Response to comments of AOAS2106-024R1

We would like to thank the editor, AE and the reviewers for their additional comments. We are glad that the reviewers found the revised paper much more clear than the previous version. Please see below our point-to-point response in BLUE to the additional comments and the changes we have made. We marked the changes in RED in the revised paper. Following one reviewer’s suggestion, we have changed the title of the paper to **Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data** to better reflect the specific data structures that we studied in the paper. We have also modified the abstract and added discussion points acknowledging the scope and limitations of our method assumptions.

**Minor Comment from Reviewer #1**: I have only one additional comment regarding to a minor typo: in Fig 3, $\mu_{F1} \neq \mu_{F2}$ should read as $\mu_{F1} \neq \mu_{F2}$

We have made this correction as suggested.

**Major comment #1 from Reviewer #2**: In the last round, I raised the question of setting $l$, the number of measurements for each subject, to be the same across all subjects. I was expecting the authors to relax this assumption as it is a very strong one. In practice, it is common to have different number of measurements from different subjects due to various reasons. In the application, the authors simply remove these subjects with fewer number of measurements from the data analysis. This is a brutal treatment and can cause bias in the result - it could be that subjects with some particular attributes are more likely to have fewer measurements. If the same number of measurements on all subjects is a requirement for applying the method, it should be made clear in the title and the abstract of the paper as this is the scope of the problem the method can handle. Some discussions on the consequences due to this limitation would also be needed.

We appreciate your further comments on this point. Our proposed method indeed assumes that the number of repeated measures is the same for different individuals since we treat the observations from each individual as a multivariate object (histogram in our case). The advantage of treating them as multivariate data is that we do not need to assume specific error structures of the repeated histogram measures over days, which could be challenging to specify for object data. For example, we do not require that the repeated measures across days have the same marginal distribution and allow them to be different from day to day. In addition, it is not clear how one should quantify the random day-to-day variation for the object data. We have stated this requirement more explicitly in the Abstract and changed title to ‘Two-sample tests for **multivariate** repeated measurements of histogram objects with applications to wearable device data’ to better reflect the data structures that our methods could handle.
However, we agree with the reviewer that there might be the cases when an unequal number of repeated observations might be informative, in which case one should interpret the results with care. We have also added more discussions on this point and pointed out the limitations of the proposed methods in the Discussion (page 25, last paragraph).

Major comment #2 from Reviewer #2: In the current setup, \( \{X_{ui}\}_{i=1}^{\ldots,l} \) can be correlated but assumed to have the same distribution \( F_1 \), and \( \{Y_{vi}\}_{i=1}^{\ldots,l} \) can also be correlated but assumed to have the same distribution \( F_2 \). In addition, the vector of \( l \)-day densities \( X_u, u = 1, \cdots, n_1 \) are i.i.d. \( \tilde{P}_1 \), and the vector of \( l \)-day densities \( Y_v, v = 1, \cdots, n_2 \) are i.i.d. \( \tilde{P}_2 \). So from the assumptions, the correlation structure between \( X_{11} \) and \( X_{12} \) must be the same as that between \( X_{21} \) and \( X_{22} \), the correlation structure between \( X_{11} \) and \( X_{13} \) must be the same as that between \( X_{21} \) and \( X_{23} \), etc. This is also a strong assumption. For example, in the daily physical activity application, it would be common that different subjects have distinct patterns, causing the correlation structure to be different across subjects. This could heavily affect \( G_{in} \) and the test statistics involving \( G_{in} \) \((T_{in}; S_R; M(\alpha, \kappa))\). Hence, it would be worthwhile to check the validity of the assumption when applying the method.

We appreciate your comment on this point. We would like to clarify that we assume that the activity object data (as a vector over \( l \) days) of different individuals are i.i.d. random samples drawn from an underlying population of multivariate object data. However, the between-day correlation from the observed densities might differ across individuals, although we CANNOT estimate such correlations for EACH individual sample separately since we treat them as a random sample from a population of multivariate objects with unspecified correlation structures. Our proposed tests are for comparisons of such population parameters. If we allow different correlation structures across individuals within group and that samples are from different populations, then it will be impossible to develop any sensible statistical tests under our model setups.

To further check the validity of such assumption in our real data, we performed sensitivity analysis by repeatedly random subsetting \( l \) days from those individuals with more than \( l \) days 1000 times, and performed each of the tests. The distribution of the -log (p-value) of each test for age group comparison under 1000 Bootstrap samples are now shown as Boxplots in Supplement F. We found that the results are robust to different choices of days used in our analysis and hence we reported the averaged pvalues through Fisher’s transformation in our paper.

Major comment #3 from Reviewer #2: I also have a question on the application: Why the histogram of activity intensities is used rather than the activity intensities over time? Using the histogram loses the time information. It could be that two histograms are similar while one has those high-activity-count periods in a short time frame while the other has those periods scattered over the whole day. These two observations should be considered
This is a great comment. There are indeed many researchers examining the time-specific activity intensities from accelerometry data using approaches such as functional data analysis. However, since an individual’s time-specific activity intensities might vary due to non-biological factors such as time shifts introduced by work schedule or life events, the activity profiles over the exact time stamps might not be directly comparable across individuals and days and hence several papers have proposed methods to register the time stamps first and make the activity profiles align (see Wrobel et al. (2019) for example). The choice of a sensible outcome from the activity tracking data will depend on the specific questions that we would like to address. In our study, activity data were collected from participants in their free-living conditions with different daily routines and schedules. Hence our paper took a different view of the activity data by examining the overall distribution of the daily activity levels, rather than the activities at a specific time. This was motivated by the need of improving upon a set of conventional physical activity markers such as proportion of time spent in moderate-to-vigorous physical activity (MVPA), which discretizes the daily activity intensities into several categories and loses the majority of the information from the original data. By directly modeling the continuous daily activity distributions, we avoid the need for registering time stamps across days and subjects from the time-specific activity intensities and wish to preserve as much information as possible from the original data. But we agree with the reviewer that the histograms does not preserve the time information and in cases where timing of activity is of major interest, one should consider studying the time-specific activity levels using appropriate statistical models.
TWO-SAMPLE TESTS FOR MULTIVARIATE REPEATED MEASUREMENTS OF HISTOGRAM OBJECTS WITH APPLICATIONS TO WEARABLE DEVICE DATA

By Jingru Zhang\textsuperscript{1}, Kathleen R. Merikangas\textsuperscript{2}, Hongzhe Li\textsuperscript{*1} and Haochang Shou\textsuperscript{*1}

\textsuperscript{1}University of Pennsylvania
\textsuperscript{2}National Institutes of Health

Repeated observations have become increasingly common in biomedical research and longitudinal studies. For instance, wearable sensor devices are deployed to continuously track physiological and biological signals from each individual over multiple days. It remains of great interest to appropriately evaluate how the daily distribution of biosignals might differ across disease groups and demographics. Hence these data could be formulated as multivariate complex object data such as probability densities, histograms, and observations on a tree. Traditional statistical methods would often fail to apply as they are sampled from an arbitrary non-Euclidean metric space. In this paper, we propose novel non-parametric graph-based two-sample tests for object data with the same structure of repeated measures. We treat the repeated measured object data as multivariate object data, which requires the same number of repeated observations per individual, but eliminates any assumptions on the errors of the repeated measures. A set of test statistics are proposed to capture various possible alternatives. We derive their asymptotic null distributions under the permutation null. These tests exhibit substantial power improvements over the existing methods while controlling the type I errors under finite samples as shown through simulation studies. The proposed tests are demonstrated to provide additional insights on the location, inter- and intra-individual variability of the daily physical activity distributions in a sample of studies for mood disorders.

1. Introduction. Repeated measures are frequently obtained to capture the within-individual variation and enhance the data reproducibility. For example, studies that examine physical activities (PA) using accelerometers often observe individuals’ 24hr activity profiles repeatedly over several days or weeks (Burton et al., 2013; Krane-Gartiser et al., 2014; De Crescenzo et al., 2017). Within each day, the physical accelerations during movement are recorded with a high frequency and processed into a time series of ac-
tivity intensity metrics such as activity counts, vector of magnitude (VM) or Euclidean norm minus one (ENMO) over certain epoch lengths (e.g., 5-, 15- or 60-seconds). Commonly extracted markers from accelerometry data include the total amount of PA such as total log activity intensities and step counts (Varma et al., 2018), and time spent in different activity intensity levels. In particular, proportion of time spent in sedentary behavior (SB), light (LPA) and moderate-to-vigorous physical activity (MVPA) have been reported to meaningfully correlated with physical and mental functioning and health (Faurholt-Jepsen et al., 2012; De Crescenzo et al., 2017; Murray et al., 2020). However, there remain several known limitations in these traditional PA endpoints. First, metric such as time spent in SB, LPA and MVPA reduces the continuous activity profiles into a composition of only three discrete categories, resulting in great loss of the rich information captured by the densely measured raw accelerometry data. In fact, MVPA might be relatively sparse in a largely sedentary population and are less sensitive to meaningful clinical differences within the population. Secondly, these variables are determined with a priori selected cutpoints. Yet there is a lack of consensus of cutpoints for data collected across study populations (e.g., children vs. adults, diseased vs. healthy individuals), type of devices, wearing positions (e.g., hip vs. wrist)(Schrack et al., 2016; Leeger-Aschmann et al., 2019). It has also been reported that the recording frequency, the choice of epoch length and wear time algorithms during processing steps could significantly vary the endpoints and potentially lead to inconsistent conclusions (Banda et al., 2016). Hence it remains challenging to compare findings across studies with these traditionally derived metrics.

Instead of relying upon a few discrete categories defined by relatively arbitrary cutpoints, recently increasing attentions have been paid to model the continuous distribution of the raw daily activity intensities (Keadle et al., 2014; Schrack et al., 2016; Yang et al., 2020). In this paper, we also take the daily activity histogram for each individual as the observed outcome and develop statistical methods that compare density objects between groups. As an illustration, we plotted the observed activity data from one individual over four days (two weekdays and two weekends) in Figure 1 from the National Institute of Mental Health (NIMH) Family Study of Spectrum Disorders (Merikangas et al., 2014, 2019). Their time-specific activity intensities at one-minute intervals are shown on the top row and the corresponding histograms of activity intensities in log-transformed scale are shown on the bottom. As Figure 1 shows, despite the overall similarity in the time-specific activity patterns across days, the evident shifts in schedules from weekdays to weekends might not be of biological interests. Hence a second advantage
of directly modeling the daily activity distributions is that it avoids the need for registering time stamps across days (Wrobel et al., 2019).

We consider the problem of testing whether the activity density functions
or distributions are significantly different between individuals from various clinical groups. As previously noted, the conventional representations of time spent in different levels of PA are derived from discretized distributions using pre-determined cutpoints. To minimize the loss of information, we will be working with the entire probability densities of the continuous daily activity intensities. Our challenges are two folds. First, probability densities, as characterized by the empirical histogram of the daily physical activity intensities, are non-Euclidean and hence many traditional two-sample test statistics are no longer applicable. Second, physical activity tracking over multiple days result in repeated probability densities. As far as we know, there are few existing method that could handle within-individual dependency in the complex object data.

While two-sample testing for multivariate objects in Euclidean space or even infinite dimensional space have been studied extensively in the statistics literature, fewer tools are available for two-sample testing when the data are samples of density or distributional functions. To deal with a wide range of data types, nonparametric tests are preferable. Friedman and Rafsky (1979) proposed the first practical test as an extension of the Wald-Wolfowitz runs test to multivariate data. This framework has been extended to other graph-based testing methods. For example, Rosenbaum (2005) used the minimum distance pairing (MDP); Schilling (1986) and Henze (1988) adopted the nearest neighbor graph (NNG); Chen and Friedman (2017) and Chen, Chen and Su (2018) proposed a generalized edge-count test and a weighted edge-count test to address the problems under scale alternatives and unequal sample sizes, respectively. Recently, an extension of analysis of variance for metric space valued data objects was proposed by Dubey and Müller (2019), where Fréchet mean and variance are used to construct the test statistic. Yang et al. (2020) proposed quantlets as basis functions to approximate the quantile function objects in a regression setting.

However, most of these existing tests for object data assume that the observations are independent, which cannot be directly applied to repeated measures of object data where within-individual observations are correlated. One simple way to deal with this issue is to apply these tests to the average of the within-individual measures and convert the problem into a standard two-sample test for independent object observations (Dawson and Lagakos, 1993). However, it is not trivial to define the average of non-Euclidean object data. In addition, taking averages oversimplifies the true complexity of data and ignores the within-individual variability that could also be clinically relevant when studying individuals’ behaviors and mood (Murray et al., 2020).
We propose a new nonparametric testing framework for density data with repeated measures. This framework builds upon graph-based two-sample testing methods that are flexible and require few assumptions (Chen and Friedman, 2017; Chen, Chen and Su, 2018). In particular, to take into account the repeated nature of the data, we consider the between-individual and within-individual similarity graphs defined via the Wasserstein distances between two density functions. Based on the constructed graph, we define several test statistics that are powerful for various possible alternatives, including difference in population Fréchet means, Fréchet variances and Fréchet covariance. A new permutation null distribution is considered using the between-individual and within-individual similarity graphs. We also derive the asymptotic null distributions of these statistics under the permutation null, facilitating their applications to large data sets.

We evaluate the proposed test statistics using simulations and compare the power with several competing tests developed for density data. Our approaches are used in an extensive analysis to evaluate the effects of age, body mass index and types of mood disorders on daily activities in the NIMH family study population.

2. Non-parametric tests for density functions with repeated measures based on a similarity graph.

2.1. A permutation null distribution for density data with repeated measures. To analyze the repeated measurements of activity data, we treat the observed activity densities over $l$ days from each individual as a vector of outcome. We assume that individuals are divided into two groups with $X_1, \ldots, X_{n_1}$ representing density objects for $n_1$ individuals from group 1 and $Y_1, \ldots, Y_{n_2}$ representing densities for $n_2$ individuals from group 2. For a given individual $u$ from group 1, we have $X_u = (X_{u1}, \ldots, X_{ul})$ representing each of the $l$ days’ activity densities. Similarly, for individual $v$ from group 2, $Y_v = (Y_{v1}, \ldots, Y_{vl})$.

We assume that each individual density $X_{ui}$ and $Y_{vj}$ ($u = 1, \ldots, n_1; v = 1, 2, \ldots, n_2; i, j = 1, 2, \ldots, l$) belongs to space $D$, where $D$ represents a class of one-dimensional densities such that $\int_R u^2 f(u) du < \infty$ for $f \in D$. For any two random densities $X, Y \in D$, we define $d_W$ to be the Wasserstein metric as

$$d_W^2(X, Y) = \int_R \{T(u) - u\}^2 X(u) du$$

where $T = F_X^{-1} \circ F_Y$ is the optimal transport, and $F_X$ and $F_Y$ are the distribution functions of $X$ and $Y$, respectively.
We further assume that $X_{ui}$ and $X_{uj}$ have identical distribution function $F_1$, but might be correlated; similarly $Y_{vi}$ and $Y_{vj}$ have the same distribution $F_2$. The vector of $l$-day densities $X_u$, $u = 1, \cdots, n_1$, however, are independently and identically distributed across individuals according to $\tilde{P}_1$. $Y_v$, $v = 1, \cdots, n_2$ are i.i.d according to $\tilde{P}_2$.

For a random density $X \in \mathcal{D}$ from group 1, we define the corresponding group-level Fréchet mean $\mu_{F_1}$ and Fréchet variance $V_{F_1}$ as

$$\mu_{F_1} = \arg \min_{x \in \mathcal{D}} E\{d^2_W(x, X)\}, \quad V_{F_1} = E\{d^2_W(\mu_{F_1}, X)\}.$$  

Similarly $\mu_{F_2}$ and $V_{F_2}$ represent the Fréchet mean and Fréchet variance for a random density $Y$ from group 2. Given a vector of random densities whose elements are dependent with each other, we define their Wasserstein covariance following the framework in Petersen and Müller (2019). Specifically, for two random densities $X_i$ and $X_j$ from group 1, the Wasserstein covariance is defined as

$$Cov_{F_1,ij} = E\left[\int_{R} \{F_X^{-1}(u) - \mu_{F_1}(u)\}\{F_X^{-1}(u) - \mu_{F_1}(u)\} du\right],$$

where the expectation was taken over the joint distribution of $X_i$ and $X_j$. Similarly, $Cov_{F_2,ij}$ denotes the Wasserstein covariance between $Y_i$ and $Y_j$ from group 2.

We are interested in testing the null hypothesis $H_0 : \tilde{P}_1 = \tilde{P}_2$, which implies that the $N = n_1 + n_2$ samples are from the same distribution. Based on our motivating examples, group differences in physical activity distributions could occur in mean $\mu_{F_1} \neq \mu_{F_2}$, between-individual variability $V_{F_1} \neq V_{F_2}$ or within-individual variability among repeated observations $Cov_{F_1,ij} \neq Cov_{F_2,ij}$ for at least one $(i, j), i \neq j$ pair. For a given test, any of such alternatives should lead to rejection of the null when the sample sizes are large enough. Instead of imposing any parametric assumptions, we propose a set of non-parametric test statistics based on a similarity graph constructed using pairwise Wasserstein distance as detailed in Section 2.2 and a permutation procedure to capture these various possible alternatives. The permutation procedure treats the repeated measures from the same individual as the permutation unit. Specifically, the permutation is done by randomly assigning $n_1$ individuals out of the total $N$ individuals to group 1 and the rest to group 2. If an individual is assigned to group 1, then the repeated measures of the individual are labeled as observations from group 1. Note that we do not require equal correlation or exchangeability within individual among repeated observations since our data are observed sequentially over time. However, to ensure the exchangeability across individuals
under the null $H_0$, we do require that the number of repeated observations is the same for all the individuals. In the following, when there is no further specification, we use $P$, $E$, $Var$ and $Cov$ to denote probability, expectation, variance and covariance, respectively under this permutation null distribution. An illustration of the data structure and distribution assumptions are presented in Figure 3.

2.2. Graph-based statistics for data with repeated measures. Our proposed test statistics are constructed from similarity graph (Chen and Friedman, 2017) that includes both a within-individual graph and a between-individual graph in order to take into account repeated measures. To construct the graph based on the Wasserstein distance $d_W$, we pool all repeated measures from a total of $N = n_1 + n_2$ individuals, and construct a similarity graph $G$ as a minimum spanning tree (MST). MST is a spanning tree that connects all observations that minimizes the sum of the total distances of the edges in the tree. In particular, a $m$-MST is the union of the 1st MST, $\cdots$, $m$th MST, where the 1st MST is the MST and the $j$th ($j > 1$) MST is a spanning tree connecting all observations that minimizes the sum of distances across edges, but individual to the constraint that this spanning tree does not contain any edge from the previous 1st MST, $\cdots$, and $(j - 1)$th MST. Since the graph-based statistics are usually more powerful under a slightly denser graph (Friedman and Rafsky, 1979), we choose 9-MST for our similarity graph $G$ in our simulation studies and real application, following the recommendation by Chen, Chen and Su (2018). Based on the similarity graph $G$ of all the observations, we further divide its edges into two parts. If an edge connects two observations from the same individual, it belongs to the within-individual similarity graph $G_{\text{in}}$, otherwise, it belongs to the between-individual similarity graph $G_{\text{out}}$ (see Figure 2 for an illustration).

Given a constructed graph $G$, we let $D = (D_{uv})_{N \times N}$ be a symmetric matrix, where $D_{uv}$ denotes the number of edges between individuals $u$ and $v$ in $G$, and let $D_u = \sum_{v \neq u} D_{uv}$ be the total number of edges connecting individual $u$ and others. The total number of edges in $G$ is denoted by $|G|$. Furthermore, let $g_i$ be an indicator function that takes value 1 when node $i$ belongs to an individual from group 1, and 2 otherwise. We denote an edge in $G$ by the indices of the nodes that are connected by the edge such as $e = (u, v)$. Define

$$R_{\text{out}, k} = \sum_{(i,j) \in G_{\text{out}}} I(g_i = g_j = k), \quad R_{\text{in}, 1} = \sum_{(i,j) \in G_{\text{in}}} I(g_i = g_j = 1).$$
Fig 2. An example of similarity graph $G$, within-individual graph $G_{in}$ and between-individual graph $G_{out}$ for three individuals with repeated measures.

Here $R_{out,k}$ is the number of between-individual edges in $G_{out}$ that connect observations belonging to the same group $k$, $k = 1, 2$. $R_{in,1}$ is the number of within-individual edges in $G_{in}$ from group 1.

To accommodate various alternatives to the null hypothesis, we consider
six different test statistics presented in Table 1. The six test statistics are defined based on different functions of \( R_{\text{out},1} \), \( R_{\text{out},2} \) and \( R_{\text{in},1} \) and their expectations and variances calculated under the permutation null. These different test statistics are developed for testing the same null hypothesis \( H_0 : \tilde{P}_1 = \tilde{P}_2 \), but their statistical power depends on specific alternatives, as summarized in Figure 3. For each of the test statistics, under \( H_0 \) and a fixed graph \( G \), one could randomly shuffle the group assignments for all individuals to estimate their corresponding null distributions.

Specifically, \( T_{\text{in}} \) builds upon the contrast of within-individual edge counts between group 1 and group 2, holding the total number of edge counts to be constant. Hence it captures the covariance among the repeatedly observed densities. Rejecting \( H_0 \) based on \( T_{\text{in}} \) implies that \( \text{Cov}_{F_1} \neq \text{Cov}_{F_2} \), suggesting that the group difference occurs in the amount of day-to-day variability in daily activity distributions.

The next three test statistics \( Z_{\text{out},w} \), \( T_{\text{out},d} \) and \( M_{\text{out}}(\kappa) \) are developed to capture the group difference in the marginal distribution of individual activity densities \( F_1 \) and \( F_2 \). In particular, \( Z_{\text{out},w} \) evaluates the mean difference between the two groups, and rejecting \( H_0 \) implies that \( \mu_{F_1} \neq \mu_{F_2} \). Similarly, \( T_{\text{out},d} \) examines the group difference in between-individual variances and rejects \( H_0 \) when \( V_{F_1} \neq V_{F_2} \). Finally, \( M_{\text{out}}(\kappa) \) combines the comparison in both mean and variance by taking the maximum of the two. Note that
these statistics are adapted from the existing formulations from Zhang and Chen (2020). However this is not a direct application from the previous work due to the existence of repeated observations per individual. Our novelty lies in expanding the similarity graph to include both $G_{out}$ and $G_{in}$ that allow more than one edges connecting between any pair of nodes. As a consequence, we construct the statistics based on the weighted edge counts and new derivations are needed to obtain the asymptomatic distributions. The two final statistics $S_R$ and $M(\alpha, \kappa)$ combine the previously defined statistics in a weighted fashion and flexibly capture differences occurred in both the between-individual distributions $F_1$ and $F_2$ as well as the within-individual covariance.

Table 1

<table>
<thead>
<tr>
<th>Proposed test statistics for the difference of two population distributions of the density functions.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within-individual test</strong></td>
</tr>
<tr>
<td>Wasserstein covariance difference</td>
</tr>
<tr>
<td>$T_{in} =</td>
</tr>
<tr>
<td>Mean difference</td>
</tr>
<tr>
<td>$Z_{out,w} = \frac{(n_2-1)R_{out,1} + (n_1-1)R_{out,2} - E((n_2-1)R_{out,1} + (n_1-1)R_{out,2})}{\sqrt{\text{Var}((n_2-1)R_{out,1} + (n_1-1)R_{out,2})}}.$</td>
</tr>
<tr>
<td>Variance difference</td>
</tr>
<tr>
<td>$T_{out,d} =</td>
</tr>
<tr>
<td>Overall difference</td>
</tr>
<tr>
<td>$M_{out}(\kappa) = \max{T_{out,d}, \kappa Z_{out,w}}$</td>
</tr>
<tr>
<td>Joint between and within-individual test</td>
</tr>
</tbody>
</table>

| **Sum-type test** |
| $S_R = \begin{pmatrix}
R_{out,1} - E(R_{out,1}) \\
R_{out,2} - E(R_{out,2}) \\
R_{in,1} - E(R_{in,1})
\end{pmatrix}^{T} \Sigma^{-1} \begin{pmatrix}
R_{out,1} - E(R_{out,1}) \\
R_{out,2} - E(R_{out,2}) \\
R_{in,1} - E(R_{in,1})
\end{pmatrix},$
where $\Sigma = \text{Var}((R_{out,1}, R_{out,2}, R_{in,1})^{T})$. |

| **Max-type test** |
| $M(\alpha, \kappa) = \max\{T_{in}, \alpha M_{out}(\kappa)\}$ |

2.3. Analytic expressions of the new statistics. In the following, we first derive the exact analytic expressions for the expectation and variance of $(R_{out,1}, R_{out,2}, R_{in,1})$, so that the proposed test statistics in Section 2.2 can be computed efficiently. The analytic expressions are provided in the following theorem. The detailed proof could be found in the Supplementary Material.

**Theorem 2.1.** Under the permutation null, the analytic expressions of
the expectation of \((R_{\text{out},k}, R_{\text{in},1})^T, k = 1, 2\) are

\[
E(R_{\text{out},k}) = |G_{\text{out}}| \frac{n_k(n_k - 1)}{N(N - 1)}, E(R_{\text{in},1}) = |G_{\text{in}}| \frac{n_1}{N},
\]

the analytic expressions of the variances

\[
\text{Var}(R_{\text{out},k}) = \frac{n_1 n_2 (n_1 - 1)(n_2 - 1)}{N(N - 1)(N - 2)(N - 3)} \times \left\{ \frac{1}{2} \sum_{u \neq v} D_{uv}^2 + \frac{n_k - 2}{N - n_k - 1} \left( \sum_u D_u^2 - \frac{4|G_{\text{out}}|^2}{N} \right) - \frac{2}{N(N - 1)} |G_{\text{out}}|^2 \right\},
\]

\[
\text{Var}(R_{\text{in},1}) = \frac{n_1 n_2}{N(N - 1)} \left( \sum_u D_u^2 - \frac{|G_{\text{in}}|^2}{N} \right),
\]

and the analytic expressions of covariance

\[
\text{Cov}(R_{\text{out},k} R_{\text{out},2}) = \frac{n_1 n_2 (n_1 - 1)(n_2 - 1)}{N(N - 1)(N - 2)(N - 3)} \times \left\{ \frac{1}{2} \sum_{u \neq v} D_{uv}^2 - \left( \sum_u D_u^2 - \frac{4|G_{\text{out}}|^2}{N} \right) - \frac{2}{N(N - 1)} |G_{\text{out}}|^2 \right\},
\]

\[
\text{Cov}(R_{\text{out},k} R_{\text{in},1}) = (-1)^{k+1} \frac{n_1 n_2 (n_k - 1)}{N(N - 1)(N - 2)} \left( \sum_{u=1}^N D_{uu} D_u - \frac{2}{N} |G_{\text{in}}| |G_{\text{out}}| \right).
\]

Using the results of Theorem 2.1, we can check that under the permutation null, \(E(Z_{\text{out},w}) = E(Z_{\text{out},d}) = E(Z_{\text{in}}) = 0\), \(\text{Var}(Z_{\text{out},w}) = \text{Var}(Z_{\text{out},d}) = \text{Var}(Z_{\text{in}}) = 1\), and \(\text{Cov}(Z_{\text{out},w}, Z_{\text{out},d}) = 0\), \(\text{Cov}(Z_{\text{out},w}, Z_{\text{in}}) = 0\). In addition,

\[
\text{Cov}(Z_{\text{out},d}, Z_{\text{in}}) = \frac{\sum_{u=1}^N D_{uu} D_u - \frac{2}{N} |G_{\text{in}}| |G_{\text{out}}|}{\sqrt{(\sum_{u=1}^N D_u^2 - \frac{4|G_{\text{out}}|^2}{N})(\sum_{u=1}^N D_{uu}^2 - \frac{|G_{\text{in}}|^2}{N})}}.
\]

It is straightforward to verify that the statistic \(S_R\) can be rewritten in the following form

\[
S_R = (Z_{\text{out},w}, Z_{\text{out},d}, Z_{\text{in}})\Omega^{-1}(Z_{\text{out},w}, Z_{\text{out},d}, Z_{\text{in}})^T,
\]

where \(\Omega = \text{Var}\{(Z_{\text{out},w}, Z_{\text{out},d}, Z_{\text{in}})^T\}\). The detailed proof is provided in the Supplementary Material.
3. Asymptotic distribution under the permutation null. The critical values of the test statistics can be determined by performing permutations of individual nodes as stated in Section 2.1. However, such a permutation procedure is often time consuming. To make the tests computationally more efficient, we have derived the asymptotic null distributions of the test statistics. In Section 4, we examine how the critical values obtained from asymptotic results agree with those obtained through permutation directly in finite sample settings.

Before stating the theorem, we need to define a few additional notations for the similarity graph $G$. Denote by $C_u$ the set of repeated measures belonging to individual $u$. For an edge $e = (i, j) \in G_{out}$, $i \in C_u$, $j \in C_v$ ($u \neq v$), let $A_{out,e}$ be the subset of edges that share nodes with $e$ as

$$A_{out,e} = \{e\} \cup \{e' = (k, l) \in G_{out} : k \in C_u \cup C_v \text{ or } l \in C_u \cup C_v\} \cup \{e'' \in G_{in} : e'' \text{ in individuals } u \text{ or } v\}.$$

For an edge $e = (i, j) \in G_{in}$, $i, j \in C_u$, let

$$A_{in,e} = \{e' \in G_{out} : \text{one endpoint of } e' \in C_u\} \cup \{e'' \in G_{in} : e'' \text{ in individual } u\}.$$

Define

$$A_e = A_{out,e}I(e \in G_{out}) + A_{in,e}I(e \in G_{in}),$$
$$B_{out,e} = \cup_{e' \in A_{out,e}} A_{e'}, B_{in,e} = \cup_{e'' \in A_{in,e}} A_{e''},$$
$$B_e = B_{out,e}I(e \in G_{out}) + B_{in,e}I(e \in G_{in}).$$

To derive the asymptotic null distribution of the proposed test statistics, we assume $n_1 = O(N)$, $n_2 = O(N)$ and $l = O(1)$. In addition, the following conditions are needed:

Condition 1 (C1): $|G_{out}|, |G_{in}| = O(N)$.
Condition 2 (C2): $\sum_u D_u^2 - 4|G_{out}|^2/N$, $\sum_u D_{uu}^2 - |G_{in}|^2/N = O(N)$.
Condition 3 (C3): $\sum_{e \in G_{out}} |A_{out,e}| |B_{out,e}| = o(N^{1.5})$.

Here, we use $a = O(b)$ to denote that $a$ and $b$ are of the same order and $a = o(b)$ to denote that $a$ is of a smaller order than $b$.

Condition C1 requires that the number of the edges in $G_{out}$ and $G_{in}$ is in the same order as $N$. Condition C2 guarantees that $(R_{out,1}, R_{out,2}, R_{in,1})^T$ does not degenerate asymptotically. Since

$$\sum_u D_u^2 - 4|G_{out}|^2/N = \sum_u \left(D_u - \frac{2|G_{out}|}{N}\right)^2,$$
$$\sum_u D_{uu}^2 - |G_{in}|^2/N = \sum_u \left(D_{uu} - \frac{|G_{in}|}{N}\right)^2,$$
if \( D_u - 2|G_{\text{out}}|/N = O(1) \) and \( D_{uu} - |G_{\text{in}}|/N = O(1) \), then Condition C2 is satisfied. Condition C3 requires the number of edges from an individual in the graph \( G \) such being not too large. A similar condition was needed for graph-based statistics for independent observations (Chen and Friedman, 2017; Chen, Chen and Su, 2018). Conditions C1 and C2 imply that \( \sum_u D_u^2, \sum_u D_{uu}^2 = O(N) \). In addition, we note that \( 2|G_{\text{out}}| = \sum_{u \neq v} D_{uv} \leq \sum_{u \neq v} D_{uv}^2 \leq \sum_u D_u^2 \) and \( |G_{\text{in}}| = \sum_u D_{uu} \leq \sum_u D_{uu} D_u \leq \sqrt{\sum_u D_{uu}^2 \sqrt{\sum_u D_u^2}} \), which leads to

\[
\sum_{u \neq v} D_{uv}^2 = O(N) \quad \text{and} \quad \sum_u D_{uu} D_u = O(N).
\]

We assume the following limits exist,

\[
\lim_{N \to \infty} \frac{|G_{\text{out}}|}{N} = b_1, \quad \lim_{N \to \infty} \frac{\sum_u D_u^2 - 4|G_{\text{out}}|^2}{N^2} = b_2, \quad \lim_{N \to \infty} \frac{\sum_{u \neq v} D_{uv}^2}{N} = b_3,
\]

\[
\lim_{N \to \infty} \frac{|G_{\text{in}}|}{N} = b_4, \quad \lim_{N \to \infty} \frac{\sum_u D_{uu}^2}{N} - \frac{|G_{\text{in}}|^2}{N^2} = b_5, \quad \lim_{N \to \infty} \frac{\sum_u D_{uu} D_u}{N} = b_6.
\]

The following theorem presents the asymptotic distribution of \((Z_{\text{out},w}, Z_{\text{out},d}, Z_{\text{in}})^T\) under the permutation null when \( N \to \infty \).

**Theorem 3.1.** Under Conditions 1-3 and under the new permutation null distribution, as \( N \to \infty \), \((Z_{\text{out},w}, Z_{\text{out},d}, Z_{\text{in}})^T\) converges to a multivariate Gaussian distribution with mean 0 and covariance matrix

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & \rho_Z \\
0 & \rho_Z & 1
\end{pmatrix},
\]

where

\[
\rho_Z = \frac{b_6 - 2b_1 b_4}{\sqrt{b_2 b_5}}.
\]

Based on Theorem 3.1, it is easy to obtain the asymptotic cumulative distribution functions (CDF) of \( T_{\text{in}}, Z_{\text{out},w}, T_{\text{out},d}, M_{\text{out}}(\kappa), S_R \) and \( M(\alpha, \kappa) \) under the permutation null. They are given in the following Corollary 3.2.

**Corollary 3.2.** Under Conditions C1-C3, and under the permutation null distribution, as \( N \to \infty \) the asymptotic CDFs for each of the test statistic are

1. \( P(T_{\text{in}} \leq x) \to 2\Phi(x) - 1 \).
2. \( P(Z_{out,w} \leq x) \rightarrow \Phi(x) \).
3. \( P(T_{out,d} \leq x) \rightarrow 2\Phi(x) - 1 \).
4. \( P(M_{out}(\kappa) \leq x) \rightarrow (1 - 2\Phi(-x))\Phi(x/\kappa) \)
5. \( P(S_R \leq x) \rightarrow \chi^2(3) \)
6. \( P(M(\alpha,\kappa) \leq x) \rightarrow \Phi(x/\alpha) \cdot P(-x/\alpha \leq Z_{out,d} \leq x/\alpha, -x \leq Z_{in} \leq x) \)

where \( \Phi(\cdot) \) denotes the CDF of a standard normal distribution.

The term \( P(-x/\alpha \leq Z_{out,d} \leq x/\alpha, -x \leq Z_{in} \leq x) \) can be calculated from function \texttt{pmvnorm()} in the \texttt{R} package \texttt{mvtnorm}, where the correlation between \( Z_{out,d} \) and \( Z_{in} \) can be estimated using finite sample estimate,

\[
\rho_{Z,N} = \frac{\sum_{u=1}^{N} D_{uu} D_u - \frac{2}{N} |G_{in}| |G_{out}|}{\sqrt{(\sum_{u=1}^{N} D_u^2 - \frac{4|G_{out}|^2}{N})(\sum_{u=1}^{N} D_{uu}^2 - \frac{|G_{in}|^2}{N})}}.
\]

It is easy to see that \( \lim_{N\to\infty} \rho_{Z,N} = \rho_Z \).

4. Simulation studies. We evaluate the performance of the proposed test statistics \( T_{in}, Z_{out,w}, T_{out,d}, M_{out}(\kappa), S_R \) and \( M(\alpha,\kappa) \) under various simulation settings. Under each setting, we compare the results with the generalized edge-count test (\( S \)) of Chen and Friedman (2017) and the Fréchet test (\texttt{Fretest}) of Dubey and Müller (2019). As far as we know, neither method allows for data with repeated measures and would rely on between-individual distance metrics from a single observation. Simply applying those tests on the individual observations without accounting for within-subject correlation leads to an inflated type 1 error (results omitted). To ensure a fair comparison, we apply these two tests on the subject-level graph constructed based on two definitions of distance metrics that respect the hierarchical structure among the repeated observations. The first distance is chosen to be the Wasserstein distance calculated from each subject’s barycenter (average distance). Alternatively, we use the integrated distance by taking the square root of the total sum square distances across all the \( l \) observations for any pair of individuals. We denote the generalized edge-count test and the Fréchet test calculated under the first distance metric by \( S_1 \) and \texttt{Fretest}1, and those calculated under the second definition as \( S_2 \) and \texttt{Fretest}2, respectively. More specifically, let \( Z_u = (Z_{u1}, \cdots, Z_{ul}) \) and \( Z_v = (Z_{v1}, \cdots, Z_{vl}) \), where \( Z_{uj}, Z_{vj} \ (j = 1, \cdots, l) \) represent the repeated measures for individuals \( u \) and \( v \), the two distances are defined as:
1. Average distance: \( d(Z_u, Z_v) = d_W(\tilde{Z}_u, \tilde{Z}_v) \), where \( \tilde{Z}_u \) and \( \tilde{Z}_v \) are the barycenters of \( Z_u \) and \( Z_v \), respectively, i.e.,

\[
\tilde{Z}_u = \arg \min_{x \in \Omega} \sum_{i=1}^{l} d_W(x, Z_{ui}), \quad \tilde{Z}_v = \arg \min_{x \in \Omega} \sum_{i=1}^{l} d_W(x, Z_{vi}).
\]

2. Integrated distance: \( d(Z_u, Z_v) = \sqrt{\sum_{i=1}^{l} d^2_W(Z_{ui}, Z_{vi})} \).

Following the recommendations from Zhang and Chen (2020), when there is no prior knowledge about the type of between-individual difference (i.e., location difference or scale difference), we choose \( \kappa = 1 \) for the statistic \( M_{out}(\kappa) \) and denote it by \( M_{out} \) for simplicity. For the statistic \( M(\alpha, \kappa) \), the parameter \( \alpha \) weights the between-individual difference. Here, we let \( \kappa = 1.14 \) and \( \alpha = 1 \), and denote the statistic by \( M \) for simplicity.

The general setup for the simulation settings is as following. We generate the observed physical activity density for individual \( u \) on day \( j \) to be equal to the density function of a \( p \)-dimensional multivariate normal distribution with mean \( \theta_{uj} \) and variance \( \omega^2_u I_p \). That is, \( X_{uj} = \psi_p(\theta_{uj}, \omega^2_u I_p), j = 1, \ldots , l \). We further assume that \( \omega_u \) is independent and identically distributed from a uniform distribution \( U(\nu_{k1}, \nu_{k2}) \), with \( k = 1, 2 \) corresponding to group label. \( \theta_{uj} \) \( (j = 1, \ldots , l) \) are sampled from another multivariate normal distribution \( N_p(a_{ku}, \sigma^2 I_p) \) with individual-specific mean \( a_{ku} \). We further assume an exchangeable correlation between \( \theta'_{uj} \)s, which leads to

\[
(\theta'^T_{u1}, \ldots , \theta'^T_{ul})^T | \mu_{ku} \sim N_p(\mu_{ku}, \sigma^2 \rho_k \otimes I_p),
\]

where \( \otimes \) denotes the Kronecker product, \( \mu_{ku} = (a^T_{ku}, \ldots , a^T_{ku})^T \) and \( \rho_k = \rho_k I_l + (1 - \rho_k) I_l \). Here, \( a_{ku} \) \( (u = 1, \ldots , n) \) \( \sim \) \( N_p(\beta_k, \epsilon^2 I_p) \). We also consider an exponentially decayed correlation between \( \theta'_{uj} \)s with the \( (s,t) \)-element of \( \rho_k \) being \( \rho_k^{[s-t]} \), \( s, t = 1, \ldots , l \). The results are similar and are given in the Supplementary Material.

We simulate unbalanced data with \( n_1 = 50, n_2 = 80 \) individuals for each group and \( l = 5 \) days for each individual. When applying the proposed statistics, we use the Wasserstein distance to measure the dissimilarity between any two density functions, which can be explicitly calculated. The similarity graph \( G \) is constructed by the procedure outlined in Section 2.2 with 9-MST.

4.1. Simulations for one-dimensional density, \( p = 1 \). We consider five different parameter settings as listed on the top rows of Table 2. All the
test statistics are compared in terms of type 1 error and power. Here, Model (A1) is the null model when there is no difference between the two groups, Models (A2)-(A4) represent the cases where the two groups differ in within-individual covariance, between-individual mean and between-individual variability, respectively. Model (A5) represents the case that differences exist in mean, variance and also in the within-individual covariance.

Table 2
Parameter values for 5 different simulation settings for comparisons. (A) one-dimensional density functions; (B) 30-dimensional density functions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Parameter Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1:</td>
<td>null model.</td>
<td>( \rho_1 = 0.6, \beta_1 = 0, \epsilon_1 = 1, \nu_{11} = 1, \nu_{12} = 2; ) ( \rho_2 = 0.6, \beta_2 = 0, \epsilon_2 = 1, \nu_{21} = 1, \nu_{22} = 2; \sigma = 1. )</td>
</tr>
<tr>
<td>A2:</td>
<td>between-individual variability difference in ( \rho. )</td>
<td>( \rho_1 = 0, \beta_1 = 0, \epsilon_1 = 1, \nu_{11} = 1, \nu_{12} = 1.2; ) ( \rho_2 = 0.8, \beta_2 = 0, \epsilon_2 = 1, \nu_{21} = 1, \nu_{22} = 1.2; \sigma = 1. )</td>
</tr>
<tr>
<td>A3:</td>
<td>between-individual mean difference in ( \beta ) and ( \nu_1 + \nu_2. )</td>
<td>( \rho_1 = 0, \beta_1 = 0, \epsilon_1 = 1, \nu_{11} = 1, \nu_{12} = 1.2; ) ( \rho_2 = 0, \beta_2 = 0.7, \epsilon_2 = 1, \nu_{21} = 0.96, \nu_{22} = 1.16; \sigma = 1. )</td>
</tr>
<tr>
<td>A4:</td>
<td>between-individual variability difference in ( \epsilon ) and ( \nu_2 - \nu_1. )</td>
<td>( \rho_1 = 0, \beta_1 = 0, \epsilon_1 = 1, \nu_{11} = 1, \nu_{12} = 1.3; ) ( \rho_2 = 0, \beta_2 = 0, \epsilon_2 = 1.1, \nu_{21} = 0.97, \nu_{22} = 1.33; \sigma = 1. )</td>
</tr>
<tr>
<td>A5:</td>
<td>between-individual variability difference in ( \rho, ) between-individual mean difference in ( \beta ) and ( \nu_1 + \nu_2, ) variance difference in ( \epsilon ) and ( \nu_2 - \nu_1. )</td>
<td>( \rho_1 = 0, \beta_1 = 0, \epsilon_1 = 1, \nu_{11} = 1, \nu_{12} = 1.3; ) ( \rho_2 = 0.35, \beta_2 = 0.5, \epsilon_2 = 1.1, \nu_{21} = 0.97, \nu_{22} = 1.36; \sigma = 1. )</td>
</tr>
</tbody>
</table>

Table 3 shows the empirical power of the proposed statistics at \( \alpha = 0.05 \) level based on 1,000 replications. Under the null model (A1), all the statistics
are able to control the type 1 errors at the nominal level.

As for detecting the group differences in the alternative Models (A2-A5), the power of $S$ and Fréchet test is uniformly lower than our proposed statistics except for $S_2$ under Model (A2). As expected, the power of different test statistics depends on the alternative hypothesis. In Model (A2), when $\rho_1$ is different from $\rho_2$, $T_{in}$ shows its superior performance of detecting group differences in covariance among repetaed measures within individuals. For Model (A3), since the difference only happens in the group mean parameters, all the proposed test statistics except $T_{in}$ and $Z_{out,w}$ yield high power. Model (A4) is designed to examine the power of the tests when the between-individual variability is different between the two groups. We observe that indeed all the proposed tests except $T_{in}$ and $Z_{out,w}$ have high power. The results for Model (A5) suggest that $T_{out,d}$ works well for detecting group difference in between-individual variability and $Z_{out,w}$ is suitable for detecting differences in the between-individual mean. Since there is a smaller difference in $\rho$'s than that under Model (A2), $T_{in}$ did not yield high power in this scenario.

4.2. Simulations for moderate-dimensional density, $p = 30$. Although we have mostly been concerned with two-sample testing for one-dimensional probability densities based on a single morality of measurements such as
physical activity intensity, it is worth noting that our proposed tests are
directly applicable to density objects from multimodal measurements as long
as there is a well defined distance metrics. In fact, many wearable devices
simultaneously collect multiple markers such as heart rate, respiratory rate
in addition to the physical movement, and there is need to compare the joint
density distributions of multivariate measures in mobile health research.
To illustrate their utility for multivariable density objects with repeated
measures, we conduct another set of simulation studies for $p = 30$. Our
simulation setups are similar to $p = 1$ case, where we simulate an unbalanced
sample with $n_1 = 50$, $n_2 = 80$ individuals in each group and $l = 5$ repeated
measures per individual. All of the statistics are assessed and compared
under five different scenarios as listed in Table 2. These 5 models parallel
the Models (A1)-(A5), except that we consider density measures for 30-
dimensional variables.
Table 3 shows the estimated empirical powers of the proposed statistics
at 0.05 significance level based on 1000 simulations. Again we observe that
all the statistics control the type 1 errors at the approximate level. However,
the type 1 errors of the Fréchet tests are slightly inflated.
For Models (B2)-(B4), the power of the proposed tests remain similar
to those under the one-dimensional setting in Section 4.1. As a compar-
ison, although tests $S_1$ and $S_2$ can detect the between-individual mean and
variance differences (B3, B4), they are not effective for detecting the within-
individual variability difference (B2). Fretest1 and Fretest2, on the other
hand, work well when only between-individual variance differ as in Model
(B4). The results for Model (B5) indicate that the proposed tests $S_R$ and $M$
perform well for the overall difference and is much better than the competing
tests $S_1$, $S_2$, Fretest1 and Fretest2.
Finally, we also perform simulations to examine whether the asymptotic
$p$-values could approximate the $p$-values obtained from 10,000 permutations.
The results show that the two $p$-values are very close and the power obtained
by the asymptotic $p$-value is similar to that based on the permutation $p$-value
for all the proposed test statistics. As sample size increases, the results are
almost identical as expected. We omit the details here and present the results
in the Supplementary Material, Section C.

5. Comparisons of physical activity distributions in mood dis-
order samples. We apply each of the six test statistics to the continuous
physical activity measures collected from a subset of the participants from
the National Institute of Mental Health (NIMH) Family Study of Spectrum
Disorders (Merikangas et al., 2014; Shou et al., 2017; Merikangas et al.,
In this study, 384 individuals were instructed to wear the Philips Activitywatch devices for about two weeks. The daily activity data were processed into 1440 minute-level intensity values each day. Meanwhile, the 384 individuals were interviewed and assessed into four clinical groups based on DSM-IV criteria as: healthy control (HC), major depressive disorders (MDD), type-I bipolar disorders (BPI) and type-II bipolar disorders (BPII). Previous research studies have consistently reported a lower average daytime motor activities among bipolar patients based on summary statistics from physical activity measures (Scott et al., 2017; Murray et al., 2020). Age and body mass index (BMI) are among the other factors are known to be associated with the mean activity levels (Schott, 2007; Varma et al., 2017). However, although there were a few papers that suggested potential links between bipolar disorder and interindividual and intraindividual variability in activity patterns, the evidence was much less robust and the extracted markers for quantifying variability was quite heterogeneous (Indic et al., 2011; Pagani et al., 2016; Robillard et al., 2015), making it even more challenging to understand the complex disease manifestations. Hence we focus on comparing the continuous physical activity profiles and testing whether mean and variability of the daily physical activity differ across disease groups or by demographic characteristics. To apply our proposed methods, we first estimate the empirical daily probability densities using the observed minute-by-minute activity intensities. Let \( Z_{u_j} = (z_{u_j}^{(1)}, \ldots, z_{u_j}^{(1440)})^T \) be the vector of ordered 1440 activity intensities for individual \( u \) on day \( j \). Here \( z_{u_j}^{(q)} \) represents the empirical \( q \)-th quantile of the probability distribution of activity intensities per day. The Wasserstein distance metric is calculated to quantify the distance between two empirical distributions based on any pairs of \( Z_{u_i} \) and \( Z_{v_j} \). Since densities are empirically estimated from the ordered values, the Wasserstein distance between densities is equivalent to the Euclidean distance between the two empirical quantiles, i.e.,

\[
d(Z_{u_i}, Z_{v_j}) = \left( \sum_{q=1}^{1440} (z_{u_i}^{(q)} - z_{v_j}^{(q)})^2 \right)^{1/2}
\]

We further construct the similarity graph \( G \) following the procedure that is introduced in Section 2.2 with 9-MST. As a sensitivity analysis, we also apply the tests using 5-MST, 15-MST and under the maximum mean discrepancy (Gretton et al., 2012). The results are similar and are provided in the Supplementary Material, Section E.

Considering the potential difference in daily routines and movement between weekdays and weekends, we apply the test statistics separately to
observations collected on weekdays and weekends with \( l = 7 \) and \( l = 3 \) days, respectively. For each analysis, the individuals with fewer than the given number of days \( l \) are excluded from the analysis. For those with more than \( l \) days of observations, a random subset of \( l \) days are included in generating the test statistics. Sensitivity analysis was conducted by repeating the random subsetting 1000 times in order to assess the variability in the test results due to choice of days (Supplementary Material, Section F). To summarize the results, we take the \( p \)-value \( p_j, j = 1, 2, \ldots, 1000 \) from each of the 1000 trials, and estimate an overall \( p \)-value as

\[
\hat{p} = 1 - \frac{2}{1 + e^{2\theta}},
\]

where

\[
\theta = \frac{1}{1000} \sum_{j=1}^{1000} \frac{1}{2} \log \left( \frac{1 + p_j}{1 - p_j} \right).
\]

5.1. Comparison of activity densities between healthy individuals and those with mood disorders. We first compare the activity densities among healthy individuals and those with histories of mood disorders in the free-living conditions during the weekdays and weekends. Figure 4 shows the \( p \)-values of the pairwise comparisons using the proposed test statistics, the generalized edge-count tests \((S_1, S_2)\) and Fréchet tests (Fretest1, Fretest2) (detailed \( p \)-values and the sample sizes for different groups are given in Table 5 of the Supplementary Material). We first observe that the differences between diagnostic groups are mostly driven by activity patterns on weekends and no significant difference is observed during the weekdays. In particular, we observe that the healthy individuals have significantly different activity distributions from those with BPI. Among the proposed statistics, \( Z_{\text{out},w} \) achieves the most significant results when comparing healthy with BPI and BPII vs BPI. While \( T_{\text{in}} \) and \( T_{\text{out},d} \) result in nonsignificant large \( p \)-values and cannot reject the null hypothesis. These results suggest that there exist significant differences in the population-level mean activity density between healthy and BPI or between BPI and BPII. This is consistent with findings from the existing literature where BPI patients were found to have lower average activity levels especially in the later of the day (Scott et al., 2017; Shou et al., 2017) and less time spent in MVPA (Chapman et al., 2017). But no significant difference is observed in the variance of activity densities or in day-to-day variability of the activity density. Since all of \( M_{\text{out}}(\kappa) \), \( S_R \) and \( M(\alpha, \kappa) \) include a \( Z_{\text{out},w} \) in their definitions, they are also effective to capture the mean difference of activity densities when \( Z_{\text{out},w} \) yields a small \( p \)-value.
5.2. Comparison of activity distributions among different age groups. It is well known that age is associated with the amount of physical activity. For example, Schrack et al. (2014) found ‘a 1.3% decrease per year’ in cumulative physical activity counts from mid-to-late life among a elderly population. Similar results have been reported in several other large cohort studies and age groups including NHANES and UK Biobank (Varma et al., 2017; Viciana, Mayorga-Vega and Martínez-Baena, 2016; Doherty et al., 2017). However, few studies has examined how inter- and intra-individual variability in physical activity differ by age. We ask whether the proposed test statistics are able to detect differences in the daily activity densities over different age categories and inform us where the difference lies. To ensure a proper power with an adequate sample size, we take the two diagnostic groups with the largest sample sizes, the HC and major depressive disorder (MDD), and stratify them into three age groups, young (age ≤ 30), middle age (30 < age ≤ 60) and older age (age > 60) groups. We also separately test activity densities from weekdays and weekends.

The p-values of the proposed test statistics, the generalized edge-count test \((S_1, S_2)\) and Fréchet test \((\text{Fretest1}, \text{Fretest2})\) are shown in Figure 5 with detailed \(p\)-values given in Table 6 of the Supplementary Material. Overall, among the healthy individuals, the proposed statistical tests find large differences in the distributions of activity intensities among the three age groups for both weekdays and weekends. Such differences are especially prominent when comparing the young age group or middle age group with the older
group during the weekdays. In contrast, Fréchet test fails to detect such differences in most of the comparisons and is only able to capture marginally significant results when comparing the young and older individuals among MDD patients. The tests S1 and S2 also show fewer significant results than our proposed tests.

To further demonstrate the possible gain of power, we note that among the patients with MDD, only the proposed \( Z_{\text{out},w} \) test shows statistically significant difference between young and older groups for both weekend and weekday activities. Fréchet test shows some difference in activity distributions between young and older groups but only for the weekdays. To confirm the detected differences in the original data, we visualize the density data in Figure 6 by projecting them onto a lower-dimensional plots using multidimensional scaling (MDS) based on the Wasserstein distances. The figure
clearly shows difference in activity densities between young and older groups for both weekdays and weekends among MDD patients.

Finally, it is also interesting that $T_{in}$ detects significant difference in day-to-day variability between the healthy young group and older groups. In fact, we obtained a negative value for $Z_{in}$ which implies that the younger subjects have larger day-to-day variability than the two groups with older ages. Lastly, $T_{out,d}$ does not yield any significant results, indicating that there is large subject heterogeneity within each age group, yet their scales are comparable.

5.3. Comparison of activity distributions among different BMI groups. A third factor that could potentially affect differential physical activity patterns is the body mass index (BMI). We apply our proposed tests to examine difference in daily activity density among individuals who are lean with BMI $\leq 25$ and obese with BMI $> 25$ among healthy individuals and those with mood disorders (OTHER). To control for the age effects, we only consider those individuals with age of 30 years or older.

The results are provided in Figure 7. Among the healthy individuals, little
difference is observed in their activity distribution patterns between lean and obese individuals during the weekdays and weekends. When assessing among patients in the OTHER group, we observe some differences in the mean of the activity distributions both during weekdays and weekends. We do not see group differences in the within-individual or between-individual variability. The generalized edge-count test and Fréchet test achieve nonsignificant large $p$-values and fail to reject the null hypothesis for all the comparisons. This further demonstrates that our proposed test statistics can detect difference in activities that could be missed by other methods.

6. Discussion. In this paper, we have extended the graph-based two-sample tests for density data and proposed several test statistics to account for repeated measure data by considering both the within-individual similarity graph $G_{in}$ and between-individual similarity graph $G_{out}$. The graph allows for more than one edge between any two individuals, which extends the existing graph-based testing methods where only one edge between any two individuals is allowed. We have proposed a list of six test statistics that capture different alternatives that are associated with distributions of density functions, including differences in mean, inter- and intra-individual variances. These statistics are constructed based on the similarity graph $G$, which is the union of $G_{in}$ and $G_{out}$. Furthermore, we have developed the asymptotic null distributions that can be used to obtain $p$-values under the permutation null. The test statistics are easy to calculate and the testing procedures are computationally efficient. Our simulation studies have shown that the proposed test statistics control the desired type 1 errors and are
more powerful than existing distance-based tests that ignore the repeated observations.

In our analysis of the physical activity measures with repeated observations, we have observed a substantial differences in the day-to-day variability within subject across disease groups and age categories. Such findings have rarely been reported previously. Our proposed tests are able to take into account such within-individual dependency and variability. Compared to the two versions of Fréchet tests, we observed increased power in detecting the differences in activity densities. In addition, by comparing results utilizing various proposed test statistics, we are able to further understand the complex data structures and decompose the source of differences between various groups.

Our proposed permutation procedure treats the entire vector of repeated observations of objects from an individual as the independent unit, which requires that we have the same number of observations for each individual. Otherwise, the within-individual Wasserstein covariance is not well defined. This approach eliminates any assumptions on the errors of the repeated measures. For example, we do not require that the repeated measures have the same marginal distribution and allow them to be different from day to day. In our analysis of the NIMH physical activity data, we noticed that the results are robust to different subsets of the observations used in our analysis and reported an average through Fisher’s transformation. However, there might be the case when an unequal number of repeated observations might be informative, in which case one should interpret the results with care. An interesting future research topic is to extend the proposed tests to allow for different numbers of repeated observations by making additional assumptions on these repeated measures.

SUPPLEMENTARY MATERIAL

Supplement to “Two-sample tests for repeated measurements of histogram objects with applications to wearable device data”: (http://www.e-publications.org/ims/support/download/imsart-ims.zip). The supplementary material contains the following:
Supplement A provides detailed proof of Theorems 2.1 and 3.1 that derive the analytic expressions and the asymptotic distributions of the proposed test statistics.
Supplement B provides simulation results for data with an exponentially decayed within-individual correlation.
Supplement C includes additional simulation results comparing the asymptotic and permutation p-value over 100 simulation runs.
Supplement D provides results of real data with 9-MST. Supplement E provides results of real data comparisons when adopting 5-MST, 15-MST as the similarity graph and adopting maximum mean discrepancy as a metric. Supplement F shows the Boxplots of -log(p-value) for each test over 1000 random subsetting of $l$ days of the weekday data as a sensitivity analysis to examine the robustness of choice of days.

Acknowledgments. This study was supported in part by the Inter-Agency Agreement from National Institute of Mental Health for Dr. Shou and NIH grants GM129781 and GM123056.

References.


ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate
distributions based on adjacency. Journal of the Royal Statistical Society: Series B
(Statistical Methodology) 67 515–530.
SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension
is large relative to the sample sizes. Computational Statistics & Data Analysis 51 6535–
6542.
SCHRACK, J. A., ZIPUNNIKOV, V., GOLDSMITH, J., BAI, J., SIMONSICK, E. M.,
CRAINICEANU, C. and FERRUCCI, L. (2014). Assessing the "physical cliff": detailed
quantification of age-related differences in daily patterns of physical activity. J Geron-
SCHRACK, J. A., COOPER, R., KOSTER, A., SHIROMA, E. J., MURABITO, J. M., RE-
in Older Adults: Unraveling the Complexity of Monitors, Measures, and Methods. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 71
1039–1048.
SCOTT, J., MURRAY, G., HENRY, C., MORKEN, G., SCOTT, E., ANGST, J., MERIKAN-
Review. JAMA Psychiatry 74 189–196.
SHOU, H., CUI, L., HICKIE, I., LAMEIRA, D., LAMERS, F., ZHANG, J., CRAINICEANU, C.,
ZIPUNNIKOV, V. and MERIKANGAS, K. R. (2017). Dysregulation of objectively assessed
24-hour motor activity patterns as a potential marker for bipolar I disorder: results of
a community-based family study. Translational Psychiatry 7 e1211.
101 102–108.
233–235.
VICIANA, J., MAYORGA-VEGA, D. and MARTÍNEZ-BAENA, A. (2016). Moderate-to-
Vigorous Physical Activity Levels in Physical Education, School Recess, and After-
1117–1123.
WROBEL, J., ZIPUNNIKOV, V., SCHRACK, J. and GOLDSMITH, J. (2019). Registration for
function on scalar regression analysis for distributional data. J Am Stat Assoc 115
90–106.
ZHANG, J. and CHEN, H. (2020). Graph-based two-sample tests for data with repeated
observations. Statistica Sinica.

DIVISION OF BIOSTATISTICS
DEPARTMENT OF BIOSTATISTICS, EPIDEMIOLOGY AND INFORMATICS
UNIVERSITY OF PENNSYLVANIA PERELMAN SCHOOL OF MEDICINE
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: jingru.zhang@pennmedicine.upenn.edu
E-MAIL: hongzhe@pennmedicine.upenn.edu
E-MAIL: hshou@pennmedicine.upenn.edu

GENETIC EPIDEMIOLOGY RESEARCH BRANCH
NATIONAL INSTITUTE OF MENTAL HEALTH
BETHESDA, MD 20892
USA
E-MAIL: merikank@mail.nih.gov