Hypothesis Testing for Matched Pairs with Missing Data by Maximum Mean Discrepancy: An Application to Continuous Glucose Monitoring

Marcos Matabuena, Paulo Félix, Marc Ditzhaus, Juan Vidal, and Francisco Gude

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain; Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik (FMA), Institut für Mathematische Stochastik (IMST), Magdeburg, Germany; Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Santiago de Compostela, Spain; Unidad de Epidemiología Clínica, Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain

ABSTRACT
A frequent problem in statistical science is how to properly handle missing data in matched paired observations. There is a large body of literature coping with the univariate case. Yet, the ongoing technological progress in measuring biological systems raises the need for addressing more complex data, for example, graphs, strings, and probability distributions. To fill this gap, this article proposes new estimators of the maximum mean discrepancy (MMD) to handle complex matched pairs with missing data. These estimators can detect differences in data distributions under different missingness assumptions. The validity of this approach is proven and further studied in an extensive simulation study, and statistical consistency results are provided. Data obtained from continuous glucose monitoring in a longitudinal population-based diabetes study are used to illustrate the application of this approach. By employing new distributional representations along with cluster analysis, new clinical criteria on how glucose changes vary at the distributional level over 5 years can be explored.

KEYWORDS
Diabetes mellitus; Distributional representations; Kernel methods; Paired missing data

ARTICLE HISTORY
Received June 2022
Accepted April 2023

1. Introduction
A common experimental design in clinical studies, specifically longitudinal ones, is the matched pair design where observations are made from the same subjects under two different conditions, often at two points in time. Testing the null hypothesis that observations are drawn from the same distribution represents an essential step before performing any modeling task. However, a typical issue while dealing with paired data is missing data.

The literature on matched pairs with missing data has primarily focused on one-dimensional, continuous, discrete, or ordinal variables, aimed at detecting changes in location/mean (Ekbohm 1976; Xu and Harrar 2012; Martinez-Camblo et al. 2013; Guo and Yuan 2017), scale/variance (Derrick et al. 2018), and distribution (Gaigall 2020). Some proposals apply multiple imputation techniques (Akritas, Kuha, and Osgood 2002; Akritas, Antoniou, and Kuha 2006; Verbeke and Molenberghs 2009); however, they often require large sample sizes to be proved correct. Other proposals rely on specific model assumptions, for example, symmetry or bivariate normality (Ekbohm 1976; Samawi, Helu, and Vogel 2011; Xu and Harrar 2012); however, they exhibit a nonrobust behavior against deviations. The common approach recently adopted in literature results from combining a nonparametric approach separate test statistics for paired and unpaired observations using weighted test statistics (Konietschke et al. 2012; Martinez-Camblo et al. 2013; Samawi and Vogel 2014; Amro and Pauly 2017; Fong et al. 2018; Gaigall 2020), a multiplication combination test (Amro, Konietschke, and Pauly 2019), or combined p-values (Yu et al. 2012; Kuan and Huang 2013; Qi, Yan, and Tian 2019; Amro, Pauly, and Ramosaj 2021).

The recent scientific and technological progress in measuring biological processes has enabled the monitoring of the conditions of patients with a growing level of detail and complexity. Thus, in addition to the ongoing identification of univariate biomarkers, new complex data structures are being incorporated into the analysis, as in the case of population ages and mortality distributions (Bigot et al. 2017), distributions of functional connectivity patterns in the brain (Petersen and Müller 2016; Dubey and Müller 2020), post-intracerebral hemorrhage hematoma densities (Petersen, Liu, and Divani 2021), and graph-based representations of connectivity and functional brain activity (Takerkart et al. 2014).

Motivated by the steady availability of new complex statistical objects as a result of modern measuring processes, this article aims to provide a statistical test for matched pairs with missing data that does not require any parametric assumptions and uses all observations available. We propose new maximum mean discrepancy (MMD) estimators to achieve the goal (Gretton et al. 2012). The energy distance and MMD are two equivalent statistical metrics with the ability to detect distributional differences between random samples (Székely and Rizzo 2013; Shen and Vogelstein 2021). Moreover, MMD-based
statistics can be considered as a natural generalization of the ANOVA test to cases where the distributions are not necessary Gaussian (Rizzo and Székely 2010). MMD overcomes Gaussian assumptions by representing distances between distributions as distances between mean embeddings in a reproducing kernel Hilbert space (RKHS), and it has been successfully applied to independence testing (Székely, Rizzo, and Bakirov 2007), two-sample testing (Gretton et al. 2012), survival analysis (Fernandez and Gretton 2019), or clustering analysis (França, Rizzo, and Vogelstein 2021).

Besides conducting an extensive simulation study, the new testing methods are applied to the AEGIS diabetes dataset, resulting from an ongoing longitudinal population-based study (Gude et al. 2017). This dataset includes data obtained from continuous glucose monitoring (CGM) performed at the beginning of the study and 5 years later. Importantly, there is a substantial loss to follow up. A distributional representation of glucose concentration summarizes several days of monitoring, providing a personal signature of glucose homeostasis (Matabuena et al. 2021). The present approach allows us to address some interesting questions related to the possible changes in CGM profile with aging or the relation between obesity and diabetes. Furthermore, adapting a previous clustering method to match pairs with missing data allows us to identify specific patient phenotypes, with potential applications in patient stratification (França, Rizzo, and Vogelstein 2021).

The remainder of this article is outlined as follows: In Section 2 we provide a motivation for the new methods based on the distributional representation of CGM data. Section 3 defines the problem in general terms and introduces the statistical methodology based on the MMD metric, providing weighted test statistics for dealing with missing data under the missing completely at random (MCAR) assumption (Section 3.1) and under the missing at random (MAR) assumption (Section 3.2). Proofs presenting theoretical guarantees of the proposed methods are delivered in the Appendix. Section 3.3 discusses the choice of kernel functions and corresponding hyperparameters. Then, we present the results of an extensive simulation study in Section 3.4. In Section 3.5 a previous clustering method is adapted to missing data under the MAR assumption. Section 4 presents some applications of hypothesis testing and clustering analysis to the AEGIS study by exploiting the distributional representation of CGM data. We close with a discussion in Section 5.

2. Motivation: A Distributional Representation of Continuous Glucose Monitoring Data

Distributional data analysis is a novel methodology that has proved successful in managing biosensor data in different settings, for example, connectivity analysis of the brain network (Petersen and Müller 2016), physical activity analysis (Ghosal et al. 2021; Matabuena and Petersen 2021), and diabetes management (Matabuena et al. 2021, 2022).

Through a new distributional representation of CGM, known as glucodensity (Matabuena et al. 2021), it is possible to obtain a functional profile of patient glucose homeostasis. Glucodensity is a natural extension of the Time in Range (TIR) metrics, which measures the proportion of time a person spends with their blood glucose levels within a healthy target range of 70–180 mg/dL. TIR has been recently validated as a clinical target for treatment decision-making and as an outcome measurement for clinical trials (Beck et al. 2019). Furthermore, there is an international consensus on specific TIR targets for different diabetes populations: adults with type 1 or type 2 diabetes, older/high-risk individuals, and women during pregnancy (Battelino et al. 2019). Moreover, in cohort studies, TIR is significant associated with the development of chronic diabetic complications, for example, retinopathy, diabetic kidney disease (Lu, Ma, and Zhou 2018; Beck et al. 2019), and all-cause and cardiovascular mortality (Lu et al. 2020).

Although very intuitive, TIR metrics have two main disadvantages: first, they result from a population-based approach and range might fit poorly depending on the characteristics of the individual being examined; second, there is a loss of information caused by the discretization of the recorded data into intervals. Instead, glucodensity effectively measures the proportion of time each individual spends at every glucose concentration level across the whole range. Previous results of glucodensities demonstrated better predictive performance than TIR metrics (Matabuena et al. 2022).

Given a series of CGM data \( \{Y_i\}_{i=1}^m \), glucodensity can be modeled as a probability density function \( f(\cdot) \) which can be approached by kernel density estimation

\[
\hat{f}(y) = \frac{1}{m} \sum_{j=1}^{m} K \left( \frac{Y_j - y}{h} \right),
\]

where \( h > 0 \) denotes the smoothing parameter and \( K(\cdot) \) denotes a nonnegative real-valued integrable function (Figure 1). Herein, we use the Gaussian function \( K(x, y) = \exp^{-\|x-y\|^2/\sigma^2} \) for all \( x, y \in \mathbb{R} \). The kernel bandwidth \( \sigma \) can be estimated through the median heuristic \( \hat{\sigma} = \text{median}(||X_i - X_j||^2 : 1 \leq i < j \leq n) \).

Let \( D \) be the space of probability density functions \( f \) such that \( \int_{\mathbb{R}} u^2 f(u) du < \infty \). To measure the difference between two glucodensities, \( f \) and \( g \), a metric on \( D \) is required. We use the 2-Wasserstein distance:

\[
d_{W_2}(f, g) = \int_0^1 \left| Q_{\hat{f}}(t) - Q_{\hat{g}}(t) \right|^2 dt, \quad f, g \in D,
\]

where \( Q_{\hat{f}} \) and \( Q_{\hat{g}} \) denote the corresponding quantile functions. The 2-Wasserstein distance in (2) depends only on quantile functions. Therefore, we can approximate it by computing empirical quantile functions \( \hat{Q}_f = \hat{F}^{-1} \) and \( \hat{Q}_g = \hat{G}^{-1} \) from the corresponding empirical cumulative distribution functions \( \hat{F} \) and \( \hat{G} \).

Figure 2 shows the glucodensity representation for continuous glucose monitoring performed on two different individuals, both in the prediabetes and later diabetes stages. This scenario immediately poses the challenge of defining new statistical methods to compare two sets of glucodensity measurements to assess whether population statistics differ. This can help compare glucose homeostasis before and after a treatment, or after a certain period.
3. Hypotheses and Statistics

Let $\mathcal{D}$ be a metric space and $(X_1, X_2) \in \mathcal{D}^2$ a random pair representing two different measurements on a subject at two different time points. Let us consider the design of a general matched pair given by iid random variables

$$X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \end{pmatrix}, \quad j = 1, \ldots, n.$$ 

To continue describing the example of our motivation, $X_{1j}$ can represent the glucodensity at the beginning of a certain study for the $j$th patient and $X_{2j}$ can represent the glucodensity after 5 years for the same patient. Let assume that $(X_{1j})_{j=1}^n$ and $(X_{2j})_{j=1}^n$ are drawn from probability measures $P_1$ and $P_2$, respectively. We are interested in testing the equality of distributions as the null hypothesis

$$H_0 : \{P_1 = P_2\},$$

against the alternative $H_1 : \{P_1 \neq P_2\}$, that is, to check whether there are systematical differences between the outcomes at different time points.

Missing data, either at the beginning of the study or at any point later, pose a potential source of bias with regard to the reliability of the testing methodology. Excluding participants with partially available data results in ignoring valuable information, specifically when the number of complete samples is small, thereby leading to incorrect statistical estimation. Three major missing data assumptions can be made in any experimental study (Rubin 1976): missing completely at random (MCAR), when the probability that an observation is missing does not depend on the values of any observed or unobserved data; missing at random (MAR), when the probability that an observation is missing can depend on the observed data but not on the unobserved data; and missing not at random (MNAR) when the probability that an observation is missing depends on unobserved data. In this last case, valuable information is lost from data and there is no way to verify the correctness of the missing mechanism based on the observed data (Tsiatis 2007). In this work we focus on MCAR and MAR assumptions.

3.1. MCAR Assumption

When some of the elements of the matched pairs are missing completely at random, the available data can be sorted as follows:

$$X = \begin{pmatrix} \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} X_{1n_1+n_2} \\ X_{2n_1+n_2} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1+n_2+n_3} \\ X_{2n_1+n_2+n_3} \end{pmatrix} \end{pmatrix}$$

Complete dataset $X^{\text{obs}}$ with $n_1$ observations

Incomplete dataset $X^{\text{inc}}$ with $n_2$ observations

Incomplete dataset $X^{\text{inc}}$ with $n_3$ observations

$$X^{\text{obs}} = \begin{pmatrix} \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} X_{1n_1+n_2} \\ X_{2n_1+n_2} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1+n_2+n_3} \\ X_{2n_1+n_2+n_3} \end{pmatrix} \end{pmatrix}$$

$$X^{\text{inc}} = \begin{pmatrix} \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} X_{1n_1+n_2} \\ X_{2n_1+n_2} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1+n_2+n_3} \\ X_{2n_1+n_2+n_3} \end{pmatrix} \end{pmatrix}$$

$$X^{\text{inc}} = \begin{pmatrix} \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} X_{1n_1+n_2} \\ X_{2n_1+n_2} \end{pmatrix} & \cdots & \begin{pmatrix} X_{1n_1+n_2+n_3} \\ X_{2n_1+n_2+n_3} \end{pmatrix} \end{pmatrix}$$
where \( n = n_1 + n_2 + n_3 \). For ease of notation, we denote \( X_{\text{com}}^n = \{X_1\}_{j=1}^{n_1}, X_{\text{com}}^n = \{X_2\}_{j=1}^{n_2}, X_{\text{inc}}^n = \{X_1\}_{j=n_1+1}^{n}, \) and \( X_{\text{inc}}^n = \{X_2\}_{j=n_1+1}^{n} \). Additionally, a missingness status variable can be defined by \( \delta_{ij} \in \{0, 1\}, i = 1, 2, j = 1, \ldots, n \), such that \( \delta_{ij} = 0 \) if the element is missing and \( \delta_{ij} = 1 \) otherwise.

A natural way of testing the equality of distributions is by measuring the distance between them. We do not assume that the observations are coming from distribution functions belonging to a certain parametric or semiparametric family. Under the MCAR assumption, we can apply a test involving the complete dataset but at the cost of not using all the information from the available data, resulting in a suboptimal solution when the complete dataset is small and consequently incurring in a potential selection bias by truncation (Amro, Pauly, and Ramosaj 2021; Tenzer, Mandel, and Zuk 2022). Thus, we propose two test statistics: \( T_1 \) for the complete dataset \( X_{\text{com}}^n \) and \( X_{\text{com}}^n \), and \( T_2 \) for the incomplete datasets \( X_{\text{inc}}^n \) and \( X_{\text{inc}}^n \), which are then combined in one weighted test statistic:

\[
T(X) = \alpha T_1(X_{\text{com}}^n, X_{\text{com}}^n) + (1 - \alpha) T_2(X_{\text{inc}}^n, X_{\text{inc}}^n),
\]

for some weighting parameter \( \alpha \in [0, 1] \). In this article, we set \( \alpha = \frac{n_1}{n_1 + n_2 + n_3} \) and \( T_1 \) and \( T_2 \) are based on the MMD to measure the empirical distance between the marginal distributions through a kernel-based approach (Gretton et al. 2012). Let \( k: \mathcal{D} \times \mathcal{D} \to \mathbb{R}^+ \) be a symmetric positive-definite kernel function. This positive definiteness ensures the existence of an inner product space \( \mathcal{H} \) and feature mapping \( \phi: \mathcal{D} \to \mathcal{H} \) such that \( k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \), where \( \langle \cdot, \cdot \rangle_{\mathcal{H}} \) denotes the inner product of \( \mathcal{H} \) (Aronszajn 1950). We can interpret \( k(x, y) \) as a linear similarity measure in the feature space between \( x \) and \( y \). Moreover, an explicit representation of \( \phi \) is not necessary, and desired similarity properties can be derived directly from the kernel (Vapnik 2000; Schölkopf and Smola 2002). A reproducing kernel of \( \mathcal{H} \) is a kernel function that satisfies the following two properties:

1. \( k(x, \cdot) \in \mathcal{H} \), for all \( x \in \mathcal{D} \);
2. \( (f, k(x, \cdot))_{\mathcal{H}} = f(x), \forall f \in \mathcal{H} \) and \( \forall x \in \mathcal{D} \).

The function \( k(x, \cdot) \) is an element of \( \mathcal{H} \) that satisfies the reproducing property through the inner product of \( \mathcal{H} \), thus, playing the role of representing the evaluation of any function in \( \mathcal{H} \) on \( x \). \( \mathcal{H} \) is then defined as a reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan 2011), and it can be proved that for every positive-definite kernel \( k \) there exists a unique RKHS with \( k \) as its reproducing kernel (Aronszajn 1950). Based on the reproducing property it follows that \( k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \) and \( \phi(x) := k(x, \cdot) \) can be regarded as a canonical feature map.

Kernel mean embedding results from extending the feature mapping \( \phi \) to the space of probability distributions (Muandet et al. 2017). Let \( \mathcal{M}(\mathcal{D}) \) be the space of probability distributions over \( \mathcal{D} \). We can map each distribution into an RKHS by means of an expectation operation, through \( \mu_P: \mathcal{M}(\mathcal{D}) \to \mathcal{H}, P \mapsto \mathbb{E}_P[f(x)] = \mathbb{E}_P[k(x, \cdot)] = \int f(x, \cdot) dP \). Moreover, it follows that \( \mathbb{E}_P[f(x)] = \langle f, \mu_P \rangle_{\mathcal{H}} \forall f \in \mathcal{H} \), a sort of reproducing property of the expectation operation in the RKHS (Smola et al. 2007). The kernel mean embedding can be empirically estimated by \( \hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} k(X_i, \cdot) \), for any \( n \in \mathbb{N} \) and \( X_1, \ldots, X_n \) observations in \( \mathcal{D} \). Then, we can measure the distance between random samples as a distance in \( \mathcal{H} \) between mean embeddings, which can be estimated as follows (Gretton et al. 2012):

\[
\left( \frac{1}{n_1} \right) T_1(X_{\text{com}}^n, X_{\text{com}}^n) = \| \hat{\mu}_1 - \hat{\mu}_2 \|^2_{\mathcal{H}}
\]

\[
= \left( \frac{1}{n_1} \sum_{i=1}^{n_1} k(X_{1i1}, \cdot) - \frac{1}{n_1} \sum_{i=1}^{n_1} k(X_{2i1}, \cdot) \right) \cdot \frac{1}{n_1} \sum_{i=1}^{n_1} k(X_{1i1}, \cdot) \right)
\]

\[
= \frac{1}{n_1^2} \sum_{i=1}^{n_1} k(X_{1i1}, X_{1j}) + \frac{1}{n_1^2} \sum_{i=1}^{n_1} k(X_{2i1}, X_{2j})
\]

\[
- \frac{2}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(X_{1i1}, X_{1j}).
\]

For the class of characteristic kernels, the embeddings are injective, and hence \( \| P - Q \|^2_{\mathcal{H}} = 0 \), if and only if \( P = Q \), \( \forall P, Q \in \mathcal{M}(\mathcal{D}) \) (Sriperumbudur et al. 2011).

To calibrate the tests under the null hypothesis, it should be pointed out that neither \( T_1 \) nor \( T_2 \) follow a free asymptotic distribution. The empirical estimate of the MMD is a one-sample V-statistic. Hence, an asymptotic distribution is difficult to obtain due to the degeneracy of the V-statistic, which incorporates a correlation structure for the complete paired observations \( X_{\text{com}}^n \) (Gretton et al. 2012). Under the assumption of exchangeable data, a permutation approach can solve the problem. Instead, we propose a wild bootstrap procedure for the first \( n_1 \) observations to address this issue. The rationale behind this is that we are assuming that the complete paired observations have been obtained according to a chronological order, and they bear witness to the same causal direction of the underlying processes. Thus, they are not exchangeable (Tawn 1988; Durante and Mesiar 2010; Bacigál, Jäger, and Mesiar 2011). That is the case with the real application provided in Section 4, where continuous glucose monitoring is performed at the beginning of the study and after 5 years, and nonexchangeability is consistent with the evidence that glucose homeostasis evolves over a long time. In contrast, the remaining \( n_2 + n_3 \) observations can be appropriately handled by using permutation-based methods, which can achieve exact Type I error control. Thus, we obtain
resampling-based \( q^b = \mathbf{1}(T > q^b_{1-\alpha}) \), where \( q^b_{1-\alpha} \) is the conditional \((1 - \alpha)\)-quantile of the resampling statistic \( T^b \) given the observations. It proceeds according to Algorithm 1.

The validity of this approach is established by the next theorem.

**Theorem 1.** Suppose that no group of observations vanishes asymptotically, that is, \( n_i/(n_1 + n_2 + n_3) \rightarrow \kappa_i \in (0, 1) \) as \( \min(n_1, n_2, n_3) \rightarrow \infty \) for \( \ell = 1, 2, 3 \). Moreover, suppose that \( \mathbb{E}_{H_j}(q^b) \rightarrow \alpha \) as \( n \rightarrow \infty \).

1. \( q^b \) is asymptotically exact under the null hypothesis \( H_0 \), that is, \( \mathbb{E}_{H_0}(q^b) \rightarrow \alpha \) as \( n \rightarrow \infty \).
2. \( q^b \) is consistent for any alternatives \( H_1 \), that is, \( \mathbb{E}_{H_1}(q^b) \rightarrow 1 \) as \( n \rightarrow \infty \).

**3.2. MAR Assumption**

We assume a specific MAR mechanism, which is commonly found in longitudinal studies, where several initial participants were later lost to follow-up; therefore, the first time point is measured, and the second one is missing. We present a real application in Section 4, which is consistent with this assumption. Furthermore, suppose that the probability of being observed (or missing) on the second time point can be assessed based on

**Algorithm 1** Calibration method for the test \( T(X) \)

**Input:** A dataset \( X \)

**Output:** A two-sided \( p \)-value for \( T(X) \) under the null hypothesis

1. for \( b = 1 \) to \( B \) do

2. For the first \( n_1 \) complete paired observations, take random weights \( w_i^b, i = 1, \ldots, n_1 \), with

\[
w_i^b = e^{-1/l_{n_1}} w_{i-1} + \sqrt{\frac{1 - e^{-2/l_{n_1}}} 1},
\]

where \( w_0, e_1, \ldots, e_n \) denote independent standard normal variables, and \( l_{n_1} \) denotes a bootstrap parameter used to mimic the dependence structure, such that \( l_{n_1} \rightarrow o(n_1) \) while \( \lim_{n_1 \rightarrow \infty} l_{n_1} = \infty \). Then,

\[
\mathcal{T}^b(X_{1}^{\text{com}}, X_{2}^{\text{com}}) = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} w_i w_j \left[ k(X_{1i}, X_{1j}) + k(X_{2i}, X_{2j}) - 2k(X_{1i}, X_{2j}) \right].
\]

3. The remaining \( n_2 + n_3 \) observations belonging to \( X_1^{\text{inc}} \) and \( X_2^{\text{inc}} \) are randomly permuted, that is, each observation is randomly assigned to new \( X_1^{\text{inc}(\pi)} \) or \( X_2^{\text{inc}(\pi)} \) sets, resulting in \( \mathcal{T}^b(X_1^{\text{inc}(\pi)}, X_2^{\text{inc}(\pi)}) \).

4. Then, calculate

\[
\mathcal{T}^b = \alpha \mathcal{T}^b(X_1^{\text{com}}, X_2^{\text{com}}) + (1 - \alpha) \mathcal{T}^b(X_1^{\text{inc}(\pi)}, X_2^{\text{inc}(\pi)})
\]

5. end for

6. Finally, compute the two-sided \( p \)-value \( = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(\mathcal{T}^b \geq T(X)) \) (the simplification is due to \( T(X) \geq 0 \)).
The same test can be applied to the case where the second time point is measured and the first one is missing, whenever the probability of being observed at the first time point can be assessed based on the information available at the second time point $P(\delta_{1j} = 1 \mid X_{1j}, X_{2j}, Z_{2j}) = P(\delta_{1j} = 1 \mid X_{2j}, Z_{2j})$.

**Theorem 2.** Suppose that $E[K^2] < \infty$, and Assumptions A.1–A.3 hold (see Appendix); then

1. $\varphi^b$ is asymptotically exact under the null hypothesis $H_0$, that is, $\mathbb{E}_{H_0}(\varphi^b) \to \alpha$ as $n \to \infty$.
2. $\varphi^b$ is consistent for any alternatives $H_1$, that is, $\mathbb{E}_{H_1}(\varphi^b) \to 1$ as $n \to \infty$.

### 3.3. Kernel Choice and Kernel Hyperparameters

We use the Gaussian kernel $k(x,y) = e^{-d^2_{W_2}(x,y)/\sigma}$ for $x, y \in \mathcal{D}$, where $\sigma > 0$. The Gaussian kernel is a characteristic kernel, and thus we can detect any difference in distribution asymptotically. The kernel bandwidth $\sigma$ is typically estimated through the median heuristic $\sigma^2 = \text{median}\{d^2_{W_2}(X_i, X_j) : 1 \leq i < j \leq n\}$ (Garreau, Jitkrittum, and Kanagawa 2017).

### 3.4. Simulation Study

We investigated the finite sample behaviors of the aforementioned methods through extensive simulations. Thus, 2000 simulations were performed for the MCAR and MAR scenarios. The methods were examined with respect to their Type-I error rate control at level 5%. A total of 2000 bootstrap runs and permutation replicas were held. The wild bootstrap parameter $l_{n1}$ was selected according to $l_{n1} = \sqrt{n_1}$.

The observations were generated by mimicking the distributional representations obtained from the CGM data. As the $L$-Wasserstein distance depends on quantile functions, observations were sampled from the following location-scale model on quantile functions (Petersen, Liu, and Divani 2021): let $Z \in \mathbb{R}^p$ be a random vector of predictor variables and $Q_0$ be a fixed quantile function; herein, we considered the age as the only predictor variable and fixed $Q_0(t) = 70 + 240t$ in the range of glucose values expected from Type-2 diabetes; let $\eta(z) = a_0 + a_1z_1$ and $\tau(z) = b_0 + b_1z_1$ be the location and scale components of the model, respectively, where $a = (a_0, a_1)$ and $b = (b_0, b_1)$ are the corresponding coefficients. We assume that $\tau(Z) > 0$ almost surely. Let $V_1$ and $V_2$ be two random variables that satisfy $\mathbb{E}(V_1|Z) = 0, \mathbb{E}(V_2|Z) = 1$, and $V_2 > 0$ almost surely; the model is defined as

$$Q(t) = V_1 + V_2 \eta(Z) + V_2 \tau(Z) Q_0(t) .$$

#### 3.4.1. MCAR Scenario

We fixed $n_1 = n_2 = n_3 = 150$. To introduce correlation structure in the quantile functions of $X^{1\text{inc}}_1$ and $X^{2\text{inc}}_2$, we sampled variables $V_1^*$ and $V_2^*$ from bivariate uniform distributions with correlation given by $\rho \in \{0.00, 0.20, 0.40, 0.60, 0.80\}$. The location-scale model is defined as $V_1 = -20 + 40V_1^*$ and $V_2 = 0.8 + 0.4V_2^*$, and the fixed parameters were $a_0 = b_0 = 0$, $a_1 = 0.3$ and $b_1 = 0.005$. The observations of $X^{1\text{inc}}_1$ and $X^{2\text{inc}}_2$ were iid generated. Then, we applied the same location-scale model.

A total of 2000 simulations were performed assuming that the age was distributed as $Z_1, Z_2 \sim \mathcal{U}_{[30,50]}$ at the beginning and end of the study, that is, for all the variables in $X$. Next, another round of 2000 simulations was performed assuming that the age was distributed as $Z_1 \sim \mathcal{U}_{[30,50]}$ at the beginning of the study, that is, for all the variables in $X^{1\text{inc}}_1$ and $X^{2\text{inc}}_2$.

#### 3.4.2. MAR Scenario

We fixed $n = 300$. The missing mechanism is given by $P(\delta_{ij} = 1 | Y_1, Y_2) = (1 + e^{-1+y_i+y_j})^{-1}, j = 1, \ldots, n$, where $Y_1, Y_2 \sim \mathcal{N}(0, 1)$ are two independent random variables. We introduced correlation structure in the quantile functions for $X^{1\text{inc}}_1$ and $X^{2\text{inc}}_2$ as we did in the MCAR scenario. We used the same location-scale model, and the same methodology to sample the age.

#### 3.4.3. Results

Table 1 lists the simulation study results. Calibration under the null hypothesis provides acceptable results. However, biases under the two missingness assumptions can be noticed. As long as the difference between the sample data and null hypothesis increases, the test rejects the null hypothesis.

#### 3.5. Paired Missing Data Clustering

Clustering is a central task of exploratory data analysis, and it has proven to be a powerful tool for discovering patterns and structure in unlabeled datasets. A recent extension to the RKHS paradigm enables new data structures to be subjected to distance measurement; thus, clustering can be performed in the associated Hilbert space (França, Rizzo, and Vogelstein 2021). Herein, we adopt this approach for paired two-sample problems where the first time point is measured and the second one can be missing, aimed at identifying valid population phenotypes in those clinical studies where a number of initial participants were later lost to follow-up. These phenotypes will enable us to gain insight into the distribution of participants in relation to the progression of their condition over time. A real data example is considered in Section 4.

### Table 1. The proportion of simulations rejecting the null hypothesis under the MCAR and MAR assumptions is shown at a confidence level of 95 percent.

| $\rho$ | $Z_1$ | $Z_2$ | MCAR | MAR |
|---|---|---|---|---|
| 0.00 | $\mathcal{U}_{[30,50]}$ | $\mathcal{U}_{[30,50]}$ | 0.03 | 0.03 |
| 0.20 | $\mathcal{U}_{[30,50]}$ | $\mathcal{U}_{[30,50]}$ | 0.04 | 0.03 |
| 0.40 | $\mathcal{U}_{[30,50]}$ | $\mathcal{U}_{[30,50]}$ | 0.05 | 0.03 |
| 0.60 | $\mathcal{U}_{[30,50]}$ | $\mathcal{U}_{[30,50]}$ | 0.03 | 0.04 |
| 0.80 | $\mathcal{U}_{[30,50]}$ | $\mathcal{U}_{[30,50]}$ | 0.04 | 0.04 |
| 0.00 | $\mathcal{U}_{[50,70]}$ | $\mathcal{U}_{[50,70]}$ | 0.98 | 0.88 |
| 0.20 | $\mathcal{U}_{[50,70]}$ | $\mathcal{U}_{[50,70]}$ | 0.99 | 0.90 |
| 0.40 | $\mathcal{U}_{[50,70]}$ | $\mathcal{U}_{[50,70]}$ | 0.99 | 0.91 |
| 0.60 | $\mathcal{U}_{[50,70]}$ | $\mathcal{U}_{[50,70]}$ | 0.99 | 0.93 |
| 0.80 | $\mathcal{U}_{[50,70]}$ | $\mathcal{U}_{[50,70]}$ | 0.99 | 0.96 |

**NOTE:** Under the null hypothesis, a nominal value of 0.05 is expected; otherwise, the nominal value will increase to 1 as $n$ grows to infinity.
Let \( X = \{(X_{ij}, Z_{ij})\}_{i=1}^{n} \) be the dataset of iid random variables obtained under a MAR mechanism where \( \pi_{ij} = \pi(X_{ij}, Z_{ij}) = P(Z_{ij} = 1|X_{ij}) \) denotes the conditional probability that the observation \( X_{ij} \) will be missing given the values of \( X_{ij} \) and \( Z_{ij} \). We associate weight \( \omega_j \) with the \( j \)th observation using an IPW estimator by applying (4). Let \( X_j = (X_{1j}, X_{2j})^\top \) and \( X_h = (X_{1h}, X_{2h})^\top \in \mathbb{X}^{\text{com}} \) be two completely observed paired samples. We define the bivariate kernel \( k(X_j, X_h) = e^{-\sum d^2_{ij}(X_j, X_h)} \sigma_j \), where \( \sigma^2 = \text{median}([d^2_{ij}(X_j, X_h)] : 1 \leq j < h \leq n) \).

Consider a disjoint partition \( \mathbb{X}^{\text{com}} = \bigcup_{l=1}^{m} C_l \), with \( C_l \cap C_l = \emptyset \), for all \( l \neq i \). Following França, Rizzo, and Vogelstein (2021), we aim to build a new partition \( \hat{C}_1, \ldots, \hat{C}_k \) by maximizing an objective function given by

\[
(\hat{C}_1, \ldots, \hat{C}_k) = \arg \max_{(C_1, \ldots, C_k)} \frac{1}{k} \sum_{l=1}^{k} \sum_{X_j \in C_l} \omega_j \hat{\omega}_l k(X_j, X_h),
\]

where \( \omega_l = \sum_{X_j \in C_l} \omega_j \). We can iterative solve this optimization problem by measuring the impact of moving each observation to another cluster. Let \( S_l = \sum_{X_j \in C_l} \hat{\omega}_j \hat{\omega}_l k(X_j, X_h) \) denote the internal similarity of cluster \( C_l \), and \( S_l(X_j) = \sum_{X_j \in C_l} \hat{\omega}_j \hat{\omega}_l k(X_j, X_h) \) the internal similarity with respect to the observation \( X_j \). By moving the observation \( X_j \) from cluster \( C_l \) to \( C_l \) we modify the result of the objective function by modifying the similarities

\[
\Delta S^{(l)\rightarrow l}(X_j) = \frac{S_l(X_j)}{\nu_l + \hat{\omega}_l} + \frac{S_l(X_j)}{\nu_l - \hat{\omega}_l} - \frac{S_l(X_j)}{\nu_l - S_l(X_j)},
\]

where \( S_l^+ = S_l + 2S_l(X_j) + \hat{\omega}_l \hat{\omega}_l k(X_j, X_h) \) is the internal similarity of the new cluster \( C_l \) after the addition of the observation \( X_j \), and \( S_l^- = S_l - 2S_l(X_j) + \hat{\omega}_l \hat{\omega}_l k(X_j, X_h) \) is the internal similarity of the new cluster \( C_l \) after removing the observation \( X_j \). Ultimately, we compute \( \rho^2 = \max_{l=1,\ldots,k, l \neq i} \Delta Q^{(l)\rightarrow l}(X_j) \), and if \( \Delta S^{(l)\rightarrow l}(X_j) > 0 \) we move \( X_j \) to cluster \( C_l \), otherwise we keep it in \( C_l \).

### 4. Illustrative Data Analysis

As a practical application, we consider an ongoing longitudinal, population-based study by Gude et al. (2017) to analyze the evolution of clinical biomarkers relating to circulating glucose in an initial random sample of 1516 patients over 10 years. In addition, a CGM is performed every five years on a randomized subset of patients. Specifically, at the beginning of the study, 581 participants were randomly selected to wear a CGM device for 3–7 days. Out of 581 participants, 68 were diagnosed with diabetes before the study and 22 during the first 5 years. Table 2 lists the baseline characteristics of these 581 patients categorized by sex. After a 5-year follow-up, only 161 participants agreed to second glucose monitoring.

The AEGIS study raises interesting questions that can be addressed with the present approach.

**Changes in CGM profile with aging.** A few recent studies explore the important role of aging in glucose dysregulation and difficulties inherent in maintaining glucose homeostasis as close to normal as possible (Chia, Egan, and Ferrucci 2018). The proposed T-test allows us to examine if statistical differences exist at a distributional level after 5 years. We estimate the missing data mechanism using logistic regression, considering the age and glycemic status (normoglycemic, prediabetes or type-2 diabetes) at the beginning of the study and sex of each participant as predictors. We applied the T-test considering glucodensities at both time points to check the null hypothesis of the equality of distributions. We obtained a p-value of 0.048, identifying significant differences at both time points.

**Obesity in diabetes.** Obesity is a critical risk factor for type-2 diabetes (Leong and Wilding 1999). To further characterize this risk subpopulation, we analyzed the normoglycemic subjects with overweight in the AEGIS dataset by examining if there exist statistical differences after 5 years at a distributional level. We applied the T-test to check the null hypothesis in the following two subgroups of the normoglycemic population: (i) individuals with a body mass index less than 22Kg/m² (low body mass index); (ii) individuals with a body mass index higher than 22Kg/m² (overweight and obesity). In the first case, we obtained a p-value of 0.36, providing no evidence against the null hypothesis. In the second case, we obtained a p-value of 0.056, which can be interpreted as borderline. Figure 3 shows the difference between the quantile curves in these two subgroups.

**Patient stratification.** Clustering analysis can identify distinctive and meaningful patient phenotypes and, consequently, guide patient stratification for delivering personalized care (Kosorok and Laber 2019). We applied clustering analysis to those individuals for whom CGM was performed at both time points. Figure 4 shows the resulting two clusters. The individuals in cluster 1 do not present significant changes between both time points, whereas some significant differences are noted in cluster 2. Table 3 lists the baseline clinical characteristics of each cluster and Figure 5 shows a summary provided by boxplots. Both groups of individuals exhibit significant differences in insulin resistance and glycemic variability metrics. In cluster 2, the average glycemic control measured in terms of glycated hemoglobin and fasting plasma glucose is consistent with prediabetes (5.7% ≤ A1c ≤ 6.4%) or 100 mg/dl ≤ FPG ≤ 125 mg/dl according to American Diabetes Association guidelines). In contrast, cluster 1 comprises normoglycemic individuals. In the end, clustering results effectively correlate with a significant change in the glycemic status.

### Table 2. Baseline characteristics of AEGIS study participants with CGM monitoring by sex.

| Age, years | Men (n = 220) | Women (n = 361) |
|------------|--------------|-----------------|
| 47.8 ± 14.8 | 48.2 ± 14.5 |                 |
| A1c, % | 5.6 ± 0.9 | 5.5 ± 0.7 |                 |
| FPG, mg/dl | 97 ± 23 | 91 ± 21 |                 |
| HOMA-IR, mg/dl | 3.97 ± 5.56 | 2.74 ± 2.47 |                 |
| BMI, kg/m² | 28.9 ± 4.7 | 27.7 ± 5.3 |                 |
| CONGA, mg/dl | 0.88 ± 0.40 | 0.86 ± 0.36 |                 |
| MAGE, mg/dl | 33.6 ± 22.3 | 31.2 ± 14.6 |                 |
| MODD | 0.84 ± 0.58 | 0.77 ± 0.33 |                 |

**NOTE:** Mean and standard deviation are given. A1c: glycated hemoglobin; FPG: fasting plasma glucose; HOMA-IR: homeostasis model assessment-insulin resistance; BMI: body mass index; CONGA: glycemic variability in terms of continuous overall net glycemic action; MAGE: mean amplitude of glycemic excursions; MODD: mean of daily differences.
Figure 3. Difference between the quantile curves (before and after) in normoglycemic individuals according to body mass status. The dispersion is more significant in the overweight and obesity subgroup and consistent with increasing glycemic risk.

Figure 4. AEGIS participants are grouped into two clusters. Both quantile curves at the beginning of the study (in blue color) and 5 years later (in orange color) are shown for each participant. The participants in cluster 2 exhibit significant changes over a period than those in cluster 1. Interestingly, cluster 1 comprises normoglycemic individuals, and the average glycemic characteristics in cluster 2 are consistent with prediabetes.

Table 3. Clinical baseline characteristics for the individuals belonging to each cluster.

| Characteristic | cluster 1       | cluster 2       |
|---------------|-----------------|-----------------|
| Age (years decimal) | 43.66 ± 12.80  | 53.11 ± 12.34  |
| A1c, %        | 5.27 ± 0.25     | 6.20 ± 1.07     |
| FPG, mg/dL    | 84.83 ± 9.93    | 108.39 ± 32.61  |
| HOMA-IR, mg/dL/μ IU/ml | 2.28 ± 1.16 | 4.97 ± 8.50    |
| BMI, kg/m²    | 27.09 ± 4.87    | 29.29 ± 4.83    |
| Waist, cm     | 87.29 ± 13.60   | 95.18 ± 14.43   |
| CONGA, mg/dL  | 0.75 ± 0.20     | 1.21 ± 0.52     |
| MAGE, mg/dL   | 26.16 ± 7.16    | 45.80 ± 24.58   |
| MODD          | 0.66 ± 0.18     | 1.05 ± 0.48     |

NOTE: Mean and standard deviation are shown.

5. Discussion

The analysis of paired data with missing values is becoming critical in longitudinal studies, particularly when comparing the conditions of participants across different time points. Previous methods are not applicable when data adopt non-vector representations, which are better suited to capture functional, structural, or other complex forms of information increasingly common in current medicine. To overcome this limitation, we have provided novel methods for hypothesis testing in the presence of complex paired missing data under the MCAR and MAR assumptions. They are not based on any parametric or semiparametric assumption. The methods are based on the MMD, a metric between mean embeddings in an RKHS that can be applied to both Euclidean and non-Euclidean data, with different structured, functional, and distributional representations, based on an appropriate design of the reproducing kernel. Specifically, the space of probability density functions has been used throughout the study to test the feasibility of this approach.

The asymptotic validity of the new methods was proven in the Appendix. In an extensive simulation study, the Type-I error rate control of the tests has been examined under the MCAR and MAR assumptions, performing well with different correlation coefficients. The sample size affects the behavior of the tests, as inference in a functional space customarily demands more data than in a vector space. Hence, a deterioration in performance is expected for very small sample sizes.

The application of these methods to a real, longitudinal, population-based diabetes study has highlighted their capabilities and advantages to explore new clinical findings by exploiting monitoring information across the continuous range of glucose values. The robustness of the results should be emphasized, even in a scenario with an important proportion of missing data. Furthermore, a complementary clustering analysis has revealed the effectiveness of this approach in providing risk identification, with the potential to enable a personalized strategy. Therefore, this study contributes to the recent debate on the identification of novel subgroups of diabetes for a proper stratification of treatment and progression to complications (Ahlqvist, Tuomi, and Groop 2019; Dennis 2020; Herder and Roden 2022), thereby emphasizing the crucial role of continuous glucose monitoring.

Finally, we performed stepwise logistic regression with forward selection to identify which baseline characteristics independently predicted the corresponding group, obtaining age, FPG and CONGA as predictive features. We checked the null hypothesis that each coefficient is equal to zero. Table 4 lists the results of this analysis, identifying FPG and CONGA as the subset of characteristics that best predicted the outcome.
Figure 5. Boxplot distribution of clinical baseline characteristics for each cluster.

Table 4. Coefficients obtained from logistic regression together with their standard deviation and statistical significance, after testing the null hypothesis that each coefficient is equal to zero. Additionally, some quality measures from different model selection criteria for the fitted model are shown.

| Coefficients | (Intercept) | age | FPG | CONGA |
|--------------|-------------|-----|-----|-------|
| Value        | -10.01 ± 1.82*** | 0.04 ± 0.02 | 0.05 ± 0.02** | 3.60 ± 0.91*** |

Quality measures
- AIC: 140.37
- BIC: 152.69
- Log Likelihood: -66.18
- Deviance: 132.37

NOTE: ***p < 0.001; **p < 0.01.

Appendix A. Proof of Theorem 1

Proof. Since $T$ in (3) is a convex combination of two independent tests $T_1$ and $T_2$, we shall consider them separately.

1. Under the null hypothesis $H_0 : \{P_1 = P_2\}$, both $T_1$ and $T_2$ are based on the following V-statistics

$$V_{n_1} = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} h_1(x_i, y_i; x_j, y_j)$$

$$V_{n_2, n_3} = \frac{1}{n_2^2 n_3} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{n_3} h_2(x_i, x_j; y_i, y_l)$$

with kernels $h_1$ and $h_2$

$$h_1(x_i, y_i; x_j, y_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$$

(8)

$$h_2(x_i, x_j; y_l, y_m) = k(x_i, x_j) + k(y_l, y_m) - k(x_i, y_l) - k(x_j, y_m)$$

(9)

$E[h_1] = 0$ and $E[h_2] = 0$ under the null hypothesis. $E[h_1^2] < \infty$ and $E[h_2^2] < \infty$ by construction. $V_{n_1}$ and $V_{n_2, n_3}$ are degenerate statistics, since $E[h_1(X_{1i}, X_{2j}; \cdot, \cdot)] = 0$, for all $X_{1i}$ and $X_{2j}$, and $E[h_2(X_{1i}; X_{2m}; \cdot, \cdot)] = 0$, for all $X_{1i}$ and $X_{2m}$. The new tests $T_1$ and $T_2$ can be defined as

$$T_1 = n_1 V_{n_1}$$

$$T_2 = \left( \frac{n_2 n_3}{n_2 + n_3} \right) V_{n_2, n_3}$$

(10)

Test $T_1$ is a one-sample test and, accordingly, we can apply the asymptotic results for degenerate V-statistics (Serfling 1981, chap. 6), and hence $T_1$ converges in distribution to a weighted sum of independent $\chi^2(1)$ random variables

$$T_1 \xrightarrow{d} \sum_{i=1}^{\infty} \chi^2_i Z_i^2,$$

(11)

where the constants $\lambda_i$'s are the eigenvalues associated with the kernel $h_1$. Test $T_2$ is a two-sample test and, accordingly, its weak convergence can be proved by applying (Neuhaus 1977, Theorem 1.1).
2. Let us consider the alternative scenario $H_1 : \{ P_1 \neq P_2 \}$. Then, we have $\mathbb{E}[h_1] > 0$ and $\mathbb{E}[h_2] > 0$, so

$$\lim_{n_1 \to \infty} V_{n_1} = c > 0, \quad (12)$$

$$\lim_{n_2, n_3 \to \infty} V_{n_2, n_3} = c' > 0, \quad (13)$$

with probability one. By the theory of degenerate V-statistics, there exist constants $c_0$ and $c'_0$ satisfying $\lim_{n_1 \to \infty} P(T_1 > c_0) = \alpha$ and $\lim_{n_2, n_3 \to \infty} P(T_2 > c'_0) = \alpha$ under the null hypothesis. Thus, under the alternative hypothesis

$$\lim_{n_1 \to \infty} P(T_1 > c_0) = 1, \quad (14)$$

$$\lim_{n_2, n_3 \to \infty} P(T_2 > c'_0) = 1. \quad (15)$$

Thus, the statistics $T_1$ and $T_2$ diverge asymptotically, and $T$ rejects the null hypothesis.

Finally, since both $T_1$ and $T_2$ are degenerate V-statistics the consistency of the wild bootstrap method immediately follows from Leucht and Neumann (2013, Theorem 3.1), and the permutation calibration strategy under random permutation of the pooled sample has been proved consistent by Székely and Rizzo (2004, Theorem 3).

**Appendix B. Proof of Theorem 2**

**B.1. Preliminaries**

**Definition 1.** (Gaussian process) A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A Gaussian process $\{X(t) : t \in [0,1]\}$ is completely specified by its mean function $\mathbb{E}[X(t)]$ and covariance function $\text{cov}(X(t), X(t')) = \mathbb{E}[(X(t) - \mathbb{E}[X(t)])(X(t') - \mathbb{E}[X(t')])]$. A Gaussian process is centered if $\mathbb{E}[X(t)] = 0$ for all $t \in [0,1]$.

**Proposition 1.** A centered Gaussian process $\{X(t) : t \in [0,1]\}$, with $X(t) \in L^2[0,1]$, satisfies the following property:

$$\|X\|^2 = \int_0^1 X^2(t)dt = \sum_{n=0}^{\infty} \lambda_n Z_n^2,$$

where $Z_n \sim \mathcal{N}(0,1)$ and the values of sequence $\{\lambda_n\}_{n=0}^{\infty}$ are strictly positive.

**Proof.** Let $\{X(t) : t \in [0,1]\}$ be a centered Gaussian process with covariance function $k(t, t') = \mathbb{E}[X(t)X(t')]$. Let $\{\phi_n\}_{n=0}^{\infty}$ be an orthonormal basis of $L^2[0,1]$, then by Parseval’s theorem any $X(t) \in L^2[0,1]$ can be represented as

$$X = \sum_{n=0}^{\infty} \langle X, \phi_n \rangle \phi_n,$$

where $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$.

Hence,

$$\|X\|^2 = \sum_{n=1}^{\infty} \langle X, \phi_n \rangle^2.$$

Let $K : L^2[0,1] \to L^2[0,1]$ denote an operator such that $Kf(t) = \int_0^1 k(s, t)f(s)ds$ and let $\{\gamma_n\}_{n=0}^{\infty}$ be the orthonormal basis induced by the spectral problem related with $K$, satisfying $\gamma_n = \lambda_n \gamma_n$, $\forall n \in \mathbb{N}$. We compute $X_n$ as the orthogonal projection of $X$ on the subspace spanned by functions $\{\gamma_n\}_{n=0}^{\infty}$

$$X_n = \langle X, \gamma_n \rangle = \int_0^1 X(t)\gamma_n(t)dt,$$

which satisfies

$$\mathbb{E}[X_n] = \int_0^1 \mathbb{E}[X(t)]\gamma_n(t)dt = 0$$

and

$$\text{var}(X_n) = \mathbb{E}[X_n^2] = \mathbb{E} \left[ \int_0^1 \int_0^1 X(s)X(t)\gamma_n(s)\gamma_n(t)dsdt \right] = \int_0^1 \int_0^1 (k(s,t)\gamma_n(s)\gamma_n(t)dt = \lambda_n \|\gamma_n\|^2 = \lambda_n.$$ (16)

Thus, $X_n \sim \mathcal{N}(0, \lambda_n)$. In addition, $\text{cov}(X_n, X_m) = 0$ for all $n \neq m$, and therefore the random variables of the sequence $\{X_n\}_{n=1}^{\infty}$ are independent Gaussian. Since $X_n \sim \mathcal{N}(0, \lambda_n)$ and finally

$$\|X\|^2 = \int_0^1 X^2(t)dt = \sum_{n=1}^{\infty} \|X_n\|^2 = \sum_{n=1}^{\infty} \lambda_n \|Z_n\|^2,$$

where $\{Z_n\}_{n=1}^{\infty}$ are independent Gaussian, that is, $Z_n \sim \mathcal{N}(0,1)$.

To adapt the maximum mean discrepancy to the MAR mechanism, we will introduce specific assumptions for the missing data mechanism based on the theory of empirical processes (Van Der Vaart and Wellner 1996). Let $\pi(Z, \theta) = P(\delta = 1|Z)$ be a missing data mechanism, with $Z \in \mathbb{Z}$ representing the covariates, and $\theta \in \Theta \subset \mathbb{R}^p$, being $\Theta$ the space of parameters. The following assumptions are made:

These assumptions are referenced as A.1, A.2 and A.3 in different sections in the article.

1. $\pi(Z, \theta) > c > 0$, for all $Z \in \mathbb{Z}$ and $\theta \in \Theta$.
2. $\pi(Z, \theta)$ is continuously differentiable with respect to $\theta \in \Theta$.
3. Let $Z_1, Z_2, \ldots, Z_n$ be independent observations; the estimator $\hat{\theta}$ admits a Bahadur representation, $\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(\delta_i, Z_i) + o_P(1)$, where $h(\cdot)$ satisfies $\mathbb{E}[h] = 0$, and $\mathbb{E}[h^2] < \infty$.

Assumption A.1 restricts the minimum value provided by the missing data mechanism to guaranteeing that it is far from zero. Such an assumption is necessary to ensure that the limit of the IPW estimator will be Gaussian (see Ma and Wang (2020) for further details). Assumption A.2 is a standard regularity condition for consistency and allows us to derive the asymptotic distribution of the parameter $\theta$ as well as to construct new statistics involving the missing data mechanism. Finally, Assumption A.3 is a simplified condition to guaranteeing that the asymptotic expansion of $\hat{\theta}$ follows a linear structure, and, as a consequence, the central limit theorem is satisfied.

Let $X = \{(X_{ij}, Z_{ij}, \delta_{ij}, Z_{ij})\}_{i,j=1}^{n}$ denote a random sample of pairs consistent with the MAR scenario described in Section 3.2. For simplicity, let $\pi(\cdot) = \pi(X_{ij}, Z_{ij}, \theta) = P(\delta_{ij} = 1|X_{ij}, Z_{ij})$ denote a missing data mechanism satisfying Assumptions A.1–A.3, for example, through logistic regression.

We need the following lemma in the proof of our results.

**Lemma 1.** Suppose a missing data mechanism $\pi(\cdot)$ satisfying Assumptions A.1 and A.2; then the test provided in Section 3.2 (5), as long as $\mathbb{E}[|k|^2] < \infty$, consistently detects deviations from the null hypothesis.
Proof. We introduce the empirical measure associated to variables $X_{ij}$, $i = 1, 2$, as
\[ P_{i,n}^{\pi} = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_{ij}} I(X_{ij}), \]
where $I(X_{ij})$ denotes the Dirac measure at the observation $X_{ij}$. The corresponding empirical process is $\sqrt{n}(P_{i,n}^{\pi} - P_i)$. Let $\mathcal{F}$ denote the Donsker class of indicator functions (Van Der Vaart and Wellner 1996).

Then $\mathcal{F} = \{ \frac{1}{\sqrt{n}} I(1), 1 \} \subseteq \mathcal{F}$ is also a Donsker class, provided that $\pi_{2j}$ is a fixed function and using Assumptions A.1 and A.2.

Since $\mathcal{F}$ is a Donsker class, by applying the central limit theorem, we obtain the following asymptotic convergence in distribution (Van Der Vaart and Wellner 1996):
\[ \sqrt{n}(P_{i,n}^{\pi} - P_i) \xrightarrow{d} \Xi_i, \]
where $\Xi_i$ is a Brownian process with a specific covariance structure shaped by function $\psi()$. More specifically, we have $\text{cov}(\Xi_i(t), \Xi_j(t)) = P_i(\min(s, t)) \left( 1 + \mathbb{E}\frac{1 - \pi_{2j}}{\pi_{2j}} \right) - P_i(s)P_i(t)$.

According to Van Der Vaart and Wellner (1996) and (Kosorok 2008, chap. 10) we can obtain the same asymptotic convergence for the corresponding bootstrap empirical processes and wild bootstrap empirical processes:
\[ \sqrt{n}(P_{i,n}^{\pi} - P_i) \xrightarrow{d} \Xi_i. \]

Since $\hat{\mu}_i = \sum_{j=1}^{n} \hat{\alpha}_j k(\cdot, X_{ij}) \in \mathcal{H}$, we can write $T(X) = \psi(\hat{\mu}_1, \hat{\mu}_2)$ where $\psi()$ is an appropriate mapping satisfying Hadamard differentiability (Van Der Vaart and Wellner 1996). According to Proposition 1, under the null hypothesis $H_0: (P_1 = P_2)$ we have
\[ nT(X) \xrightarrow{d} \sum_{n=1}^{\infty} \lambda_n Z_n, \quad (16) \]
where $(Z_n)_{n=1}^{\infty}$ are independent Gaussian, that is, $Z_n \sim \mathcal{N}(0, 1)$. Moreover, the centered bootstrap process satisfies the same property.

Let us note that by definition, $nT(X) \geq 0$, where the equality holds if and only if $P_1 = P_2$. Thus, under the alternative hypothesis $T(X)$ diverges asymptotically due to statistics consistency, following a similar reasoning to that of Theorem 1. In addition, since the bootstrap test calibration strategy under the null hypothesis mimics the limit distribution, the bootstrap consistency is guaranteed (Janssen and Pauls 2003).

B.2. Proof of Theorem 2

Proof. Consider the following simple decomposition:
\[ \begin{align*}
\| \hat{\mu}^\pi_1 - \hat{\mu}^\pi_2 \|_{\mathcal{H}}^2 &- \| \mu_1 - \mu_2 \|_{\mathcal{H}}^2 = \| \hat{\mu}^\pi_1 - \hat{\mu}^\pi_2 \|_{\mathcal{H}}^2 - \| \mu_1 - \mu_2 \|_{\mathcal{H}}^2 \\
\| \hat{\mu}^\pi_1 - \hat{\mu}^\pi_2 \|_{\mathcal{H}}^2 &- \| \mu_1 - \mu_2 \|_{\mathcal{H}}^2
\end{align*} \]
Assumptions A.1–A.3 ensure the consistency of $V$-statistics indexed by the function $\pi()$ and its estimator $\hat{\pi}()$. By the continuous mapping theorem applied to the function $\psi(\hat{\pi}) = \| \hat{\mu}^\pi_1 - \hat{\mu}^\pi_2 \|_{\mathcal{H}}$, we have
\[ \| \hat{\mu}^\pi_1 - \hat{\mu}^\pi_2 \|_{\mathcal{H}}^2 - \| \mu_1 - \mu_2 \|_{\mathcal{H}}^2 \rightarrow 0. \]

Finally, by invoking the Slutsky theorem and using the results established in Lemma 1, we prove asymptotic convergence to $\chi^2$ distribution and test consistency under the null hypothesis. By replicating the same arguments as before, we obtain the counterpart for the bootstrap process (Kosorok 2008).

Funding

This study was supported by the Spanish Ministry of Science and Innovation (Grants no.: PID2021-123152OB-C21, TEC2021-130374B-C21, PID2020-112623GB-I00, PDC2021-121072-C21), the Carlos III Institute of Health (Grant no.: ISCIII/P120/01069) (co-funded by the European Union), and the Network for Research on Chronicity, Primary Care, and Health Promotion (Grant no.: ISCIII/RD21/0016/002) (also co-funded by the European Union).

References

Ahlqvist, E., Tuomi, T., and Groop, L. (2019), “Clusters Provide a Better Holistic View of Type 2 Diabetes than Simple Clinical Features,” The Lancet Diabetes & Endocrinology, 7, 668–669. [364]
Akritas, M. G., Antoniou, E. S., and Kuha, J. (2006), “Nonparametric Analysis of Factorial Designs with Random Missingness: Bivariate Data,” Journal of the American Statistical Association, 101, 1513–1526. [357]
Amro, L., and Pauly, M. (2017), “Permuting Incomplete Paired Data: A Novel Exact and Asymptotic Correct Randomization Test,” Journal of Statistical Computation and Simulation, 87, 1148–1159. [357]
Amro, L., Pauly, M., and Ramosaj, B. (2021), “Asymptotic-based Bootstrap Approach for Matched Pairs with Missing Data,” Sociological Methods & Research, 30, 425–454. [357]

Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., Bosi, E., Buckingham, B. A., Cefalu, W. T., Close, K. L. et al. (2019), ‘Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations from the International Consensus on Time in Range,” Diabetes Care, 42, 1593–1603. [358]
Beck, R. W., Bergenstal, R. M., Riddellworth, T. D., Kolman, C., Li, Z., Brown, A. S., and Close, K. L. (2019), “Validation of Time in Range as an Outcome Measure for Diabetes Clinical Trials,” Diabetes Care, 42, 400–405. [358]
Berliner, A., and Thomas-Agnan, C. (2011), Reproducing Kernel Hilbert Spaces in Probability and Statistics, New York: Springer. [360]

Biggerstaff, B., Gouet, R., Klein, T., and Lopez, A. (2017), “Geodesic PCA in the Wasserstein Space by Convex PCA,” Annales de l’Institut Henri Poincaré, 53, 1–26. [357]
Chi, C. W., Egan, J. M., and Ferrucci, L. (2018), “Age-Related Changes in Glucose Metabolism, Hyperglycemia, and Cardiovascular Risk,” Circulation Research, 123, 886–904. [363]

Dennis, J. (2020), “Precision Medicine in Type 2 Diabetes: Using Individualized Prediction Models to Optimize Selection of Treatment,” Diabetes, 69, 2075–2085. [364]
Derrick, B., Ruck, A., Toher, D., and White, P. (2018), “Tests for Equality of Variances Between Two Samples which Contain Both Paired Observations and Independent Observations,” Journal of Applied Quantitative Methods, 13, 36–47. [357]
Dubey, P., and Müller, H.-G. (2020), “Functional Models for Time-Varying Random Objects,” Journal of the Royal Statistical Society, Series B, 82, 275–327. [357]

Durante, F., and Mesiar, R. (2010), “L-Sets and S-Approximations and Independent Observations,” Journal of Mathemati-...
Van Der Vaart, A. W., and Wellner, J. A. (1996), “Weak Convergence,” in *Weak Convergence and Empirical Processes*, pp. 16–28, New York: Springer. [366,367]

Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, New York: Springer. [360]

Verbeke, G., and Molenberghs, G. (2009), *Linear Mixed Models for Longitudinal Data*, New York: Springer. [357]

Xu, J., and Harrar, S. W. (2012), “Accurate Mean Comparisons for Paired Samples with Missing Data: An Application to a Smoking-Cessation Trial,” *Biometrical Journal*, 54, 281–295. [357]

Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., and Jang, W. (2012), “Permutation Test for Incomplete Paired Data with Application to cDNA Microarray Data,” *Computational Statistics & Data Analysis*, 56, 510–521. [357]