Power and sample size calculation for Two-Sample projection-based testing of sparsely observed functional data

Journal Title
XX(X):2–37

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Salil Koner[1] and Sheng Luo[1]

Prepared using sagej.cls

Abstract

Projection-based testing of the typical trajectory for two or multiple groups of irregular and sparsely observed (multi-dimensional) data from underlying a continuous stochastic process has increased attention in the literature due to its applicability in wide range of (non-stationary) covariance structures. We aspire that this article will enhance the applicability of this powerful inference mechanism, to many biostatistical applications, especially from the perspective of designing new clinical trials. We derive the theoretical power function and construct sample size calculation toolbox for this projection-based test under a general structure of group difference and broad class of covariance structures of the underlying process. Numerical studies demonstrate that the power of the test does not degrade when the percentage of missing observations remain within a certain range, making the test *missing-immune*. Applicability of the test is demonstrated for several real data examples. A user-friendly R package `fPASS` with an elaborately explained vignette is created to enhance its applicability in practical applications.

Keywords

Power, Projection-based test, Functional principal component analysis.

1 Introduction

Many real world applications involve comparing two groups based on a primary outcome observed over time. For examples, in clinical trials

[1]Duke University, Durham NC USA

Corresponding author:
Salil Koner, Department of Biostatistics, Duke University School of Medicine, Durham NC USA.
Email: salil.koner@duke.edu

testing the efficacy of a study drug is usually concluded by means of a statistically validated two-sample testing procedure. In modern clinical trials, the outcomes are collected longitudinally, thereby two-sample inference techniques designed for longitudinal data are the most frequently adopted in this front[4]. These techniques have many advantages, including being well-studied in the literature and having efficient statistical software readily available for implementation. Additionally, power and sample size calculation formula for these procedures are readily available in the standard software, making them a valuable tool in clinical trial design[1].

The standard state-of-the-art two-sample testing procedures for repeated measures design as well as the commonly used power and sample size calculation (PASS) softwares such[9], `proc mixed`[11], assumes certain restrictive assumptions on the evolution of the mean trajectory over time. For example, in clinical trial design phase, it is a standard practice to assume that the the mean function grows linearly, or quadratically over time. Also, the power function formula requires to specify a certain stationary covariance structure of the longitudinal response, such as Compound symmetry (CS) or Auto-regressive (AR), or Banded. Although the standard random effect model induces a non-stationary covariance structure, the implementation becomes quite burdensome when the number of random effects grows more than three. Moreover, many of the tests require that the number of observation times for each subject are uniform and the interval between two subsequent visits are regular.

Over the past few years, several attempts have been made to develop powerful testing procedures for longitudinal data by encompassing the longitudinal design under a sparse and irregular functional design. These procedures allows the mean functions to evolve non-parametrically over time, and they do not impose any structural assumptions on the covariance. Inference problems involving sparse and irregularly observed functional

data are more challenging than dense regularly observed functional data because a few number of observations are observed for each subject. These procedure rely on smoothness of temporal mean and covariance function and borrow information from all subjects to make inference about the mean.

The first genre of tests developed for sparse functional data are the likelihood ratio tests[2,3] and its various extension to more complicated covariance structures[12,14]. The other genre is the projection-based testing procedure that projects the data onto a suitable set of data-driven leading directions and use the projection-scores to further test for the difference between the two groups.[10] employed an Anderson-Darling test on projection scores to test the equality of distributions, whereas[13] used a Hotelling's T^2 type statistic to test the difference in the mean between the scores. The advantage of Hotelling T^2 type statistic over the nonprameteric distribution test is that it considers the scores as multivariate vectors and subsequently, it tests the null hypopthesis of group difference based on a single value, whereas the non-parameteric distribution test requires multiple testing. This facilitates extension of this test to multivariate sparsely observed functional setting, as developed by[8]. Moreover, the Hotelling T^2 type test can be easily extended to test difference between multiple groups.

Unfortunately, applicability of all these powerful tests developed for sparse and irregularly observed functional data is still confined to researchers interested in developing statistical methodology. Although these tests make much less restrictive assumption on the mean and covariance structure of the response, allows irregularly observed data, they still did not attract the attention of clinical design practitioners. This is primarily driven by two obstacles – firstly, owing to the theoretical intricacies involved, finite sample power function formula is yet to be developed for any of these tests. Subsequently, sample size calculation

softwares based on these test to achieve a certain degree of power is not yet available to the end-user. Therefore, although these tests can be applied to test for group difference after collection of data in clinical trials, they still fail to take the place of the primary statistical approach in the statistical analysis plan (SAP) document because the optimal sample size and the corresponding power analysis cannot be presented if one decide to use these tests as the primary testing procedure.

In this article, we derive a PASS formula for the projection-based Hotelling T^2 type test developed for sparse and irregularly observed functional data by[13]. We first derive the asymptotic power function of the test and present a user-friendly formula to derive the minimum sample size required to achieve a certain power by inverting the power function. We show that the non-null distribution of the test-statistic follows an approximate non-central χ^2 distribution with non-centrality parameter comprising of the projection of the true difference in the mean on the principal directions. The PASS formula have several nodes of flexibilities, First it can accommodate any smooth structure of the group difference, it does not need to be linear, quadratic or piece-wise linear. Second, the user can specify any arbitrary non-stationary smooth covariance structure for the response, in addition to standard parametric structural covariance assumed for longitudinal data. Third, the number of observations for each subject can vary (randomly) and the user can specify any irregular design for the observation points.

For fixed sample size, the power function of the test fundamentally relies on how well the covariance structure as well as the principal data-driven directions are estimated from the data. If the sample size is reasonably large enough so that the covariance structure is well-estimated, via several numerical studies we demonstrate the test's power does not deteriorate even if the percentage of missing observations for each subject increases until within a certain range, rendering the test *missing-immune*. This

feature is highly desirable in real-world applications where missing data is often encountered. We demonstrate how the test can be applied in two real clinical trial examples involving patients with Parkinson's Disease (PD), and illustrate how such applying such a powerful testing procedure would possibly lead to a recruitment of significantly lower number of subjects to achieve the same power compared to the one specified in the respective SAPs.

To enhance the practical applicability of this testing mechanism, we develop a user-friendly R package, "fPASS", hosted on Github at `https://github.com/SalilKoner/fPASS`. We hope that the availability of this package will make the projection-based test more accessible to practitioners in the field of biostatistics, especially in the design and analysis of clinical trials.

The rest of the article is presented as follows. In Section 2, we describe the testing procedure briefly and present the non-null distribution of the statistic when the projection functions are known. Section 3 is dedicated to the asymptotic non-null distribution of the test when the projection functions are estimated from the data and how the power can be calculated in practical situations. Sample size calculation formula is presented in Section 4. Validation of the PASS formula is presented via several numerical studies are presented in Section 5 with real data applications are presented in Section 6.

2 Projection-based testing framework

In this we will briefly describe the testing framework of[13] to preface its PASS formula. Let $Y_{ij,g}$ denotes the jth observation for the ith subject in the gth group, for $g = 1, 2$, $i = 1, \ldots, n_g$, and $j = 1, \ldots, N_{i,g}$. Let $T_{ij,g}$ is the associated timepoint when $Y_{ij,g}$ is observed. We posit the following

model for the response,

$$Y_{ij,g} = X_{i,g}(T_{ij,g}) + \epsilon_{ij,g}, \tag{1}$$

where $\epsilon_{ij,g}$ are independent and identically distributed (i.i.d) measurement error with common variance $\tau^2 > 0$. For each $g = 1, 2$, we assume that the latent process $\{X_{i,g} : i = 1, \ldots, n_g;\}$ are i.i.d copies of an underlying stochastic process $X_g(t) \in \mathcal{L}^2(\mathcal{T})$, the space of all square integrable random functions in the compact domain $\mathcal{T}$. The associated inner product of two functions $f_1(t)$ and $f_2(t)$ in the space $\mathcal{L}^2(\mathcal{T})$ is defined as $\langle f_1, f_2 \rangle := \int_{\mathcal{T}} f_1(t) f_2(t)\, dt$. We assume that $X_g(t)$ have a group specific mean function $\mu_g(t)$ and a (common) covariance operator $(\Xi\ f)(t) = \int \Sigma(t, s) f(s) ds$, induced by the covariance kernel $\Sigma(t, s) = \text{Cov}(X_g(t), X_g(s))$ for both $g = 1, 2$. We also assume that the $X_{i,g}(t)$ are independent across the groups as well. We focus on sparse random design here[16], i.e. we assume that the number of observations for each subject, $N_{i,g}$ is finite with $\sup_{i,g} N_{i,g} < \infty$ and the observation points $T_{ij,g} : j = 1, \ldots, N_{i,g}$ are i.i.d copies of a random variable with bounded density function within the domain $\mathcal{T}$. Lastly, let $n := n_1 + n_2$ be the total number of subjects combing the groups. We are interested to test the hypothesis

$$\begin{aligned} H_0 &: \mu_1(t) = \mu_2(t) \ \ \forall \ \ t \in \mathcal{T} \\ &\text{vs} \\ H_1 &: \mu_1(t) \neq \mu_2(t) \ \ \text{for some} \ \ t \in \mathcal{T}. \end{aligned} \tag{2}$$

We will briefly describe the projection-based testing framework here. Let $\{\psi_k(t)\}_{k \geq 1}$ be the orthonormal eigenfunctions of the covariance operator Ξ, corresponding to the sequence of ordered eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Throughout Section 2 we assume that the eigenfunctions are known.[13] considered a dimension reduction approach to project the

response trajectories onto a set of orthogonal basis functions $\{\psi_k(t)\}_{k\geq 1}$ for $\mathcal{L}^2(\mathcal{T})$ to represent the latent process as

$$X_{i,g}(t) = \mu_0(t) + \sum_{k=1}^{\infty} \zeta_{ik,g}\psi_k(t), \quad g = 1, 2, \tag{3}$$

where $\mu_0(t)$ is the pooled mean function, common to both the groups, and $\zeta_{ik,g} = \langle X_{i,g} - \mu_0, \psi_k \rangle$ is the group-specific projection that captures the group specific difference. Specifically, $\{\zeta_{ik,g}\}$ are uncorrelated across k, has a mean $\langle \mu_g - \mu_0, \psi_k \rangle$ and variance λ_k, $g = 1, 2$. We set $\mu_g(t) - \mu_0(t) = 0$ for either of $g = 1$ or $g = 2$ to ensure identifiability. In this projection-based framework, the null hypothesis in (2) is equivalent to testing

$$H_0 : \{\mathbb{E}(\zeta_{ik,1})\}_{k\geq 1} = \{\mathbb{E}(\zeta_{ik,2})\}_{k\geq 1}$$

vs

$$H_1 : \{E(\zeta_{ik,1})\}_{k\geq 1} \neq \{E(\zeta_{ik,2})\}_{k\geq 1}.$$

See[13] for detailed insights for the above statement.

For sparsely observed functional data, the projections $\{\zeta_{ik,g}\}$ can not be consistently estimated, since the entire trajectories are not observed in a dense grid[16]. Hence, in the sparse design, the scores are obtained via best linear unbiased prediction (BLUP) under a mixed model with a "working" Gaussian assumption[16],

$$\mathbf{Y}_{i,g} = \boldsymbol{\mu}_{0i,g} + \boldsymbol{\Psi}_{i,g}\boldsymbol{\zeta}_{i,g} + \boldsymbol{\epsilon}_{i,g}, \tag{4}$$

where $\mathbf{Y}_{i,g} = (Y_{i1,g}, \ldots, Y_{iN_{i,g},g})^\top$ be the $N_{i,g}$-length stacked vector of response for the ith subject, $\boldsymbol{\mu}_{0i,g} = (\mu_0(T_{i1,g}), \ldots, \mu_0(T_{iN_{i,g},g}))^\top$ and $\boldsymbol{\Psi}_{i,g} = (\boldsymbol{\psi}_{i1,g}, \ldots, \boldsymbol{\psi}_{iK,g})$ be the column-stacked version of $\{\boldsymbol{\psi}_{ik,g}\}_{k=1}^K$'s with $\boldsymbol{\psi}_{ik,g} = (\psi_k(T_{i1,g}), \ldots, \psi_k(T_{iN_{i,g},g}))^\top$ are the mean vector and the eigenvector matrix for the ith subject, and $\boldsymbol{\epsilon}_{i,g}$ are the measurement error. Under the Gaussian assumption of the scores and the measurement error,

the best linear unbiased predictor (BLUP) of $\boldsymbol{\zeta}_{i,g}$ under the model (4) is of the form

$$\begin{aligned}\widetilde{\boldsymbol{\zeta}}_{i,g} &:= \mathbb{E}(\boldsymbol{\zeta}_{i,g}|\mathbf{Y}_{i,g}) \\ &= \text{diag}(\lambda_1,\ldots,\lambda_K)\boldsymbol{\Psi}_{i,g}^{\top}\mathbf{G}_{\mathbf{Y}_{i,g}}^{-1}(\mathbf{Y}_{i,g}-\boldsymbol{\mu}_{0i,g}),\end{aligned} \tag{5}$$

where $\mathbf{G}_{\mathbf{Y}_{i,g}} = \text{Cov}(\mathbf{Y}_{i,g}) = \{\Sigma(T_{ij,g},T_{ij',g}) + \tau^2\mathbb{I}(j=j')\}_{1\le j,j'\le N_{i,g}}$ be the covariance matrix of $\mathbf{Y}_{i,g}$. The quantity $\widetilde{\boldsymbol{\zeta}}_{i,g}$ is termed as "shrinkage" scores, and it is rational choice for estimating the true score $\boldsymbol{\zeta}_{i,g}$ because

$$\mathbb{E}(\widetilde{\boldsymbol{\zeta}}_{i,g}) = \mathbb{E}[\mathbb{E}(\boldsymbol{\zeta}_{i,g} \mid \mathbf{Y}_{i,g})] = \mathbb{E}(\boldsymbol{\zeta}_{i,g})$$

The BLUP estimator $\widetilde{\boldsymbol{\zeta}}_{i,g}$ is a consistent estimator of the unobserved true scores $\boldsymbol{\zeta}_{i,g}$ as the number of observations per subject grows and the measurement error gets small. It is also important to note that unlike the true scores, $\{\widetilde{\zeta}_{ik,g}\}_{k=1}^{K}$ are also not uncorrelated across k.

Let $\widetilde{\boldsymbol{\zeta}}_{i,g} = (\widetilde{\zeta}_{i1,g},\ldots,\widetilde{\zeta}_{iK,g})^{\top}$ be the K-dimensional multivariate scores for the ith subject. Define, $\widetilde{\boldsymbol{\zeta}}_{1+} = n_1^{-1}\sum_{i=1}^{n_1}\widetilde{\boldsymbol{\zeta}}_{i,1}$ and the $\widetilde{\boldsymbol{\zeta}}_{2+} = n_2^{-1}\sum_{i=1}^{n_2}\widetilde{\boldsymbol{\zeta}}_{i,2}$ are the average of the scores for the two groups. Similarly define, $\widetilde{\boldsymbol{\Lambda}}_1 = (n_1-1)^{-1}\sum_{i=1}^{n_1}(\widetilde{\boldsymbol{\zeta}}_{i,1}-\widetilde{\boldsymbol{\zeta}}_{1+})(\widetilde{\boldsymbol{\zeta}}_{i,1}-\widetilde{\boldsymbol{\zeta}}_{1+})^{\top}$ and $\widetilde{\boldsymbol{\Lambda}}_2 = (n_2-1)^{-1}\sum_{i=2}^{n_2}(\widetilde{\boldsymbol{\zeta}}_{i,2}-\widetilde{\boldsymbol{\zeta}}_{2+})(\widetilde{\boldsymbol{\zeta}}_{i,2}-\widetilde{\boldsymbol{\zeta}}_{2+})^{\top}$ are the sample variance of the 'shrinkage' scores for the two groups. Further, define the pooled sample covariance as $\widetilde{\boldsymbol{\Lambda}} = \{(n_1-1)\widetilde{\boldsymbol{\Lambda}}_1 + (n_2-1)\widetilde{\boldsymbol{\Lambda}}_2\}/(n-2)$. To test for H_0, a Hotelling T^2 random variable[6] using the 'shrinkage' scores can be constructed as

$$T_n = \frac{n_1 n_2}{n_1+n_2}(\widetilde{\boldsymbol{\zeta}}_{1+}-\widetilde{\boldsymbol{\zeta}}_{2+})^{\top}\widetilde{\boldsymbol{\Lambda}}^{-1}(\widetilde{\boldsymbol{\zeta}}_{1+}-\widetilde{\boldsymbol{\zeta}}_{2+}), \tag{6}$$

As[13] derived, the asymptotic null distribution of the test statistic T_n is χ^2 distributed with K degrees of freedom. However, for fixed sample size n, by relating the Hotelling T^2 distribution to the F-distribution[7] Chapter

5, the test rejects H_0 at a specified significance level $\alpha \in (0, 1)$ if

$$\frac{(n-K-1)T_n}{(n-2)K} > F_\alpha(K, n-K-1), \tag{7}$$

where K is the dimension of fPC scores estimated from the data, and $F_\alpha(a, b)$ is the $100(1-\alpha)\%$ quantile of F-distribution with a and b degrees of freedom.

Non-null distribution of the test

Having the testing framework discussed briefly in the previous section, we are now ready to demonstrate the power of the test to reject the null hypothesis H_0 when the functional form of the group difference is specified. We assume that the data are observed under a sparse design with the underlying model (1). In the following proposition, we present the non-null distribution of T_n when the group difference $\mu_1(t) - \mu_2(t)$ is different from zero and the eigenfunctions are known.

Notice that the Hotelling T^2 statistic is constructed under the fact that the population covariance of the 'shrinkage' scores is the same for the two groups when the null hypothesis is true. However, under the alternate hypothesis, the covariance of the 'shrinkage' scores for the two groups differ from each other, which is clear from Equation (2.11) of[13]. This poses a significant challenge in the derivation of the non-null distribution of the test statistic because the non-null distribution of the test must be derived under the assumption of unequal variance of the ''shrinkage' scores between the two groups. The next proposition specifies the fixed sample non-null distribution of the test under the assumption of unequal variance.

Theorem 1. *Assume that model (1) for the observed response* $\{\mathbf{Y}_{ij,g} : j = 1, \ldots, N_{i,g}\}_{i=1}^{n_g}$ *is true, and* $\sup_{i,g} N_{i,g} < \infty$ *almost surely. Assume that the scores* $\{\zeta_{ik,g}\}$ *are independent across* i *and* g *and they are*

Gaussian. Further assume that the true mean functions $\mu_g(t)$ and the true eigencomponents $\{\lambda_k, \psi_k(t)\}_{k\geq 1}$ are known. Then, conditional on the truncation parameter K,

$$\frac{n_2(1+1/\kappa)T_n}{n_1+n_2-2} \stackrel{d}{=} \frac{\sum_{k=1}^{K} d_k^{-1}\chi_1^2\left(n_1(\boldsymbol{u}_k^\top \boldsymbol{\Lambda}^{\dagger -1/2}\boldsymbol{\Delta})^2\right)}{\chi^2_{\nu-K+1}/\nu} \tag{8}$$

where $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_K)^\top$ with $\delta_k = \langle \mu_1 - \mu_2, \psi_k\rangle$ and $\boldsymbol{\Lambda}^\dagger = \boldsymbol{\Lambda}_1 + \kappa\boldsymbol{\Lambda}_2$, with $\boldsymbol{\Lambda}_g : g = 1, 2$ are the population covariance of the 'shrinkage' scores, $\{\widetilde{\boldsymbol{\zeta}}_{i,g}\}$, for $g = 1, 2$, defined in (5). Moreover, $\{d_k, \boldsymbol{u}_k\}_{k=1}^K$ are the eigenvalue and the eigenvectors of the spectral decomposition $\boldsymbol{\Omega}^\dagger = \sum_{k=1}^K d_k \boldsymbol{u}_k\boldsymbol{u}_k^\top$, with $\boldsymbol{\Omega}^\dagger := \kappa(\kappa - 1/n_2)\boldsymbol{\Omega} + (1-1/n_2)(\mathbf{I}_K - \boldsymbol{\Omega})$, with $\boldsymbol{\Omega} := \boldsymbol{\Lambda}^{\dagger -1/2}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}^{\dagger -1/2}$. Finally, ν is the degrees of freedom of the chi-squared distribution under the denominator which is of the form

$$\begin{aligned}\nu = n_2\left\{\mathrm{tr}(\boldsymbol{\Omega}^{\dagger 2}) + \mathrm{tr}^2(\boldsymbol{\Omega}^\dagger)\right\}\big[\kappa^2(\kappa - n_2^{-1})\{\mathrm{tr}(\boldsymbol{\Omega}^2) + \mathrm{tr}^2(\boldsymbol{\Omega})\} \\ + (1-n_2^{-1})\{\mathrm{tr}(\mathbf{I}_K - \boldsymbol{\Omega})^2 + \mathrm{tr}^2(\mathbf{I}_K - \boldsymbol{\Omega})\}\big]^{-1}.\end{aligned} \tag{9}$$

Theorem 1 gives the exact distribution of the test-statistic T_n when the eigenfunctions are known and the scores have a Gaussian distribution. The distribution is not an exact F distribution, but we notice that it is very close to a non-central F distribution, as the numerator is a linear combination of K independent non-central chi-squared random variables, each with one degree of freedom, and the denominator is a $\chi^2_{\nu-K+1}$. Although the form of the non-null distribution is non-trivial, however, we can efficiently simulate random samples from the distribution, and compute the power function of the test. In the next section, we extend the statement of the above proposition when the eigenfunctions are unknown and explicitly discuss how the power function of the test can be calculated in practical situations.

3 Calculation of power function in practical situations

In practical situation, the true eigenfunctions are unknown. Hence, the shrinkage scores $\widetilde{\zeta}_{ik,g}$ are estimated from the data by first estimating the mean and the covariance function of the data, followed by conducting an functional principal component analysis (fPCA)[16] of the estimated covariance function to get a consistent estimate of the eigenfunctions. It is important to remark that even for the sparse design, a consistent estimate of the eigenfunctions can be obtained as long the number of subjects in each groups are reasonably high[17]. Plugging in the estimator of common mean $\mu_0(t)$ and the eigenfunctions in (5), we can obtain the estimator of the 'shrinkage' scores. Let $\widehat{\boldsymbol{\zeta}}_{i,g} = (\widehat{\zeta}_{i1,g}, \ldots, \widehat{\zeta}_{iK,g})^\top$ be the estimated scores for the ith subject, define $\widehat{\boldsymbol{\zeta}}_{1+} = n_1^{-1}\sum_{i=1}^{n_1}\widehat{\boldsymbol{\zeta}}_{i,1}$ and the $\widehat{\boldsymbol{\zeta}}_{2+} = n_2^{-1}\sum_{i=1}^{n_2}\widehat{\boldsymbol{\zeta}}_{i,2}$ are the average of the estimated scores for the two groups. Similarly define, $\widehat{\boldsymbol{\Lambda}}_1 = (n_1-1)^{-1}\sum_{i=1}^{n_1}(\widehat{\boldsymbol{\zeta}}_{i,1} - \widehat{\boldsymbol{\zeta}}_{1+})(\widehat{\boldsymbol{\zeta}}_{i,1} - \widehat{\boldsymbol{\zeta}}_{1+})^\top$ and $\widehat{\boldsymbol{\Lambda}}_2 = (n_2-1)^{-1}\sum_{i=2}^{n_2}(\widehat{\boldsymbol{\zeta}}_{i,2} - \widehat{\boldsymbol{\zeta}}_{2+})(\widehat{\boldsymbol{\zeta}}_{i,2} - \widehat{\boldsymbol{\zeta}}_{2+})^\top$ are the sample variance of the scores. Defining $\widehat{\boldsymbol{\Lambda}} = \{(n_1-1)\widehat{\boldsymbol{\Lambda}}_1 + (n_2-1)\widehat{\boldsymbol{\Lambda}}_2\}/(n-2)$ to be the pooled sample covariance, in practical situations, we test the null hypothesis H_0 based on the statistic

$$\widehat{T}_n = \frac{n_1 n_2}{n_1+n_2}(\widehat{\boldsymbol{\zeta}}_{1+} - \widehat{\boldsymbol{\zeta}}_{2+})^\top \widehat{\boldsymbol{\Lambda}}^{-1}(\widehat{\boldsymbol{\zeta}}_{1+} - \widehat{\boldsymbol{\zeta}}_{2+}). \tag{10}$$

Observed that the test statistic $\widehat{T}_n$ is the observed counterpart of the unobserved quantity T_n defined in (6). In the next proposition, we will derive the asymptotic non-null distribution of the test-statistic $\widehat{T}_n$, which serves as the fundamental result to the derivation of the asymptotic power of the projection-based test.

Proposition 2. *Assume that model (1) for the observed response* $\{\mathbf{Y}_{ij,g} : j = 1, \ldots, N_{i,g}\}_{i=1}^{n_g}$ *is true, and* $\sup_{i,g} N_{i,g} < \infty$ *almost surely. Suppose that the alternate hypothesis is true and that the true mean difference is*

characterized by

$$H_1 : \mu_1(t) - \mu_2(t) = n^{-\varrho/2}\eta(t),$$

with $\eta(t) \neq 0$*, is a fixed known function of* t*, for some* $\varrho \in [0, 1]$*. Assume that the mean functions for both the groups and the covariance components are estimated consistently, i.e.* $\|\widehat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}_g\| = o_p(1)$ *for both* $g = 1, 2$*,* $\|\widehat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\| = o_p(1)$*,* $\|\widehat{\lambda}_k - \lambda_k\| = o_p(1)$ *for all* $k = 1, \ldots, K$*, and* $|\widehat{\tau}^2 - \tau^2| = o_p(1)$*, and that* $\lim n_1/n_2 \to \kappa \in (0, \infty)$*. Then, conditional on the truncation parameter* K*,*

$$\widehat{T}_n \xrightarrow{d} \begin{cases} \kappa \sum_{k=1}^{K} d_k^{-1} \chi_1^2 \left((\boldsymbol{u}_k^\top \boldsymbol{\Lambda}^{\dagger^{-1/2}} \boldsymbol{\Delta})^2 \right) & \text{if } \varrho = 1 \\ \infty & \text{if } \varrho < 1 \end{cases}$$

where $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_K)^\top$ *with* $\delta_k = \langle \eta, \psi_k \rangle$ *and* $\boldsymbol{\Lambda}^\dagger = \boldsymbol{\Lambda}_1 + \kappa \boldsymbol{\Lambda}_2$*, with* $\boldsymbol{\Lambda}_g : g = 1, 2$ *are the population covariance of the 'shrinkage' scores,* $\{\widetilde{\boldsymbol{\zeta}}_{i,g}\}$*, for* $g = 1, 2$*, defined in (5). Moreover,* $\{d_k, \boldsymbol{u}_k\}_{k=1}^K$ *are the eigenvalue and the eigenvectors of the spectral decomposition* $\boldsymbol{\Omega}^\dagger = \sum_{k=1}^K d_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$*, with* $\boldsymbol{\Omega}^\dagger := \mathbf{I}_K + (\kappa^2 - 1)\, \boldsymbol{\Lambda}^{\dagger^{-1/2}} \boldsymbol{\Lambda}_1 \boldsymbol{\Lambda}^{\dagger^{-1/2}}$.

Proposition 2 provides the asymptotic distribution of the projection-based test under the local alternative, or Pitman type of alternative hypothesis. When the true group difference decays at a rate slower than $n^{-1/2}$, then the test rejects the null hypothesis with probability 1. The asymptotic non-null distribution of the test statistic does not require a Gaussian assumption on the scores. It is important to remark that the asymptotic distribution of $\widehat{T}_n$ is contingent on the consistent estimation of the eigenfunctions from the data.

Asymptotic power function formula

To investigate the power of the test as a function as function of the sample size, as well as to find the minimal sample size calculation (Section 4),

we relate the non-null distribution of the test-statistic with a non-central F distribution as in Theorem 1, under a working Gaussian assumption of the scores. Therefore, for fixed sample sizes, (n_1, n_2) of the two groups, the theoretical power function of the test can be calculated as

$$\mathcal{P}_{n_1,n_2}(\eta) = \mathbb{P}\left(F^* > \frac{Kn_2F_\alpha(K, n-K-1)}{(1+1/\kappa)^{-1}(n-K-1)}\right) \tag{11}$$

where F^* is identically distributed as the random variable on the right hand side of the expression (8). One can empirically compute the above probability with a high precision by generating a desirably large number of random samples from the distribution of F^*.

Equation (11) provides the working formula for the calculation of the theoretical power of the projection-based test. This means, from a practitioner's perspective, if the true group difference $\eta(t)$, and the eigenfunctions $\{\psi_k(t)\}_{k=1}^K$ are specified, then one can compute the non-centrality parameter of the F distribution in $\mathcal{P}_{n_1,n_2}(\eta)$ for any n, and the power function of the test for any (n_1, n_2) provided that we have a certain knowledge about the unknown covariance parameters $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. From Eq. (2.11) of[13], we can follow that $\mathbf{\Lambda}_g : g = 1, 2$ is a complicated function of the mean and the eigenfunctions of the stochastic process, and the marginal density of the observation points $\{T_{ij,g} : j = 1, \ldots, N_{i,g}\}$. In the next part, we will present a computationally efficient algorithm for the estimation of the power function $\mathcal{P}_{n_1,n_2}(\eta)$ where the unknown quantities $\mathbf{\Lambda}_g : g = 1, 2$ can be reliably estimated from a representative large sample.

3.1 Data-driven estimation of theoretical power function

So far, the power function formula of the test presented in (11) requires the knowledge about the true eigenfunctions of the data. That means, the power function of the test is primarily dependent upon the covariance structure of the response through its eigenfunctions. However, having

a scientific or informed knowledge about the covariance structure of the observation process, $\Sigma(t,t')$ rather than having a knowledge about the eigenfunctions is more pragmatic in real life situation. Therefore, a mechanism to extract the true eigenfunctions from the covariance function $\Sigma(t,t')$ is necessary to calculate the power function.

A consistent estimate of the eigenfunctions can be obtained by spectral decomposition of the covariance matrix evaluated a fine grid of points. To be specific, first set a large grid of size R (say 100) and generate a sequence of points of between the compact interval $\mathcal{T}$ to evaluate the $R \times R$ covariance matrix with elements $\mathbf{\Sigma} = ((\Sigma(t_r, t_{r'})))_{1\leq r,r'\leq R}$, followed by a spectral decomposition of $\mathbf{\Sigma}$ to obtain a reliable estimate of the true eigenfunctions $\{\widetilde{\psi}_k(t) : t = t_1, \ldots, t_R\}_{k=1}^K\}$ with K chosen by a pre-specified percentage of variation explained (PVE) (e.g. 90%) to extract only the leading eigenfunctions of the covariance. Then, the projection δ_k can be obtained numerically by $\delta_k = R^{-1}\sum_{r=1}^R \eta(t_r)\widehat{\psi}_k(t_r)$, for $k = 1, \ldots, K$. However, we still have to obtain a reliable estimate of unknown $\mathbf{\Lambda}$ to compute the power function.

As an one-stop solution for both obtaining a consistent estimator of the eigenfunctions as well the covariance parameters $\{\mathbf{\Lambda}_g : g = 1, 2\}$, we empirically generate a data with (possibly) large number of subjects (say 5000) under the model (1) with $\mu_1(t) = 0$ and $\mu_2(t) = \eta(t)$ and the specified covariance structure. The number of observations and observation points for each subject are also generated following the pre-specified sampling design. Then, we can employ the standard statistical softwares available for fPCA under sparse design (such as R package `face`[15]) to internally estimate the smooth covariance and the eigenfunctions. We can further use the scoring technique available in the software to estimate the shrinkage scores for each of the 5000 subjects and obtain a highly reliable estimate of the $\mathbf{\Lambda}_g = \text{Cov}(\boldsymbol{\zeta}_{i,g})$ by computing the sample covariance of 'shrinkage' scores based on 5000 subjects. Since, the

eigenfunctions and the covariance of the shrinkage scores are estimated based on a quite large number of subjects, we expect that them to be reasonably close enough to the true eigenfunctions. We specify these details concisely in Algorithm 3.1 below.

Input: Data information: True difference in mean function between two groups, i.e. $\eta(t) := \mu_1(t) - \mu_2(t)$, covariance of the latent process $\Sigma(t, t')$, measurement error variance τ^2, significance level $\alpha \in (0, 1)$, sample sizes (n_1, n_2), and a pre-specified PVE to estimate the eigencomponents of the covariance;

1 Generate a dataset with large number of subjects, say 5000, using the model (1) by setting $\mu_1(t) = 0$ and $\mu_2(t) = \eta(t)$ and covariance of the $X_g(t)$ as $\Sigma(t, t')$ and the measurement error variance $\tau^2 > 0$. ;

2 Preset a large grid of size R (say 100) and generate a sequence of points $t_1 < t_2, \ldots < t_R$ in $\mathcal{T}$;

3 Conduct fPCA on the generated dataset to obtain a highly reliable estimate of the eigenfunctions $\{\widetilde{\psi}_k(t) : k = 1, \ldots, K\}$ at $t = t_1, \ldots, t_R$. The number of eigenfunctions K are chosen by the specified PVE ;

4 Calculate the fPC scores for each subject and obtain the sample covariances $\{\widetilde{\mathbf{\Lambda}}_g : g = 1, 2\}$ of the scores for the two groups. This will serve as a reliable estimated of the true covariances $\{\mathbf{\Lambda}_g : g = 1, 2\}$. ;

5 **for** $k \in \{1, \ldots, K\}$ **do**

6 Calculate the projection $\widetilde{\delta}_k = \int \eta(t)\widetilde{\psi}_k(t)dt \approx R^{-1} \sum_{r=1}^{R} \eta(t_r)\widetilde{\psi}_k(t_r)$

7 **end**

8 Construct the vector $\widetilde{\mathbf{\Delta}} = (\widetilde{\delta}_1, \ldots, \widetilde{\delta}_K)^\top$;

9 Return power function $\mathcal{P}_{n_1,n_2}(\eta)$ as in Equation (11) by replacing $\mathbf{\Delta}$ and $\{\mathbf{\Lambda}_g\}$ with $\widetilde{\mathbf{\Delta}}$ and $\{\widetilde{\mathbf{\Lambda}}_g\}$ respectively;

Algorithm 3.1: Algorithm for power function for projection-based test

Noe that the Step 1 - Step 4 of Algorithm 3.1 is only done to obtain a highly reliable estimate of the true eigenfunction of the covariance parameter of the 'shrinkage' scores, which is an essential requirement for reliable estimation of the power function of the test, as specified in Proposition 2. Once the eigenfunctions and the optimal number of functions K are consistently estimated, we can substitute the true eigenfunctions with the estimated ones and employ the formula in (11) to compute the power function of the test.

Usually, in the longitudinal data analysis literature, the commonly assumed knowledge of covariance are primarily stationary, such as compound symmetric, autoregressive (AR) etc. However, we remark that Algorithm 3.1 is capable of computing the power function of the test for any general covariance structure of the stochastic process $X_g(t)$.

There is a certain degree of caution we must point out. A vivid statistical enthusiasist might want to check how does the theoretical power function of the test $\mathcal{P}_{n_1,n_2}(\eta)$ calibrates with the empirical power function of the test. By empirical power, we mean that the user generates a large replication (say $B = 1000$) of the data using the specified mean difference and the covariance structure and conduct the test (7) in each of the datasets, and record the percentage of the time the test rejects H_0. Then, for a fixed sample size n, we remark that the theoretically calculated power obtained via Algorithm 3.1 and the empirically calculated power (based on large number of replications) will be close to each other only if the sample size n is large enough so that the eigenfunctions can be consistently estimated from the data in each of the B replications for the calculation of empirical power function. If the sample size n is so small that the eigenfunctions can not be consistently estimated, then the asymptotic null distribution of the test statistic $\widehat{T}_n$ is no more a χ^2_K. Therefore, the power function computed empirically computed using the test-rule (7) based on B datasets is not a reliable one. Hence, it should not compared to the power function

obtained from Algorithm 3.1 when n is small, because $\mathcal{P}_{n_1,n_2}(\eta)$ requires the eigenfunctions to be consistently estimated, as stated in Proposition 2. This is numerically demonstrated in Section 5.2.

Another important remark is that if the eigenfunctions are known or they are well-estimated from data, and the number of sampling points for each subjects are reasonable enough that the scores as well as the population covariance of the scores can be consistently estimated, the power function of the test is not affected by the increase percentage of missing observations for each subjects (until a certain level), since the calculation of power function at Step 5 - 9 do not require any information of the sampling design. Therefore, for large sample size, the asymptotic efficiency of the test is not deprecated by the missingness in the observation points, making the test 'missing-immune'. We numerically demonstrate this in Section 5.3.

4 Sample size calculation

Now that we have presented the formula for power function of the projection-based test in Section 3, we are now ready to present the algorithm for the minimum sample size required for our projection-based test to achieve a stipulated power of $100(1-\gamma)\%$ for any $\gamma \in (0,1)$.

Fix any $\gamma \in (0,1)$. Then the minimum sample size required for the projection-based test to detect a non-null group difference specified by $\eta(t) := \mu_1(t) - \mu_2(t)$ with a power of $100(1-\gamma)\%$ can be obtained by inverting the power function formula specified in (11), which is presented in the next proposition.

Proposition 3. *Assume that the conditions of the Proposition 1 is true. Suppose that the difference of the mean function between the two groups are specified by $\mu_1(t) - \mu_2(t)$. Assume that the eigenfunctions of the covariance of $X_g(t)$ is given, so that we have the projection-vector $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_K)^\top$ with $\delta_k = \langle \mu_1 - \mu_2, \psi_k \rangle$ and the population covariance of the 'shrinkage' scores $\{\boldsymbol{\Lambda}_g : g = 1, 2\}$ is known. Let $\kappa := n_1/n_2$ be the*

allocation ratio of the samples in the group. Then the minimum sample size required for the test to achieve a power of $100(1-\gamma)\%$ *is* $(\kappa n^*, n^*)$ *where* n^* *is the minimum positive integer that satisfies*

$$\mathbb{P}\left(\frac{\sum_{k=1}^{K} d_k^{-1}\chi_1^2\left(\kappa n^*(\boldsymbol{u}_k^\top \boldsymbol{\Lambda}^{\dagger\,-1/2}\boldsymbol{\Delta})^2\right)}{\chi^2_{\nu(n^*)-K+1}/\nu(n^*)} > \frac{Kn^*F_\alpha(K,(1+\kappa)n^*-K-1)}{(1+1/\kappa)^{-1}((1+\kappa)n^*-K-1)}\right) > \gamma, \tag{12}$$

where $\boldsymbol{\Lambda}^\dagger = \boldsymbol{\Lambda}_1 + \kappa\boldsymbol{\Lambda}_2$, *and* $\{d_k, \boldsymbol{u}_k\}_{k=1}^K$ *are the eigenvalue and the eigenvectors of the spectral decomposition* $\boldsymbol{\Omega}^\dagger = \sum_{k=1}^K d_k \boldsymbol{u}_k\boldsymbol{u}_k^\top$, *with* $\boldsymbol{\Omega}^\dagger := \kappa(\kappa - 1/n^*)\boldsymbol{\Omega} + (1-1/n^*)(\mathbf{I}_K - \boldsymbol{\Omega})$, *with* $\boldsymbol{\Omega} := \boldsymbol{\Lambda}^{\dagger\,-1/2}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}^{\dagger\,-1/2}$. *Finally,* $\nu(n^*)$ *is the degrees of freedom of the chi-squared distribution under the denominator which is of the form*

$$\nu(n^*) = n^*\left\{\mathrm{tr}(\boldsymbol{\Omega}^{\dagger 2}) + \mathrm{tr}^2(\boldsymbol{\Omega}^\dagger)\right\}\left[\kappa^2(\kappa - 1/n^*)\{\mathrm{tr}(\boldsymbol{\Omega}^2) + \mathrm{tr}^2(\boldsymbol{\Omega})\} + (1-1/n^*)\{\mathrm{tr}(\mathbf{I}_K - \boldsymbol{\Omega})^2 + \mathrm{tr}^2(\mathbf{I}_K - \boldsymbol{\Omega})\}\right]^{-1}.$$

Note that the unknown quantity n^* is at the both side of the above equation. One can find the minimum value of n^* by iterative calculating the left hand side of the above equation over all n^* starting with an initial value of 3 and stop when the left hand side is greater than γ. Alternative, one can use the standard root-solver algorithm in any statistical software (e.g. `uniroot()` function in R) to find the value of n^* so that the left hand side probability is equal to γ.

As discussed earlier, the true eigenfunctions as well as the covariance of the 'shrinkage' scores, $\boldsymbol{\Lambda}$, are unknown in practice. For a specified covariance structure of the data, we must have a consistent estimate of the eigenfunctions and the $\boldsymbol{\Lambda}$ in order to provide a reliable of estimate of the minimum sample size using the formula (12). We will replicate the data-driven approach described in Section 3.1 to obtain a highly reliable estimate of the unknown eigenfunction and the covariance parameter $\boldsymbol{\Lambda}$.

compute the sample size. We reuse the Step 1 - 4 of Algorithm 3.1 to provide a computationally efficient sample size calculation algorithm for our projection-based test below.

Input: Data information: True difference in mean function between two groups, i.e. $\eta(t) := \mu_1(t) - \mu_2(t)$, covariance of the latent process $\Sigma(t, t')$, measurement error variance τ^2, significance level $\alpha \in (0, 1)$, target power $\gamma \in (0, 1)$, ratio of the sample size between two groups, $\kappa = n_1/n_2 > 0$, and a pre-specified PVE to estimate the eigencomponents of the covariance;

1 Generate a dataset with large number of subjects, say 5000, using the model (1) by setting $\mu_1(t) = 0$ and $\mu_2(t) = \eta(t)$ and covariance of the $X_g(t)$ as $\Sigma(t, t')$ and the measurement error variance $\tau^2 > 0$. ;

2 Preset a large grid of size R (say 100) and generate a sequence of points $t_1 < t_2 < \ldots < t_R$ in $\mathcal{T}$;

3 Conduct fPCA on the generated dataset to obtain a highly reliable estimate of the eigenfunctions $\{\widetilde{\psi}_k(t) : k = 1, \ldots, K\}$ at $t = t_1, \ldots, t_R$. The number of eigenfunctions K are chosen by the specified PVE ;

4 Calculate the fPC scores for each subject and obtain the sample covariance $\widetilde{\mathbf{\Lambda}}$ of the scores. This will serve as a reliable estimated of the true covariance $\mathbf{\Lambda}$;

5 **for** $k \in \{1, \ldots, K\}$ **do**

6 Calculate the projection $\widetilde{\delta}_k = \int \eta(t)\widetilde{\psi}_k(t)dt \approx R^{-1} \sum_{r=1}^{R} \eta(t_r)\widetilde{\psi}_k(t_r)$

7 **end**

8 Construct the vector $\widetilde{\mathbf{\Delta}} = (\widetilde{\delta}_1, \ldots, \widetilde{\delta}_K)^\top$;

9 Replace $\mathbf{\Delta}$ and $\mathbf{\Lambda}$ with $\widetilde{\mathbf{\Delta}}$ and $\widetilde{\mathbf{\Lambda}}$ respectively in (12) to calculate the minimum value of n^* such that equation holds ;

10 Return the minimum sample size required for the two groups as $(\kappa n^*, n^*)$.

Algorithm 4.1: Sample size calculation algorithm

Algorithm 4.1 requires the true group difference, covariance function of the process, the design of the observation points and the allocation ratio of samples in the two groups to compute the minimum sample size required for the test to achieve a power of $100(1-\gamma)\%$. The most time consuming part of the algorithm are the first four steps, which generates a large data and conduct an fPCA to compute the eigenfunctions and the covariance of 'shrinkage' scores. This is a one time task, and it is unavoidable as we must obtain a consistent estimate of the true eigenfunctions and $\mathbf{\Lambda}$ to be able to provide an adequate sample size. Once these estimates are obtained, the rest of the steps of the algorithm takes no time to provide the minimum sample size.

Computational software

A user-friendly R package fPASS implementing the Algorithm 3.1 and 4.1 is hosted in Github at `https://github.com/SalilKoner/fPASS`. The R package is capable of taking any form of covariance structure of the data, allows for different schedule of observations for each subjects, along with option of specification of missing percentages, and can take any form of the mean difference between two groups. The overall functionality of the package resembles to that of widely used PASS software. The software efficiently computes the power function and sample size of the test for any specific experimental design.

5 Numerical studies

5.1 Data generation

In this section, we empirically validate the power function of the test as a function of sample size and also validate the correctness of the sample size calculation of the test as an inverse problem of the power function validation. To exercise our task, we perform several numerical

experiments to demonstrate that the power function formula is indeed valid. Specifically, we calculate the power and sample size of our projection-based test under a factorial design. For the first factor, we take the difference of the mean function between the two groups as $\mu_1(t) - \mu_2(t) = \eta t^3$, where we take different values of η greater than zero, to ensure departure from the null hypothesis. Secondly, we consider three different (common) covariance structure of the stochastic process $X_g(t)$, with first two being stationary and the last one being non-stationary. The covariance structures are:

Case 1: *Compound symmetry (CS).* $\Sigma(t,t') = \sigma^2\{\rho + (1-\rho)\mathbb{I}(t=t')\}$,

Case 2: *Autoregressive of order 1, AR(1).* $\Sigma(t,t') = \sigma^2 0.5^{|t-t'|}$ with $\sigma^2 = 1$, and

Case 3: *Spectral.* $\Sigma(t,t')$ is equal to the covariance of error process $\xi_{i1}\sqrt{2}\sin(2\pi t) + \xi_{i2}\sqrt{2}\cos(2\pi t)$ where $\xi_{i1} \overset{\text{iid}}{\sim} N(0,1)$ and $\xi_{i2} \overset{\text{iid}}{\sim} N(0,0.5)$.

This induces a non-stationary covariance structure of the data. Thirdly, we consider two different random sampling designs, for the stationary covariance structures (Case 1. and 2.), we take $m_i = 5$ (*high* sparsity) or 8 (*medium* sparsity) sampling points randomly taken between $[0,1]$, whereas, for non-stationary covariance (Case 3.), we take m_i randomly sampled either between $\{4,5,6,7\}$ (*high* sparsity) or between $\{8,9,\ldots,12\}$ (*medium* sparsity). We consider the measurement error variance $\sigma_e^2 = 0.001$.

5.2 Power function

We apply Algorithm 3.1 for each of the above specified setting to calculate the power function of the test for different values of total sample size, n and η. The asymptotic power function of the test, $\mathcal{P}_n(\eta)$, is presented in Table 1-3 for $n = 200, 400, 800$ and 1600 and $\eta = 0.5, 0.75$ and 1, across different sparsity levels of the design, and covariance structure of the data.

We assume equal allocation of samples in each of the two groups. As we can see that the power function of the test increases to one as the number of the samples increases and/or the difference between the groups increases. Also, it is important to observe that for fixed η and n, the power function of the test is different across the difference covariance structure of the data. However, it seems that the sparsity level of the do not have a significant influence on the power function of the test. This is primarily driven by the fact that as long as the eigenfunctions and the population covariance of the scores scores are estimated with high-precision, the sparsity level do not have significant impact on the power function. However, for small sample size case, having a more dense design will lead to better estimation of the eigenfunction, leading to accurate calculation of the asymptotic power function.

We validate the theoretically calculated asymptotic power function of the test empirically by Monte-Carlo simulation. Specifically, for a fixed sample size n, mean difference specified by η, the covariance structure and the random sampling design, we generate $B = 1000$ replications and then calculate the percentage of times the test rejects the null hypothesis using the test rule (7). This number is denoted as $\widehat{\mathcal{P}}_n(\eta)$, and presented alongside the theoretically calculated power function $\mathcal{P}_n(\eta)$, in each of Table 1-3. For small sample size situations, while empirically validating the power of the test, the eigenfunctions may not be consistently estimated, hence the empirically calculated power function of the test for small sample size cases are unreliable. However, as the sample size n increases, the theoretically calculated power function of the test matches with the empirically calculated power. This is justified by the fact that for large sample size cases, the estimated eigenfunctions for each of the $B = 1000$ datasets are close to the true eigenfunctions of the covariance, and therefore the numbers match, and the empirically validated power functions are reliable.

Table 1. Power function across different sample sizes.

Covariance : *Compound symmetric*

# observations per subject: *low*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.28	0.30	0.54	0.54	0.78	0.74
400	0.49	0.46	0.83	0.75	0.97	0.92
800	0.79	0.7	0.98	0.93	0.99	0.99
1600	0.97	0.92	1	1	1	1
# observations per subject: *medium*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.27	0.36	0.51	0.60	0.76	0.77
400	0.47	0.52	0.81	0.80	0.96	0.94
800	0.76	0.71	0.98	0.95	0.99	0.99
1600	0.97	1	1	1	1	1

Table 2. Power function across different sample sizes.

Covariance : *Auto-regressive of order 1*

# observations per subject: *low*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.14	0.4	0.26	0.56	0.42	0.64
400	0.24	0.55	0.47	0.68	0.70	0.81
800	0.42	0.59	0.75	0.84	0.94	0.96
1600	0.71	0.76	0.96	0.96	1	1
# observations per subject: *medium*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.14	0.55	0.26	0.62	0.42	0.69
400	0.24	0.52	0.46	0.71	0.70	0.82
800	0.42	0.57	0.75	0.85	0.94	0.95
1600	0.71	0.74	0.96	0.97	1	1

5.3 Missing-immunity of the power

In the previous subsection, via empirical validation, we have established that the algorithm for the fast calculation of the asymptotic power function of the test is correct. The empirical validation is only justified when n is large, because for small n, the eigenfunctions are not estimated consistently when one conducts the test empirically for each of $B = 1000$ replications. Noting that the empirically calculated power matches with the asymptotic power when n is large, we present the empirically

Table 3. Power function across different sample sizes.

Covariance : *Non-stationary*

# observations per subject: *low*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.11	0.36	0.20	0.44	0.33	0.53
400	0.18	0.43	0.37	0.54	0.60	0.68
800	0.33	0.43	0.65	0.72	0.89	0.91
1600	0.60	0.62	0.92	0.92	0.99	0.99
# observations per subject: *medium*						
n	$\eta = 0.5$		$\eta = 0.75$		$\eta = 1$	
	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_n(\delta)$	$B_n(\delta)$	$\widehat{B}_{n^*}(\delta)$
200	0.11	0.36	0.20	0.40	0.33	0.51
400	0.18	0.31	0.37	0.44	0.59	0.66
800	0.33	0.34	0.65	0.62	0.89	0.89
1600	0.59	0.57	0.92	0.92	1	1

Table 4. Power function across different missing percentages.

Covariance : *Compound symmetric*

# observations per subject : *low*								
Sample	$\delta = 0.2$				$\delta = 0.4$			
size	0%	10%	20%	40%	0%	10%	20%	40%
500	0.15	0.16	0.17	0.17	0.46	0.47	0.49	0.54
800	0.21	0.22	0.22	0.24	0.63	0.64	0.65	0.73
1200	0.28	0.29	0.30	0.32	0.79	0.80	0.82	0.84
# observations per subject : *medium*								
Sample	$\delta = 0.2$				$\delta = 0.4$			
size	0%	10%	20%	40%	0%	10%	20%	40%
500	0.15	0.15	0.15	0.16	0.43	0.44	0.46	0.50
800	0.19	0.20	0.20	0.21	0.59	0.60	0.62	0.65
1200	0.26	0.27	0.28	0.29	0.76	0.76	0.77	0.81

calculated power of the test for large values of n and difference percentage of missing observations for each subjects in Table 4-6. As we can see, although the number of missing observations for each subjects increases from 0% to as high as 40%, the empirical power of the test remains the same. This is due to the fact that even if the number of missing observations for each subject increases, the eigencomponents are still consistently estimated due to large sample size, hence the empirical power of the test is unaffected, validating the 'missing-immunity' of our test in numerical studies as well.

Table 5. Power function across different missing percentages.

Covariance : *Auto-regressive of order* 1								
# observations per subject : *low*								
Sample size	$\delta = 0.2$				$\delta = 0.4$			
	0%	10%	20%	40%	0%	10%	20%	40%
500	0.09	0.09	0.09	0.09	0.22	0.21	0.21	0.21
800	0.11	0.11	0.11	0.11	0.31	0.30	0.31	0.30
1200	0.14	0.14	0.14	0.14	0.42	0.42	0.42	0.41
# observations per subject : *medium*								
Sample size	$\delta = 0.2$				$\delta = 0.4$			
	0%	10%	20%	40%	0%	10%	20%	40%
500	0.09	0.09	0.09	0.09	0.21	0.22	0.22	0.21
800	0.11	0.11	0.11	0.11	0.30	0.30	0.31	0.30
1200	0.14	0.14	0.14	0.14	0.41	0.42	0.42	0.42

Table 6. Power function across different missing percentages.

Covariance : *Non-stationary*								
# observations per subject : *low*								
Sample size	$\delta = 0.5$				$\delta = 1$			
	0%	10%	20%	40%	0%	10%	20%	40%
500	0.25	0.25	0.25	0.25	0.80	0.77	0.80	0.77
800	0.35	0.35	0.35	0.35	0.91	0.91	0.92	0.92
1200	0.49	0.49	0.49	0.49	0.98	0.98	0.98	0.98
# observations per subject : *medium*								
Sample size	$\delta = 0.5$				$\delta = 1$			
	0%	10%	20%	40%	0%	10%	20%	40%
500	0.22	0.22	0.22	0.22	0.71	0.72	0.71	0.71
800	0.33	0.33	0.33	0.33	0.90	0.90	0.90	0.90
1200	0.60	0.60	0.60	0.60	1.00	1.00	1.00	1.00

5.4 Sample size validation

Table 7-9 tabulates the minimum sample size required for the test to achieve a power of $100(1-\gamma)\%$ when the group difference, parametrized by η increases from 0.5 to 2, across all the three covariance structures, assuming an equal allocation $(\kappa = 1)$ of samples in each group. The total sample size combining the two groups are obtained by implementing Algorithm 4.1, and reported in the column with heading $n_{\min}$ in the three tables. To ensure the correctness of the calculation of minimum sample size, we generate 500 synthetic data of size equal to the minimum sample size computed by the algorithm with the difference between the mean of the two groups specified by δ and the underlying covariance structure, and

Table 7. Sample size justification across different target power.

Covariance : *Compound symmetric*

# observations per subject: *low*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	642	0.64	825	0.73	1099	0.81
1	164	0.68	208	0.76	273	0.85
1.5	74	0.64	94	0.76	124	0.84
2	44	0.66	54	0.75	72	0.85
# observations per subject: *medium*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	690	0.67	876	0.77	1179	0.86
1	174	0.75	222	0.83	296	0.87
1.5	80	0.75	100	0.81	134	0.89
2	46	0.73	58	0.82	76	0.9

Table 8. Sample size justification across different target power.

Covariance : *Auto-regressive of order 1*

# observations per subject: *low*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	1577	0.79	2022	0.83	2679	0.91
1	396	0.81	504	0.87	678	0.93
1.5	178	0.84	228	0.87	302	0.94
2	102	0.81	128	0.87	170	0.93
# observations per subject: *medium*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	1596	0.75	2021	0.80	2714	0.92
1	400	0.82	510	0.88	677	0.92
1.5	180	0.87	228	0.91	304	0.94
2	102	0.85	128	0.90	172	0.96

conduct our projection-based test for each of simulated dataset to compute the empirical power, as tabulated in the column 4. As the empirical power is close to the target power with the sample size increases, we can strongly advocate for the correctness of the sample size calculation algorithm of our test.

Table 9. Sample size justification across different target power.

Covariance : *Non-stationary*						
# observations per subject: *low*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	1993	0.70	2454	0.81	3254	0.88
1	496	0.76	619	0.85	812	0.91
1.5	222	0.77	280	0.86	363	0.90
2	127	0.78	158	0.81	206	0.90
# observations per subject: *medium*						
δ	$\gamma = 0.7$		$\gamma = 0.8$		$\gamma = 0.9$	
	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$	n^*	$\widehat{B}_{n^*}(\delta)$
0.5	2002	0.70	2509	0.80	3274	0.90
1	498	0.71	626	0.79	824	0.90
1.5	224	0.76	283	0.86	369	0.91
2	128	0.79	160	0.87	208	0.92

6 Application on real data

6.1 Azillect study

Azillect clinical trial is the first large randomized clinical trials in Japan that uses the changes in the total of Part II and Part II MDS-UPDRS scores (which is an updated and improved scale than UPDRS) as the primary endpoints for patients with early PD. In this phase-3 trial, the participants with an age ranging from 30-79 years with a diagnosis of PD within five years are randomized 1:1 to receive the study drug rasagaline or placebo up to 26 weeks. A complete description about the study design is presented in Hattori. We will demonstrate how the findings obtained from the Azillect study can be used to design a new trial.

At the end of the study (week 26), the mean change in the part II + III total score from the baseline was 1.87 for the placebo and -4.52 for the rasagaline group. Figure 1 presents the typical change in the total of MDS-UPDRS part II and III scores from the baseline for the rasagaline and the placebo group, based on fitting model (1) with `gam()` function in the `mgcv` package in R. Figure 2 presents the typical change in the in the total of MDS-UPDRS part II and III scores from the baseline

between the rasagaline and the placebo group. The solid line. (in red) in the figure 2 is the estimate of the treatment effect, i.e. a smooth estimate of $\widehat{\mu}_1(t) - \widehat{\mu}_2(t)$, along with the 95% pointwise confidence interval, shaded in blue.

The projection-based test also renders modelling the covariance structure of the data in smooth non-parametric manner. Figure 3 presents the two leading eigenfunctions obtained from conducting fPCA on the smoothed estimated covariance. The eigenvalues corresponding to the two eigenfunctions are 31.8 (red line) and 2.6 (blue line) respectively, meaning that the first eigenfunctions explains 91% of the total variability in the data. The pattern of the eigenfunctions conveys important insights into the nature of variations in the total score; the primary direction (in red) depicts a gradual variation throughout the course of the study, whereas, the change in sign of the secondary direction (in blue) reflects the variations before and after the midway (around week 15 approximately) of the study.

Using the estimated eigenfunctions, we use the estimated subject-specific fPC-scores between the two groups to conduct the Hotelling-T^2 test. The confidence band (does not cover 0) in Figure 2 along with a low p-value ($< 3 \times 10^{-6}$) from the projection-based test justifies that rasagline is able to significantly reducing the MDS-UPDRS part II III total scores over the 26 weeks, comparing to the placebo.

If we want to design a new study with the target treatment effect as plotted in Figure 2 and the covariance specified by the two leading eigenfunctions presented in Figure 3, then we need a minimum sample size of 74 to achieve a power of 80%, and a minimum sample of 96 to achieve a power of 90%, assuming an equal allocation ratio of the subjects in the two groups, which is less than half of the minimum sample size calculated based on t test (about 240) as stated in Section 2.4 of[5]. Therefore, calculating the sample size using the projection-based test enlarges the scope of reducing the minimum sample size required

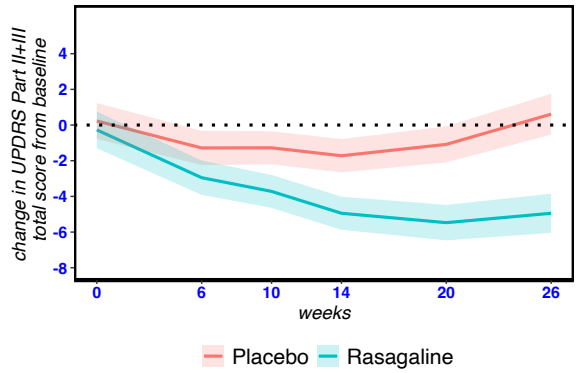

Figure 1. Change from the baseline in MDS-UPDRS part II + III total score for Rasagaline and Placebo group. The shaded region reflects the 95% pointwise confidence interval.

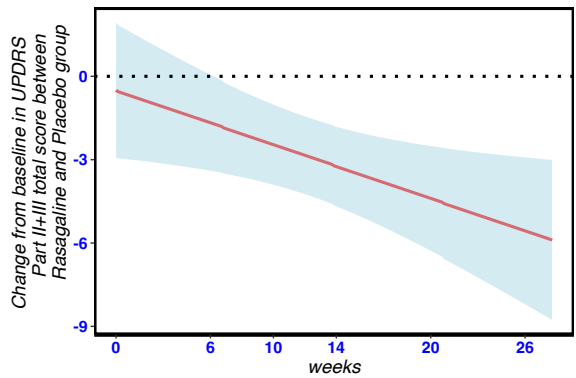

Figure 2. Effect of the Rasagaline on MDS-UPDRS Part II + III score over the course of the study with 95% pointwise confidence interval (shaded in blue).

to achieve the same power compared to traditional methods used in the clinical trials.

6.2 SURE-PD3 study

The primary endpoints for this study is the change from the baseline in the total of Part I, II and III MDS-UPDRS scores.

Figure 4 presents the estimated change in the total UPDRS scores from the baseline for the Inosine (in red) and the placebo group (in blue), along

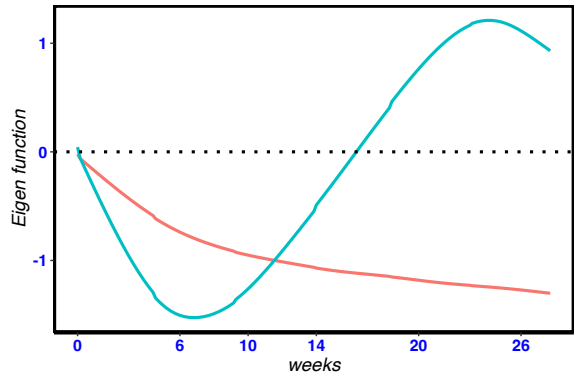

Figure 3. Estimated eigenfunction from the Azillect study, based on a PVE of 95%. The red curve corresponds the primary direction covering 91% of the total variability, and the blue one for the secondary direction of variability.

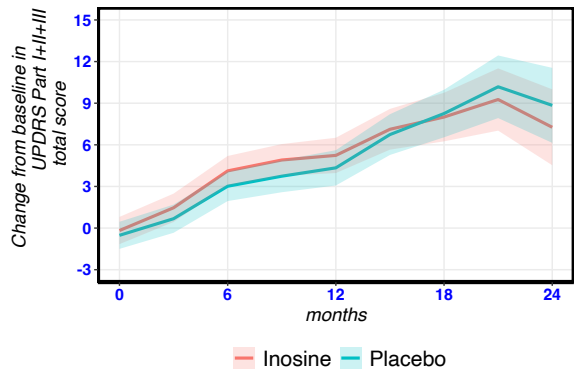

Figure 4. Change from the baseline in MDS-UPDRS part I + II + III total score for Inosine and Placebo group in the SURE-PD3 study. The shaded region reflects the 95% pointwise confidence interval.

with the 95% pointwise confidence band. There does not seem to be a significant difference in the change of total scores from baseline between the two groups.

Figure 6 presents the two leading eigenfunctions obtained from conducting fPCA on the smoothed estimated covariance. The eigenvalues corresponding to the two eigenfunctions are 48.5 (red line) and 4.7 (blue line) respectively, meaning that the first eigenfunctions explains 91% of

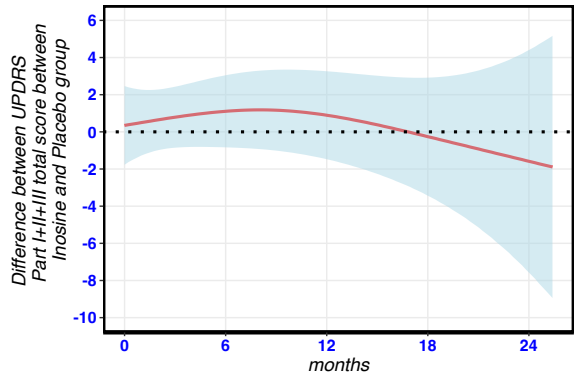

Figure 5. Effect of the Inosine on MDS-UPDRS Part I + II + III score over the course of the study with 95% pointwise confidence interval (shaded in blue).

the total variability in the data. As we have in the case of Azillect study, the pattern of the eigenfunctions conveys similar insights into the nature of variations in the total score; the primary direction (in red) depicts a gradual variation throughout the course of the 24 months of the study, whereas, the change in sign of the secondary direction (in blue) reflects the variations before and after the midway (around month 13 approximately) of the study.

After extracting the eigencomponents by FPCA, using subject-specific fPC-scores between the two groups we conduct the Hotelling-T^2 test. The p-value of the Hotelling T^2 test is 0.42 suggesting that Inosine is not able to improve the total MDS-UPDRS scores significantly compared to the placebo, over the course of 24 months. This conclusion is also corroborated by the factor 95% pointwise confidence interval of the mean difference obtained from the `gam()` contains zero (Figure 5), and the length of the band increases at the later months, because of considerable amount of early-withdrawals.

If we want to design a new study whether the effect of the study drug is the of the form of the red line in Figure 5, i.e. initially it goes above zero and then gradually going away from zero after around 12 months,

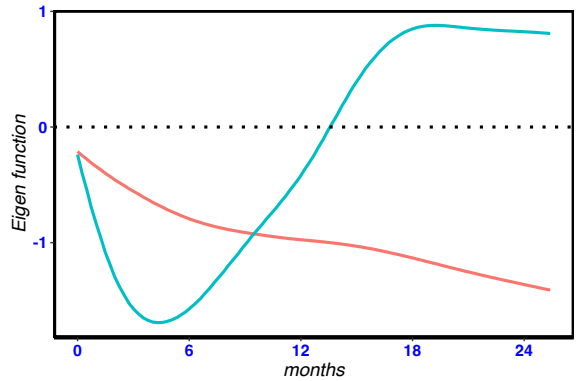

Figure 6. Estimated eigenfunction from the SURE-PD3 study, based on a PVE of 95%. The red curve corresponds the primary direction covering 91% of the total variability, and the blue one for the secondary direction of variability.

then assuming the covariance structure of the data as characterized by the two principal eigenfunctions (Figure 6 with the specified eigenvalues, we need about approximately 145 subjects in each group in order to achieve a power of 80%, whereas about 190 subjects in each groups is required to achieve a power of 90%.

Funding

To typeset a "Funding" section.

Acknowledgements

Sheng Luo would like to acknowledge the support of his research grant.

Supplemental material

To typeset a "Supplemental material" section.

7 Alternate distribution of the Hotelling T^2 statistic

We want to find the alternate distribution of the Hotelling T^2 random variable

$$T_n = \frac{n_1 n_2}{n_1 + n_2}(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+})^\top \widetilde{\boldsymbol{\Lambda}}^{-1}(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+}),$$

as defined in (6), when the true mean difference $\eta(t) = \mu_1(t) - \mu_2(t)$ is different from zero. From the formula of the mean and the covariance of the *shrinkage* scores (equation 2.11 of [13]), it is clear that the population covariance of the *shrinkage* scores are different when the true mean function of the two groups are different. On the other hand, under the null, the covariance matrices of the scores of the two groups are the same. Therefore, under the alternative we have to find the distribution of the Hotelling T^2 statistic under the assumption that the covariances of the scores between the two groups are different. Therefore, the alternate distribution of the test-statistic no more follows a non-central F distribution. Here, we will establish the alternate distribution of T_n assuming that the covariances of the scores between the two groups, Λ_1 and Λ_2 are different. Under the assumption of normality of the *shrinkage* scores,

$$\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+} \sim \mathrm{N}_K(\boldsymbol{\Delta}, \boldsymbol{\Lambda}_1/n_1 + \boldsymbol{\Lambda}_2/n_2),$$

and

$$(n_g - 1)\widetilde{\boldsymbol{\Lambda}}_g \sim \mathbf{W}_K(n_g - 1, \boldsymbol{\Lambda}_g) \qquad g = 1, 2$$

To scale things properly as a function of the sample size, we want to represent the distribution of T_n in terms of the allocation ratio of the sample size between the two groups. Let $\kappa = n_1/n_2$ be the allocation ratio, then define, $\boldsymbol{\Lambda}^\dagger = \boldsymbol{\Lambda}_1 + \kappa\boldsymbol{\Lambda}_2$, so that $\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+} \sim \mathrm{N}_K(\boldsymbol{\Delta}, n_1^{-1}\boldsymbol{\Lambda}^\dagger)$. We can represent $(1/n_1 + 1/n_2)T_n$ equivalently as

$$\begin{aligned}(1/n_1 + 1/n_2)T_n = n_1(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+})^\top \boldsymbol{\Lambda}^{\dagger\,-1/2} \\ (n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}^{\dagger\,-1/2})^{-1}\boldsymbol{\Lambda}^{\dagger\,-1/2}(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+}).\end{aligned} \tag{13}$$

Define $\mathcal{Z} = \sqrt{n_1}\boldsymbol{\Lambda}^{\dagger\,-1/2}(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+})$ and $\mathcal{S} := n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}^{\dagger\,-1/2}$, so that $(1/n_1 + 1/n_2)T_n = \mathcal{Z}^\top\mathcal{S}^{-1}\mathcal{Z}$. Now, we derive the asymptotic distribution of $\mathcal{S}$. Note that,

$$\begin{aligned}(n_1 + n_2 - 2)\mathcal{S} = n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\{(n_1 - 1)\widetilde{\boldsymbol{\Lambda}}_1\}\boldsymbol{\Lambda}^{\dagger\,-1/2} \\ + n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\{(n_2 - 1)\widetilde{\boldsymbol{\Lambda}}_2\}\boldsymbol{\Lambda}^{\dagger\,-1/2}\end{aligned}$$

By the property of Wishart distribution,

$$\begin{aligned}n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\{(n_1 - 1)\widetilde{\boldsymbol{\Lambda}}_1\}\boldsymbol{\Lambda}^{\dagger\,-1/2} &\sim \mathbf{W}_K(n_1 - 1, n_1\boldsymbol{\Omega}) \\ n_1\boldsymbol{\Lambda}^{\dagger\,-1/2}\{(n_2 - 1)\widetilde{\boldsymbol{\Lambda}}_2\}\boldsymbol{\Lambda}^{\dagger\,-1/2} &\sim \mathbf{W}_K(n_2 - 1, n_2(\mathbf{I}_K - \boldsymbol{\Omega})),\end{aligned}$$

where $\mathbf{\Omega} = {\mathbf{\Lambda}^{\dagger}}^{-1/2}\mathbf{\Lambda}_1{\mathbf{\Lambda}^{\dagger}}^{-1/2}$, and using the fact that $\mathbf{\Lambda}_2 = \kappa^{-1}(\mathbf{\Lambda}^{\dagger} - \mathbf{\Lambda}_1)$. By the result of sum of two Wishart distributions proved by[?], $\mathcal{S}$ has an approximate Wishart distribution as follows,

$$(n_1 + n_2 - 2)\mathcal{S} \overset{\text{a.d}}{=} \mathbf{W}_K(\nu, \nu^{-1}\mathbf{\Omega}^*), \tag{14}$$

where

$$\begin{aligned}\mathbf{\Omega}^* &= n_1(n_1-1)\mathbf{\Omega} + n_2(n_2-1)(\mathbf{I}_K - \mathbf{\Omega}).\\ &= n_2^2\left\{\kappa(\kappa - 1/n_2)\mathbf{\Omega} + (1-1/n_2)(\mathbf{I}_K - \mathbf{\Omega})\right\}\\ &= n_2^2\mathbf{\Omega}^{\dagger}\end{aligned}$$

and ν is the approximate degree of freedom of the Wishart distribution specified by

$$\begin{aligned}\nu = n_2\left\{\text{tr}(\mathbf{\Omega}^{\dagger 2}) + \text{tr}^2(\mathbf{\Omega}^{\dagger})\right\}\big[\kappa^2(\kappa - n_2^{-1})\{\text{tr}(\mathbf{\Omega}^2) + \text{tr}^2(\mathbf{\Omega})\}\\ + (1-n_2^{-1})\{\text{tr}(\mathbf{I}_K - \mathbf{\Omega})^2 + \text{tr}^2(\mathbf{I}_K - \mathbf{\Omega})\}\big]^{-1}.\end{aligned}$$

Since $\widetilde{\mathbf{\Lambda}}$ is independently distributed to $\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+}$, combining Equation (14) into the representation of the test-statistic in (13), and applying by Theorem 3.2.12 of[?],

$$\frac{\nu n_2^{-2}\mathcal{Z}^{\top}{\mathbf{\Omega}^{\dagger}}^{-1}\mathcal{Z}}{\mathcal{Z}^{\top}\{(n_1+n_2-2)\mathcal{S}\}^{-1}\mathcal{Z}} \sim \chi^2_{\nu-K+1}$$

This implies

$$\frac{\nu(n_1+n_2-2)\mathcal{Z}^{\top}{\mathbf{\Omega}^{\dagger}}^{-1}\mathcal{Z}}{n_2^2(1/n_1+1/n_2)T_n} \sim \chi^2_{\nu-K+1} \tag{15}$$

By the property of multivariate normal distribution,

$$\mathcal{Z} = \sqrt{n_1}{\mathbf{\Lambda}^{\dagger}}^{-1/2}(\widetilde{\boldsymbol{\zeta}}_{1+} - \widetilde{\boldsymbol{\zeta}}_{2+}) \sim \text{N}_K(\sqrt{n_1}{\mathbf{\Lambda}^{\dagger}}^{-1/2}\mathbf{\Delta}, \mathbf{I}_K).$$

Suppose, that the full rank matrix $\mathbf{\Omega}^{\dagger}$ admits a spectral decomposition $\mathbf{\Omega}^{\dagger} = \sum_{k=1}^{K} d_k\boldsymbol{u}_k\boldsymbol{u}_k^{\top}$, with $\boldsymbol{u}_k^{\top}\boldsymbol{u}_j = \mathbb{I}(k=j)$ by Theorem 1 of[?],

$$\mathcal{Z}^{\top}{\mathbf{\Omega}^{\dagger}}^{-1}\mathcal{Z} \sim \sum_{k=1}^{K} d_k^{-1}\chi_1^2\left(n_1(\boldsymbol{u}_k^{\top}{\mathbf{\Lambda}^{\dagger}}^{-1/2}\mathbf{\Delta})^2\right) \tag{16}$$

Dividing the left hand side of (16) with that of (15) we can see that our test statistic is approximately distributed as

$$T_n \stackrel{d}{=} \left\{ \frac{\sum_{k=1}^{K} d_k^{-1} \chi_1^2 \left(n_1 (\boldsymbol{u}_k^\top \boldsymbol{\Lambda}^{\dagger\, -1/2} \boldsymbol{\Delta})^2 \right)}{\chi^2_{\nu-K+1}/\nu} \right\} \times (\kappa + 1 - 2/n_2)(1/\kappa + 1)^{-1}$$

8 Appendix

References

1. Brown H and Prescott R (2015) *Applied mixed models in medicine*. John Wiley & Sons.
2. Crainiceanu C, Ruppert D, Claeskens G and Wand MP (2005) Exact likelihood ratio tests for penalised splines. *Biometrika* 92(1): 91–103.
3. Crainiceanu CM and Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B* 66(1): 165–185.
4. Diggle P, Diggle PJ, Heagerty P, Liang KY, Zeger S et al. (2002) *Analysis of longitudinal data*. Oxford university press.
5. Hattori N, Takeda A, Takeda S, Nishimura A, Kitagawa T, Mochizuki H, Nagai M and Takahashi R (2019) Rasagiline monotherapy in early parkinson's disease: A phase 3, randomized study in japan. *Parkinsonism & Related Disorders* 60: 146–152.
6. Hotelling H (1992) The generalization of student's ratio. In: *Breakthroughs in statistics*. Springer, pp. 54–65.
7. Johnson RA and Wichern DW (2002) *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.
8. Koner S and Luo S (2023) Projection-based two-sample inference for sparsely observed multivariate functional data.
9. PASS (2023) Power Analysis and Sample Size Software. URL ncss.com/software/pass.
10. Pomann GM, Staicu AM and Ghosh S (2016) A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65(3): 395–414.
11. SAS Institute Inc (2015) SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

12. Staicu A, Li Y, Crainiceanu CM and Ruppert D (2014) Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics* .
13. Wang Q (2021) Two-sample inference for sparse functional data. *Electronic Journal of Statistics* 15(1): 1395–1423.
14. Wiencierz A, Greven S, Küchenhoff H et al. (2011) Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic Journal of Statistics* 5: 1718–1734.
15. Xiao L, Li C, Checkley W and Crainiceanu C (2022) *face: Fast Covariance Estimation for Sparse Functional Data*. URL https://CRAN.R-project.org/package=face. R package version 0.1-7.
16. Yao F, Müller HG and Wang JL (2005) Functional data analysis for sparse longitudinal data. *Journal of the American statistical association* 100(470): 577–590.
17. Zhang X and Wang JL (2016) From sparse to dense functional data and beyond .