A Modified Neighborhood Hypothesis Test for Population Mean in Functional Data

Dhanamalee BANDARA, Leif ELLINGSON, Souparno GHOSH, and Ranadip PAL

When dealing with very high-dimensional and functional data, rank deficiency of sample covariance matrix often complicates the tests for population mean. To alleviate this rank deficiency problem, Munk et al. (J Multivar Anal 99:815–833, 2008) proposed neighborhood hypothesis testing procedure that tests whether the population mean is within a small, pre-specified neighborhood of a known quantity, M. How could we objectively specify a reasonable neighborhood, particularly when the sample space is unbounded? What should be the size of the neighborhood? In this article, we develop the modified neighborhood hypothesis testing framework to answer these two questions. We define the neighborhood as a proportion of the total amount of variation present in the population of functions under study and proceed to derive the asymptotic null distribution of the appropriate test statistic. Power analyses suggest that our approach is appropriate when sample space is unbounded and is robust against error structures with nonzero mean. We then apply this framework to assess whether the near-default sigmoidal specification of dose-response curves is adequate for widely used CCLE database. Results suggest that our methodology could be used as a pre-processing step before using conventional efficacy metrics, obtained from sigmoid models (for example: IC_{50} or AUC), as downstream predictive targets.

Key Words: Neighborhood hypothesis test; Rank deficiency; Dose-response curves; Cancer cell line encyclopedia.

1. INTRODUCTION

Advancement in instrumentation and growth of computing power over past few decades have brought about a proliferation of very high-dimensional and functional data that

D. Bandara, Department of Mathematics and Statistics, University of Wisconsin-Green Bay, 2420 Nicolet Dr, Green Bay, WI 54115, USA (E-mail: *bandarad@uwgb.edu*).
L. Ellingson, Department of Mathematics and Statistics, Texas Tech University, 2500 Broadway, Lubbock, TX 79409, USA (E-mail: *leif.ellingson@ttu.edu*).
S. Ghosh, Department of Statistics, University of Nebraska-Lincoln, 1400 R Street, Lincoln, NE 68588, USA (E-mail: *sghosh5@unl.edu*).
R. Pal, Department of Electrical and Computer Engineering, Texas Tech University, 2500 Broadway, Lubbock, TX 79409, USA (E-mail: *ranadip.pal@ttu.edu*).

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-023-00549-y

Published online: 04 June 2023

needs to be analyzed. It has been well-documented that traditional approaches for working with multivariate, but low-dimensional, data often do not scale to these very high-dimensional/functional situations (see, for example, Ramsay and Silverman 2005; Wainwright Wainwright 2019 and references therein). This is true even for the classical problem of testing hypotheses about a population mean of the form $H_0 : \mu = M$ because traditional procedures for doing so face mathematical complications due to the rank deficiency of the sample covariance matrix when either the dimension of the data is larger than the sample size in the multivariate setting or, in general, for functional data. Even when Hotelling T^2-style statistics are altered to accommodate this deficiency, the asymptotic results are often only applicable with prohibitively large sample sizes (See Xu 2014; Kuelbs and Vidyashankar 2010). To alleviate these problems, Munk et al. (2008) proposed an alternative procedure, they referred to as a neighborhood hypothesis test, that avoids the rank deficiency problem by first redefining the null hypothesis to be that μ is within a small, pre-specified neighborhood of M. The specification of neighborhood was motivated by the fact that, for high-dimensional and functional data, it becomes impractical to expect for a mean to exactly follow a prescribed form.

This type of imprecise specification of null hypothesis could be traced back to the works of Hodges and Lehmann (1954) who described how some categories of frequentist statistical tests could be modified to account for hypotheses in which the null parameter space is not a single point. Berger and Delampady (1987) formalized the relationships between tests of null hypotheses of equality and what we now refer to as neighborhood hypotheses within the scope of Bayesian inference. About a decade later, Dette and Munk (1998), Dette (2003) utilized this idea as a model validation technique in nonparametric regression settings. Finally, Munk et al. (2008) formally developed neighborhood hypothesis tests for functional data for applications in projective shape analysis. Ellingson et al. (2013) then further generalized this neighborhood hypothesis test framework for testing hypothesis on means of random objects lying on Hilbert manifolds. We note that the principle focus of all research following from Dette and Munk (1998), Dette (2003) was to bolster the theoretical foundation of neighborhood hypothesis tests. None of these works were aimed at illustrating practical performance of this inferential framework—which turns out to be important because, in its original form, this methodology required precise specification of neighborhood *a priori*—which restricted its application to some unique settings (for example, similarity shape analysis considered in Ellingson et al. (2013)) where the parameter space for the mean is compact and, as such, the size of the neighborhood could be naturally interpreted with respect to the maximum possible distance between points in the space. In general, however, the disadvantage associated with precise specification of neighborhood becomes pronounced when the sample space is unbounded, as is often the case for both high-dimensional and functional data. In these situations objective specification of the neighborhood is typically not feasible and often not interpretable from a scientific perspective.

Our goal here is to develop a theoretically sound technique that could bring more objectivity in the neighborhood specification thereby making this testing procedure applicable to a wide variety of fields. To that end, we propose a novel modified neighborhood test that defines a neighborhood as a proportion of the total amount of variation present in the population of functions under study. In essence, our approach can also be viewed as shifting

the hypothesis to be about an effect size rather than purely about the mean itself- thereby increasing the interpretability of the conclusion. In addition to proposing the modified version of the neighborhood test, we study the properties of our posited methodology using both asymptotic analyses and simulations. We then implement our method to assess the adequacy of a model assumption commonly imposed on dose-response data obtained from the Cancer Cell Line Encyclopedia (CCLE) database.

Broadly speaking, CCLE provides experimentally observed dose-response curves for 24 anti-cancer drugs administered on 479 representative cancer cell-lines (Barretina et al. 2012). This data has been widely used for generating predictive models for drug efficacy (Ma et al. 2021; De Niz et al. 2016; Wan and Pal 2014). Several of these studies had utilized customary drug efficacy metrics (for example: AUC, EC_{50}, IC_{50} etc.) as predictive targets. These metrics were obtained by fitting a parametric sigmoid curve to the observed dose-response data. However, if the foregoing sigmoidal formulation is itself an inadequate model for the dose-response data, utilizing this model output for downstream modeling, without explicit accounting for misspecification error induced by sigmoidal model, can adversely impact the predictive reliability of drug efficacy. We demonstrate that our modified neighborhood hypothesis test offers a screening procedure to assess the adequacy of sigmoidal specification over a population of drug-specific dose-response curves. Our procedure does not require researchers to test the goodness-of-fit for each dose-response curve. Rather, given the drug-specific curves, observed over a set of cell-lines, our procedure conducts a single hypothesis test whether the population mean of the observed set of curves is within an objectively chosen, interpretable, neighborhood of the posited parametric sigmoidal curve. Failure to reject the neighborhood hypothesis would indicate lack of statistical evidence in support of misspecifcation error thereby allowing researchers to utilize the foregoing model-based drug efficacy metrics with greater confidence. Rejection of neighborhood hypothesis, on the other hand, would alert the researchers about existence of potential misspecifcation errors in drug efficacy metrics that must be taken into account for any downstream modeling and inference.

The remainder of the paper is organized as follows. In Sect. 2, we describe the core methodology for the neighborhood hypothesis test of Munk et al. (2008), highlight how it avoids the rank deficiency problem faced by classical procedures and provide commentary about some practical considerations. Then, in Sect. 3, we present our modifications to that methodology and the asymptotic results for the new procedure. In Sect. 4, we offer simulation studies to assess the power of our testing procedure, including cases where the sample size is not considerably larger than the dimension of the data. Section 5 deals with an illustrative application of our method on dose-response data extracted from CCLE database. We conclude with Sect. 6, giving a brief summary and discussion of future work.

2. THE PREVIOUS NEIGHBORHOOD HYPOTHESIS TEST

Let $X_1, \ldots, X_n$ be independent and identically distributed random elements in a Hilbert space $\mathbb{H}$ with population mean $\mu \in \mathbb{H}$ and covariance operator $\Sigma : \mathbb{H} \to \mathbb{H}$ that satisfy the

condition that $E\left(||X||^4\right) < \infty$. Estimators of μ and Σ are, respectively,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \;\; S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \otimes (X_i - \overline{X}) \tag{1}$$

If we wish to test only whether the population mean is approximately equal to a hypothesized quantity rather than exactly equal to it, we can express this mathematically as testing whether μ is within a small neighborhood of M. Munk et al. (2008) stated this formally as:

$$H_0 : \rho^2(\mu, M) \leq \delta^2 \text{ versus } H_1 : \rho^2(\mu, M) > \delta^2, \tag{2}$$

where $\rho^2(x, y) = \|x - y\|^2$ is the squared distance between any $x, y \in \mathbb{H}$ and $\delta > 0$.

Munk et al. (2008) showed that, under the null hypothesis, if $\varphi_M(X) = \rho^2(X, M)$, then

$$\sqrt{n}(\varphi_M(\overline{X}) - \varphi_M(\mu)) \rightarrow_d N(0, \tau^2), \text{ as } n \rightarrow \infty, \tag{3}$$

where

$$\tau^2 = 4\left\langle \mu - M, \Sigma\ (\mu - M)\right\rangle.$$

If we have exact equality of μ and M, then $\tau^2 = 0$, which results in a degenerate distribution with a point mass at 0. In practice τ^2 is an unknown parameter and must be estimated in order to construct a valid test statistic. A consistent estimator of τ^2 is given by

$$\hat{\tau}^2 = 4\left\langle \overline{X} - M, S\ (\overline{X} - M)\right\rangle.$$

As such, this procedure replaces the singularity problem faced in the Hotelling T^2 test whenever S is not of full rank to the single point $\overline{X} = M$, which results in $\hat{\tau}^2$ equaling 0. Fortunately, though, this point is a set of measure 0, so it does not cause problems in practice. Consequently, it can be proved that

$$\frac{\sqrt{n}(\varphi_M(\overline{X}) - \varphi_M(\mu))}{\hat{\tau}} \rightarrow_d N(0, 1), \text{ as } n \rightarrow \infty. \tag{4}$$

This asymptotic null distribution obtained in (4) leads to the test statistic

$$T_0 = \frac{\sqrt{n}(\varphi_M(\overline{X}) - \delta^2)}{\hat{\tau}} \tag{5}$$

For $0 < \alpha < 1$, if we let z_α denote the $100(1 - \alpha)$-th percentile of the standard normal distribution, then H_0 is rejected at asymptotic level α if $T_0 > z_\alpha$ (Munk et al. 2008). As a further practical consideration, even if $\overline{X} \rightarrow M$, in which case the null hypothesis would be true, the term $\varphi_M(\overline{X})$ in the numerator converges to 0 faster than $\hat{\tau}$ does in the denominator, so the test statistic will approach $-\infty$. Thus, the null hypothesis would not be rejected, as desired in that case.

Note that, we need to define δ *a priori* to apply the above testing procedure. However, in most practical situations it is difficult to define δ, especially when the sample space is unbounded. When the support is compact, the radius of this space can provide a natural upper bound for δ, but, even then, the possibly abstract distances may be difficult to distill into an easily interpretable neighborhood. One may attempt to elicit information about δ from domain experts, but there does not exist, to our knowledge, an objective way to choose δ. Hence, we propose a modified version of the original neighborhood hypothesis that relaxes our dependence on a completely subjective choice for δ.

3. THE MODIFIED NEIGHBORHOOD HYPOTHESIS TEST

We begin by modifying the original neighborhood hypotheses as follows:

$$H_0 : \rho^2(\mu, M) \leq \gamma \mathrm{v}_F \text{ and } H_1 : \rho^2(\mu, M) > \gamma \mathrm{v}_F, \tag{6}$$

where $\gamma \in (0, 1)$ is user specified and $\mathrm{v}_F = E(\rho^2(X, \mu)) = \int \rho^2(X, \mu) dQ$ is the total Fréchet variation, where Q is a probability measure defined on the sample space. When ρ^2 is as defined as in the previous section, we have the further result that $\mathrm{v}_F = \mathrm{Tr}(\Sigma)$. As such, the null hypothesis now makes a statement about how far μ is from the function M with respect to a proportion (γ) of the total amount of variation present in the population of curves. While, in principle, other summaries of the variation in the population, such as the generalized variance $|\Sigma|$, could instead be used to define the size of the neighborhood, the total Fréchet variation is a natural choice for two immediate reasons. First, as presented in Patrangenaru and Ellingson (2015), the nonparametric definition of the mean is $\mu = \mathrm{argmin}_{c \in \mathbb{H}} E(\rho^2(X, c))$ and v_F, as defined above, is $\min_{c \in \mathbb{H}} E(\rho^2(X, c))$, so the two quantities are inherently linked together through the use of the distance ρ. Secondly, given this perspective, the null hypothesis can be restated as $\rho^2(\mu, M) \leq \gamma E(\rho^2(X, \mu))$, making both sides of the inequality statements about squared distances to μ, which will not typically be the case for other scalar descriptors of the variation. Further details about the interpretation and practical selection of γ can be found in Sect. 6.

In an ideal situation, we can simply replace δ^2 in T_0 (5) with $\gamma \mathrm{v}_F$ to get a test statistic of the form

$$T_1 = \frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma \mathrm{v}_F\right)}{\hat{\tau}}, \tag{7}$$

but, in practice, the unknown parameter v_F needs to be estimated. A reasonable choice is the consistent estimator $\hat{\mathrm{v}}_F = \mathrm{Tr}(S)$. If we simply replace v_F in T_1 with this estimate, Slutsky's Theorem cannot be directly applied to (7) to derive the asymptotic distribution of the resultant quantity

$$\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma \hat{\mathrm{v}}_F\right)}{\tau}.$$

However, asymptotic results for $\hat{\mathrm{v}}_F$ described in Patrangenaru and Ellingson (2015) allow us to obtain the asymptotic distribution of the above quantity. Hence, we posit the following lemmas that lead us to the ultimate test statistic derived in Theorem 3.1. Proofs of Lemma 3.1, Lemma 3.2 and Theorem 3.1, which utilize the aforementioned results, are relegated to the appendix.

Lemma 3.1. *If $X_1, \ldots, X_n$ are independent and identically distributed random elements in a Hilbert space $\mathbb{H}$ with population mean $\mu \in \mathbb{H}$ and covariance operator $\Sigma : \mathbb{H} \to \mathbb{H}$ such that $E\left(||X||^4\right) < \infty$, then*

$$\sigma_1^2 = Var\left(\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau}\right) = 1 - \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) + \frac{\gamma^2}{\tau^2}\left[E[\rho^4(\mu, X)] - {\mathrm{v}_F}^2\right]. \tag{8}$$

From this, we arrive at the asymptotic distribution of $\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau}$.

Lemma 3.2. *Under the conditions of Lemma 3.1, then*

$$\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau} \to_d N\left(0, 1 - \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) + \frac{\gamma^2}{\tau^2}\left[E[\rho^4(\mu, X)] - {\mathrm{v}_F}^2\right]\right) \tag{9}$$

However, this asymptotic result cannot be applied immediately in practice since there are a number of unknown parameters in the asymptotic variance. Instead, though, we can use the following plug-in estimator for σ_1^2:

$$\hat{\sigma}_1^2 = 1 - \frac{2\gamma n}{\hat{\tau}^2}\widehat{\mathrm{Cov}}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) + \frac{\gamma^2}{\hat{\tau}^2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho^4(\overline{X}, X_i) - \hat{\mathrm{v}}_F^2\right], \tag{10}$$

where the estimate for $\widehat{\mathrm{Cov}}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right)$ can be obtained via nonparametric bootstrap. From this result, we can now arrive at our test statistic for the modified neighborhood hypotheses.

Theorem 3.1. *Under the conditions of Lemma 3.1 and the mild assumption that $\hat{\sigma}_1^2 > 0$, we arrive at the following asymptotic result:*

$$T_2 = \frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\hat{\tau}\hat{\sigma}_1} \to_d N(0, 1). \tag{11}$$

As a result, we reject H_0 at asymptotic level α when $T_2 > z_\alpha$.

4. SIMULATION EXAMPLES

We perform three simulation studies to assess how well our testing procedure can detect departure from hypothesized mean. In the first example, we simulate functions from a standard Gaussian Process (GP). In the second example, we simulate from a kernel convolution of scales χ^2 noises. In the final simulation study, we illustrate how our methodology can be used to assess adequacy of parametric models assumed for functional data.

4.1. SIMULATION 1

We simulate functions using the model

$$X(t) = \mu(t) + \epsilon_1(t) \tag{12}$$

where $\mu(t) = 0.5t - 0.5t^2 + 1.9t^3 - 2.2\sin(t)$ and $\epsilon_1(t)$ is a zero-mean GP with squared exponential covariance function given by $C(t, t') = \sigma^2 \exp(-\frac{(t-t')^2}{2\theta^2})$. We fix $\sigma^2 = 2$ and $\theta = 1$. Instead of specifying M exactly as described in (12), we simulate 2000 observations from the above simulation model and fix their sample mean as M. As a result, we are testing whether the population mean is in the neighborhood of the Monte Carlo approximation of the true mean.

To assess the power associated with our test, we fix $\alpha = 0.05$, $\gamma = 0.01$ and simulate $n = 30, 100, 500$, and 1000 observations. Due to the dimensionality of the data, it would be impossible to exhaustively evaluate the power function over an extended neighborhood around M. Instead, we evaluate the power function along the first principal component of M. We then calculate the test's power over a fine grid of points along this line and plot the resulting power functions in Fig. 1.

On each plot, we denote the boundaries of the hypothesized neighborhood using vertical lines to illustrate the null and alternative parameter spaces. The interval between the vertical lines is the projection of the null parameter space along the set of M considered. The area in between two cutoffs is equivalent to the null parameter space. The scale for the x-axis is the distance each candidate mean function is along the specified line from M. A better understanding of the size of the neighborhood can be obtained by looking at Fig. 2. It shows the population mean and the functions M corresponding to the cutoffs represented by the vertical lines in Fig. 1 that denote the two boundaries of the neighborhood in positive and negative directions along the first principal component. This plot clearly shows that the functions at the boundaries of the neighborhood are nearly visually indistinguishable from μ, which suggests that the choice of the radius of the neighborhood was not too large to substantially differ from the traditional null hypothesis of exact equality.

The power curves consistently remain below the significance level for all values of M we evaluated along the specified line that are in the null parameter space. Empirically, it appears that our modified neighborhood testing procedure maintains the desired level of significance even with relatively small sample sizes and has the desirable property of increased power outside the acceptance region with increase in sample size.

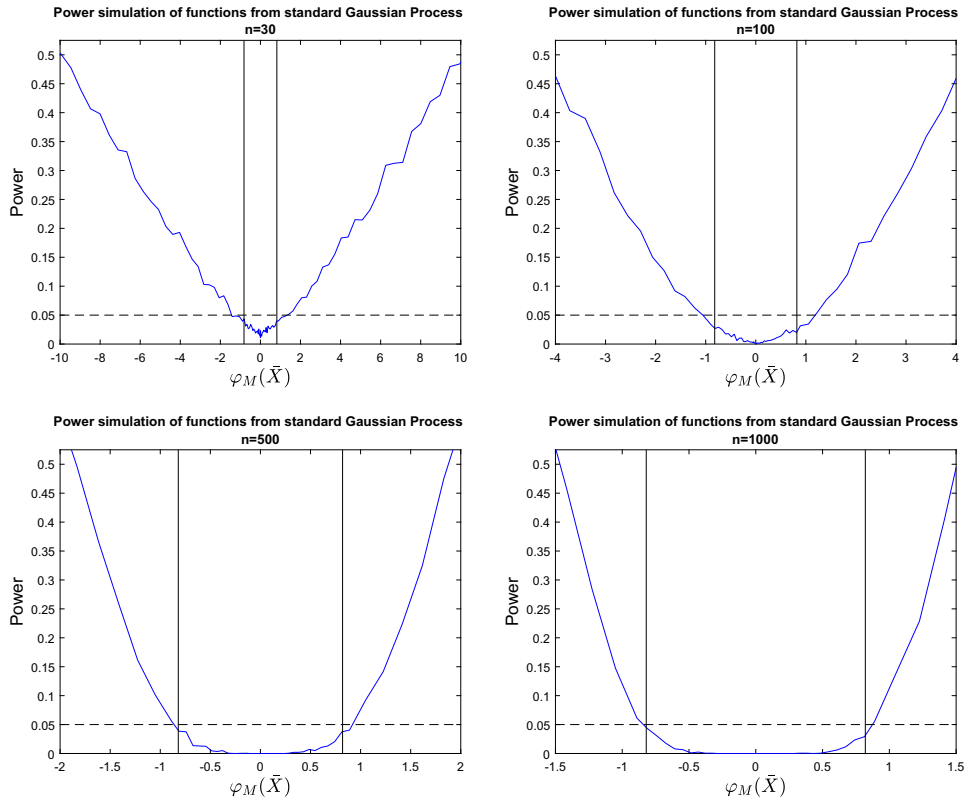

Figure 1. Power simulation of functions sampled from (12) with $\gamma = 0.01$ and $T = 1000$ where T represents the number of replications.

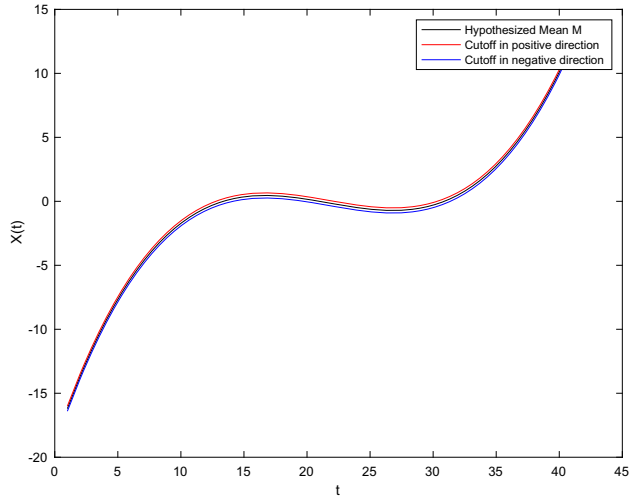

Figure 2. Boundary between null and alternative parameter space of power simulation study for functional data.

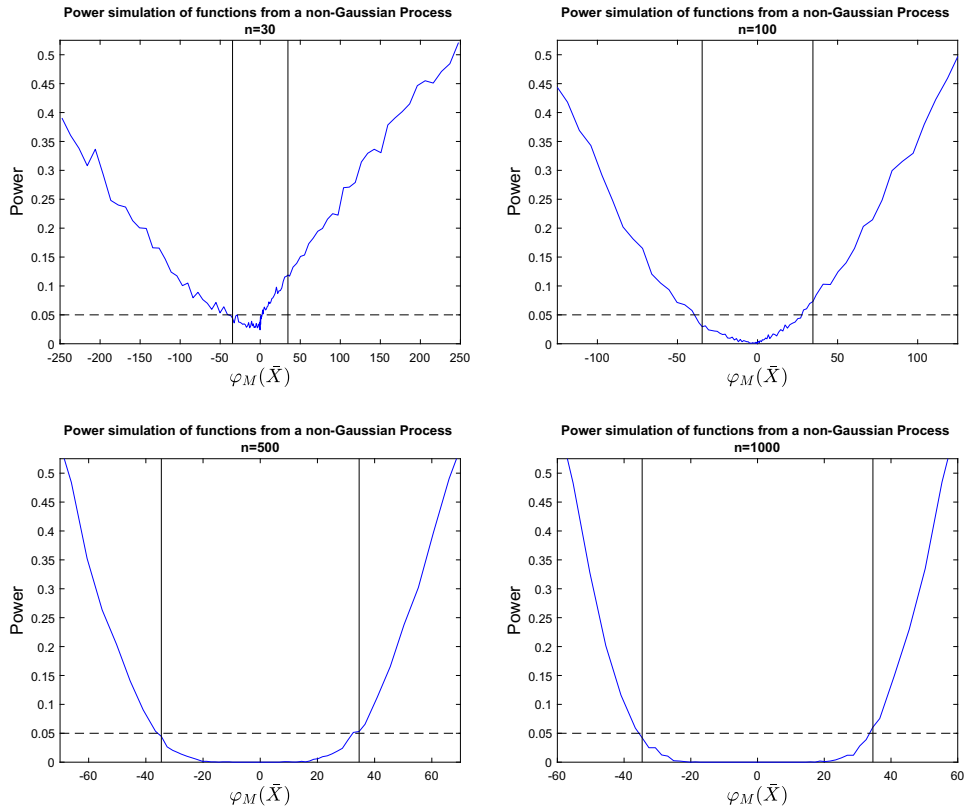

Figure 3. Power simulation of functions samples from (13) case with $\gamma = 0.01$ and $T = 1000$ where T represents the number of replications.

4.2. SIMULATION 2

In this example we investigate the power of our asymptotic test when the functions are sampled from a non-Gaussian process. We generate the functions using the model

$$X(t) = \mu(t) + \epsilon_2(t) \tag{13}$$

where $\mu(t)$ remains same as describe in Sect. 4.1. We define $\epsilon_2(t)$ as a discrete convolution of iid $scaled\chi^2$ random variates, i.e. we choose $t_1, t_2, \ldots t_m$ as specific set of locations in the domain of $X(t)$ and define $\epsilon_2(t) = \sum_{i=1}^{m} k(t_i - t)\eta^2(t_i)$, where $k(t)$ is smoothing kernel (we fix $k(t)$ as Normal(0, 0.1) density) and $\eta(t_i)$ are iid Normal(0,1). Clearly, the true population mean is no longer $\mu(t)$ and the noise process in (13) is highly skewed. Once again, we specify M as the Monte Carlo approximation of $\mu(t)$, fix $\alpha = 0.05$, simulate $n = 30, 100, 500, 1000$ observations from (13) and plot the power functions in Fig. 3.

According to Fig. 3, the power curve is higher than the significance level for most of the values of M in the null parameter space for the small sample sizes like $n = 30$. Also, for the same sample size, the monotonicity of the power is violated in outside of acceptance region. However, the power curve tends to adhere to the significance level in the acceptance region

and monotonicity in outside of it as the sample size increases. In addition, the time taken by the power curve to converge once it is outside the acceptance region is less for large sample sizes.

4.3. SIMULATION 3

In the first two simulations, we had tested whether the Monte Carlo approximation of population mean is in the neighborhood of true population mean. Consequently, we did not explicitly estimate the parameters of mean function from the set of random curves. In this simulation, however, we tackle the adequacy of a hypothesized mean function more directly. Consider a set of random functions generated from the model $X_i(t) = f(t, \delta) + \epsilon_i(t), , i = 1, 2, \ldots, n$, over a grid of eight equi-spaced epochs, i.e., $t \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ with

$$f(t, \delta) = [1 + \exp(-t^{1+\delta})]^{-1}, \tag{14}$$

$\epsilon_i \sim N_8(0_8, \Sigma)$, and a fixed value for the decay parameter δ. Instead of a covariance function, we assume the following form of Σ

$$\Sigma = \begin{pmatrix} 1 & .7 & .4 & .1 & 0 & 0 & 0 & 0 \\ .7 & 1 & .7 & .4 & .1 & 0 & 0 & 0 \\ .4 & .7 & 1 & .7 & .4 & .1 & 0 & 0 \\ .1 & .4 & .7 & 1 & .7 & .4 & .1 & 0 \\ 0 & .1 & .4 & .7 & 1 & .7 & .4 & .1 \\ 0 & 0 & .1 & .4 & .7 & 1 & .7 & .4 \\ 0 & 0 & 0 & .1 & .4 & .7 & 1 & .7 \\ 0 & 0 & 0 & 0 & .1 & .4 & .7 & 1 \end{pmatrix},$$

Suppose we wished to test whether the population mean curve is given by $f(t, 0)$. One option was to fit the model (14) for each i and test $H_0 : \delta = 0$. However, instead of performing n tests, our methodology could be used as an initial screen to assess whether $f(t, 0)$ adequately captured the population mean of $X(t)$ curves. But what should M be? Of course, we could resort to Monte Carlo to extract M—as was done in the first two simulations. Instead, we followed a more direct—differential equation approach—that offered an alternative way to extract M (Ramsey & Silverman, 2nd ed, pp. 7).

Consider applying linear differential operator on the responses of the form $LX_i(t) = dX_i(t)/dt - X_i(t)(1 - X_i(t))$. Observe that, under the hypothesis that the population mean curve is given by $f(t, 0)$, we have $f'(t, 0) - f(t, 0)[1 - f(t, 0)] = 0$. Thus, if the mean function was indeed, $f(t, 0)$, then $LX \approx 0$ for any function of form (14). Defining $Y_i(t) = LX_i(t)$, we would expect that the population mean of Y_i (μ_Y, say) should be approximately be equal to the zero function. Hence, the null hypothesis, for our modified neighborhood test, could be written as $H_0 : \rho^2(\mu_Y, 0) < \gamma \mathrm{v}_F$. Now that we have an explicit formulation of H_0 in terms of population mean curve and an hypothesized value of the same, our modified neighborhood hypothesis testing framework could be deployed to assess the adequacy of $f(t, 0)$ specification using a single test. We therefore proceed to assess the power of our test

to detect departure from $f(t, 0)$ by simulating $X(t)$ curves over a grid of δ values using the procedure described below.

First, observe that, despite the assumption of continuity of $X(t)$, the observations were sampled at finite number of functional epochs. To offset this sparse sampling over t, we linearly interpolated the observed X_is over a dense grid of t resulting in a more densely sampled equi-spaced piecewise linear functions that were conducible to application of the foregoing differential operator L. Subsequently, Y curves were computed as follows:

$$Y_i(t) = \frac{X_i(t) - X_i(t-h)}{h} - X_i(t)(1 - X_i(t)).$$

The test statistic (11) was computed using the sample mean of the Y curves and γ was determined empirically via simulating large number of observations under the foregoing H_0.

The power curves shown in Fig. 4 were obtained by repeating the above process by varying the decay parameter δ from -0.05 to 0.05 with an increment of 0.001. For each value of δ, we computed the power curves for sample sizes $n = \{10, 100, 500, 1000\}$. In each case, the power decreases when δ is close to 0 and shows an increasing trend when δ is further from 0 in both the negative and positive directions. While monotonicity of the power curve is violated for small sample sizes, such as $n = 10$, the power curve is monotonic in each direction for larger sample sizes, as is desired.

We do note that the conventional approach would have been to test whether residuals have a mean of zero, but that would require an assumption whether the noise is driven by a smooth stochastic process or iid. We recommend the differential operator approach because it does not require such assumption and directly test the mean specification.

5. APPLICATION TO CCLE DATASET

In the era of precision medicine, it is essential to generate genomics informed personalized therapeutic regimes with higher efficacy. Accurate prediction of sensitivity of an individual tumor to a drug is fundamental in designing highly precise cancer therapy treatments. Large-scale pharmacogenomics studies are conducted in order to identify in-vitro effect of several drugs on specific cancers against panels of molecularly characterized cancer cell lines (Safikhani et al. 2017). Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012) is one such publicly available pharmacogenomic database which provides *in-vitro* experimental pharmacological sensitivities of 24 drug compounds across 479 cancer cell-lines. The data consists of an 8-point drug concentration scale that ranged from 2.5 to $8\mu M$ and relative cell growth rate measured after 72–84 h from the application of the drug compound. This yielded an 8-point dose-response assay for each cell line. Additionally, Barretina et al. (2012) fitted 4-parameter Hill curves to each of the aforementioned dose-response curves to estimate maximal effect level and concentration at half maximal effect of the drug. More precisely, the following parametric curves were fitted:

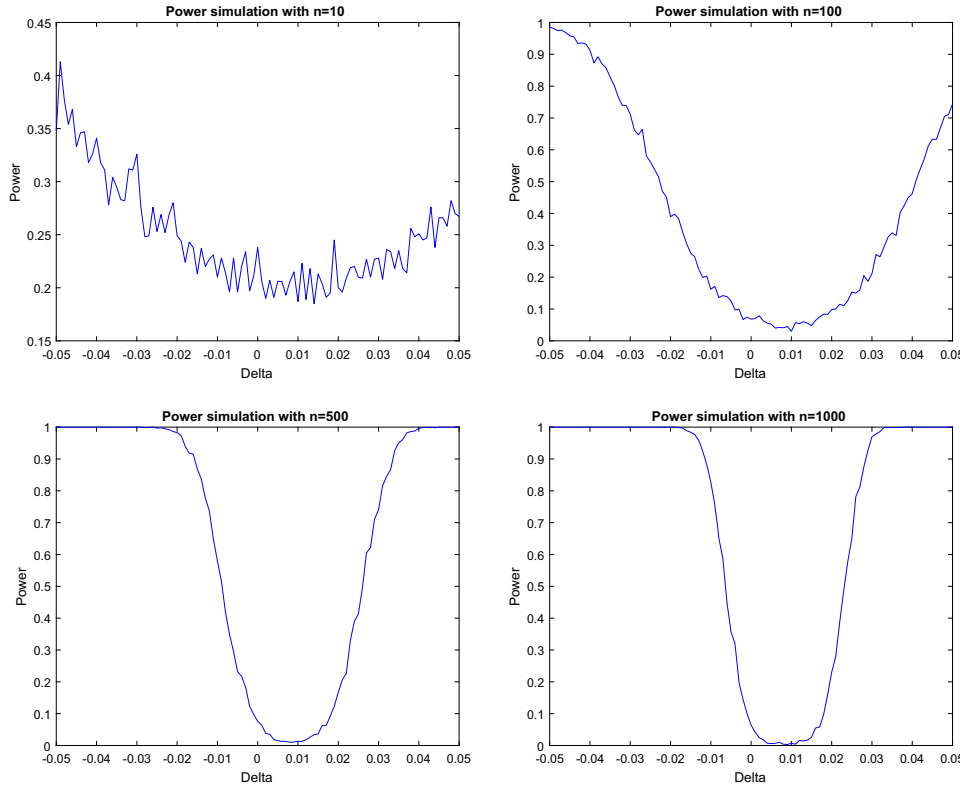

Figure 4. Power simulation from the differential equation approach (14) with estimated $\gamma = 0.008$ and T = 1000 where T represents the number of replications.

$$X_i(t) = \beta_{1,i} + \frac{\beta_{2,i} - \beta_{1,i}}{1 + \left(\frac{t}{\beta_{3,i}}\right)^{\beta_{4,i}}} \tag{15}$$

where $X_i(t)$ denotes the observed pharmacological sensitivity of a compound at dose level t for cell-line i. The parameters could be interpreted as follows: $\beta_{1,i}$ and $\beta_{2,i}$ are, respectively, the lower and upper asymptotes of the dose-response curve for the ith cell-line. $\beta_{3,i}$ is the point on the sigmoidal curve halfway between $\beta_{1,i}$ and $\beta_{2,i}$. The Hill slope, i.e., the slope at the steepest point of the sigmoidal curve is given by $\beta_{4,i}$.

An obvious way to ascertain if the parametric form (15) is an adequate model for $X(t)$ would be to fit (15) to every dose-response curve and perform a goodness-of-fit test for each curve. We, on the other hand, deployed our modified neighborhood hypothesis to perform an omnibus test to assess whether the population mean of $X(t)$ curves was within a specified neighborhood of (15). Rejection of the null hypothesis associated with out procedure would necessitate measuring the goodness-of-fit for each curve.

Following the arguments developed in Simulation 3, we defined $X_i(t) = f_{Hill}(t, \boldsymbol{\beta}_i) + \epsilon_i(t)$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$, $f_{Hill}(t, \boldsymbol{\beta})$ was the RHS of (15) and $\epsilon(t)$ was a zero-mean stochastic process. Next, we defined a linear differential operator of the form

$$Lf(t) = f'(t) - c_1 t^{c_2}(f(t) - c_3)^2 \tag{16}$$

where c_1, c_2, c_3 were, potentially unknown, constants. Define $f'_{Hill}(t, \boldsymbol{\beta}) = \frac{\delta}{\delta t} f_{Hill}(t, \boldsymbol{\beta})$. Since, $f'_{Hill}(t, \boldsymbol{\beta}) - \frac{\beta_4 t^{\beta_4 - 1}}{(\beta_2 - \beta_1)\beta_3^{\beta_4}}(f_{Hill}(t, \boldsymbol{\beta}) - \beta_1)^2 = 0$, if the population mean was indeed $f_{Hill}(t, .)$ then we posited that $LX(t) \approx 0$. Therefore, setting $U(t) = \frac{\delta X(t)}{\delta t} - \frac{\beta_4 x^{\beta_4 - 1}}{(\beta_2 - \beta_1)\beta_3^{\beta_4}}(X(t) - \beta_1)^2$, we can formally state our null hypothesis $H_0 : \rho(\mu_U, 0) < \gamma v_F$.

However, unlike Simulation 3, values of $\boldsymbol{\beta}$ remained unspecified in the above H_0. So, to obtain the $U(t)$ curves from our observed sample of $X(t)$ curves, we proceeded as follows:

- We fitted the sigmoidal model (15) to every observed dose-response curve in our sample using nonlinear least squares and obtained the estimates $\hat{\boldsymbol{\beta}}_i, i = 1, 2, \ldots, n$.
- We then defined U

$$\hat{U}_i(t) = \frac{X_i(t) - X_i(t-h)}{h} - \frac{\hat{\beta}_{4,i} t^{\hat{\beta}_{4,i} - 1}}{(\hat{\beta}_{2,i} - \hat{\beta}_{1,i})\hat{\beta}_{3,i}^{\hat{\beta}_{4,i}}}(X_i(t) - \hat{\beta}_{1,i})^2$$

- If X curves were sparsely sampled, we followed the linear interpolation technique outlined in Simulation 3 to augment the sampling density of $X(t)$. U curves were then computed from the interpolated X curves.

As an illustrative example of application of our methodology, we considered the drug 'Nutlin-3' - experimentally shown to increase radio-sensitivity of laryngeal squamous cell carcinoma (Arya et al. 2010)—in CCLE database. Since CCLE contained dose-response information on 240 cell-lines for Nutlin-3, we anticipated that the large sample size would justify our asymptotic inferential scheme. Figure 5 shows the cell viability scores for one randomly selected cell-line measured at 8-point log(dose) scale. Given the sampling sparsity of observed dose-response curves ($X(t)$), we first perform the aforesaid linear interpolation and then generate the $\hat{U}(t)$ curves. Figure 6 illustrates the two functions to be compared in the test $H_0 : \rho(\mu_U, 0) < \gamma v_F$.

Using $\gamma = 0.05$, the value of the test statistics $T_2 \approx -3.1280$. We, therefore, failed to reject the null hypothesis and concluded that at the 5% level of significance, there was no evidence to suggest that a sigmoidal model (15) for the population mean of the dose-response curves, associated with the drug 'Nutlin-3' reported in the CCLE database, was not adequate. Quite obviously, since we failed to reject the null hypothesis of our omnibus test, we do not need to proceed to test the goodness-of-fit of each $X_i(t), i = 1, 2, \ldots, n$.

In the simulation studies, we were able to choose the value for γ via Monte Carlo simulation so that the asymptotic power of the test was exactly equal to α on the boundary of the neighborhood. Here, though, we cannot do the same since we are not making assumptions about the distribution or covariance structure of the $\epsilon_i(t)$. As such, a completely objective selection of δ is not possible without making more assumptions. Therefore, we chose the value of γ with respect to the interpretation of it as the percentage of total variation present in the population. We chose γ to be larger than it was in our simulation studies in order

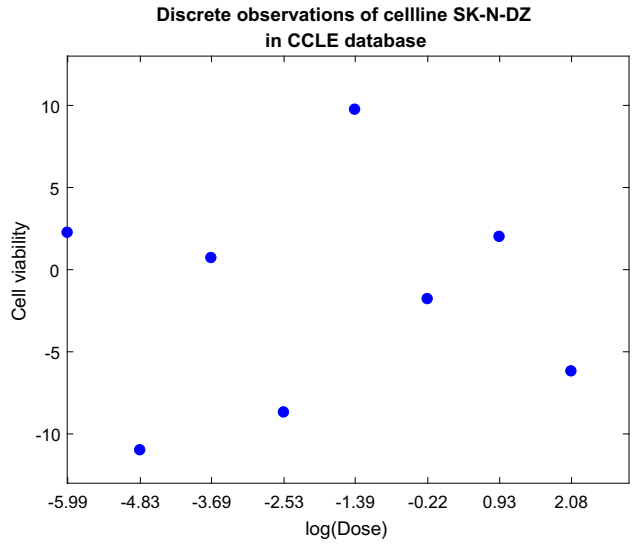

Figure 5. Cell viability calculated over 8 log(Dose) levels.

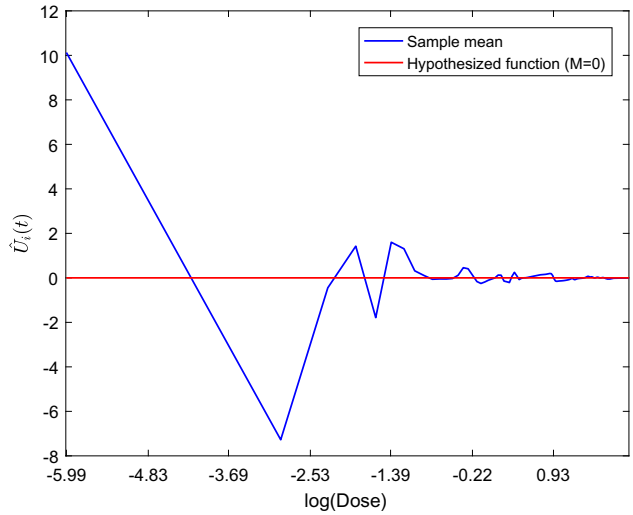

Figure 6. A plot the sample mean of $\hat{U}$ and the hypothesized mean M across log(Dose) levels.

to accommodate the extra uncertainty involved with estimating the parameters of the Hill curves that was not present in those studies.

6. DISCUSSION

The neighborhood hypothesis testing methodology, as defined in Munk et al. (2008), provided a solid theoretical foundation for testing hypotheses for the mean of functional data where sample covariance matrix suffer from rank deficiency. Rather than try to regularize the covariance matrix or project the data to a lower dimensional space, their method made

a clever alteration to the hypothesis itself that allowed for a near avoidance of the rank deficiency altogether. However, one drawback of their proposed procedure was that the radius δ of the neighborhood remained difficult to define objectively *a priori* and interpret that parameter in a meaningful way since the sample space for functional data is typically modeled as a non-compact Hilbert space. As such, it remained an open problem to determine a way to state the null hypothesis developed in Munk et al. (2008) with greater objectivity.

In this paper, we addressed this issue theoretically and offered a practical and interpretable way of selecting δ. First, since the Fréchet total variance, v_F, is a fixed quantity, we modified the null hypothesis to define the neighborhood in terms of a proportion γ of v_F. This allows for a clearer interpretation of the neighborhood and makes it easier to define its radius in a more objective manner. The revised hypothesis required the test statistic to be modified, as well, since v_F is a nuisance parameter. We proved that our modified test statistic is asymptotically normally distributed, in agreement with the asymptotic distribution of the test statistic associated with the original version of neighborhood hypothesis test.

Although we have utilized the above interpretation of γ throughout the manuscript—since that is how it arose from the decomposition of δ^2 and how it appeared in the calculation of the test statistic—we note that there exists another, equally appealing, interpretation of γ. Rewriting the modified hypotheses as

$$H_0 : \frac{\rho^2(\mu, M)}{\mathrm{v}_F} \leq \gamma \text{ and } H_1 : \frac{\rho^2(\mu, M)}{\mathrm{v}_F} > \gamma, \tag{17}$$

we observed that $\frac{\rho^2(\mu,M)}{\mathrm{v}_F}$ could be interpreted as the square of a one-population analogue of Cohen's d measure of effect size (Cohen 1988) in terms of Euclidean distances between vectors. Consequently, Sawilowsky (2009) guidelines for interpreting magnitudes of Cohen's d might provide reasonable guides for selecting $\sqrt{\gamma}$ and, thus, γ. With this in mind, our modified neighborhood hypotheses, with a pre-specified $\gamma = 0.01$, could be interpreted as testing claims about whether the effect size for the deviation between the population mean μ and claimed mean M was, using Sawilowsky's definition, "very small".

Turning to application of our framework, our simulation studies highlighted the performance of our test, in practice, for densely sampled functional data. The simulated power functions in Figs. 1 and 3 showed that the Type I error rate at the boundary of the projected null parameter space was approximately the same as the asymptotic significance level of the test even when n was comparable to the number p of sampling points used to specify the function. In the illustrative example, we applied our methodology on a widely used publicly available pharmacogenomic database and demonstrated that a four-parameter sigmoidal model can adequately capture the population level dose-response relationship for a drug Nutlin-3. Instead of performing a goodness-of-fit test for each dose-response curve (since each curve admitted individual-level parameters), our testing procedure offered an omnibus screening test even when population parameters remained unspecified. At a broader level, several researchers have used EC_{50}, AUC values obtained by fitting the foregoing sigmoidal model to the CCLE data for downstream modeling and inference purposes (Ma et al. 2021; De Niz et al. 2016; Wan and Pal 2014). Statistical evidence of a mis-specified sigmoidal model could potentially negatively impact any downstream estimation and predic-

tion. Consequently, we submit that, before utilizing estimated EC_{50}, AUC, etc. for downstream modeling, our screening test procedure could be used right at the outset to assess the empirical validity of these drug efficacy measures.

On the limitation side of our procedure, we note the presence of singularity when M approaches μ. The test statistic is not well defined when $M = \mu$, and its null distribution is degenerate. In essence, then, it can be said that this test reduces the singularities formed when $p > n$ to the single point where μ is *exactly* equal to the hypothesized mean. It was unclear from the theoretical results whether the test would perform well near this singularity. However, limited investigation (Figs. 1 and 3) indicate that the power function does not blow up as the distance between M and μ approaches 0. We are actively investigating a robust way to handle this singularity. Future work will offer more detailed guidelines on how to modify our testing framework when M is arbitrarily close to μ. Developing a modified neighborhood ANOVA procedure is also an exciting future research direction. Additionally, adapting this procedure in non-Euclidean setting would not only be of theoretical importance but will have many practical applications. Since similarity shape data are often non-Euclidean, developing a modified neighborhood hypothesis test for data on manifolds would allow us to compare planar shapes from different populations. To that end, we are currently investing modifying the non-Euclidean version of the hypothesis developed in Ellingson et al. (2013) to get more interpretable neighborhood.

Funding The Funding was provided by National Science Foundation (CCF-2007418, CCF-2007903).

[Received November 2022. Revised April 2023. Accepted May 2023.]

A PROOFS FOR SECTION 3 THE MODIFIED NEIGHBORHOOD HYPOTHESIS TEST

Lemma 3.1. *If $X_1, \ldots, X_n$ are independent and identically distributed random elements in a Hilbert space $\mathbb{H}$ with population mean $\mu \in \mathbb{H}$ and covariance operator $\Sigma : \mathbb{H} \to \mathbb{H}$ such that $E\left(||X||^4\right) < \infty$, then*

$$\sigma_1^2 = \mathrm{Var}\left(\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau}\right) = 1 - \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) + \frac{\gamma^2}{\tau^2}\left[E[\rho^4(\mu, X)] - {\mathrm{v}_F}^2\right]. \tag{8}$$

Proof. The test statistic T_1 can be decomposed as follows:

$$T_1 = \frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F + \gamma\hat{\mathrm{v}}_F - \gamma\mathrm{v}_F\right)}{\tau} = \frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau} + \frac{\gamma\sqrt{n}\left(\hat{\mathrm{v}}_F - \mathrm{v}_F\right)}{\tau}. \tag{18}$$

From Patrangenaru and Ellingson (2015, pg. 179), we also know that

$$\sqrt{n}(\hat{\mathrm{v}}_F - \mathrm{v}_F) \to_d N\left(0, E\left[\rho^4(\mu, X)\right] - {\mathrm{v}_F}^2\right).$$

As such,

$$\sigma_2^2 = \mathrm{Var}\left(\frac{\gamma}{\tau}\sqrt{n}(\hat{\mathrm{v}}_F - \mathrm{v}_F)\right) = \frac{\gamma^2}{\tau^2}\left(E\left[\rho^4(\mu, X)\right] - {\mathrm{v}_F}^2\right) \tag{19}$$

From Sect. 2, we know that $Var(T_1) = 1$. Combining this with the above results yields

$$\begin{aligned}
1 = \mathrm{Var}(T_1) &= \sigma_1^2 + \sigma_2^2 + 2\mathrm{Cov}\left(\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau}, \frac{\gamma}{\tau}\sqrt{n}(\hat{\mathrm{v}}_F - \mathrm{v}_F)\right) \\
&= \sigma_1^2 + \sigma_2^2 + \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F, \hat{\mathrm{v}}_F - \mathrm{v}_F\right) \\
&= \sigma_1^2 + \sigma_2^2 + \frac{2\gamma n}{\tau^2}\left[\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) - \mathrm{Cov}\left(\varphi_M(\overline{X}), \mathrm{v}_F\right)\right. \\
&\left. -\gamma\mathrm{Cov}\left(\hat{\mathrm{v}}_F, \hat{\mathrm{v}}_F\right) + \gamma\mathrm{Cov}\left(\hat{\mathrm{v}}_F, \mathrm{v}_F\right)\right] \\
&= \sigma_1^2 + \sigma_2^2 + \frac{2\gamma n}{\tau^2}\left[\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) - \gamma Var\left(\hat{\mathrm{v}}_F\right)\right] \\
&= \sigma_1^2 + \sigma_2^2 + \frac{2\gamma n}{\tau^2}\left[\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) - \gamma\frac{\tau^2}{\gamma^2 n}\sigma_2^2\right] \\
&= \sigma_1^2 + \sigma_2^2 + \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) - \frac{2\gamma n}{\tau^2}\frac{\tau^2}{\gamma n}\sigma_2^2 \\
&= \sigma_1^2 - \sigma_2^2 + \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right)
\end{aligned} \tag{20}$$

Solving for σ_1^2 combined with (19) yields

$$\sigma_1^2 = 1 - \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right) + \frac{\gamma^2}{\tau^2}\left[E[\rho^4(\mu, X)] - {\mathrm{v}_F}^2\right]. \tag{21}$$

□

Lemma 3.2. *Under the conditions of Lemma 3.1, then*

$$\begin{aligned}
\frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\tau} \to_d N\left(0, 1 - \frac{2\gamma n}{\tau^2}\mathrm{Cov}\left(\varphi_M(\overline{X}), \hat{\mathrm{v}}_F\right)\right. \\
\left. + \frac{\gamma^2}{\tau^2}\left[E[\rho^4(\mu, X)] - {\mathrm{v}_F}^2\right]\right)
\end{aligned} \tag{9}$$

Proof. This follows immediately from (18) to (21). □

Theorem 3.1. *Under the conditions of Lemma 3.1 and the mild assumption that $\hat{\sigma}_1^2 > 0$, we arrive at the following asymptotic result:*

$$T_2 = \frac{\sqrt{n}\left(\varphi_M(\overline{X}) - \gamma\hat{\mathrm{v}}_F\right)}{\hat{\tau}\hat{\sigma}_1} \to_d N(0, 1). \tag{11}$$

Proof. From the proof of Lemma 3.1 and results from nonparametric bootstrap theory, then if $\hat{\sigma}_1^2 > 0$, then it is a consistent estimator of σ_1^2. We can then apply Slutsky's Theorem to the result of Lemma 3.2, yielding this result. □

REFERENCES

Arya AK, El-Fert A, Devling T, Eccles RM, Aslam MA, Carlos P, Vlatkovi'c N, Fenwick J, Lloyd BH, Sibson DR et al (2010) Nutlin-3, the small-molecule inhibitor of MDM2, promotes senescence and radiosensitises laryngeal carcinoma cells harbouring wild-type p53. Br J Cancer 103(2):186–195

Barretina B, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D et al (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603–607

Berger JO, Delampady M (1987) Testing precise hypotheses. Stat Sci 2:317–352

Cohen J (1988) Statistical power analysis for the behavioral sciences. Routledge, Milton Park

De Niz C, Rahman R, Zhao X, Pal R (2016) Algorithms for drug sensitivity prediction. Algorithms 9(4):77

Dette H, Munk A (1998) Validation of linear regression models. Ann Stat 26:778–800

Dette H, Munk A (2003) Some methodological aspects of validation of models in nonparametric regression. Stat Neerl 57:207–244

Ellingson L, Patrangenaru V, Ruymgaart FH (2013) Nonparametric estimation of means on Hilbert manifolds and extrinsic analysis of mean shapes of contours. J Multivar Anal 122:317–333

Hodges L, Lehmann L (1954) Testing the approximate validity of statistical hypotheses. J Roy Stat Soc B 16(2):261–268

Kuelbs J, Vidyashankar A (2010) Asymptotic inference for high-dimensional data. Ann Stat 38(2):836–869

Ma J, Fong SH, Yunan Luo, Bakkenist CJ, Shen JP, Mourragui S, Wessels LFA, Hafner M, Sharan R, Jian Peng et al (2021) Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. Nat Cancer 2(2):233–244

Munk A, Paige R, Pang J, Patrangenaru V, Ruymgaart F (2008) The one and multi sample problem for functional data with application to projective shape analysis. J Multivar Anal 99:815–833

Patrangenaru V, Ellingson L (2015) Nonparametric statistics on manifolds and their applications to object data analysis. Chapman & Hall/CRC, London

Ramsay JO, Silverman BW (2005) Functional data analysis. In: Springer series in statistics. Springer

Safikhani Z, Smirnov P, Thu KL, Silvester J, El-Hachem N, Quevedo R, Lupien M, Mak TW, Cescon D, Haibe-Kains B (2017) Gene isoforms as expression-based biomarkers predictive of drug response in vitro. Nat Commun 8:1126

Sawilowsky S (2009) New effect size rules of thumb. J Modern Appl Stat Methods 8(2):467–474. https://doi.org/10.22237/jmasm/1257035100

Wainwright Martin J (2019) High-dimensional statistics: a non-asymptotic viewpoint, vol 48. Cambridge University Press, Cambridge

Wan Q, Pal R (2014) An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge. PLoS ONE 9(6):e101183

Xu M, Zhang D, Wu W (2014) L2 asymptotics for high-dimensional data. arXiv:1405.7244