^{G}\sum_{r''=1}^{n_g}\sum_{i''=1}^{n_{gr}} I(\|\boldsymbol{\zeta}_{gri} - \boldsymbol{\zeta}_{g''r''i''}\| \leq R_0)I(g'' = g^*).$$

Additionally, denote by $A_{i^*\cdot}$ and $A_{\cdot g^*}$ the row and column sums. Lastly, denote by $T(gri;g'r'i')$ the Pearson's score for this partition; that is,

$$T(gri;g'r'i') = \sum_{i^*=1}^{2}\sum_{g^*=1}^{G}\frac{(A_{i^*g^*} - E_{i^*g^*})^2}{E_{i^*g^*}}$$

$$\text{where } E_{i^*g^*} = \frac{A_{i^*\cdot}A_{\cdot g^*}}{\sum_{g=1}^{G}\sum_{r=1}^{n_g} n_{gr}}.$$

The overall test statistic for the sample can be found by summing over all pairs; that is, $T_{HHG} = \sum_{g,g'=1}^{G}\sum_{r,r'=1}^{n_g}\sum_{i,i'=1;i\neq i'}^{n_{gr}} T(gri;g'r'i')$.

Heller et al. (2013) developed the null distribution, and considered an approximation based on random permutations of the group assignments, of the classical HHG test under the assumption that the observations within a group are independent and identically distributed. This assumption does not hold in our case, where recall we only assume independence of $X_{gri}(\cdot,\cdot)$ over r; applying the testing procedure while ignoring the dependence results in an inflated type I error. We propose a bootstrap-based

Algorithm 1 Resampling of the unit level data

1: **for** $b \in \{1, \ldots, B\}$ **do**
2: Re-sample the group-unit index pairs from $\{(1,1), \ldots, (1,n_1), \ldots, (G,n_G)\}$ with replacement. Let $R^{(b)}$ be the resulting sample
3: Define the bth bootstrap data by:
$\text{data}^{(b)} = [\{Y_{g^*r^*i,k\ell}\}_{(k,\ell)} : (g^*, r^*) \in R^{(b)}, i = 1, \ldots, n_{g^*r^*}]$
4: Reassign the group indices by unit, so that $g = 1$ for the first n_1 units, $g = 2$ for the next n_2 units, and so on. Re-define the bth bootstrap data accordingly.
5: Using $\text{data}^{(b)}$, estimate the model components and recover the estimated coefficient vectors, $\widehat{\zeta}^{(b)}_{gri}$, as described in Section 3.2
6: Calculate the HHG test statistic and denote it by $T^{(b)}_{HHG}$
7: **end for**
8: Calculate the p-value as $\sum_{b=1}^{B} I(T^{(b)}_{HHG} > T_{HHG})/B$

approach to approximate the null distribution of the HHG test by modifying the permutation procedure to account for the hierarchical structure of the data; see Algorithm 1. For each permutation iteration, the observed data are resampled with replacement by unit-level identifier. We briefly comment on step 5 of Algorithm 1; reestimation of model components with each iteration is computationally burdensome, although it is necessary to prevent the results of the test from being conditional on the estimated eigenfunctions and better accounts for the uncertainty of that estimation step. The test statistic calculated using the observed data is then compared to the distribution of test statistics after resampling; a p-value is estimated by the sample proportion of observing a value of the test statistic as large or larger in the bootstrap set of statistics.

5 | SIMULATION STUDY

In this section, we present simulation studies to illustrate the performance of our proposed approach described in Section 3. We consider two distinct frameworks. First, we generate multivariate data to isolate the performance of the resampling method in a simpler setting. Then we generate functional data to assess the performance of the entire method as presented. For both frameworks, we describe the scenarios used to assess the performance of our method, introduce comparative approaches, and present the results.

5.1 | Framework 1: Multivariate data

5.1.1 | Generation of multivariate data

We evaluate the performance of the proposed approach by first generating data such that the null hypothesis of interest, that the distributions of the responses are the same across groups, is true to evaluate the type I error rate; we also generate data under different forms of deviations from the null hypothesis to assess power.

We generate data $Y_{gijk} \in \mathbb{R}^P$ as

$$Y_{gijk,p} = \alpha_{gi,p} + \beta_{gij,p} + \epsilon_{gijk,p}, \quad p = 1, \ldots P, \tag{3}$$

where $g = 1, \ldots, 3$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, and $k = 1, \ldots, 3$ are indices for the observations that induce the hierarchical structure. The model components are generated according to $\alpha_{gi,p} \sim N(0, \sigma^2_{\alpha,p})$, $\beta_{gij,p} \sim N(0, \sigma^2_{\beta,p})$, and $\epsilon_{gijk,p} \sim N(0, \sigma^2_{\epsilon,p})$. Further, $\alpha_{gi,p}$, $\beta_{gij,p}$, and $\epsilon_{gijk,p}$ are mutually independent and are independent across all indices. The hierarchical structure induced by this model is analogous to the structure of the motivating dataset described in Section 2: g denotes estrous cycle phase, i denotes rat, j denotes neuron, and k denotes replicate. The sample sizes included for the simulations are reflective of the size of the motivating dataset; $n_1 = 7, 10$ and $n_2 = 3, 5$.

For the dimensions of Y_{gijk}, we consider $P = 2$ and 5. When $P = 2$, we borrow from Scenario 1 in Xu et al. (2018) to specify the variances of the components in (3). Thus, we let $(\sigma^2_{\alpha,1}, \sigma^2_{\alpha,2}) = (1, 0.25)$, $(\sigma^2_{\beta,1}, \sigma^2_{\beta,2}) = (0.5, 0.25)$, and $(\sigma^2_{\epsilon,1}, \sigma^2_{\epsilon,2}) = (5, 0.5)$. When considering $P = 5$, we instead let $(\sigma^2_{\alpha,1}, \ldots, \sigma^2_{\alpha,5}) = (1, 0.5, 0.33, 0.25, 0.2)$ and $\sigma^2_{\beta,p} = \sigma^2_{\epsilon,p} = \sigma^2_{\alpha,p}$.

To assess power performance, we generate data from three types of alternative hypotheses. First, we consider a shift in the mean: $\widetilde{Y}_{gijk,p} = \mu_g + Y_{gijk,p}$ where $Y_{gijk,p}$ is as in (3), $\mu_1 = 0$, $\mu_2 = \delta$, and $\mu_3 = -\delta$, and $\delta > 0$ controls the difference in the element-wise mean between groups. Second, we consider a shift in the second moment which we do in two ways. We slightly modify model (3) by generating $\alpha_{1i,p} \sim N(0, \sigma^2_{\alpha,p} + \delta)$. Alternatively, we instead modify model (3) by generating $\beta_{1ij,p} \sim N(0, \sigma^2_{\beta,p} + \delta)$. In the first setting, δ controls the difference in the *intersubject* variance between the first group and the remaining

two groups, whereas in the second setting, δ controls the difference in the *intrasubject* variance. Finally, we consider a shift in the third moment, modifying model (3) by first generating $\alpha_{1i,p} \sim \chi^2_\delta$ and $-\alpha_{2i,p} \sim \chi^2_\delta$ and then standardizing these variables, so they are mean 0 with variance $\sigma^2_{\alpha,p}$, that is so the mean and variance are the same across groups. Concurrently, we make analogous changes for $\beta_{gij,p}$ and $\epsilon_{gijk,p}$. Overall, for this setting, the coefficients for one group are generated to be positively skewed, the coefficients for another are to be negatively skewed, and the coefficients for the final group are to directly follow model (3) and thus are not skewed while δ controls the difference in the skewness between groups.

5.1.2 | Competing methods and metrics

To evaluate the resampling-based testing methodology when using multivariate data, we implement the proposed method, denoted by HHG-CB, using the `hhg.test.k.sample()` function in the R package *HHG* to calculate the HHG test statistic (Brill & Kaufman, 2019).

As comparative methods, we also consider the classic multivariate ANOVA (MANOVA), implemented using the `manova()` function in the R package *stats* (R Core Team, 2019) with Pillai's trace statistic for its robustness properties. MANOVA relies on independence across observations, which is obviously violated in this setting. Thus, we also consider an extension of the MANOVA by approximating the null distribution using the same resampling-based approach used for the primary method (denote by MANOVA-CB). We also borrow from the approach described in Pomann et al. (2016) and use an element-wise Anderson–Darling (AD) test with a Bonferroni adjustment for multiple comparisons; this is a conservative adjustment, and we consider it due to the small number of comparisons. The AD test is implemented using the `ad.test()` function in the R package *kSamples* (Scholz & Zhu, 2019). Lastly, we also consider the clustered Wilcoxon rank sum test (CW), separately for each element in the random vector (Datta & Satten, 2005). As with the AD test, we use a Bonferroni adjustment for multiple comparisons. This is implemented using the `clusWilcox.test()` function in the R package *clusrank* (Jiang, 2018).

The performance of each method is evaluated using the estimated type I error rates and power, each calculated as the average proportion of rejections of the null hypothesis across Monte Carlo replicates. When assessing type I error rates, we use 5000 Monte Carlo replicates; when assessing power, we use 1000 Monte Carlo replicates. When necessary, we use 1000 replicates to approximate the null distribution by resampling.

5.1.3 | Results

In the interest of space, all tables and figures for this section are provided in the Supporting Information; the results and interpretation are very similar to those presented in Section 5.2.3 when considering functional data.

5.2 | Framework 2: Functional data

5.2.1 | Generation of functional data

We generate hierarchically clustered functional data according to the model

$$Y_{gijk}(t,u) = \sum_{p=1}^{2} \sum_{q=1}^{Q_p} \zeta_{gijk,pq} \phi_p(t) \psi_{pq}(u) + \varepsilon_{gijk}(t,u), \quad (4)$$

where $Q_1 = 3$ and $Q_2 = 2$ and the indices are as described in the multivariate setting. The vector of coefficients $\boldsymbol{\zeta}_{gijk} = (\zeta_{gijk,11}, \zeta_{gijk,12}, \zeta_{gijk,13}, \zeta_{gijk,21}, \zeta_{gijk,22})$ is generated according to model (3) under the null hypothesis. The functions $\phi_p(t)$ and $\psi_{pq}(u)$ are taken to be the leading eigenfunctions estimated using the motivating dataset so that the data used in the simulations mimic the data in the motivating dataset. Additionally, $\varepsilon_{gijk}(t,u)$ is a white noise process, independent of $\boldsymbol{\zeta}_{gijk}$, with zero mean and variance equal to σ^2_{WN}; σ^2_{WN} is chosen to correspond to a signal-to-noise ratio (SNR) of 5. We use an equispaced grid of 100 locations for $t \in [0,1]$ and an equispaced grid of 10 locations for $u \in [0,1]$; the domains $\mathcal{T}$ and $\mathcal{U}$ are rescaled after estimating the eigenfunctions for simplicity.

As in the multivariate setting, we also generate functional data when the null hypothesis is not true to assess statistical power. We generate these functional data by modifying the generation of the coefficients as described in Section 5.1.1 for the multivariate setting. We again consider differences in the mean, inter- and intrasubject variance, and skewness.

In this functional data framework, we also consider the performance of the proposed method under a nonadditive data-generating mechanism. In this setting, we again consider functional data generated using model (4); however, we modify the generative model for $\boldsymbol{\zeta}_{gijk}$. In lieu of model (3), we instead generate the coefficients as $\zeta_{gijk,p} = \alpha_{gi,p} + \beta_{gij,p} + \epsilon_{gijk,p} + \alpha_{gi,p}\beta_{gij,p}$, where $\alpha_{gi,p}$, $\beta_{gij,p}$, and $\epsilon_{gijk,p}$ are as defined above. To assess power in

TABLE 1 Estimated type I error rates for the methods described in Section 5.2.2 when applied to functional data based on 5000 replicates. Nominal type I error rate is 0.05. Standard errors < 0.004.

n_1	n_2	HHG-CB	MANOVA-CB	CW
7	3	0.042	0.058	0.002
10	3	0.042	0.055	0.004
7	5	0.040	0.058	0.003
10	5	0.043	0.055	0.005

this setting, we consider differences in the mean similar to those described above.

5.2.2 | Competing methods and implementation

For the purposes of testing, after estimation of the coefficients $\{\zeta_{gijk,pq}\}$, we utilize the multivariate testing methods and implement them as described in Section 5.1.2; we again assess the type I error rates and power. We now discuss the implementation of the modeling step to estimate the necessary eigenfunctions and coefficients using functions available in the R package *refund* (Goldsmith et al., 2018). We first estimate the common bivariate mean function $\mu(\cdot,\cdot)$ using the `fbps()` function and center the data. Then, we estimate $\{\phi_p(\cdot)\}_{p\geq 1}$ and coefficients $\{\xi_{gijk,p}(u)\}$ using the `fpca.face()` function. We select the truncation parameter P using a 95% PVE threshold. Finally, we estimate $\{\psi_{pq}(\cdot)\}_{q\geq 1}$ and coefficients $\{\zeta_{gijk,pq}\}$ using the `fpca.sc()` function. The truncation parameters Q_p are also selected using a 95% PVE threshold.

5.2.3 | Results

We first consider the simulation results when data are generated so that there are no distributional differences across groups. The estimated type I error rates are shown in Table 1. The estimated type I error rates for the HHG-CB and MANOVA-CB methods are close to the nominal rate, whereas the CW method is much more conservative.

We next consider the estimated power of each method; the estimated power curves for each method are shown in Figure 3. In the interest of space, we do not include the estimated power curves for the CW method as the power is significantly lower than with the other methods; see the Supporting Information for these figures. We start by considering a difference in the mean. Both the HHG-CB and MANOVA-CB approaches perform similarly, although the MANOVA-CB method has moderately higher power. That this approach performs well in this setting is not surprising as MANOVA is designed to detect differences in the mean. We also see from this figure that increasing the sample size, either by increasing the number of rats or neurons, resulted in an increase in power. It does appear that adding additional rats, and therefore adding observations that are independent from the rest of the data, is of greater benefit than adding additional neurons per rat, which is to be expected.

The interpretation of the results when considering other forms of group distributional differences is generally similar to those discussed above, with the key exception being that only the proposed approach is shown to detect these differences. Neither of the other two approaches is sensitive to second-order moment differences between groups. When considering differences in skewness, all methods perform poorly, although the proposed approach does exhibit the highest, albeit still small, power. In preliminary simulations, we saw substantial improvement by the proposed method to detect differences in skewness when the data are generated without noise, which is consistent with results seen in the multivariate framework.

6 | ANALYSIS OF A WHPC EXPERIMENT

We now discuss the analysis of the motivating dataset described in Section 2, the objective of which is to test for distributional differences in medium spiny neuron electrophysiological properties between phases of the estrous cycle. First, we look at the estimated mean function for each estrous cycle phase, as shown in Figure 4; the estimated bivariate mean functions and the univariate trajectories conditional on different frequencies and currents are displayed. To estimate the group specific mean, we use the sandwich smoother estimate (Xiao et al., 2013) described in Section 3.2 on the data separately by group. Generally, we see that for a fixed current, the mean log-periodogram decreases with increasing frequency. Further, for a fixed phase of the estrous cycle, the rate of decay decreases as current increases. The mean log-periodogram for the three phases is very similar when no current is injected. Differences between the phases become apparent as the amount of current applied increases. The mean log-periodogram for the diestrus phase is noticeably larger than the mean for the other two phases when a low (e.g., +0.05 nA) amount of current is applied. This suggests that an important difference between phases is the amount of current necessary to generate an action potential. As current increases, the mean log-periodogram increases until it plateaus, the magnitude of which depends on the frequency but not the phase of the estrous cycle. While neurons in the diestrus phase differ from those in the other

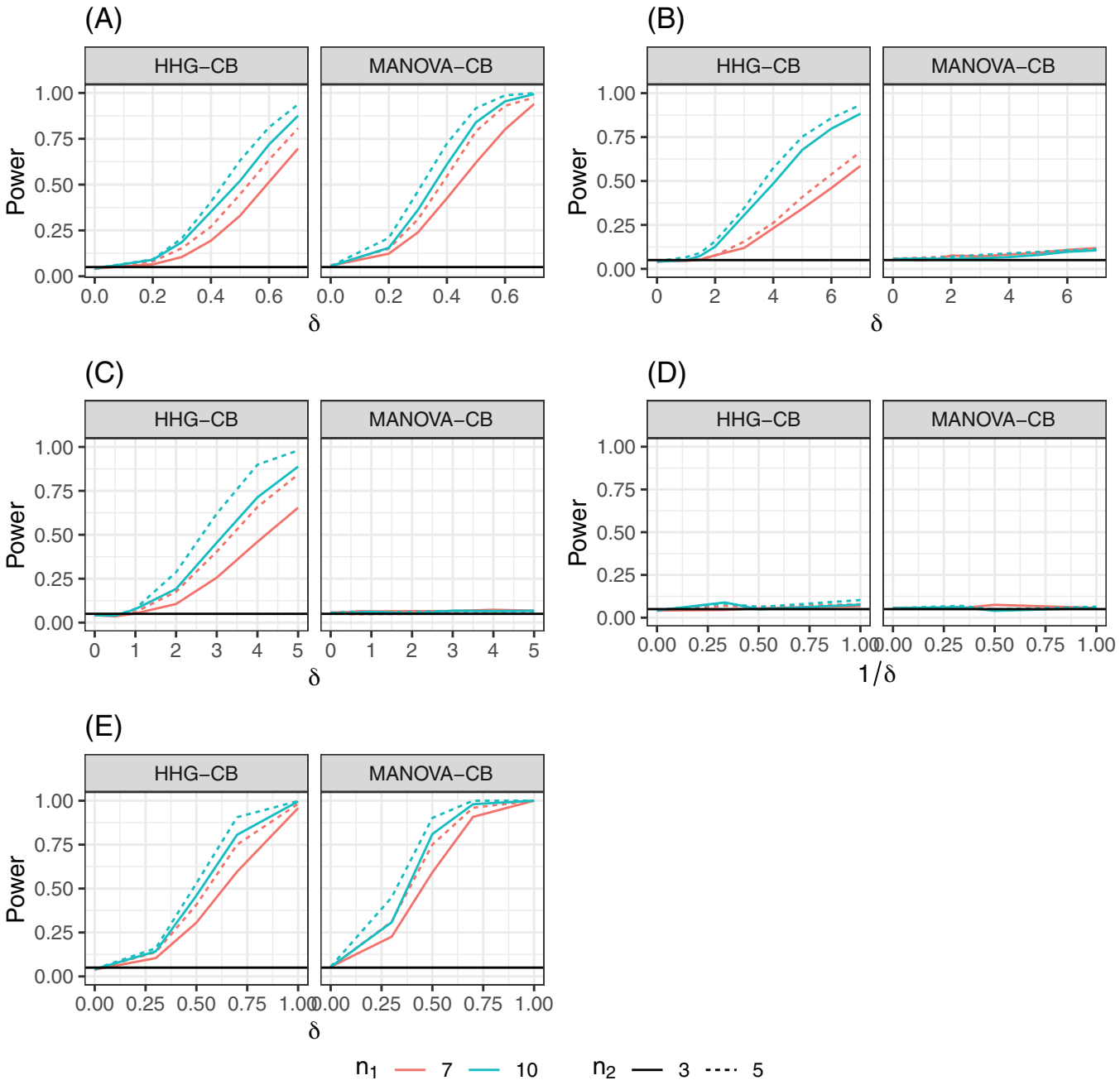

FIGURE 3 Estimated power curves for the HHG-CB and MANOVA-CB methods for detecting differences in (A) the mean, (B) the variance of the rat-level effect, (C) the variance of the neuron-level effect, and (D) the skewness when applied to functional data. Estimated power curves for both methods for detecting differences in the mean under a nonadditive model (E). All estimates based on 1000 replicates. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

two phases, neurons in the estrus and proestrus phases of the cycle appear to behave similarly. We note that for the majority of the data, the applied current is less than +0.2 nA; this explains why the estimated mean functions are less smooth for higher current. Also, due to the comparatively few observations at currents higher than +0.2 nA, it is likely that the estimated mean (and other estimates) has larger standard errors than for lower currents. However, because we are ultimately focused on hypothesis testing rather than estimation, and because the observations are projected onto a common set of estimated eigenfunctions that do not differ by group, we consider the relative uncertainty caused by few observations for high current to not have a meaningful impact on the proposed method or the eventual results.

To estimate the eigenfunctions of the marginal covariance function, we first estimate the mean function using the sandwich smoother and then use it to obtain the

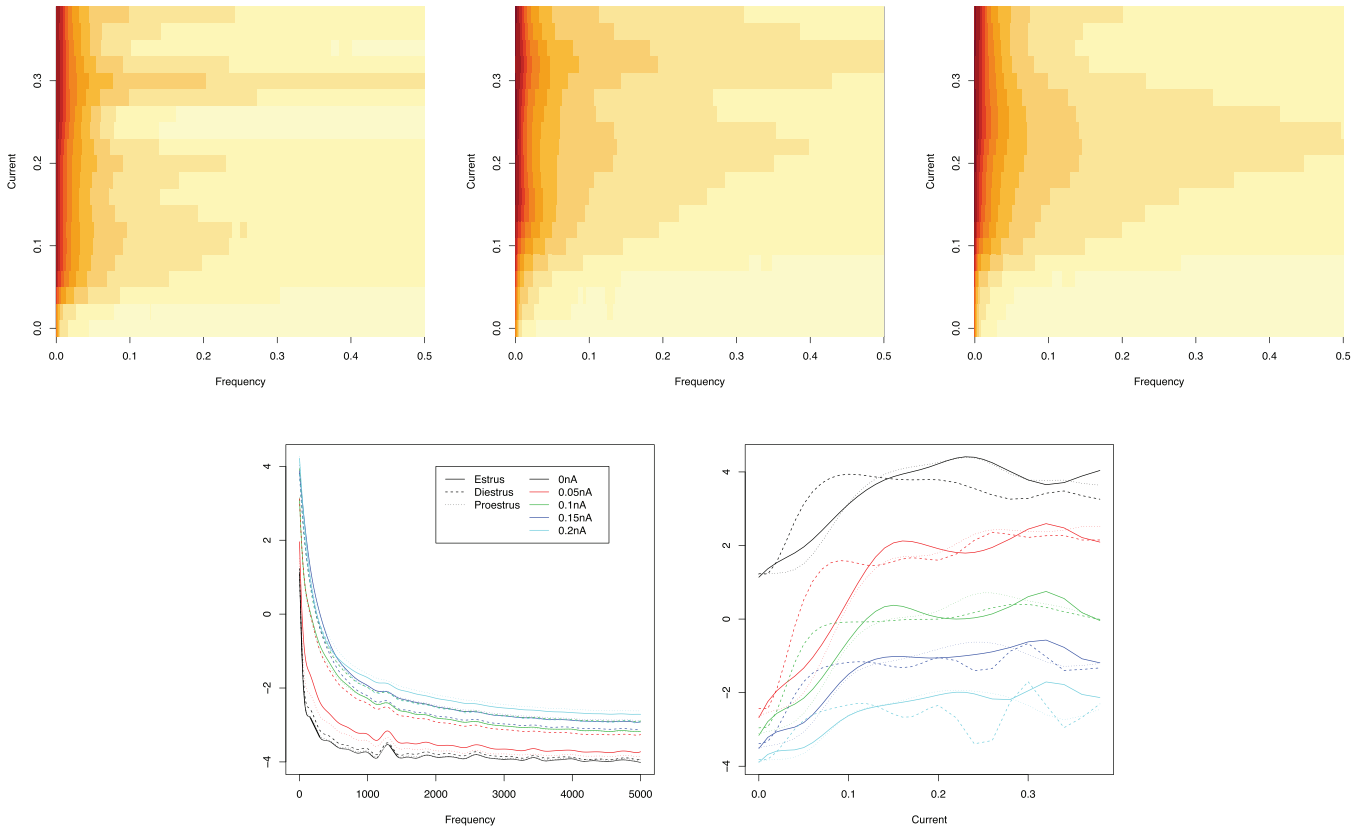

FIGURE 4 The estimated mean of the log-periodogram of the membrane potential for each phase in the estrous cycle. (Top) The estimated bivariate mean function for the diestrus (left), estrus (center), and proestrus (right) phases. (Bottom left) The different phases of the cycle are indicated by line style. The mean trajectories corresponding to different amounts of current are indicated by color. (Bottom right) The estimated mean of the log-periodogram for fixed frequency and changing current is shown for each phase in the estrous cycle. The mean trajectories corresponding to different frequencies are indicated by color. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

centered data. We then estimate the eigenfunctions as described in Section 3.2. To select the truncation parameters, P and Q_p, we use a 0.95 PVE threshold, separately for each parameter. This choice results in selecting $P = 2$, $Q_1 = 4$, and $Q_2 = 4$. The estimates of $\{\phi_p(\cdot)\}_{p=1}^{P=2}$, the eigenfunctions of the marginal covariance function of frequency from the initial decomposition are shown in the left-most panel in Figure 5. The leading eigenfunction indicates a deviation from the mean that is approximately constant across frequency. The second eigenfunction indicates a large positive deviation from the mean at low frequencies and a small negative deviation from the mean at higher frequencies. A large positive coefficient of this second eigenfunction would likely have action potentials occurring with higher frequency than an average observation. The estimates of $\{\psi_{1q}(\cdot)\}_{q=1}^{Q_1=4}$ and $\{\psi_{2q}(\cdot)\}_{q=1}^{Q_2=4}$ are shown in the middle and right-most panel of Figure 5, respectively. The estimate of the leading eigenfunction of the covariance of coefficients of the leading eigenfunction from the initial decomposition, $\widehat{\psi}_{1,1}(\cdot)$, is shown in black in the middle panel of Figure 5. This eigenfunction indicates a large negative deviation from the mean at all frequencies when the current applied to the neuron is low. When the current is high, this eigenfunction indicates little deviation from the mean across all frequencies.

We use the HHG test with the proposed bootstrap-based procedure to test for differences in the log-periodogram across phases of the estrous cycle, while accounting for the hierarchical structure of the observed data. As in the simulations, we use 1000 bootstrap samples to estimate the null distribution of the test statistic. The p-value of the proposed testing method is 0.003, indicating that there is a significant difference in the log-periodogram across phase. As a sensitivity analysis, we considered the impact of the number of components selected using PVE on the results; across all considered scenarios, the results are robust to changes in the number of components with the p-value always ≤ 0.005 (see Supporting Information).

Since we detect a significant difference in the log-periodogram between phases of the estrous cycle, we can use the estimated coefficients of the eigenfunctions to explore how the groups differ. We do this by testing each

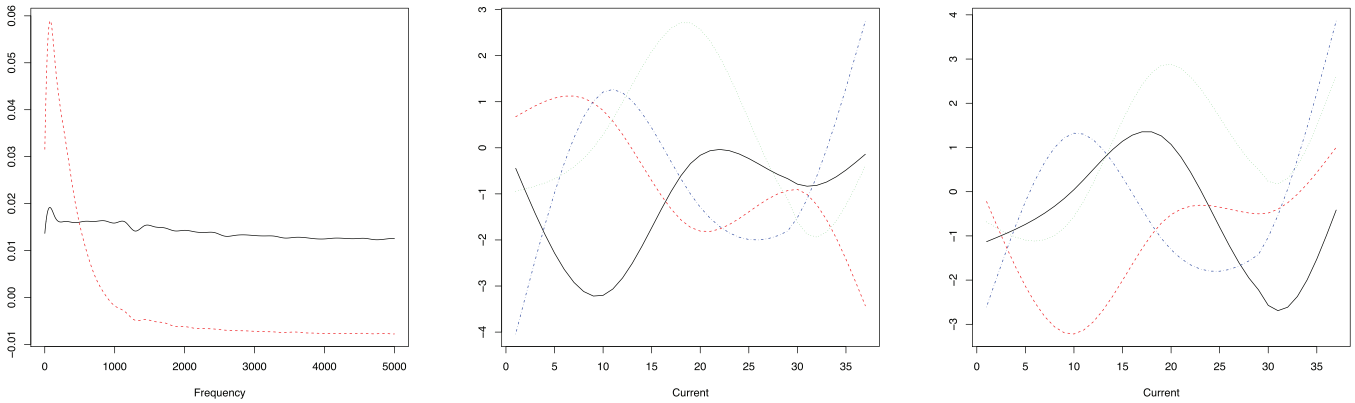

FIGURE 5 (Left) The estimated eigenfunctions in frequency from the initial decomposition of the data; $\widehat{\phi}_1(\cdot)$ in black, $\widehat{\phi}_2(\cdot)$ in red. (Middle) The estimated eigenfunctions $\widehat{\psi}_{1q}(\cdot)$, for $q = 1$ (black), 2 (red), 3 (green), and 4 (blue). (Right) The estimated eigenfunctions $\widehat{\psi}_{2q}(\cdot)$, identified similar to $\widehat{\psi}_{1q}(\cdot)$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

coefficient one at a time. Rather than resampling the functional data to account for the hierarchical structure, we instead resample the coefficients themselves to approximate the null distribution of the test statistic. From this analysis, we see that the differences in the log-periodogram across phase of the estrous cycle are explained by different coefficients for the leading eigenfunctions. For example, we see that the coefficient $\zeta_{1,1}$ differs significantly across phase. By examining the distribution of coefficient estimates by estrous cycle phase, we see that observations from the diestrus phase tend to have large negative values for this coefficient. This indicates that neurons in the diestrus phase exhibit above-average log-periodogram values across all frequencies when the current applied is low; this is similar to what was seen from the plots of the mean functions in Figure 4.

Our analysis provides new insights into the understanding of the neurophysiological properties originally described in Proaño et al. (2018). From our new analysis of this dataset, we found that neurons from rats in the diestrus phase required less current to generate an action potential than those in either of the other two phases of the estrous cycle. This was one of the properties of interest in the original analysis of this dataset; our findings are consistent with those described in Proaño et al. (2018). Despite the similar results between the two analyses, with our novel method, we did not have to prespecify this parameter of interest and process the data accordingly.

7 | DISCUSSION

In this paper, we propose a testing procedure to detect group distributional differences in hierarchically clustered functional data. We applied this method to the motivating dataset to show that the electrophysiological properties of certain neurons in adult female rats differ across phases of the estrous cycle. While the focus of this paper was on this specific application, the proposed method can be applied in other settings in which hierarchically clustered functional data are observed. To that point, we evaluated the performance of the proposed method in various simulations and illustrated the advantages against alternative methods. A limitation of the proposed method is that the resampling method to approximate the null distribution of the test statistic can be computationally intensive, particularly with larger datasets. However, this step can easily be done in parallel to shorten the necessary runtime.

ACKNOWLEDGMENTS

The authors thank Dr. Stephanie Proaño who was crucial in generating the motivating dataset.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information section of this paper.

ORCID

Brian J. Reich https://orcid.org/0000-0002-5473-120X

REFERENCES

Abramovich, F. & Angelini, C. (2006) Testing in mixed-effects FANOVA models. *Journal of Statistical Planning and Inference*, 136(12), 4326–4348.

Antoniadis, A. & Sapatinas, T. (2007) Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, 51(10), 4793–4813.

Arnegard, M.E., Whitten, L.A., Hunter, C. & Clayton, J.A. (2020) Sex as a biological variable: a 5-year progress report and call to action. *Journal of Women's Health*, 29(6), 858–864. PMID: 31971851.

Bathke, A.C., Harrar, S.W. & Madden, L.V. (2008) How to compare small multivariate samples using nonparametric tests. *Computational Statistics & Data Analysis*, 52(11), 4951–4965.

Brill, B. & Kaufman, S. (2019) *HHG: Heller-Heller-Gorfine tests of independence and equality of distributions*. R package version 2.3.2.

Cao, J., Dorris, D.M. & Meitzen, J. (2018) Electrophysiological properties of medium spiny neurons in the nucleus accumbens core of prepubertal male and female Drd1a-tdTomato line 6 BAC transgenic mice. *Journal of Neurophysiology*, 120(4), 1712–1727.

Chen, K., Delicado, P. & Müller, H.G. (2017) Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1), 177–196.

Chen, K. & Müller, H.G. (2012) Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500), 1599–1609.

Cuesta-Albertos, J.A. & Febrero-Bande, M. (2010) A simple multiway ANOVA for functional data. *TEST*, 19(3), 537–557.

Datta, S. & Satten, G.A. (2005) Rank-Sum Tests for Clustered Data. *Journal of the American Statistical Association*, 100(471), 908–915.

Di, C.Z., Crainiceanu, C.M., Caffo, B.S. & Punjabi, N.M. (2009) Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1), 458–488.

Druckmann, S., Hill, S., Schürmann, F., Markram, H. & Segev, I. (2012) A hierarchical structure of cortical interneuron electrical diversity revealed by automated statistical analysis. *Cerebral Cortex*, 23(12), 2994–3006.

Fan, J. & Zhang, W. (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4), 715–731.

Fremdt, S., Steinbach, J.G., Horváth, L. & Kokoszka, P. (2013) Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1), 138–152.

Galea, L.A., Choleris, E., Albert, A.Y., McCarthy, M.M. & Sohrabji, F. (2020) The promises and pitfalls of sex difference research. *Frontiers in Neuroendocrinology*, 56, 100817.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C. & Reiss, P.T. (2018) *refund: Regression with Functional Data*. R package version 0.1-17.

Heller, R., Heller, Y. & Gorfine, M. (2013) A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2), 503–510.

Hernáth, F., Schlett, K. & Szücs, A. (2019) Alternative classifications of neurons based on physiological properties and synaptic responses, a computational study. *Scientific Reports*, 9(1), 13096.

Horváth, L., Kokoszka, P. & Reeder, R. (2013) Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1), 103–122.

Jiang, Y. (2018) *clusrank: Wilcoxon rank sum test for clustered data*. R package version 0.6-2.

Li, H., Kozey Keadle, S., Staudenmayer, J., Assaad, H., Huang, J.Z. & Carroll, R.J. (2015) Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, 16(4), 754–771.

Mamlouk, G.M., Dorris, D.M., Barrett, L.R. & Meitzen, J. (2020) Sex bias and omission in neuroscience research is influenced by research model and journal, but not reported NIH funding. *Front Neuroendocrinol*, 57, 100835.

Paparoditis, E. & Sapatinas, T. (2016) Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, 103(3), 727–733.

Park, S.Y. & Staicu, A.M. (2015) Longitudinal functional data analysis. *Stat (International Statistical Institute)*, 4(1), 212–226.

Pomann, G.M., Staicu, A.M. & Ghosh, S. (2016) A two sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 65(3), 395–414.

Proaño, S.B., Morris, H.J., Kunz, L.M., Dorris, D.M. & Meitzen, J. (2018) Estrous cycle-induced sex differences in medium spiny neuron excitatory synaptic transmission and intrinsic excitability in adult rat nucleus accumbens core. *Journal of Neurophysiology*, 120(3), 1356–1373.

R Core Team (2019) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J.O. & Silverman, B.W. (2005) *Functional data analysis*. Springer, New York.

Scheffler, A., Telesca, D., Li, Q., Sugar, C.A., Distefano, C., Jeste, S. & Şentürk, D. (2018) Hybrid principal components analysis for region-referenced longitudinal functional EEG data. *Biostatistics*, 21(1), 139–157.

Scholz, F. & Zhu, A. (2019) *kSamples: K-sample rank tests and their combinations*. R package version 1.2-9.

Staicu, A.M., Lahiri, S.N. & Carroll, R.J. (2015) Significance tests for functional data with complex dependence structure. *Journal of Statistical Planning and Inference*, 156, 1–13.

Staicu, A.M., Li, Y., Crainiceanu, C.M. & Ruppert, D. (2014) Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, 41(4), 932–949.

Tannenbaum, C., Ellis, R.P., Eyssel, F., Zou, J. & Schiebinger, L. (2019) Sex and gender analysis improves science and engineering. *Nature*, 575(7781), 137–146.

Wynne, G. & Duncan, A.B. (2020) A kernel two-sample test for functional data.

Xiao, L., Li, Y. & Ruppert, D. (2013) Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(3), 577–599.

Xiao, L., Zipunnikov, V., Ruppert, D. & Crainiceanu, C. (2016) Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26(1), 409–421.

Xu, Y., Li, Y. & Nettleton, D. (2018) Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, 113(522), 593–606.

Yao, F., Müller, H.G. & Wang, J.L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.

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, J.T. & Liang, X. (2014) One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics*, 41(1), 51–71.

Zhu, H., Li, R. & Kong, L. (2012) Multivariate varying coefficient model for functional responses. *The Annals of Statistics*, 40(5), 2634–2666.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 3–6, the data analyzed in Section 6, and R code for the analyses and simulations described in Sections 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Long, A.S., Reich, B.J., Staicu, A.-M. & Meitzen, J. (2023) A nonparametric test of group distributional differences for hierarchically clustered functional data. *Biometrics*, 1–14. https://doi.org/10.1111/biom.13846