Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Kernel-Based Partial Permutation Test for Detecting Heterogeneous Functional Relationship

Xinran Li, Bo Jiang & Jun S. Liu

To cite this article: Xinran Li, Bo Jiang & Jun S. Liu (2022): Kernel-Based Partial Permutation Test for Detecting Heterogeneous Functional Relationship, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.2000867

To link to this article: https://doi.org/10.1080/01621459.2021.2000867

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–19: Theory and Methods
https://doi.org/10.1080/01621459.2021.2000867

Kernel-Based Partial Permutation Test for Detecting Heterogeneous Functional Relationship

Xinran Li[a], Bo Jiang[b], and Jun S. Liu[c]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Two Sigma Investments LP, New York, NY; [c]Department of Statistics, Harvard University, Cambridge, MA

ABSTRACT

We propose a kernel-based partial permutation test for checking the equality of functional relationship between response and covariates among different groups. The main idea, which is intuitive and easy to implement, is to keep the projections of the response vector $\boldsymbol{Y}$ on leading principle components of a kernel matrix fixed and permute $\boldsymbol{Y}$'s projections on the remaining principle components. The proposed test allows for different choices of kernels, corresponding to different classes of functions under the null hypothesis. First, using linear or polynomial kernels, our partial permutation tests are exactly valid in finite samples for linear or polynomial regression models with Gaussian noise; similar results straightforwardly extend to kernels with finite feature spaces. Second, by allowing the kernel feature space to diverge with the sample size, the test can be large-sample valid for a wider class of functions. Third, for general kernels with possibly infinite-dimensional feature space, the partial permutation test is exactly valid when the covariates are exactly balanced across all groups, or asymptotically valid when the underlying function follows certain regularized Gaussian processes. We further suggest test statistics using likelihood ratio between two (nested) Gaussian process regression models, and propose computationally efficient algorithms utilizing the EM algorithm and Newton's method, where the latter also involves Fisher scoring and quadratic programming and is particularly useful when EM suffers from slow convergence. Extensions to correlated and non-Gaussian noises have also been investigated theoretically or numerically. Furthermore, the test can be extended to use multiple kernels together and can thus enjoy properties from each kernel. Both simulation study and application illustrate the properties of the proposed test.

ARTICLE HISTORY
Received December 2019
Accepted October 2021

KEYWORDS
Gaussian kernel; Gaussian process regression; Permutation test; Polynomial kernel; Regression discontinuity design

1. Introduction

Testing whether the same functional relationship between response and covariates holds across different groups is a challenging and important problem. For example, in clinical trial studies, people want to compare effects of several treatments conditional on some important covariates of patients such as age, gender, and genetic information. Traditional methods assume parametric forms of these functional relationships, such as linear or quadratic with unknown coefficients. When such assumptions cannot be supported by prior knowledge, nonparametric tests for the equality of functional relationships were recommended, especially in the exploratory stage of data analysis. Most methods in the nonparametric setting focus on univariate functions and use kernel estimator for estimating regression curves. For example, Pardo-Fernández, Van Keilegom, and González-Manteiga (2007) proposed empirical process based procedures for testing the equality of multiple regression curves. A comprehensive review on this topic can be found in Neumeyer and Dette (2003).

Testing the equality of functions has also been studied from a Bayesian perspective. Behseta and Kass (2005) proposed two methods for testing the equality of two univariate functions using the Bayesian adaptive regression splines. Benavoli and Mangili (2015) used Gaussian processes for Bayesian hypothesis testing on the equality of two functions, as well as the monotonicity and periodicity of a function. Behseta, Kass, and Wallstrom (2005) applied hierarchical Gaussian processes to study the variability among multiple functions, where they assumed an independent Gaussian process prior for each function and focused on the estimation of the variance component.

A closely related study for comparing two regression functions, but with slightly different focus, is the regression discontinuity design (Thistlethwaite and Campbell 1960), under which there can be no overlap between covariate distributions for the two groups in comparison. In this case, testing equality of two functions essentially reduces to testing whether two functions can be smoothly connected at the boundary. Various frequentist approaches have been proposed, including nonparametric kernel regression methods and local linear regression (Hahn, Todd, and der Klaauw 2001); for a comprehensive review, see Imbens and Lemieux (2008). Most existing methods focus mainly on the case with univariate covariates. Recently, Branson et al. (2019) and Rischard et al. (2018) proposed Bayesian approaches using Gaussian processes, and extended the regression discontinuity design to multivariate settings with spatial covariates.

CONTACT Jun S. Liu jliu@stat.harvard.edu Harvard University, 1 Oxford Street, Cambridge, MA 2138.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In this article, we first propose a partial permutation test for linear functional relationships, and then generalize it to handle nonlinear relationships via kernel methods. We demonstrate the exact validity of the partial permutation test when the kernel corresponds to a finite-dimensional feature mapping whose linear span contains the underlying true function, or when the covariates are exactly balanced across all groups. We further establish the asymptotic validity of the partial permutation test for a general smooth functional relationship when we choose the kernel adaptively with the sample size, or when the underlying function is from some Gaussian process. Note that the Gaussian process regression (GPR) model has received much attention recently for modeling functional relationships (see, e.g., Rasmussen and Williams 2006; Shi and Choi 2011) and is closely related to the kernel regression, which minimizes a squared loss with penalization on the functional norm in a reproducing kernel Hilbert space (RKHS) characterized by a kernel. Intuitively, we can understand p-values under the GPR model in an averaging sense over the Gaussian process prior on the underlying function. As Meng (1994) suggested, uniformity under parameters following prior is a useful criterion for the evaluation of any proposed p-value. We also investigate the power of the test when there exists functional heterogeneity across different groups, and extend the test to cases with correlated noises across individuals.

The article proceeds as follows. Section 2 introduces notations, model assumptions and the partial permutation test based on the linear or polynomial kernel, and proves its finite-sample validity when the underlying function is linear or polynomial. Section 3 first studies the partial permutation test using general kernels for general underlying functions under the null hypothesis with homogeneous functional relationship across all groups, and then shows its finite-sample or asymptotic validity under additional conditions on the kernel, the underlying function and the covariate distribution. Section 4 studies the power of the partial permutation test under the alternative hypotheses with heterogeneous function relations across all groups. Section 5 discusses practical implementation of the partial permutation test. Section 6 extends the partial permutation test to correlated noises. Section 7 conducts simulation study, and Section 8 applies the proposed test to a real dataset. Section 9 concludes with a short discussion.

2. Notations, Hypotheses, Kernels, and Permutation Tests

2.1. Notations and Problem Formulation

Let $Y_i \in \mathbb{R}$, $\boldsymbol{X}_i \in \mathbb{R}^d$ and $Z_i \in \{1, 2, \ldots, H\}$ denote the response variable, covariates of dimension d, and the group indicator for the ith ($1 \le i \le n$) observation, respectively, and let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)^\top$ and $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^\top$ be the corresponding vectors of all the n units. Given observations from multiple groups, we want to test whether they share the same (unknown) functional relationship. Specifically, given a response variable Y and a vector of covariates $\boldsymbol{X}$, the null hypothesis assumes that individuals from H ($H \ge 2$) groups have the same relationship $\mathbb{E}(Y|\boldsymbol{X}) = f_0(\boldsymbol{X})$ plus a Gaussian noise with constant variance, where f_0 is an unknown function in a given class (e.g., linear or polynomial functions), that is,

$$H_0 : Y_i = f_0(\boldsymbol{X}_i) + \varepsilon_i, \quad \varepsilon_i|\mathbf{X}, \boldsymbol{Z} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \qquad (1 \le i \le n) \tag{1}$$

where $\mathbf{X}$ and $\boldsymbol{Z}$ can be either fixed or random, and noises ε_i's are independent and identically distributed (iid) conditional on $\mathbf{X}$ and $\boldsymbol{Z}$. The alternative hypothesis allows different groups to have different (unknown) functions $f_1, \ldots, f_H$:

$$H_1 : Y_i = f_{Z_i}(\boldsymbol{X}_i) + \varepsilon_i, \quad \varepsilon_i|\mathbf{X}, \boldsymbol{Z} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \qquad (1 \le i \le n) \tag{2}$$

or even different noise variances in different groups:

$$H_1' : Y_i = f_{Z_i}(\boldsymbol{X}_i) + \varepsilon_i, \quad \varepsilon_i|\mathbf{X}, \boldsymbol{Z} \sim \mathcal{N}(0, \sigma_{Z_i}^2), \qquad (1 \le i \le n). \tag{3}$$

2.2. Partial Permutation Test for Linear Functional Relationship

We first consider a special case in which the relationship between the response and covariates under H_0 is linear, that is, $f_0(\boldsymbol{x}) = \beta_0 + \sum_{k=1}^d \beta_k x_k$ in (1) for some $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^{d+1}$. Let $K_{\text{Linear}}(\boldsymbol{x}, \boldsymbol{x}') = 1 + \boldsymbol{x}^\top \boldsymbol{x}'$ denote the linear kernel function with $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$. We write the corresponding sample kernel matrix as $\boldsymbol{K}_n \in \mathbb{R}^{n \times n}$, with its (i,j)th element being $[\boldsymbol{K}_n]_{ij} = K_{\text{Linear}}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ and its eigen-decomposition denoted as $\boldsymbol{\Gamma C \Gamma}^\top$. Here, $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n) \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $\boldsymbol{C} = \text{diag}(c_1, \ldots, c_n) \in \mathbb{R}^{n \times n}$ has nonnegative diagonal elements in descending order.

The linear kernel can be equivalently written as $K_{\text{Linear}}(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, an inner product in a feature space defined by the feature mapping $\phi : \boldsymbol{x} \to (1, \boldsymbol{x}^\top)^\top \in \mathbb{R}^{d+1}$. Let $\boldsymbol{\Phi} \equiv (\phi(\boldsymbol{X}_1), \ldots, \phi(\boldsymbol{X}_n))^\top \in \mathbb{R}^{n \times (d+1)}$ be the matrix of all the observed covariates mapped into the feature space, and let $\boldsymbol{f}_0 \equiv (f_0(\boldsymbol{X}_1), \ldots, f_0(\boldsymbol{X}_n))^\top$ be the vector of function values evaluated at these covariates. Under the null model (1), we can verify that $\boldsymbol{f}_0 = \boldsymbol{\Phi\beta}$ lies in the column space of $\boldsymbol{\Phi}$, or equivalently the column space of kernel matrix $\boldsymbol{K}_n = \boldsymbol{\Phi\Phi}^\top$. Because $\boldsymbol{K}_n$ has at most rank $d+1$, the eigenvectors $(\boldsymbol{\gamma}_{d+2}, \ldots, \boldsymbol{\gamma}_n)$ must be orthogonal to the column space of $\boldsymbol{K}_n$, as well as the vector $\boldsymbol{f}_0$ in this column space. As a result, under H_0, we have

$$\boldsymbol{\Gamma}^\top \boldsymbol{Y} = \left(\boldsymbol{\gamma}_1^\top \boldsymbol{f}_0, \boldsymbol{\gamma}_2^\top \boldsymbol{f}_0, \ldots, \boldsymbol{\gamma}_{d+1}^\top \boldsymbol{f}_0, 0, \ldots, 0\right)^\top + \left(\boldsymbol{\gamma}_1^\top \boldsymbol{\varepsilon}, \boldsymbol{\gamma}_2^\top \boldsymbol{\varepsilon}, \ldots, \boldsymbol{\gamma}_n^\top \boldsymbol{\varepsilon}\right)^\top,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. Therefore, $\boldsymbol{\gamma}_i^\top \boldsymbol{Y} = \boldsymbol{\gamma}_i^\top \boldsymbol{\varepsilon}$ for $i = d+2, \ldots, n$, and are iid conditional on $\mathbf{X}$ and $\boldsymbol{Z}$. Consequently, given any test statistic, we can perform permutation tests by randomly permuting $\boldsymbol{\gamma}_i^\top \boldsymbol{Y}$ for $i = d+2, \ldots, n$. Note that this procedure takes advantage of the fact that projections of $\boldsymbol{Y}$ onto the eigenvectors corresponding to zero eigenvalues are just random noises. Intuitively, this observation may be generalized so that one can treat projections of $\boldsymbol{Y}$ onto eigenvectors with small eigenvalues as Gaussian noises (i.e., $\boldsymbol{\varepsilon}$) instead of signals (i.e., $\boldsymbol{f}_0$), which are then exchangeable and permit permutation tests.

For the convenience of presentation, we summarize in Algorithm 1 a general discrete or continuous partial permutation test procedure with a given kernel function K, permutation size b_n, and test statistic T. For linear functional relationships, we have the following theorem on the validity of the p-value from either the discrete or continuous partial permutation test using the linear kernel.

Algorithm 1 Discrete and continuous partial permutation tests with kernel function K, permutation size b_n and test statistic T for $\{\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$

1) Perform eigen-decomposition for kernel matrix $\boldsymbol{K}_n = \boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{\Gamma}^\top$, where $[\boldsymbol{K}_n]_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $\boldsymbol{\Gamma}$ is an orthogonal matrix and $\boldsymbol{C}$ is a diagonal matrix with diagonal elements in descending order.
2) Let $\boldsymbol{W} = \boldsymbol{\Gamma}^\top \boldsymbol{Y} \equiv (W_1, \ldots, W_n)^\top$.
   (a) For the *discrete* partial permutation test, we define the permutation set $\mathcal{S}_y$ as follows:

$$\mathcal{S}_y = \{\boldsymbol{Y}_\psi : \boldsymbol{Y}_\psi = \boldsymbol{\Gamma}\boldsymbol{W}_\psi, \boldsymbol{W}_\psi \in \mathcal{S}_w\}, \text{ with}$$
$$\mathcal{S}_w = \{\boldsymbol{W}_\psi : \boldsymbol{W}_\psi = (W_{\psi(1)}, W_{\psi(2)}, \ldots, W_{\psi(n)}), \psi \in \mathcal{M}(n, b_n)\},$$

   where $\mathcal{M}(n, b_n)$ is defined to be the set of permutations of $\{1, 2, \ldots, n\}$ that keep the first $n - b_n$ elements invariant, that is,

$$\mathcal{M}(n, b_n) = \big\{\psi : \psi(i) = i, \text{ for } i = 1, 2, \ldots, n - b_n, \text{ and } \{\psi(n - b_n + 1), \ldots, \psi(n)\} \text{ is a permutation of } \{n - b_n + 1, \ldots, n\}\big\}.$$

   Note that we allow the sets $\mathcal{S}_y$ and $\mathcal{S}_w$ to have members of identical value. For example, if $\psi \neq \psi' \in \mathcal{M}(n, b_n)$ but $\boldsymbol{W}_\psi = \boldsymbol{W}_{\psi'}$ (which may happen if some W_j's take on the same value), then they are treated as two elements in $\mathcal{S}_w$.
   (b) For *continuous* partial permutation test, define the permutation set $\mathcal{S}_y$ as follows:

$$\mathcal{S}_y = \{\boldsymbol{Y}^* : \boldsymbol{Y}^* = \boldsymbol{\Gamma}\boldsymbol{W}^*, \boldsymbol{W}^* \in \mathcal{S}_w\}, \text{ with}$$
$$\mathcal{S}_w = \Big\{\boldsymbol{W}^* : W_i^* = W_i, i = 1, 2, ..., n - b_n, \sum_{i=n-b_n+1}^{n} (W_i^*)^2 = \sum_{i=n-b_n+1}^{n} W_i^2\Big\}.$$

3) Draw $\boldsymbol{W}^{\mathrm{p}} \in \mathcal{S}_w$ uniformly, and let $\boldsymbol{Y}^{\mathrm{p}} = \boldsymbol{\Gamma}\boldsymbol{W}^{\mathrm{p}}$. Naturally, $\boldsymbol{Y}^{\mathrm{p}}$ is uniformly distributed on $\mathcal{S}_y$, where both $\mathcal{S}_w$ and $\mathcal{S}_y$ can be viewed as a function of $\mathbf{X}$ and $\boldsymbol{Y}$.
4) The resulting partial permutation p-value with test statistic T is then defined as

$$p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \Pr\{T(\mathbf{X}, \boldsymbol{Y}^{\mathrm{p}}, \boldsymbol{Z}) \geq T(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) | \mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}\}.$$

Theorem 1. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \leq i \leq n}$ denote samples from the model under H_0 in (1), where the functional relationship $f_0(\boldsymbol{x})$ is linear in $\boldsymbol{x}$. Then, the p-value obtained by either the discrete or continuous partial permutation test described in Algorithm 1 with kernel K_{Linear}, permutation size $b_n \leq n - (d + 1)$, and any test statistic T is valid, that is, $\forall \alpha \in (0, 1)$, $\Pr_{H_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \leq \alpha | \mathbf{X}, \boldsymbol{Z}\} \leq \alpha$.

Remark 1. When the matrix $\boldsymbol{\Phi}$ consisting of the covariates mapped into the feature space is not of full rank, we can relax the constraint to be $b_n \leq n - \mathrm{rank}(\boldsymbol{\Phi})$. Similar relaxations also hold for Theorem 2 and Corollaries 1 and 2.

Theorem 1 suggests that we can use any test statistic to conduct a valid permutation test as long as the underlying functional relationship between the response and the covariates is linear. To achieve a high power when the null hypothesis is false, we suggest to use the likelihood ratio statistics with respect to alternative hypotheses that are of particular interest. For example, we may choose either (2) or (3) as the alternative hypothesis, where we assume that the functions $f_1, \ldots, f_H$ are still linear in the covariates but can have different coefficients across the H groups.

Under the Gaussian linear regression model, Algorithm 1 is able to generate permutation samples that change only the responses but keep both the covariates and group indicators fixed, which means that our partial permutation test is an exact conditional test—conditioning on the covariates and group indicators $(\mathbf{X}, \boldsymbol{Z})$. This is important since it avoids imposing any distributional assumption on $(\mathbf{X}, \boldsymbol{Z})$. As a side note, simply permuting the group indicators Z_i's may not lead to a valid permutation test since such a permutation does not maintain the joint distribution of the covariates and the group indicator. An analogous approach is the classic bootstrap procedure based on residual resampling (Freedman and Peters 1984; Hinkley 1988), which generates new data similar to the observed ones but keeps $(\mathbf{X}, \boldsymbol{Z})$ fixed. In general, the residual bootstrap can help relax the Gaussianity assumption on the noises, but loses the finite-sample exact validity. Moreover, the parametric F-test, whose test statistic is equivalent to a likelihood ratio statistic, is finite-sample valid and is also regarded as most powerful under the linear model with Gaussian noises. As demonstrated both theoretically and empirically in Sections 4 and 7, our partial permutation test can have almost the same power as the F-test.

2.3. Partial Permutation Test for Polynomial Functional Relationship

We consider here a more general case in which the relationship between the response and covariates under H_0 is a polynomial of degree p (or smaller), where p is a positive integer. Specifically, under H_0, we assume that $f_0(\boldsymbol{x}) = \sum_{j_1+j_2+\cdots+j_d \leq p} \beta_{j_1 j_2 \cdots j_d} x_1^{j_1} x_2^{j_2} \ldots x_d^{j_d}$. Let $K_{\mathrm{Poly}}(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^p$ denote a degree-p polynomial kernel function. We again let $\boldsymbol{K}_n$ denote the corresponding sample kernel matrix with entries $[\boldsymbol{K}_n]_{ij} = K_{\mathrm{Poly}}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, and let $\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{\Gamma}^\top$ be the eigen-decomposition of $\boldsymbol{K}_n$, where $\boldsymbol{C} = \mathrm{diag}(c_1, \ldots, c_n)$ is the diagonal matrix with nonnegative eigenvalues c_i in descending order, and $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n) \in \mathbb{R}^{n \times n}$ is an orthogonal matrix.

As with the linear case, we can rewrite the kernel matrix's entry as an inner product in a feature space defined by the feature mapping ϕ, that is, $K_{\mathrm{Poly}}(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, where $\phi(\boldsymbol{x})$ consists of all the monomials $x_1^{j_1} x_2^{j_2} \ldots x_d^{j_d}$ with $j_1 + j_2 + \cdots + j_d \leq p$ up to some positive coefficients. Let $\boldsymbol{\Phi} = (\phi(\boldsymbol{X}_1), \ldots, \phi(\boldsymbol{X}_n))^\top$ be the $n \times \binom{d+p}{d}$ matrix consisting of the observed covariates mapped into the feature space, and let $\boldsymbol{f}_0 = (f_0(\boldsymbol{X}_1), \ldots, f_0(\boldsymbol{X}_n))^\top$ denote the vector of function values evaluated at these covariates. Under the null model (1), we can verify that $\boldsymbol{f}_0$ must lie in the column space of $\boldsymbol{\Phi}$ or equivalently the column space of $\boldsymbol{K}_n = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$. Because the rank of $\boldsymbol{K}_n$ is at most $\binom{d+p}{d}$ provided that $\binom{d+p}{d} < n$, $(\boldsymbol{\gamma}_{\binom{d+p}{d}+1}, \ldots, \boldsymbol{\gamma}_n)$ must

be orthogonal to the column space of $\boldsymbol{K}_n$, as well as $\boldsymbol{f}_0$ in the column space. Therefore, under H_0, we have

$$\boldsymbol{\Gamma}^\top \boldsymbol{Y} = \left(\boldsymbol{\gamma}_1^\top \boldsymbol{f}_0, \ldots, \boldsymbol{\gamma}_{\binom{d+p}{d}}^\top \boldsymbol{f}_0, 0, \ldots, 0\right)^\top + \left(\boldsymbol{\gamma}_1^\top \boldsymbol{\varepsilon}, \ldots, \boldsymbol{\gamma}_n^\top \boldsymbol{\varepsilon}\right)^\top,$$

recalling that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. So $\{\boldsymbol{\gamma}_i^\top \boldsymbol{Y},\ i = \binom{d+p}{d}+1, \ldots, n\} = \{\boldsymbol{\gamma}_i^\top \boldsymbol{\varepsilon},\ i = \binom{d+p}{d}+1, \ldots, n\}$ are iid conditional on $(\mathbf{X}, \boldsymbol{Z})$, and we can perform permutation test by permuting $\{\boldsymbol{\gamma}_i^\top \boldsymbol{Y},\ i = \binom{d+p}{d}+1, \ldots, n\}$.

Theorem 2. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from the model under H_0 in (1), where the functional relationship $f_0(\boldsymbol{x})$ is polynomial in $\boldsymbol{x}$ with degree at most p. Then, the p-value obtained by either the discrete or continuous partial permutation test with kernel K_{Poly}, permutation size $b_n \le n - \binom{d+p}{d}$, and any test statistic T is valid, that is, $\forall \alpha \in (0,1)$, $\Pr_{H_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

Similar to that for Theorem 1, we recommend to use a likelihood ratio statistic with carefully chosen alternative hypothesis of interest in order to achieve a good power. For example, we may hypothesize that under the alternative hypothesis (2) or (3) the functions $f_1, \ldots, f_H$ are still polynomial up to degree p but can have different coefficients across different groups.

3. Partial Permutation Test for General Functional Relationship

3.1. Partial Permutation Test under the Null Hypothesis

Inspired by the partial permutation test based on linear and polynomial kernels, we generalize it to arbitrary kernels. Let K be any kernel that is symmetric, positive definite and continuous, and let $\boldsymbol{K}_n \in \mathbb{R}^{n\times n}$ be the corresponding sample kernel matrix with $[\boldsymbol{K}_n]_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$. Similar to the previous sections, we define eigen-decomposition on the kernel matrix $\boldsymbol{K}_n = \boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{\Gamma}^\top$, where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n) \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\boldsymbol{C} = \operatorname{diag}(c_1, \ldots, c_n) \in \mathbb{R}^{n\times n}$ has nonnegative diagonal elements in descending order. Recall that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and $\boldsymbol{f}_0 = (f_0(\boldsymbol{X}_1), \ldots, f_0(\boldsymbol{X}_n))^\top$. Then, we have

$$\boldsymbol{\Gamma}^\top \boldsymbol{Y} = \left(\boldsymbol{\gamma}_1^\top \boldsymbol{f}_0, \boldsymbol{\gamma}_2^\top \boldsymbol{f}_0, \ldots, \boldsymbol{\gamma}_n^\top \boldsymbol{f}_0\right)^\top + \left(\boldsymbol{\gamma}_1^\top \boldsymbol{\varepsilon}, \boldsymbol{\gamma}_2^\top \boldsymbol{\varepsilon}, \ldots, \boldsymbol{\gamma}_n^\top \boldsymbol{\varepsilon}\right)^\top. \tag{4}$$

Different from linear or polynomial kernel for linear or polynomial functions, the kernel matrix $\boldsymbol{K}_n$ can be of full rank, and $\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0 = 0$ may not hold exactly for any i. However, the kernel matrix $\boldsymbol{K}_n$ often has its eigenvalues decreasing quickly and is effectively rank-deficient (see, e.g., Hastie and Zhu 2006), and $\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0$ is often close to 0 for sufficiently large i when f_0 is relatively smooth with respect to kernel K. Below we give some intuition for this.

Assume that covariates $\boldsymbol{X}_i$'s are iid with respect to probability measure μ. By Mercer's theorem, the kernel function has the following eigen-decomposition: $K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}')$, where $\lambda_1 \ge \lambda_2 \ge \ldots$ are the eigenvalues, and the eigenfunctions ψ_i's are orthonormal bases for the class of square-integrable functions. The cross-product $\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0$ can be intuitively understood as an approximation (or sample analog) of the inner product $\int f_0 \psi_j \mathrm{d}\mu$ between the function f_0 and the $i$the eigenfunction ψ_i after proper scaling (see, e.g., Braun, Buhmann, and Müller 2008). Note that ψ_i becomes more and more non-smooth with respect to kernel K as i increases. When the underlying function f_0 is relatively smooth with respect to K, the inner product between f_0 and ψ_i, and thus the sample version $\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0$, diminishes quickly as i increases. Consequently, the projection of $\boldsymbol{Y}$ onto the space spanned by the $\boldsymbol{\gamma}_i$'s for large i is mostly dominated by the Gaussian noise, based on which we can then conduct permutation tests.

Unlike Theorems 1 and 2, with a general kernel K and a general function f_0, the partial permutation test is not finite-sample valid. This motivates us to investigate how to adjust the partial permutation test. It turns out that the correction needed for the partial permutation p-value depends crucially on

$$\omega(b_n, \sigma_0^{-1} f_0) = \sigma_0^{-2} \sum_{i=n-b_n+1}^{n} (\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0)^2, \tag{5}$$

which can be intuitively understood as the *left-over signal-proportion* (LOSP) among the components used for the partial permutation test of size b_n. Let Q_{b_n} denote the quantile function of the χ^2-distribution with degrees of freedom b_n, and for any $0 < \alpha_0 < 1$, define

$$\nu(b_n, \sigma_0^{-1} f_0, \alpha_0) = \frac{1}{2} \exp\left\{2\sqrt{2\omega(b_n, \sigma_0^{-1} f_0)} \cdot \sqrt{Q_{b_n}(1-\alpha_0) + \omega(b_n, \sigma_0^{-1} f_0)}\right\} - \frac{1}{2} \tag{6}$$

as a function of permutation size b_n, standardized function $\sigma_0^{-1} f_0$ and $\alpha_0 \in (0,1)$. The following theorem shows that, by adding the correction term (6) to the p-value, the partial permutation test becomes valid under H_0.

Theorem 3. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model H_0 in (1). Given $1 \le b_n \le n$ and $\alpha_0 \in (0,1)$, we define the corrected partial permutation p-value as

$$p_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) = p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) + \nu(b_n, \sigma_0^{-1} f_0, \alpha_0) + \alpha_0,$$

where $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is the p-value from either the discrete or continuous partial permutation test (as in Algorithm 1) with kernel K, permutation size b_n, and any test statistic T; and $\nu(b_n, \sigma_0^{-1} f_0, \alpha_0)$ is defined as in (6). Then the corrected partial permutation p-value is valid under model H_0, that is, $\forall \alpha \in (0,1)$, $\Pr_{H_0}\{p_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

In Theorem 3, the correction term $\nu(b_n, \sigma_0^{-1} f_0, \alpha_0)$ is increasing in b_n, that is, the larger the permutation size, the larger the correction for the p-value will be. Note that the corrected permutation p-value $p_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$ depends on the unknown true function f_0 and cannot be calculated directly. Besides, the asymptotic validity of the uncorrected partial permutation p-value, $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$, requires that $b_n \cdot \omega(b_n, \sigma_0^{-1} f_0)$ converges to zero in probability as $n \to \infty$, which may or may

not hold depending on the complexity of function f_0 as well as the choice of the permutation size. Nevertheless, Theorem 3 helps us understand the bias of this p-value for finite samples and provides insights on how to correct for it. In Sections 3.2–3.4, we will consider special cases under which the LOSP $\omega(b_n, \sigma_0^{-1} f_0)$ defined in (5) can be exactly or asymptotically zero and the partial permutation test can itself be finite- or large-sample valid without requiring any correction. Moreover, in Section 3.5, we consider GPR models, under which the LOSP can be bounded stochastically and the permutation test becomes asymptotically valid.

3.2. Special Case: Kernels With Finite-Dimensional Feature Space

We first consider the case in which kernel K has only a finite number of nonzero eigenvalues, or equivalently, the corresponding feature space is finite-dimensional, that is, $K(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$ with $\phi(\boldsymbol{x}) \in \mathbb{R}^q$ for some $q < \infty$. Following the same arguments as with the linear and polynomial kernels discussed in Sections 2.2 and 2.3, which are special cases of the current setting, we decompose the sample kernel matrix $\boldsymbol{K}_n$ as $\boldsymbol{\Gamma C \Gamma}^\top$, with $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_n)$ being the orthogonal matrix of eigenvectors and $\boldsymbol{C}$ the diagonal matrix of eigenvalues in descending order. If function $f_0(\boldsymbol{x})$ is linear in $\phi(\boldsymbol{x})$, then $\boldsymbol{\gamma}_i^\top \boldsymbol{f}_0 = 0$ for $i > q$ and thus the partial permutation test is finite-sample valid when the permutation size b_n is no larger than $n - q$. We summarize the results in the following corollary. Although it is a straightforward extension of Theorem 2, this result provides us some intuition and a bridge to general kernels.

Corollary 1. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \le i \le n}$ denote samples from model H_0 in (1). Suppose that the kernel function K has the decomposition $K(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$ with $\phi(\boldsymbol{x}) \in \mathbb{R}^q$ for some $q < \infty$, and the underlying function $f_0(\boldsymbol{x})$ is linear in $\phi(\boldsymbol{x})$. Then, the p-value obtained by either the discrete or continuous partial permutation test with kernel K, permutation size $b_n \le n - q$, and any test statistic T is valid, that is, $\forall \alpha \in (0, 1)$, $\Pr_{H_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

3.3. Special Case: Kernels With Diverging-Dimensional Feature Space

We extend Section 3.2 to consider kernels whose feature space dimensions can increase with the sample size, under which the partial permutation test can be (asymptotically) valid for a wider class of underlying functional relationships. Specifically, let $\{e_j : j = 1, 2, \ldots\}$ be a given series of basis functions of the covariate. For each integer $q > 0$, we define kernel $K_q(\boldsymbol{x}, \boldsymbol{x}') \equiv \phi_q(\boldsymbol{x})^\top \phi_q(\boldsymbol{x}') \equiv \sum_{j=1}^q e_j(\boldsymbol{x}) e_j(\boldsymbol{x}')$, where $\phi_q(\boldsymbol{x}) = (e_1(\boldsymbol{x}), e_2(\boldsymbol{x}), \ldots, e_q(\boldsymbol{x}))^\top$ denotes the corresponding feature mapping based on the first q basis functions. Motivated by Corollary 1, intuitively, the partial permutation test using kernel K_q is approximately valid if the underlying function $f_0(\cdot)$ can be approximated well by a linear combination of the first q basis functions. Moreover, we can increase the feature space dimension q at a proper rate as the sample size increases, and render the partial permutation test asymptotically valid provided that $f_0(\cdot)$ lies in the space generated by the infinite series of basis functions $\{e_1(\boldsymbol{x}), e_2(\boldsymbol{x}), \ldots\}$, as characterized more precisely in the following corollary. For any function f of the covariate, we introduce

$$\begin{aligned} \mathrm{r}(f; q) &= \min_{\boldsymbol{b} \in \mathbb{R}^q} \int (f - \boldsymbol{b}^\top \phi_q)^2 \mathrm{d}\mu \\ &= \min_{b_1, \ldots, b_q \in \mathbb{R}} \int (f - \sum_{j=1}^q b_j e_j)^2 \mathrm{d}\mu \end{aligned}$$

to denote the squared distance between f and its best linear approximation using the first q basis functions, where μ denotes the probability measure for the covariate. The limiting behavior of $\mathrm{r}(f; q)$ as q goes to infinity then characterizes how well f can be linearly approximated by this infinite series of basis functions. Note that here we implicitly assume that both f and e_j's are square-integrable.

Corollary 2. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \le i \le n}$ denote samples from model H_0 in (1), and assume that $\boldsymbol{X}_i$'s are identically distributed from some probability measure μ. Suppose that kernel function K_q has the form $K_q(\boldsymbol{x}, \boldsymbol{x}') \equiv \phi_q(\boldsymbol{x})^\top \phi_q(\boldsymbol{x}') \equiv \sum_{j=1}^q e_j(\boldsymbol{x}) e_j(\boldsymbol{x}')$ for $q \ge 1$ and some series of basis functions $\{e_j\}_{j=1}^\infty$. If there exists a sequence $\{q_n\}_{n=1}^\infty$ such that $q_n < n$ for all n and $n(n - q_n)\mathrm{r}(f_0; q_n) \to 0$ as $n \to \infty$, then the resulting p-value obtained by either the discrete or continuous partial permutation test (as in Algorithm 1) with kernel K_{q_n}, permutation size $b_n \le n - q_n$, and any test statistic T is asymptotically valid, that is, $\forall \alpha \in (0, 1)$, $\limsup_{n \to \infty} \Pr_{H_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha\} \le \alpha$.

From Corollary 2, the existence and construction of a valid large-sample partial permutation test depends crucially on how well the underlying functional relationship f_0 can be linearly approximated by the basis functions $\{e_j\}_{j=1}^\infty$. Below we consider constructing $\{e_j\}_{j=1}^\infty$ based on a general kernel K with infinite-dimensional feature space. Recall the discussion in Section 3.1. Suppose we have the eigen-decomposition of the kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^\infty \lambda_j \psi_j(\boldsymbol{x}) \psi_j(\boldsymbol{x}')$, where $\lambda_1 \ge \lambda_2 \ge \ldots$ are the eigenvalues, and $\psi_1, \psi_2, \ldots$ are the eigenfunctions and form an orthonormal basis for the space of square-integrable functions. If we choose $e_j = \lambda_j^{1/2} \psi_j$ for all $j \ge 1$, then kernel K_q based on the first q basis functions converges to K as q goes to infinity. Moreover, if the underlying function f_0 belongs to the RKHS $\mathcal{H}_K$ corresponding to kernel K, then we have $f_0 = \sum_{j=1}^\infty \alpha_j \lambda_j^{1/2} \psi_j = \sum_{j=1}^\infty \alpha_j e_j$ for some coefficients α_j's with $\sum_{j=1}^\infty \alpha_j^2 < \infty$, under which $\mathrm{r}(f_0; q) = \sum_{j>q} \alpha_j^2 \lambda_j \le \lambda_q \sum_{j>q} \alpha_j^2 = o(\lambda_q)$ as $q \to \infty$. As discussed later in Section 5.1, the eigenvalues λ_j's often decays at a polynomial rate with power greater than 1, in the sense that $\lambda_q = O(q^{-\kappa})$ for some $\kappa > 1$. This then implies that $\mathrm{r}(f_0; q) = o(\lambda_q) = o(q^{-\kappa})$.

Intuitively, we prefer a larger permutation size and thus a smaller q_n, which can generally lead to a more powerful test. However, conditions in Corollary 2 require us to be more considerate in selecting q_n. Let us focus on the case where the approximation error for f_0 decays polynomially, that is, $\mathrm{r}(f_0; q) = o(\lambda_q) = o(q^{-\kappa})$ for some $\kappa > 1$. When $\kappa \in (1, 2)$, we can choose $q_n = n - c_n n^{\kappa - 1}$ with c_n being of constant order;

in this case, the permutation size can be $n - q_n \asymp n^{\kappa-1}$ and $n(n-q_n)\mathrm{r}(f_0; q_n)$ must be of order $o(1)$. When $\kappa \geq 2$, we can choose $q_n = c_n n^{2/\kappa}$ with c_n being of constant order; in this case, the permutation size can be $n - q_n = n - c_n n^{2/\kappa} \asymp n$ and $n(n-q_n)\mathrm{r}(f_0; q_n)$ must be of order $o(1)$.

3.4. Special Case: Exactly Balanced Covariates across All Groups

Assume that the design matrix $\mathbf{X}$ enjoys a balancing property that the empirical distributions of covariates are exactly the same across all H groups, that is,

$$\begin{aligned}\{\boldsymbol{X}_i : Z_i = 1, 1 \leq i \leq n\} &= \{\boldsymbol{X}_i : Z_i = 2, 1 \leq i \leq n\} \\ &= \ldots \\ &= \{\boldsymbol{X}_i : Z_i = H, 1 \leq i \leq n\}. \end{aligned} \tag{7}$$

Let r denote the number of distinct covariate values. Obviously, $r \leq n/H$, and the equality holds if and only if the covariates within each group are all distinct. We can verify that the rank of kernel matrix $\boldsymbol{K}_n$ for all units is the same as that for the r distinct covariate values. Thus, $\mathrm{rank}(\boldsymbol{K}_n) \leq r$; moreover, the equality generally holds when kernel function K corresponds to an infinite-dimensional feature space, for example, the Gaussian kernel. When $\boldsymbol{K}_n$ is indeed of rank r, as demonstrated in the supplementary material, for any underlying function f_0, the LOSP $\omega(b_n, \sigma_0^{-1} f_0)$ in (5) is exactly zero as long as the permutation size b_n is no larger than $n - r$, under which the correction term $\nu(b_n, \sigma_0^{-1} f_0, \alpha_0)$ in (6) also reduces to zero. Consequently, the partial permutation p-value $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$ must be valid under model H_0, as summarized in the following corollary.

Corollary 3. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \leq i \leq n}$ denote samples from model H_0 in (1). If the design matrix is exactly balanced in the sense that (7) holds and the kernel matrix for all the $r \leq n/H$ distinct covariate values in each group is of full rank (or equivalently $\mathrm{rank}(\boldsymbol{K}_n) = r$), then the partial permutation p-value from either the discrete or continuous partial permutation test with kernel K, permutation size $b_n \leq n - r$, and any test statistic T is valid under model H_0, that is, $\forall\, \alpha \in (0,1)$, $\Pr_{H_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \leq \alpha | \mathbf{X}, \boldsymbol{Z}\} \leq \alpha$.

The validity of the test in Corollary 3 is closely related to that of the usual permutation test, which permutes the group indicators of samples with the same covariate value. Corollary 3 is more general in the sense that it allows for more general rotations (instead of purely switching) of the responses, utilizing the Gaussianity of the noises.

3.5. Special Case: Gaussian Process Regression Model

In this subsection, instead of treating f_0 in (1) under H_0 as a fixed unknown function as in previous sections, we here assume that the function follows a Gaussian process and show that $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is asymptotically valid under such a GPR model. We note that the GPR model has been widely used in functional analysis.

3.5.1. The Model Formulation

Given a symmetric, positive definite, and continuous kernel K, the GPR model assumes that

$$\begin{aligned} &\tilde{H}_0 : Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad \varepsilon_i | \mathbf{X}, \boldsymbol{Z} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \\ &f \sim \mathrm{GP}\left(0, \frac{\delta_0^2}{n^{1-\gamma}} K\right), \end{aligned} \tag{8}$$

where f is independent of $\mathbf{X}, \boldsymbol{Z}$ and the ε_i's, and $\delta_0^2/n^{1-\gamma}$, which depends on the sample size n, represents our belief on the smoothness of the underlying function. The GPR model is closely related to kernel regression, which minimizes a penalized mean squared loss over a RKHS $\mathcal{H}_K$ corresponding to kernel K. Specifically, the kernel regression estimator $\hat{f}_{n,\tau_n}$ is given by

$$\hat{f}_{n,\tau_n} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \tau_n ||f||^2_{\mathcal{H}_K}, \tag{9}$$

where τ_n is a regularization parameter penalizing the $\mathcal{H}_K$ norm of f. This estimator is identical to the posterior mean of f under the GPR model in (8) when $\tau_n = \sigma_0^2/(n^\gamma \delta_0^2)$. Christmann and Steinwart (2007) studied sufficient conditions on τ_n to guarantee the consistency of kernel regression estimator $\hat{f}_{n,\tau_n}$, which then provides us some guidance on the choice of the smoothness parameter $\delta_0^2/n^{1-\gamma}$. The following proposition is a direct corollary of Christmann and Steinwart (2007, Theroem 12).

Proposition 1. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n$ be random samples from model H_0 in (1). If the $\boldsymbol{X}_i$'s are iid with a compact support $\mathcal{X}$, f_0 is a measurable function, $\mathbb{E}_{H_0}(Y^2) < \infty$, K is a universal kernel, and $0 < \gamma < 1/4$, then the posterior mean $\tilde{f}$ induced by model $\tilde{H}_0$ in (8) is consistent for the underlying true f_0 in (1), that is, $\mathbb{E}_{H_0} |\tilde{f}(\boldsymbol{X}) - f_0(\boldsymbol{X})|^2 \xrightarrow{\Pr} 0$ as $n \to \infty$.

In Proposition 1, the universal kernel was introduced by Micchelli, Xu, and Zhang (2006), which has the property that the corresponding RKHS is dense in $\mathcal{C}(\mathcal{X})$, the space consisting of all continuous functions on $\mathcal{X}$ with the infinity norm. We can intuitively summarize conditions on the Gaussian process prior of f and the relation between its variance parameter and sample size as follows. First, f should be almost surely continuous. Second, if two realizations of f fit observations equally well, it is preferable to give the smoother one more weight. Third, as the sample size increases, the posterior mean and mode of f should increasingly concentrate around the true functional relationship. All the requirements above can be satisfied by the GPR model with an appropriate choice of the kernel function and by letting the variance parameter decrease at a proper rate as the sample size increases.

3.5.2. Large-Sample Valid Partial Permutation Test

The following theorem shows that the partial permutation p-value is asymptotically valid under the GPR model $\tilde{H}_0$ in (8) under certain conditions.

Theorem 4. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \leq i \leq n}$ denote samples from model $\tilde{H}_0$ in (8). If the $\boldsymbol{X}_i$'s are iid from a compact support $\mathcal{X}$ with some probability measure μ, the eigenvalues $\{\lambda_k\}$ of kernel K on

$(\mathcal{X}, \mu)$ satisfy $\lambda_k = O(k^{-\rho})$ for some $\rho > 1$, and $\gamma < 1 - \rho^{-1}$, then for sequence $\{b_n\}$ satisfying $b_n = O(n^{\kappa})$ with $0 < \kappa < 1 - \rho^{-1} - \gamma$, the partial permutation p-value from either the discrete or continuous partial permutation test with kernel K, permutation size b_n, and any test statistic T is asymptotically valid under $\tilde{H}_0$, that is, $\forall \alpha \in (0,1)$, $\limsup_{n\to\infty} \Pr_{\tilde{H}_0}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha\} \le \alpha$.

In Section 5.1, we will show that there exist universal kernels with polynomially decaying eigenvalues. Coupled with a choice of the smoothness parameter γ that satisfies the conditions in Proposition 1 and Theorem 4, the partial permutation test is asymptotically valid under a GPR model that imposes a reasonable amount of regularization on the underlying function. Furthermore, we emphasize that the asymptotic validity of the partial permutation test essentially requires that the ratio between the variance parameter for the Gaussian process prior and the variance of observation noises is of order $n^{-(1-\gamma)}$ for some $\gamma < 1 - \rho^{-1}$. Thus, even if the Gaussian process prior on f does not follow the regularized form as in (8), we can still perform asymptotically valid partial permutation test by adding noises to the responses.

Theorem 4 proves the large-sample validity of the partial permutation test. Below we investigate its finite-sample performance in analogous to Section 3.1. Let $\xi_n = (\delta_0^2/n^{1-\gamma})/\sigma_0^2$ denote the variance ratio for the function and noise. For any given b_n, we define

$$\tilde{\omega}(b_n, \xi_n) = \frac{\delta_0^2/n^{1-\gamma}}{\sigma_0^2} \cdot c_{n-b_n+1} = \xi_n \cdot c_{n-b_n+1}$$

to denote the LOSP for the components of $\boldsymbol{Y}$ used for the partial permutation test of size b_n, recalling that c_{n-b_n+1} is the $(n - b_n + 1)$th largest eigenvalue of $\boldsymbol{K}_n$. Note that here the LOSP $\omega(b_n, \sigma_0^{-1} f)$ defined in Section 3.1 can be bounded by $b_n \cdot \tilde{\omega}(b_n, \xi_n)$ in expectation under the GPR in (8). Recall that Q_{b_n} is the quantile function of the χ^2-distribution with degrees of freedom b_n. For $1 \le b_n \le n$ and $\alpha_0 \in (0,1)$, we define

$$\tilde{\nu}(b_n, \xi_n, \alpha_0) = \frac{1}{2}\exp\left[\frac{1}{2}\tilde{\omega}(b_n, \xi_n) \cdot Q_{b_n}(1-\alpha_0)\right] - \frac{1}{2}. \quad (10)$$

The following theorem shows that, by adding a correction term, the partial permutation p-value becomes finite-sample valid under $\tilde{H}_0$.

Theorem 5. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model $\tilde{H}_0$ in (8). Given $1 \le b_n \le n$ and $0 < \alpha_0 < 1$, we define the corrected partial permutation p-value as follows:

$$\tilde{p}_c(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) = p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) + \tilde{\nu}(b_n, \xi_n, \alpha_0) + \alpha_0,$$

where $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is the p-value from either the discrete or continuous partial permutation test with kernel K, permutation size b_n, and any test statistic T, and $\tilde{\nu}(b_n, \xi_n, \alpha_0)$ is as defined in (10). Then the corrected partial permutation p-value is valid under model $\tilde{H}_0$, that is, $\forall \alpha \in (0,1)$, $\Pr_{\tilde{H}_0}\{\tilde{p}_c(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

Note that in Theorem 5, the correction term $\tilde{\nu}(b_n, \xi_n, \alpha_0)$ is monotone increasing in the permutation size b_n. This is intuitive since the larger the permutation size, the larger the correction for the partial permutation p-value is needed. As discussed shortly in Section 5.2, Theorem 5 provides us some guidance on the choice of permutation size in finite samples.

4. Partial Permutation Test Under Alternative Hypotheses

4.1. Kernels With Finite-Dimensional Feature Space

While previous discussions focused on the validity of partial permutation tests under the null hypothesis that the samples share the same functional relationship across all groups, we here investigate how such tests behave under alternative hypotheses. As the permutation test allows for a flexible choice of test statistics, which can be tailored based on the alternative hypotheses of interest, we study a special class of test statistics that are linked to a certain form of likelihood ratio statistics under a general kernel with finite-dimensional feature space. That is, the kernel function can be decomposed as $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}')$ with $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^q$ for some $q < \infty$.

As demonstrated in Corollary 1, the partial permutation test is exactly valid under model H_0 in (1) when $f_0(\boldsymbol{x})$ is linear in the transformed covariates $\boldsymbol{\phi}(\boldsymbol{x})$. It is then straightforward to hypothesize that, under the alternative model specified in (2), the functional relationship between the response and covariates is also linear in the transformed covariates, but the coefficients can vary across groups, that is, for $1 \le i \le n$,

$$Y_i = \sum_{h=1}^{H} \mathbb{1}(Z_i = h)\boldsymbol{\beta}_h^\top \boldsymbol{\phi}(\boldsymbol{X}_i) + \varepsilon_i, \quad \varepsilon_i | \mathbf{X}, \boldsymbol{Z} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \quad (11)$$

where $\boldsymbol{\beta}_h$ denotes the regression coefficient vector for samples in the hth group. This motivates us to use the F statistic for testing $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_H$ as our test statistic, which is equivalent to the likelihood ratio statistic up to a monotone transformation. Let $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ denote the projection matrices onto the column spaces of the transformed covariates for regression model (11) under the null model that $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_H$ and the full model without any constraint on the parameters, respectively, and let p_0 and p_1 denote the matrices' ranks. Then, the F statistic for testing $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_H$ has the form

$$F(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \frac{\boldsymbol{Y}^\top(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y}^\top/(p_1 - p_0)}{\boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}^\top/(n - p_1)}. \quad (12)$$

It turns out that the permutation distribution of the F statistic in (12) under our continuous partial permutation test with kernel K and permutation size $b_n = n - p_0$ is F distributed with degrees of freedom $p_1 - p_0$ and $n - p_1$, which matches the repeated sampling distribution of the F statistic when model (11) holds with $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_H$. Therefore, with the same choice of the test statistic (i.e., F statistic or equivalently the likelihood ratio statistic), the partial permutation test is equivalent to the usual F-test or likelihood ratio test for nested regression models. We summarize the results in the following theorem. Let F_{d_1,d_2} denote the distribution function of the F distribution with degrees of freedom d_1 and d_2.

Theorem 6. Consider any samples $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ and any kernel K of form $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}')$ with $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^q$ and $q < \infty$. The permutation distribution of the F statistic in (12) under the continuous partial permutation test with kernel K and permutation size $n - p_0$ is an F distribution with degrees of freedom $p_1 - p_0$ and $n - p_1$, and the corresponding partial permutation p-value is $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}) = 1 - F_{p_1-p_0, n-p_1}(F(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}))$.

In Theorem 6, if the transformed covariates are linearly independent within each group (i.e., the matrix whose rows consist of $\phi(\boldsymbol{X}_i)^\top$ for samples in group h is of full column rank, $1 \le h \le H$), then $p_0 = q$ and $p_1 = Hq$. The equivalence between the partial permutation test and F-test in Theorem 6 has two implications. First, it confirms the finite-sample validity of the partial permutation test when the null hypothesis (i.e., model (11) with $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_H$) holds. Second, it shows that the partial permutation test using the F statistic is most powerful when the the alternative hypothesis is indeed of form (11) with possibly unequal $\boldsymbol{\beta}_h$'s. Furthermore, as demonstrated in Corollary 1, the partial permutation test allows for a more flexible choice of test statistics, which can be tailored toward any alternative hypothesis of interest, since the test uses partial permutation to get the valid null distribution. Finally, although Theorem 6 considers only kernels with a finite-dimensional feature space, it sheds light on general kernels as well since we can always view a general kernel as the limit of kernels with finite-dimensional feature spaces.

4.2. Kernels With Diverging-Dimensional Feature Space

We now extend the discussion in Section 4.1 to kernels with diverging-dimensional feature spaces as the sample size increases, similar to that in Section 3.3. Let $\{e_j\}_{j=1}^\infty$ be a given series of basis functions of the covariate, and let $K_q(\boldsymbol{x}, \boldsymbol{x}') = \phi_q(\boldsymbol{x})^\top \phi_q(\boldsymbol{x}') = \sum_{j=1}^q e_j(\boldsymbol{x}) e_j(\boldsymbol{x}')$ be the kernel with feature mapping $\phi_q(\boldsymbol{x}) = (e_1(\boldsymbol{x}), \ldots, e_q(\boldsymbol{x}))^\top$ consisting of the first q basis functions. We consider partial permutation test based on kernel K_{q_n} whose feature space dimension q_n can vary with the sample size n, and study its power using the F statistic as in (12) with ϕ replaced by ϕ_{q_n}. Analogously, we let $\boldsymbol{P}_{n0}$ and $\boldsymbol{P}_{n1}$ denote the projection matrices on to the column spaces of the transformed covariates under the null and the full models, and let p_{n0} and p_{n1} denote the matrices' ranks, respectively. Moreover, since we will investigate the power of the test under local alternatives, we allow the functional relationship between response and covariates under model H_1 in (2) to also vary with the sample size, and write them explicitly as $f_{n1}, f_{n2}, \ldots, f_{nH}$. Throughout this subsection, we assume that the covariates $\boldsymbol{X}_i$'s are identically distributed from some probability measure μ, and use $\mathrm{r}(f_{nh}; q) = \min_{\boldsymbol{b} \in \mathbb{R}^q} \int (f_{nh} - \boldsymbol{b}^\top \phi_q)^2 \mathrm{d}\mu$ to denote the squared error for the best linear approximation of f_{nh} using the first q basis functions.

Theorem 7. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1 \le i \le n}$ denote samples from model H_1 in (2), and assume that the $\boldsymbol{X}_i$'s follow probability measure μ. If, as $n \to \infty$ and for some $\theta \ge 0$,

$$p_{n1} - p_{n0} \to \infty, \quad \frac{p_{n1} - p_{n0}}{n - p_{n1}} \to 0, \quad \frac{n \sum_{h=1}^H \mathrm{r}(f_{nh}; q_n)}{\sqrt{p_{n1} - p_{n0}}} \to 0,$$
$$\frac{\boldsymbol{f}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{n0}) \boldsymbol{f}}{\sqrt{p_{n1} - p_{n0}}} = (\text{or} \ge)\, \theta + o_{\Pr}(1), \tag{13}$$

where $\boldsymbol{f} = (f_{nZ_1}(\boldsymbol{X}_1), f_{nZ_2}(\boldsymbol{X}_2), \ldots, f_{nZ_n}(\boldsymbol{X}_n))^\top$, then, $\forall \alpha \in (0, 1)$, the p-value from the continuous partial permutation test with kernel K_{q_n}, permutation size $n - p_{n0}$, and F test statistic as in (12) must satisfy that

$$\Pr\left(p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \le \alpha\right) = (\text{or} \ge)\, \Phi\left(z_\alpha + \theta/\sqrt{2}\right) + o(1),$$

where $\Phi(\cdot)$ denotes the distribution function of standard Gaussian distribution and z_α denotes the αth quantile of the standard Gaussian distribution.

From the discussion after Theorem 6, we generally expect that $p_{n0} \asymp q_n$ and $p_{n1} - p_{n0} \asymp q_n$, under which the first two conditions in (13) reduce to that $q_n \to \infty$ and $q_n/n \to 0$ as $n \to \infty$. Below we assume these are true and discuss two implications from Theorem 7.

First, we consider the case where the null hypothesis holds, that is, $f_{n1} = f_{n2} = \cdots = f_{nH} = f_0$ for some f_0 that depends neither on the group index nor the sample size. We can then demonstrate that $\boldsymbol{f}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{n0}) \boldsymbol{f} = n\mathrm{r}(f_0; q_n) \cdot O_{\Pr}(1)$. From Theorem 7 with $\theta = 0$, the partial permutation test will be asymptotically valid when $n\mathrm{r}(f_0; q_n) = o(q_n^{1/2})$. Suppose the approximation error for f_0 decays polynomially, that is, $\mathrm{r}(f_0; q) = o(q^{-\kappa})$ for some $\kappa > 0$. Then a sufficient condition for the large-sample validity of the partial permutation test will be $nq_n^{-\kappa} = O(q_n^{1/2})$, under which we can choose $q_n \asymp n^{2/(2\kappa+1)}$. Compared to Corollary 2, Theorem 7 imposes weaker conditions on $\{q_n\}$ for ensuring the validity of the test. This is not surprising since Corollary 2 allows for an arbitrary choice of test statistics while Theorem 7 concerns only the F statistic.

Second, we consider the case where the alternative hypothesis holds, and assume that the underlying functions have the form $f_{nh} = f_0 + \delta_n \zeta_h$ $(1 \le h \le H)$, for some constant sequence $\delta_n = O(1)$ and some functions $f_0, \zeta_1, \ldots, \zeta_H$ that do not vary with the sample size. Intuitively, $\{\delta_n\}$ and $\tau_{hh'} \equiv \int (\zeta_h - \zeta_{h'})^2 \mathrm{d}\mu$ measure the functional heterogeneity across the H groups. For simplicity, we further assume that the covariates in the H groups are exactly balanced and the covariates within each group are iid, under which we can bound $\boldsymbol{f}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{n0}) \boldsymbol{f}$ from below by $(2H)^{-1} \delta_n^2 \sum_{i=1}^n (\zeta_h(\boldsymbol{X}_i) - \zeta_{h'}(\boldsymbol{X}_i))^2 = (2H)^{-1} n \delta_n^2 \{\tau_{hh'} + o_{\Pr}(1)\}$ for all $1 \le h, h' \le H$. From Theorem 7, if $n\mathrm{r}(f_0; q_n) = o(q_n^{1/2})$, $n\mathrm{r}(\zeta_h; q_n) = o(q_n^{1/2})$ for all h, and $n\delta_n^2 \ge \theta \sqrt{8H^3 q_n}$ for sufficiently large n and some $\theta \ge 0$, then asymptotically the power of the level-α partial permutation test is at least $\Phi(z_\alpha + \theta \max_{h,h'} \tau_{hh'})$; see the supplementary material for details. Suppose that the approximation errors for functions $f_0, \zeta_1, \ldots, \zeta_H$ all decay polynomially, that is, $\mathrm{r}(f_0; q) = o(q^{-\kappa})$ and $\mathrm{r}(\zeta_h; q) = o(q^{-\kappa})$ as $q \to \infty$, and that $\tau_{h,h'} > 0$ for at least one pair of $h \ne h'$. From the discussion before, we can then choose $q_n \asymp n^{2/(2\kappa+1)}$ to ensure Type I error control. Consequently, if $\delta_n \gg (\sqrt{q_n}/n)^{1/2} = n^{-\kappa/(2\kappa+1)}$, then the power of the level-α partial permutation test must converge to 1 as the sample size n goes to ∞. Recall the discussion in Section 3.3 and note that the mth-order Sobolev space on $[0, 1]$ corresponds to a RKHS with eigenvalue λ_j decaying polynomially at rate j^{-2m} (Xing et al. 2020). The derived rate with $\kappa = 2m$ actually matches the minimax distinguishable rate $n^{-2m/(4m+1)}$ in Xing et al. (2020) for testing whether two functions in the mth-order Sobolev space are parallel; see also Shang and Cheng (2013).

5. Implementation of Partial Permutation Test

5.1. Choice of the Kernel Function

We first show there exist kernel functions with polynomially decaying eigenvalues as discussed in Sections 3.3 and 3.5. Indeed, as demonstrated by Kühn (1987), such a property holds for a general kernel as long as it is sufficiently smooth. For any set $\mathcal{D} \subset \mathbb{R}^d$ and $0 \le s \le 1$, define $\mathcal{C}^{s,0}(\mathcal{D}, \mathcal{D})$ as the set consisting of all continuous functions $G : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ such that

$$||G||_{\mathcal{C}^{s,0}(\mathcal{D},\mathcal{D})} \equiv \max \left\{ \sup_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in \mathcal{D}} |G(\boldsymbol{x}_1, \boldsymbol{x}_2)|, \sup_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3 \in \mathcal{D}, \boldsymbol{x}_1 \neq \boldsymbol{x}_2} \frac{|G(\boldsymbol{x}_1, \boldsymbol{x}_3) - G(\boldsymbol{x}_2, \boldsymbol{x}_3)|}{||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2^s} \right\} < \infty. \tag{14}$$

The following proposition is a direct corollary of Kühn (1987).

Proposition 2. For any compact set $\mathcal{X} \subset \mathbb{R}^d$ with any probability measure μ and any positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, if there exists b such that (i) $\mathcal{X} \subset \overline{\mathcal{X}} \equiv [-b, b]^d$, (ii) the kernel function K can be extended to domain $\overline{\mathcal{X}} \times \overline{\mathcal{X}}$, and (iii) $K \in C^{s,0}(\overline{\mathcal{X}}, \overline{\mathcal{X}})$, then the corresponding eigenvalues of K, $\{\lambda_k\}_{k \ge 1}$, satisfy that $\lambda_k = O(k^{-1-s/d})$.

From Proposition 2 and by the definition in (14), if a symmetric and positive definite kernel function is continuously differentiable on $\mathbb{R} \times \mathbb{R}$ and the covariate support $\mathcal{X}$ is compact, then the eigenvalue λ_k of the kernel must decay at least in an order of $k^{-1-1/d}$, under which the condition in Theorem 4 holds with $\rho = 1 + 1/d$. Two examples of continuously differentiable kernels are the Gaussian kernel and rational quadratic kernel, which have the following forms:

$$K_{\mathrm{G}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_{k=1}^{d} \omega_k (x_k - x_k')^2 \right\},$$
$$K_{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{x}') = \left\{ 1 + \sum_{k=1}^{d} \omega_k (x_k - x_k')^2 \right\}^{-\eta}, \tag{15}$$

where ω_j's and η are arbitrary positive numbers.

Moreover, both kernels in (15) are also universal (Micchelli, Xu, and Zhang 2006). Thus, if we use any of them for model (8) and let the smoothness parameter be any constant between between 0 and $\min\{1/4, 1/(d+1)\}$, then the conditions in both Proposition 1 and Theorem 4 hold. Consequently, we are able to conduct asymptotically valid partial permutation test under the GPR model, with a certain regularized but still flexible prior for the underlying functional relationship. As discussed shortly in the next subsection, the choice of γ is not crucial in practice. However, the choice of parameters for the kernel function, for example, the ω_j's for the Gaussian kernel in (15), does play an important role.

Parameters in the kernel function play an important role in controlling the smoothness of the underlying functional relationship. For instance, for the Gaussian kernel in (15), smaller ω_j's imply wider, flatter kernels and a suppression of wiggly and rough functions (Hastie and Zhu 2006). In contrast, larger ω_j's indicate a more wiggly functional relation and thus generally lead to a smaller permutation size. Theoretical investigation for the optimal choice of kernel parameters for testing is challenging, and it may differ from that for the optimal estimation (Shang and Cheng 2013; Xing et al. 2020). In the literature, various approaches have been proposed to choose kernel parameters, or more generally kernel functions, adaptively based on the data, such as cross validation and maximizing marginal likelihood (Rasmussen and Williams 2006). We here opt to use the maximum marginal likelihood approach, choosing the kernel parameter to be the one that maximizes the marginal likelihood of $\boldsymbol{Y}$ given $\mathbf{X}$ and $\boldsymbol{Z}$ under the GPR model $\tilde{H}_0$ in (8).

When the data follow an alternative hypothesis model in which the functional relationships for different groups are different, the marginal likelihood for the null, which is based on a common model built using the pooled data, tends to suggest kernels that can tolerate more erratic functions, that is, large values of ω_j's for the Gaussian kernel. This may be due to the fact that, when the data contain multiple functional relationships between the response and covariates, enforcing a common functional relationship necessarily results in an overly volatile function, which then reduces the partial permutation size and damages the power of the test. To avoid this potential power loss, we also obtain kernel parameters that maximize the marginal likelihood using samples from each group separately. If all of them suggest smoother functional relationships (e.g., smaller ω_j's for Gaussian kernels) than the pooled data, we require the smoothness of the shared functional relationship to be no worse than the most non-smooth one among those obtained within each group (e.g., choosing the maximum ω_j's estimated from individual groups).

5.2. Choice of Permutation Size

Both Theorems 3 and 5 provide us with guidance on the choice of permutation size b_n: we want b_n to be large and the correction terms ν in (6) (or $\tilde{\nu}$ in (10)) and α_0 to be small in order to have a good power for the test. Note that either $\nu(b_n, \sigma_0^{-1} f_0, \alpha_0)$ in (6) or $\tilde{\nu}(b_n, \xi_n, \alpha_0)$ in (10) depends on unknown functional relation f_0 and noise level σ_0 or the unknown variance ratio ξ_n. Therefore, we first estimate f_0 and σ_0 (or ξ_n) and then use a plug-in approach to compute ν or $\tilde{\nu}$ under model H_0 or $\tilde{H}_0$. To be more specific, we choose α_0 and b_n in the following way:

1. for model (1) of H_0, $\alpha_0 = 10^{-4}\alpha$, and $b_n = \max\{b_n : \nu(b_n, \hat{\sigma}_0^{-1}\hat{f}_0, \alpha_0) + \alpha_0 \le 10^{-3}\alpha\}$;
2. For model (8) of $\tilde{H}_0$, $\alpha_0 = 10^{-4}\alpha$, and $b_n = \max\{b_n : \tilde{\nu}(b_n, \hat{\xi}_n, \alpha_0) + \alpha_0 \le 10^{-3}\alpha\}$.

There is a tradeoff for the choice of α_0 and b_n: a larger permutation size b_n can lead to a larger power for detecting violation of the null hypothesis while at the same time requires a larger correction to avoid Type I error inflation. Here we consider an intuitive scheme that requires only a small correction for the partial permutation p-value. For model (8), the estimate $\hat{\xi}_n$ can be obtained by using the maximum likelihood estimates for $\delta_0^2/n^{1-\gamma}$ and σ_0^2. For model (1), we can estimate f_0 based on the penalized regression of form (9) or other regularization method such as early stopping (Raskutti, Wainwright, and Yu

2014; Liu and Cheng 2018). Here, for simplicity, we first obtain the posterior mean of f under $\tilde{H}_0$, denoted by $\hat{f}$, after plugging in the maximum likelihood estimates, and then use $\hat{f}$ as an estimator for f_0 and $n^{-1}\sum_{i=1}^{n}(Y_i-\hat{f}(\boldsymbol{X}_i))^2$ as an estimator for the variance of noise. Finally, the corrected p-value is simply the p-value from partial permutation plus the correction term $10^{-3}\alpha$.

5.3. Choice of the Test Statistic

One advantage of the permutation test is that it allows for a flexible choice of test statistics, for which we can use permutations, instead of a complicated and often unreliable asymptotic analysis, to get its reference null distribution. Moreover, we can choose the test statistic tailored to the alternative hypothesis of interest so as to gain power.

For a general kernel function, we first consider test statistics based on kernel regression of form (9). Specifically, we perform kernel regression both to fit a common function using all samples and to fit group-specific functions using samples from each group separately, with, say, cross-validation or marginal likelihood maximization for choosing the regularization parameter τ_n in (9). Motivated by the likelihood ratio test for nested regression models, we compare the mean squared errors from the pooled and group-specific kernel regressions to construct test statistics. For example, we can consider test statistic of the following form:

$$T(\mathbf{X},\boldsymbol{Y},\mathbf{Z}) = n\log(\mathrm{MSE}) - \sum_{h=1}^{H} n_h \log(\mathrm{MSE}_h), \tag{16}$$

where $n_1,\ldots,n_H$ are the group sizes, $\mathrm{MSE} = n^{-1}\sum_{i=1}^{n}(Y_i-\hat{f}(\boldsymbol{X}_i))^2$ with $\hat{f}$ being the kernel regression estimate using all the samples, and $\mathrm{MSE}_h = n_h^{-1}\sum_{i:Z_i=h}(Y_i-\hat{f}_h(\boldsymbol{X}_i))^2$ with $\hat{f}_h$ being the kernel regression estimate using only the samples in group h. Due to the flexibility of the permutation method, we can use loss functions other than the squared loss in (9) to conduct kernel regression, such as the epsilon-intensive loss and Huber loss (see, e.g., Wang 2005; Cavazza and Murino 2016).

We then consider test statistics based on GPR models. We introduce two general alternative models and compute their likelihood ratios against $\tilde{H}_0$. Specifically, we model functions in different groups as dependent Gaussian processes under the alternative hypothesis, and decompose each function into two components, a shared component and a group-specific component, assuming that these components follow independent Gaussian processes with the same general kernel but different variances:

$$\begin{aligned}&\tilde{H}_1: Y_i = f_{Z_i}(\boldsymbol{X}_i)+\varepsilon_i,\quad \varepsilon_i|\boldsymbol{X}_i,Z_i \overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma_0^2),\quad f_h = f+\bar{f}_h,\\ &f\sim \mathrm{GP}\left(0,\frac{\delta_0^2}{n^{1-\gamma}}K\right),\quad \bar{f}_h\sim \mathrm{GP}\left(0,\frac{\delta_h^2}{n^{1-\gamma}}K\right),\end{aligned} \tag{17}$$

where $f,\bar{f}_1,\ldots,\bar{f}_H$ and $\{(\boldsymbol{X}_i,Z_i,\varepsilon_i)\}_{i=1}^n$ are jointly independent. We can further extend the above homoscedastic model to allow noises to have different conditional variances in different groups as follows:

$$\tilde{H}_1': \text{ same as } \tilde{H}_1 \text{ in (17) except that } \varepsilon_i|\boldsymbol{X}_i,Z_i\sim\mathcal{N}(0,\sigma_{Z_i}^2). \tag{18}$$

We define the test statistic based on the likelihood ratio of $\tilde{H}_1$ in (17) (or $\tilde{H}_1'$ in (18)) versus $\tilde{H}_0$ in (8), that is,

$$T(\mathbf{X},\boldsymbol{Y},\mathbf{Z}) = \frac{\max f(\boldsymbol{Y}|\mathbf{X},\mathbf{Z},\tilde{H}_1)}{\max f(\boldsymbol{Y}|\mathbf{X},\mathbf{Z},\tilde{H}_0)}\left(\text{or }\ \frac{\max f(\boldsymbol{Y}|\mathbf{X},\mathbf{Z},\tilde{H}_1')}{\max f(\boldsymbol{Y}|\mathbf{X},\mathbf{Z},\tilde{H}_0)}\right). \tag{19}$$

In the supplementary material, we discuss different ways to compute (19) including the EM algorithm (Dempster, Laird, and Rubin 1977), Newton's method, the Fisher scoring, and quadratic programming.

Here we briefly comment on hypothesis testing of $\tilde{H}_0$ against $\tilde{H}_1$ or $\tilde{H}_1'$. Note that under $\tilde{H}_0$, the variance parameters δ_h^2's are zero and thus are at their boundaries. Therefore, the classical likelihood ratio testing procedure using the chi-squared approximation for the null distribution does not work here. This also suggests the importance and nontriviality of Theorem 4. To reduce the computational cost, we further introduce the following "pseudo" alternative model, which may not contain the null model $\tilde{H}_0$ as a submodel:

$$\begin{aligned}&\tilde{H}_{\text{pseudo}}: Y_i = f_{Z_i}(X_i)+\varepsilon_i,\ \ \varepsilon_i|\boldsymbol{X}_i,Z_i\sim\mathcal{N}(0,\sigma_{Z_i}^2),\\ &f_h\sim \mathrm{GP}\left(0,\frac{\delta_h^2}{n^{1-\gamma}}K\right).\end{aligned} \tag{20}$$

As discussed in the supplementary material, the likelihood ratio of $\tilde{H}_{\text{pseudo}}$ versus $\tilde{H}_0$ can be efficiently computed using the EM algorithm.

6. Extension to Correlated Noises

In the following discussion, we assume that the noises ε_i's are correlated instead of iid as in models H_0 in (1) and $\tilde{H}_0$ in (8), and the covariance matrix of $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^\top$ is known up to a certain positive scale, unless otherwise stated. For example, when the residuals have equal variance, we essentially require that the correlation matrix of $\boldsymbol{\varepsilon}$ is known. In practice, we suggest to first estimate the covariance matrix for $\boldsymbol{\varepsilon}$ based on all the structure information we have (e.g., equal correlations or blockwise independence), and then plug in the estimate to conduct the partial permutation tests described below.

We extend the regression model H_0 in (1) to allow for correlated noises:

$$\begin{aligned}&H_0^{\mathrm{C}}: Y_i = f_0(\boldsymbol{X}_i)+\varepsilon_i,\qquad (1\le i\le n)\\ &\boldsymbol{\varepsilon} = (\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)^\top\sim\mathcal{N}(\mathbf{0},\ \sigma_0^2\boldsymbol{\Sigma}),\end{aligned} \tag{21}$$

where we use the supscript in H_0^{C} to emphasize that the noises under model (21) are allowed to be correlated. Moreover, we assume that $\boldsymbol{\Sigma}$ is known and positive definitive but σ_0^2 can be unknown, that is, the covariance matrix of $\boldsymbol{\varepsilon}$ is known up to a positive scale. Recall that $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\top$ and $\boldsymbol{f}_0 = (f_0(\boldsymbol{X}_1),\ldots,f_0(\boldsymbol{X}_n))^\top$. Under H_0^{C} in (21), we have $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{f}_0 + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}$, where $\boldsymbol{\Sigma}^{-1/2}$ is the inverse of the positive

definitive square root of $\boldsymbol{\Sigma}$. By our model assumption, it is easy to see that the elements of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}$ are iid Gaussian with mean zero and variance σ_0^2. This then motivates us to consider a partial permutation test based on response vector $\boldsymbol{Y}^{\mathrm{C}} \equiv \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}$ and sample "kernel" matrix $\boldsymbol{K}_n^{\mathrm{C}} \equiv \boldsymbol{\Sigma}^{-1/2}\boldsymbol{K}_n\boldsymbol{\Sigma}^{-1/2}$. More precisely, in Algorithm 1, we replace $\boldsymbol{Y}$ and $\boldsymbol{K}_n$ by $\boldsymbol{Y}^{\mathrm{C}}$ and $\boldsymbol{K}_n^{\mathrm{C}}$, and denote the resulting p-value by $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma})$, which depends crucially on the noise covariance structure $\boldsymbol{\Sigma}$.

By the same logic as Theorem 3, we can derive a finite-sample valid partial permutation test with a certain correction on the permutation p-value. For $1 \le b_n \le n$ and $0 < \alpha_0 < 1$, we define $\omega_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma}) = \sigma_0^{-2}\sum_{i=n-b_n+1}^{n}(\boldsymbol{\gamma}_i^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{f}_0)^2$ to denote the LOSP for the components used for partial permutation, and

$$v_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma}, \alpha_0) = \frac{1}{2}\exp\left\{2\sqrt{2\omega_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma})}\cdot \sqrt{Q_{b_n}(1-\alpha_0) + \omega_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma})}\right\} - \frac{1}{2}, \tag{22}$$

where Q_{b_n} denotes the quantile function of the $\chi^2_{b_n}$-distribution.

Theorem 8. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model H_0^{C} in (21). Given $1 \le b_n \le n$ and $\alpha_0 \in (0,1)$, we define the corrected partial permutation p-value as

$$p_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) = p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) + v_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma}, \alpha_0) + \alpha_0,$$

where $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma})$ is the p-value from either the discrete or continuous partial permutation test based on kernel K, permutation size b_n, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$, and $v_{\mathrm{C}}(b_n, \sigma_0^{-1}f_0, \boldsymbol{\Sigma}, \alpha_0)$ is as defined in (22). Then the corrected partial permutation p-value is valid under model H_0^{C}, that is, $\forall\alpha \in (0,1)$, $\Pr_{H_0^{\mathrm{C}}}\{p_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

Again, it is generally difficult to show the asymptotic validity of the p-value $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma})$ for a general kernel under general underlying function and noise covariance structure, and its correction term in (22) depends on the unknown f_0 and σ_0. In practice, we can adopt similar strategies as discussed in Section 5. Below we consider four special cases, in parallel to Sections 3.2–3.5, under which we can demonstrate the exact or asymptotic validity of the partial permutation test that takes into account the noise covariance structure.

6.1. Special Case: Kernels with Finite-Dimensional Feature Space

When the kernel has a finite-dimensional feature space and the underlying function is linear in features mapped to this space, the partial permutation test is exactly valid.

Corollary 4. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model H_0^{C} in (21). Suppose kernel function K has the decomposition $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\phi}(\boldsymbol{x}')$ with $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^q$ for some $q < \infty$, and the underlying function $f_0(\boldsymbol{x})$ is linear in $\boldsymbol{\phi}(\boldsymbol{x})$. Then, the p-value obtained by either the discrete or continuous partial permutation test with kernel K, permutation size $b_n \le n - q$, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$ is valid, that is, $\forall\alpha \in (0,1)$, $\Pr_{H_0^{\mathrm{C}}}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

6.2. Special Case: Kernels With Diverging-Dimensional Feature Space

Similar to Sections 3.3 and 4.2, we consider kernels with diverging-dimensional feature space, that is, $K_q(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}_q(\boldsymbol{x})^\top\boldsymbol{\phi}_q(\boldsymbol{x}')$ for $q \ge 1$ with $\boldsymbol{\phi}_q(\boldsymbol{x}) = (e_1(\boldsymbol{x}), e_2(\boldsymbol{x}), \ldots, e_q(\boldsymbol{x}))^\top$ and $\{e_j\}_{j=1}^{\infty}$ being a series of basis functions. We assume that the covariates are identically distributed from some probability measure μ, and use $\mathrm{r}(f; q) = \min_{\boldsymbol{b}\in\mathbb{R}^q}\int(f - \boldsymbol{b}^\top\boldsymbol{\phi}_q)^2\mathrm{d}\mu$ to denote the squared error for the best linear approximation of f using the first q basis functions. The following corollary shows that the partial permutation test is asymptotically valid when the underlying functional relationship can be well approximated by the basis functions and the smallest eigenvalue of the noise covariance matrix $\lambda_{\min}(\boldsymbol{\Sigma})$ decays not too fast.

Corollary 5. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model H_0^{C} in (21), and assume that $\boldsymbol{X}_i$'s are identically distributed from some probability measure μ. Suppose that the kernel function K_q has the form $K_q(\boldsymbol{x}, \boldsymbol{x}') \equiv \boldsymbol{\phi}_q(\boldsymbol{x})^\top\boldsymbol{\phi}_q(\boldsymbol{x}') \equiv \sum_{j=1}^{q} e_j(\boldsymbol{x})e_j(\boldsymbol{x}')$ for $q \ge 1$ and some series of basis functions $\{e_j\}_{j=1}^{\infty}$. If there exists a sequence $\{q_n\}_{n=1}^{\infty}$ such that $q_n < n$ for all n and $n(n-q_n)\mathrm{r}(f_0; q_n)/\lambda_{\min}(\boldsymbol{\Sigma}) \to 0$ as $n \to \infty$, then the resulting p-value obtained by either the discrete or continuous partial permutation test with kernel K_{q_n}, permutation size $b_n \le n - q_n$, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$ is asymptotically valid, that is, $\forall\alpha \in (0,1)$, $\limsup_{n\to\infty}\Pr_{H_0^{\mathrm{C}}}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha\} \le \alpha$.

6.3. Special Case: Exactly Balanced Covariates across All Groups

In the case that the covariates are exactly balanced across all groups as in (7) and the kernel matrix for distinct covariates within each group is of full rank (which generally holds when the kernel has an infinite-dimensional feature space, for example, the Gaussian kernel), the following corollary shows that the partial permutation test is exactly valid under a general functional relationship.

Corollary 6. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model H_0^{C} in (21). If the design matrix is exactly balanced in the sense that (7) holds and the kernel matrix for the $r \le n/H$ distinct covariates within each group is of full rank (or equivalently $\mathrm{rank}(\boldsymbol{K}_n) = r$), then the partial permutation p-value from either the discrete or continuous partial permutation test with kernel K, permutation size $b_n \le n - r$, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$ is valid under model H_0^{C}, that is, $\forall\alpha \in (0,1)$, $\Pr_{H_0^{\mathrm{C}}}\{p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha | \mathbf{X}, \boldsymbol{Z}\} \le \alpha$.

Furthermore, if all covariates within each group are distinct and the covariance among noises enjoys the following structure: (i) the noises have equal variances, (ii) the noises for samples with different covariates are uncorrelated, and (iii) the noises for samples with the same covariates are equally correlated with correlation ρ, then the partial permutation test is always valid even if we use a correlation matrix with incorrect correlation $\tilde{\rho} \ne \rho$. This means that, with this special covariance structure, we are able to conduct valid permutation tests even if the true

correlation matrix is unknown. Such covariance structure is reasonable when the same covariate corresponds to the same individual and the response within each group corresponds to measurement at different time periods. As a side note, the usual permutation test that switches group indicators of samples with the same covariates is valid in more general setting, as long as the noises for samples with different covariate values are mutually independent and the noises for samples with the same covariate values are exchangeable. The partial permutation test allows for more general permutation or rotation, but with a stronger Gaussianity assumption on the noises.

6.4. Special Case: Gaussian Process Regression Model

Finally we extends the GPR model $\tilde{H}_0$ in (8) to allow for correlated noises:

$$\tilde{H}_0^{\mathrm{C}} : Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad \boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}),$$
$$f \sim \mathrm{GP}\left(0, \frac{\delta_0^2}{n^{1-\gamma}} K(\cdot,\cdot)\right). \tag{23}$$

The following theorem extends Theorem 4 and demonstrates the asymptotic validity of the partial permutation test after taking into the account the noise covariance structure.

Theorem 9. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model $\tilde{H}_0^{\mathrm{C}}$ in (23). If the covariates $\boldsymbol{X}_i$'s are iid from a compact support $\mathcal{X}$ with probability measure μ, the eigenvalues $\{\lambda_k\}$ of kernel K on $(\mathcal{X}, \mu)$ satisfy $\lambda_k = O(k^{-\rho})$ with $\rho > 1$, the smallest eigenvalue of $\boldsymbol{\Sigma}$ for the noises satisfies $\lambda_{\min}(\boldsymbol{\Sigma}) \ge cn^{-\zeta}$ for some positive c and $\zeta < 1 - \rho^{-1}$, and γ is a constant less than $1 - \rho^{-1} - \zeta$, then for sequence $\{b_n\}$ satisfying $b_n = O(n^{\kappa})$ with $0 < \kappa < 1 - \rho^{-1} - \zeta - \gamma$, the partial permutation p-value from either the discrete or continuous partial permutation test with kernel K, permutation size b_n, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$ is asymptotically valid under $\tilde{H}_0^{\mathrm{C}}$, that is, $\forall \alpha \in (0,1)$, $\limsup_{n\to\infty} \Pr_{\tilde{H}_0^{\mathrm{C}}}\{p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha\} \le \alpha$.

Theorem 4 proves the large-sample validity of the partial permutation test. Below we investigate its finite-sample performance. Analogous to Theorem 5, let $\xi_n = (\delta_0^2/n^{1-\gamma})/\sigma_0^2$ denote the variance ratio, and $\tilde{\omega}_{\mathrm{C}}(b_n, \xi_n, \boldsymbol{\Sigma}) = \xi_n \cdot \zeta_{n-b_n+1}$ denote the LOSP, where ζ_{n-b_n+1} denotes the $(n - b_n + 1)$th largest eigenvalue of $\boldsymbol{K}_n^{\mathrm{C}}$. We then define

$$\tilde{\nu}_{\mathrm{C}}(b_n, \xi_n, \boldsymbol{\Sigma}, \alpha_0) = \frac{1}{2} \exp\left[\frac{1}{2}\tilde{\omega}_{\mathrm{C}}(b_n, \xi_n, \boldsymbol{\Sigma}) \cdot Q_{b_n}(1-\alpha_0)\right] - \frac{1}{2}, \tag{24}$$

recalling that Q_{b_n} is the quantile function of the $\chi^2_{b_n}$-distribution. The following theorem shows that the partial permutation p-value can be finite-sample valid under $\tilde{H}_0^{\mathrm{C}}$ after an adjustment.

Theorem 10. Let $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{1\le i\le n}$ denote samples from model $\tilde{H}_0^{\mathrm{C}}$ in (23). Given $1 \le b_n \le n$ and $0 < \alpha_0 < 1$, we define the corrected partial permutation p-value as follows,

$$\tilde{p}_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) = p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) + \tilde{\nu}(b_n, \xi_n, \boldsymbol{\Sigma}, \alpha_0) + \alpha_0,$$

where $p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma})$ is the p-value from either the discrete or continuous partial permutation test with kernel K, permutation size b_n, any test statistic T, and covariance matrix $\boldsymbol{\Sigma}$, and $\tilde{\nu}(b_n, \xi_n, \boldsymbol{\Sigma}, \alpha_0)$ is as defined in (10). Then the corrected partial permutation p-value is valid under model $\tilde{H}_0^{\mathrm{C}}$, that is, $\forall \alpha \in (0,1)$, $\Pr_{\tilde{H}_0^{\mathrm{C}}}\{\tilde{p}_{\mathrm{c}}(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \le \alpha | \mathbf{X}, \mathbf{Z}\} \le \alpha$.

By the same logic as Section 5.2, we can then use Theorem 10 to guide the choice of permutation size in finite samples.

7. Simulation Study

In this section, we present simulation results based on various choices of the kernels and discrete partial permutation tests described in Algorithm 1. Specifically, in Sections 7.1–7.3, we investigate Type I error control under the null hypothesis, and in Sections 7.4 and 7.5 we compare powers of the partial permutation test and some other methods. We also conduct simulations with non-Gaussian or correlated noises, which are relegated to supplementary material. Moreover, for simulation under the null hypothesis, we focus mainly on the Gaussian kernel and choose the tuning parameters, permutation size b_n and test statistic T, as follows: (i) we standardize both the response and covariates, and consider Gaussian kernel K_{G} in (15) with the maximum marginal likelihood estimates for parameters ω_k's as discussed in Section 5.1; (ii) we choose the permutation size b_n as suggested in Section 5.2 based on model $\tilde{H}_0$ with significance level $\alpha = 0.05$; (iii) we choose the likelihood ratio of $\tilde{H}_{\mathrm{pseudo}}$ versus $\tilde{H}_0$ as the test statistic T due to its lower computation cost, unless otherwise stated.

7.1. Simulation Under the Null Hypothesis With Scalar Covariate

We first consider partial permutation test under H_0 with a scalar covariate and two groups. We generate data as iid samples from the following model:

$$\begin{aligned} \text{Scenario 1:} \quad & Y = f_0(X) + \varepsilon, \quad \varepsilon|X, Z \sim \mathcal{N}(0, \sigma_0^2), \\ & X|Z \sim a_Z \cdot \mathrm{Unif}(-1, 0) + (1 - a_Z) \cdot \mathrm{Unif}(0, 1), \\ & \Pr(Z = h) = p_h, \quad h = 1, 2, \end{aligned} \tag{25}$$

where $\mathrm{Unif}(-1,0)$ and $\mathrm{Unif}(0,1)$ refer to uniform distributions on $(-1,0)$ and $(0,1)$, (a_1, a_2) control the mixture weights for covariate distributions in two groups, and (p_1, p_2) denote the fractions of observations (in expectation) from two groups. We consider the five cases in Table 1 that vary both the proportions of units and the covariate distributions in two groups. Specifically, in case (e), covariates from the two groups do not overlap at all. Therefore, case (e) resembles the regression discontinuity design, under which we can interpret the null hypothesis H_0 in (1) as that the underlying functions for the two groups can be smoothly connected at the boundary. Finally, we fix $\sigma_0^2 = 0.1$ for all cases in Table 1, and consider the following six choices of the underlying function f_0, all in the range of $[-1, 1]$:

$$\begin{aligned} &\text{(i) } f_0 = x, && \text{(ii) } f_0 = 2x^2 - 1, \\ &\text{(iii) } f_0 = 4x^3/3 - x/3, && \text{(iv) } f_0 = 4/(1 + x^2) - 3, \\ &\text{(v) } f_0 = \sin(4x), && \text{(vi) } f_0 = \sin(6x). \end{aligned} \tag{26}$$

Table 1. Cases with varying balancedness of group sizes and covariate distributions between the two groups in comparison.

Case	Groups	Covariates	(p_1, p_2)	(a_1, a_2)
(a)	Balanced	Balanced	(0.5, 0.5)	(0.5, 0.5)
(b)	Unbalanced	Balanced	(0.2, 0.8)	(0.5, 0.5)
(c)	Balanced	Unbalanced	(0.5, 0.5)	(0.8, 0.2)
(d)	Unbalanced	Unbalanced	(0.2, 0.8)	(0.8, 0.2)
(e)	Balanced	Non-overlap	(0.5, 0.5)	(1, 0).

Figure 1 shows the empirical distribution functions of the partial permutation p-values under all cases with sample size $n = 200$, showing that all are very close to Unif(0, 1) and demonstrating the validity of the partial permutation test.

7.2. Simulation Under the Null Hypothesis With Two-Dimensional Covariates

We generate data as iid samples from the following two-dimensional covariates model:

$$\text{Scenario 2:} \quad Y = f_0(X_1, X_2) + \varepsilon, \quad \varepsilon | \boldsymbol{X}, Z \sim \mathcal{N}(0, \sigma_0^2),$$
$$X_1 | Z \sim a_Z \cdot \text{Unif}[-1, 0] + (1 - a_Z) \cdot \text{Unif}[0, 1],$$
$$X_2 | Z \sim a_Z \cdot \text{Unif}[-1, 0] + (1 - a_Z) \cdot \text{Unif}[0, 1],$$
$$X_1 \perp\!\!\!\perp X_2 | Z, \quad P(Z = h) = p_h, \quad h = 1, 2, \quad (27)$$

where the choice of (a_1, a_2) and (p_1, p_2) is the same as in Table 1. We again fix $\sigma_0^2 = 0.1$ and consider the following six choices of the underlying function f_0, all in the range of $[-1, 1]$:

$$\begin{aligned} &\text{(i) } f_0 = (x_1 + x_2)/2, && \text{(ii) } f_0 = x_1 x_2, \\ &\text{(iii) } f_0 = 2(x_1 + x_2)^3/15 \\ &\qquad -(x_1 + x_2)/30, && \text{(iv) } f_0 = 3/(1 + x_1^2 + x_2^2) - 2, \\ &\text{(v) } f_0 = \sin(6x_1) + x_2, && \text{(vi) } f_0 = \sin(6x_1 + 6x_2). \end{aligned} \quad (28)$$

Figure 2 shows the empirical distribution functions of the partial permutation p-values, which are close to Unif(0, 1) for all cases.

7.3. Simulation Under the Null Hypothesis With Non-Smooth Functions

In the previous two subsections, we focus on null hypothesis with smooth functions. Here, we consider the following contin-

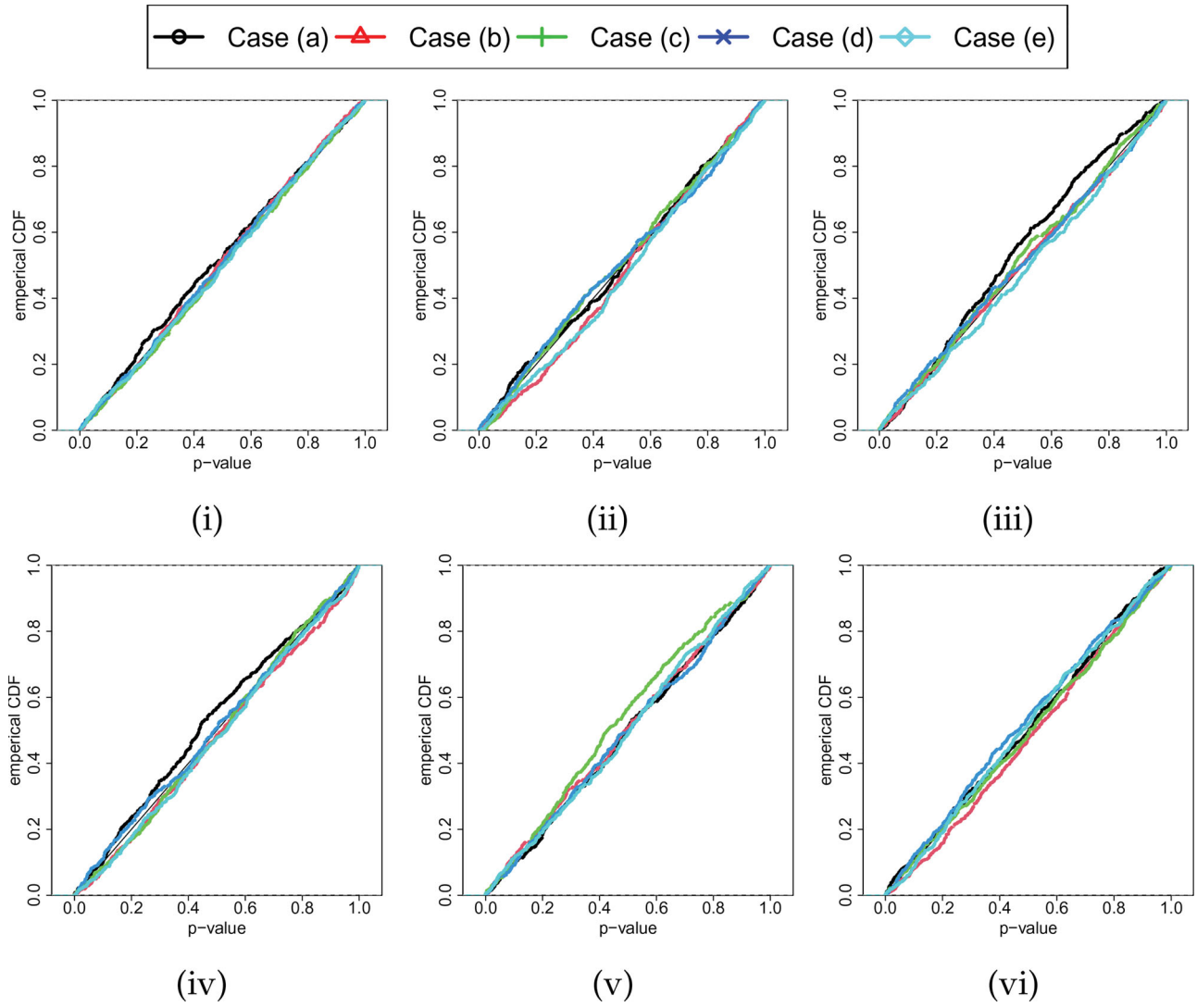

Figure 1. Empirical distributions of the partial permutation p-values when data are generated from Scenario 1 in (25) under all cases in Table 1 with sample size $n = 200$. The six figures correspond to six choices of the underlying function f_0 in (26).

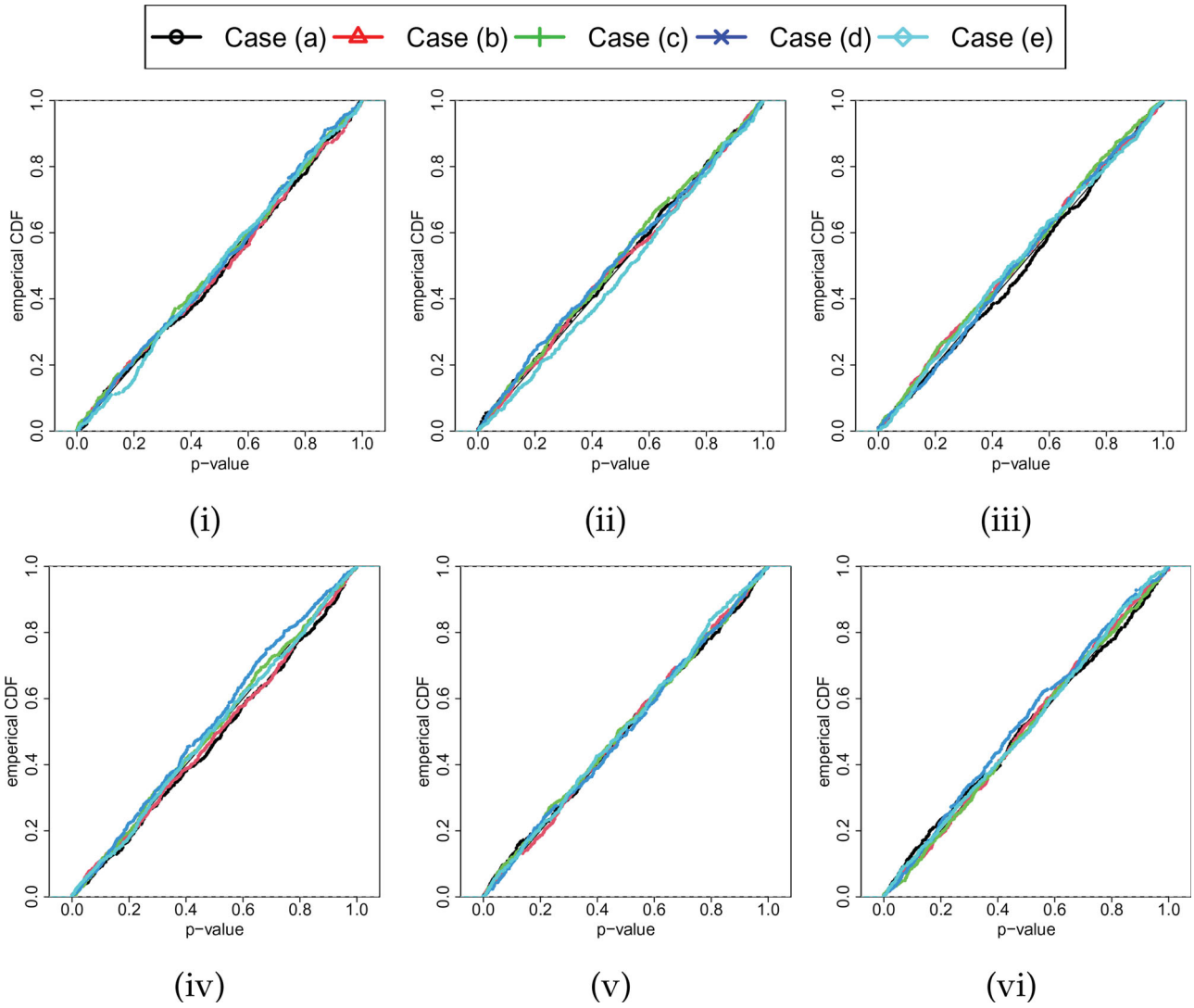

Figure 2. Empirical distributions of the partial permutation p-values when data are generated from Scenario 2 in (27) under all cases in Table 1 with sample size $n = 200$. The six figures correspond to six choices of the underlying function f_0 in (28).

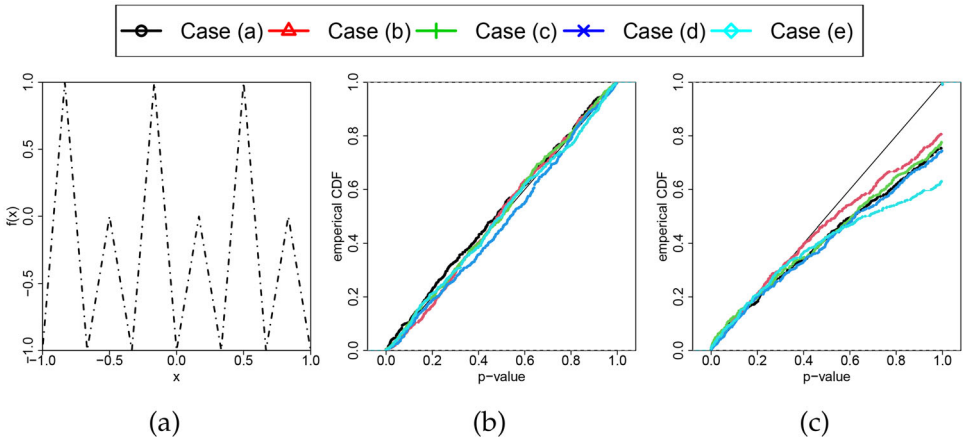

Figure 3. (a) Plot of function $g_0(x)$ in (29). (b) and (c): Empirical distributions of the partial permutation p-values when data are generated from models (25) with underlying function $g_0(x)$, and model (27) with the underlying function $g_0(x_1)g_0(x_2)$, respectively.

uous but non-differentiable univariate function:

$$g_0(x) = 2 * \min\{|3x - \lfloor 3x \rfloor|, |3x - \lfloor 3x \rfloor - 1|\} \cdot (\lfloor 3x \rfloor \bmod 2 + 1) - 1, \tag{29}$$

where $\lfloor 3x \rfloor$ denotes the largest integer less than or equal to $3x$ and $(\lfloor 3x \rfloor \bmod 2)$ denotes the remainder of $\lfloor 3x \rfloor$ divided by 2. Figure 3(a) shows the shape of $g_0(x)$.

We consider simulations from model (25) with a single covariate and function $f_0(x) = g_0(x)$, and from model (27) with two covariates and function $f_0(x_1, x_2) = g_0(x_1)g_0(x_2)$, with sample size $n = 200$. Figures 3(b) and (c) show the empirical distributions of the partial permutation p-values, for models (25) and (27), respectively, under the five cases with varying imbalance in covariate distributions and group sizes as shown in Table 1, which demonstrate that the Type I error is still approximately controlled. Note that, with two-dimensional covariates, the distributions of the partial permutation p-values are quite different from Unif(0, 1), and the p-values appear to be slightly conservative at significance levels higher than 0.3. The reason is that, over all simulations, about 25% of the time the partial permutation test has permutation size 1 and thus results in p-value equal to 1. Such extreme permutation size is due to the non-smoothness of the underlying functional relationship, under which we lack enough permutation size as well as power for rejecting the null hypothesis. This is also intuitive as it is difficult to distinguish whether the multiple groups in comparison share the same functional relation if the underlying function is very nonsmooth. In such cases, a conservative p-value is preferred so as to avoid inflating the Type I error.

7.4. Power Comparison With the Classical F-Test Under the Alternative Hypotheses

We generate iid samples from the following two data generating scenarios (under alternative hypotheses) with one- and two-dimensional covariates:

$$\begin{aligned} \text{Scenario 3:}\quad & Y = f_Z(X) + \varepsilon,\ \ \varepsilon|X, Z \sim \mathcal{N}(0, \sigma_0^2), \\ & X|Z \sim \text{Unif}(-1, 1), \\ & P(Z = h) = p_h, \quad h = 1, 2, \end{aligned} \tag{30}$$

and

$$\begin{aligned} \text{Scenario 4:}\quad & Y = f_Z(X_1, X_2) + \varepsilon,\ \ \varepsilon|\boldsymbol{X}, Z \sim \mathcal{N}(0, \sigma_0^2), \\ & X_k|Z \sim \text{Unif}(-1, 1), \quad k = 1, 2, \\ & X_1 \perp\!\!\!\perp X_2|Z,\ \ P(Z = h) = p_h, \quad h = 1, 2. \end{aligned} \tag{31}$$

For Scenario 3, we consider the following three choices of (f_1, f_2):

$$\begin{aligned} &\text{(i) } f_1 = 1 + x, && f_2 = 2 + 3x, \\ &\text{(ii) } f_1 = 1/3 + x/2, && f_2 = (x + 1)^2/4, \\ &\text{(iii) } f_1 = 1/3 + x/2, && f_2 = 1/5 + x/2 - x^4 + x^2; \end{aligned} \tag{32}$$

For Scenario 4, we consider the following three choices of (f_1, f_2):

$$\begin{aligned} &\text{(iv) } f_1 = 1 + x_1 + x_2, && f_2 = 2 + 3x_1 + x_2, \\ &\text{(v) } f_1 = 1/3 + x_1/2 + x_2/2, && f_2 = (x_1 + 1)^2/4 \\ & && \quad +(x_2 + 1)^2/4 - 1/3, \\ &\text{(vi) } f_1 = 1/3 + x_1/2 + x_2/2, && f_2 = 1/3 + x_1/2 + x_2/2 \\ & && \quad + \sin(\pi x_1) \cdot \sin(\pi x_2). \end{aligned} \tag{33}$$

We conduct partial permutation test using either the Gaussian or polynomial kernels. For the Gaussian kernel, we consider three choices of test statistics, the likelihood ratio (19) of $\tilde{H}_1$ against $\tilde{H}_0$, the pseudo likelihood ratio of $\tilde{H}_{\text{pseudo}}$ against $\tilde{H}_0$, and (16) based on the mean squared errors from the pooled and group-specific kernel regression, and choose the permutation size based on $\tilde{H}_0$ as discussed in Section 5.2. For polynomial kernels, we consider degree p of 1, 2 and 3, use the likelihood ratio of the model where the underlying functions are polynomial of degree up to p and can vary across groups against that with the same polynomial function of degree up to p across all groups, and choose the permutation size based on Theorem 2. We also consider the classical F-test or equivalently the likelihood ratio test for whether the functions for different groups are the same polynomial function of degree p, for $p = 1, 2, 3$. Here, the F-test is considered to be most powerful as long as the polynomial regression model is true within each group and does not include unnecessary higher order terms.

Figure 4 shows the power of different tests. Since the partial permutation tests using polynomial kernels have almost the same power as the corresponding F-tests, which is not surprising given Theorem 6, they are omitted in Figure 4. As shown in Figures 4(i), (ii), (iv), and (v), when the underlying functions are indeed polynomial, the F-test with the correct degrees of freedom is the most powerful one. However, as suggested by Figures 4(ii) and (v), if we fail to include some higher order terms, it is possible that the F-tests have almost no power to detect the functional heterogeneity across two groups. Furthermore, the powers of the partial permutation test using the Gaussian kernel with either test statistic (16) or the pseudo likelihood ratio statistic are similar and are also close to that of the corresponding most powerful F-test, although the gap seems to increase with the dimension of the covariates. They both performed better than that with the likelihood ratio statistic of $\tilde{H}_1$ versus $\tilde{H}_0$, partly because the former two consider different noise variances in different groups. Finally, as shown in Figures 4(iii) and (vi), when the underlying functions contain either higher-order or non-polynomial terms, the partial permutation test using Gaussian kernel can have a much higher power than the classical F-test.

7.5. Power Comparison With Other Nonparametric Methods Under Balanced Covariates

Our partial permutation test focuses on whether samples from different groups share the same functional relationship. This is closely related to the literature focusing on whether different groups share parallel functional relationship (Degras et al. 2012; Xing et al. 2020). Specifically, with exactly balanced covariates as in (7) and centered response within each group (assuming the true average function values within each group is known), the groups in comparison shares parallel functional relation if and only if they share the same functional relation. Following Xing et al. (2020), we generate data from the following model:

$$\begin{aligned} \text{Scenario 5:}\quad & Y_i = f_{Z_i}(X_i) + \varepsilon_i, \quad \varepsilon_i|X_i, Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2), \\ & X_i \equiv X_{n/2+i} \overset{\text{iid}}{\sim} \text{Unif}(0, 1), \quad (1 \le i \le n/2) \\ & Z_1 = \cdots = Z_{n/2} = 1, \quad Z_{n/2+1} = \cdots = Z_n = 2. \end{aligned} \tag{34}$$

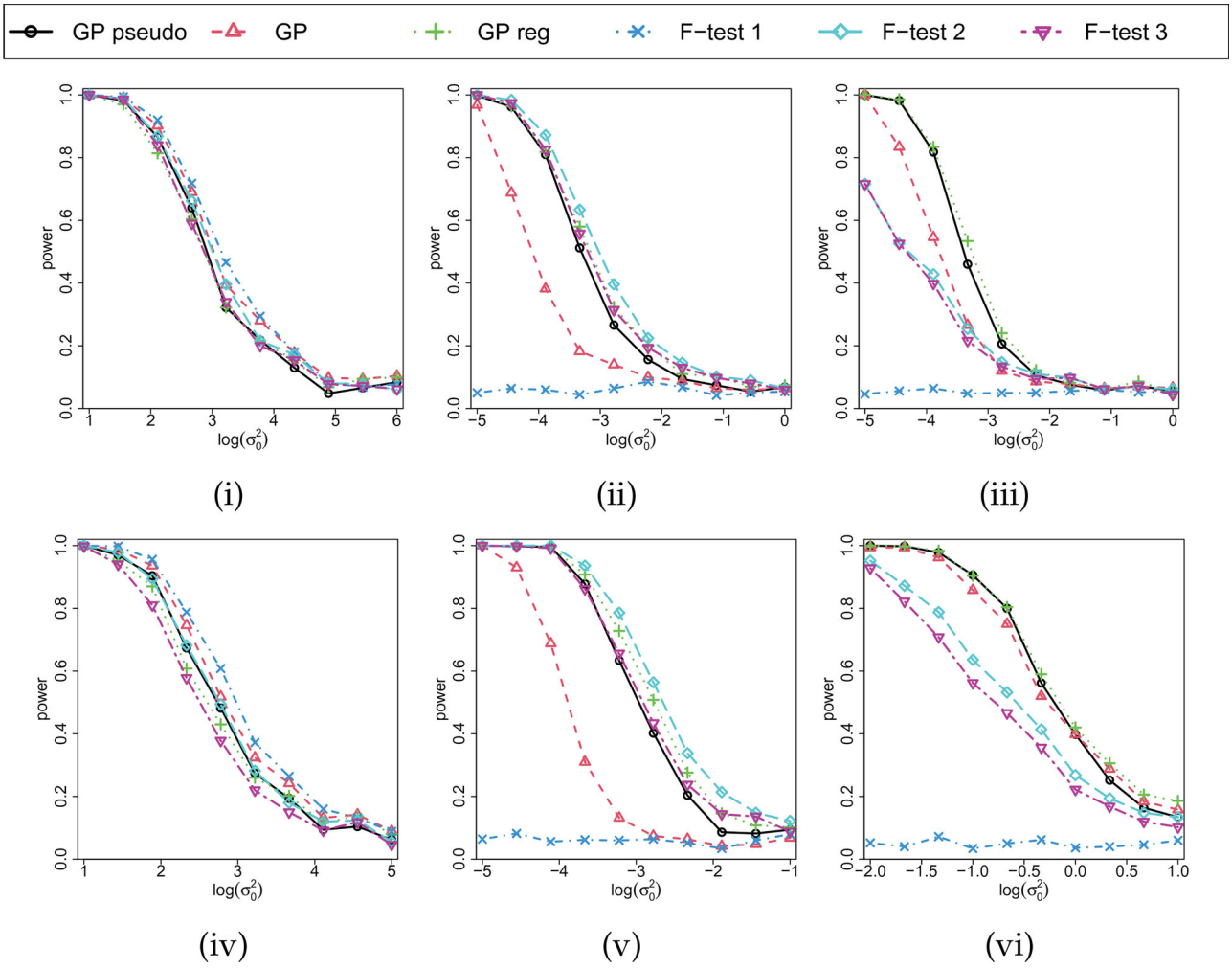

Figure 4. Power of the partial permutation tests when data are generated from Scenario 3 in (30) and Scenario 4 in (31) with sample size $n = 200$. The six figures correspond to six choices of the underlying functions f_1 and f_2 in (32) and (33). The partial permutation tests here use Gaussian kernel with three choices of test statistics, the likelihood ratio (19) of $\tilde{H}_1$ against $\tilde{H}_0$ (denoted by GP), the pseudo likelihood ratio of $\tilde{H}_{\text{pseudo}}$ against $\tilde{H}_0$ (denoted by GP pseudo), and (16) based on the mean squared errors (denoted by GP reg). The F-tests test whether the functions for different groups are the same polynomial functions of degree p (denoted by F-test p), for $p = 1, 2, 3$.

Table 2. Comparison between the parallelism and partial permutation tests.

Result	Method	0.01	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50
Power	Parallel	1.000	0.920	0.624	0.484	0.378	0.312	0.300	0.270	0.244	0.196
under	PPT	1.000	0.824	0.482	0.342	0.280	0.216	0.200	0.156	0.172	0.152
H_1	PPT + Parallel	1.000	0.836	0.490	0.352	0.256	0.188	0.206	0.168	0.162	0.114
Type I	Parallel	0.146	0.130	0.138	0.090	0.112	0.070	0.110	0.092	0.112	0.078
error	PPT	0.044	0.056	0.058	0.046	0.046	0.036	0.044	0.042	0.050	0.042
under H_0	PPT + Parallel	0.062	0.062	0.054	0.040	0.064	0.040	0.036	0.040	0.040	0.032
Corrected	Parallel	1.000	0.838	0.476	0.376	0.242	0.238	0.230	0.204	0.186	0.140
Power	PPT	1.000	0.812	0.468	0.346	0.296	0.266	0.216	0.192	0.172	0.166
under H_1	PPT + Parallel	1.000	0.798	0.484	0.394	0.242	0.248	0.230	0.202	0.176	0.150

NOTES: Data are generated from model in (34) with functions in (35) (i) and sample size $n = 200$. The heading row indicates various noise levels. PPT and PPT+Parallel refer to the partial permutation tests using the pseudo likelihood ratio and the minus p-values from the parallelism test, respectively, as test statistics.

and consider the following two choices in which the two groups share neither the same nor parallel functional relationships:

$$\begin{aligned} &\text{(i)}\ f_1 = 2.5 \cdot \sin(3\pi x) \cdot (1-x) - m_1, \\ &\qquad f_2 = 3.5 \cdot \sin(3\pi x) \cdot (1-x) - m_2, \\ &\text{(ii)}\ f_1 = 2.5 \cdot \sin(3\pi x) \cdot (1-x) - m_1, \\ &\qquad f_2 = 2.5 \cdot \sin(3.4\pi x) \cdot (1-x) - m_3, \end{aligned} \tag{35}$$

To make the comparison fairer, we choose constants m_1, m_2 and m_3 such that each function has mean zero, that is, $\mathbb{E}(f_k(X)) = 0$ with $X \sim \text{Unif}(0,1)$, which helps avoid the partial permutation test to gain additional power by the mean shift. Tables 2 and 3 show the power of the partial permutation test using the pseudo likelihood ratio as the test statistic and that of the minimax nonparametric parallelism test in Xing et al. (2020), which was shown to be superior to other tests in the literature under similar simulation settings. For Tables 2 and 3, we let the sample size $n = 200$, the noise level σ_0^2 vary in [0.01, 4.5], and the significance level be fixed at 0.05. Besides, we also

Table 3. Comparison between the parallelism and partial permutation tests.

Result	Method	0.01	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50
Power	Parallel	1.000	0.886	0.642	0.440	0.354	0.286	0.262	0.218	0.208	0.188
under	PPT	1.000	0.764	0.482	0.288	0.202	0.160	0.178	0.098	0.122	0.092
H_1	PPT + Parallel	1.000	0.764	0.486	0.290	0.232	0.170	0.156	0.122	0.126	0.102
Type I	Parallel	0.144	0.128	0.140	0.092	0.108	0.076	0.112	0.096	0.110	0.082
error	PPT	0.046	0.056	0.050	0.050	0.044	0.034	0.044	0.036	0.050	0.048
under H_0	PPT + Parallel	0.062	0.062	0.056	0.042	0.058	0.042	0.036	0.040	0.038	0.038
Corrected	Parallel	1.000	0.764	0.480	0.308	0.200	0.198	0.172	0.144	0.146	0.138
Power	PPT	1.000	0.752	0.482	0.288	0.224	0.206	0.190	0.126	0.122	0.100
under H_1	PPT + Parallel	1.000	0.730	0.474	0.320	0.192	0.210	0.176	0.136	0.142	0.156

NOTES: Data are generated from (34) with functions in (35)(ii) and sample size $n = 200$. The description of the table is the same as that of Table 2.

Table 4. Comparison between the parallelism and partial permutation tests with sample sizes $n = 500$ and 1000.

n	Result	Method	0.01	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50
500	Power	Parallel	1.000	1.000	0.958	0.838	0.688	0.624	0.534	0.452	0.392	0.384
	under H_1	PPT	1.000	1.000	0.892	0.718	0.580	0.492	0.424	0.350	0.300	0.262
	Type I error	Parallel	0.078	0.086	0.098	0.082	0.086	0.082	0.096	0.068	0.092	0.09
	under H_0	PPT	0.044	0.054	0.042	0.046	0.058	0.070	0.064	0.050	0.038	0.05
	Corrected Power	Parallel	1.000	0.998	0.936	0.786	0.612	0.500	0.462	0.35	0.290	0.322
	under H_1	PPT	1.000	0.998	0.898	0.732	0.542	0.442	0.402	0.35	0.336	0.282
1000	Power	Parallel	1.000	1.000	1.000	0.988	0.946	0.886	0.836	0.724	0.716	0.634
	under H_1	PPT	1.000	1.000	1.000	0.964	0.892	0.812	0.752	0.652	0.604	0.494
	Type I error	Parallel	0.064	0.078	0.074	0.072	0.080	0.078	0.074	0.088	0.074	0.066
	under H_0	PPT	0.034	0.036	0.066	0.064	0.046	0.050	0.042	0.062	0.036	0.044
	Corrected Power	Parallel	1.000	1.000	1.000	0.982	0.918	0.858	0.820	0.652	0.674	0.586
	under H_1	PPT	1.000	1.000	1.000	0.962	0.894	0.812	0.786	0.604	0.648	0.520

NOTES: Data are generated from (34) with functions in (35)(i). The description of the table is the same as that of Table 2, except that here we do not consider "PPT+Parallel."

consider cases where the functions in both groups are the same as f_1 to investigate the Type I error of these tests.

Tables 2 and 3 show that, although the parallelism test has a better power, its Type I error is significantly inflated. In contrast, the partial permutation test controls its Type I errors well at the nominal level. After correcting the Type I error by using the 0.05 quantile of the null distribution (i.e., the functions in both groups are the same as f_1 in (35)) of the p-value as the threshold, the power of the two tests becomes similar. We further increase the sample size to $n = 500$ and 1000. As shown in Table 4, Type I errors of the partial permutation test are always well controlled, whereas those of the parallelism test are still inflated but are closer to the nominal level as the sample size increases. The two tests always have similar powers after the Type I error correction.

Note that the partial permutation test allows for an arbitrary choice of the test statistic. As shown in Tables 2 and 3, we also use the minus p-value from the parallelism test as our test statistic. From Tables 2 and 3, the resulting Type I error is well controlled and the power is similar to the original parallelism test after correcting the inflated Type I error. In practice, however, such Type I error corrections cannot be easily achieved since the underlying true functions are unknown. We may use the distribution from the partial permutation as a reference null distribution to calibrate the p-value from the parallelism test.

Similar to other permutation-based method, our partial permutation test relies on permutations to generate the reference distribution instead of a closed-form asymptotic approximation, and thus requires more computation. Averaging over all simulations for Tables 2 and 3 with $n = 200$, the parallelism test, "PPT," and "PPT+Parellel" took 0.39, 34.57, and 269.17 seconds, respectively. For Table 4 with sample size $n = 500$ and 1000, on average, the parallelism test took 3.05 and 21.29 seconds, while the "PPT" took 61.23 and 404.22 seconds. The issue of computational cost for the permutation method can be mitigated by parallelizing the calculation of the test statistic over permutations.

Table 5. Partial permutation test p-values for comparing relationships between the expenditure on food and the total expenditure among households with different numbers of members.

Test statistic	Comparison				Comparison after truncation			
	(2, 3, 4)	(2, 3)	(3, 4)	(2, 4)	(2, 3, 4)	(2, 3)	(3, 4)	(2, 4)
(19) with $\tilde{H}_1$ vs $\tilde{H}_0$	0.002	0.050	0.908	0	0.006	0.048	0.780	0
(19) with $\tilde{H}'_1$ vs $\tilde{H}_0$	0	0.030	0.664	0.002	0.004	0.064	0.860	0
(19) with $\tilde{H}_{\text{pseudo}}$ vs $\tilde{H}_0$	0.002	0.006	0.498	0.002	0	0.006	0.532	0
(16)	0	0.002	0.346	0	0	0	0.440	0

8. Application

We apply the partial permutation test to a dataset analyzed in Pardo-Fernández, Van Keilegom, and González-Manteiga (2007), which consists of monthly expenditures of several Dutch households and the numbers of members in each households. The dataset includes accumulated expenditures on food and total expenditures over the year (October 1986 to September 1987) for households with two members (159 in total), three members (45 in total) and four members (73 in total).

Let Y be the logarithm of the expenditure on food, X be the logarithm of the total expenditure, and Z be the number

of house members minus one (indicating the size of a family). To compare the relationship between Y and X among the three groups defined by Z, we use the partial permutation test with Gaussian kernel after standardizing both covariates and outcomes, and choose the permutation size based on model $\tilde{H}_0$ as suggested in Section 5.2 at the significance level $\alpha = 0.05$. We first test whether the same functional relationship between Y and X holds across all three groups, and then perform pairwise comparisons. Table 5 shows the resulting p-values using different test statistics, including the likelihood ratio statistics in (19) of $\tilde{H}_1, \tilde{H}_1'$ and $\tilde{H}_{\text{pseudo}}$ against $\tilde{H}_0$, and the test statistic (16) based on mean squared errors from pooled and group-specific kernel regression. It is very interesting to observe from Table 5 that the relationship between X and Y differs significantly between "no-kid" households (size=2) and larger-sized ones. However, between the households of size 3 and those of size 4, the relationships between X and Y are not significantly different. To avoid potential sensitivity to heavy-tailed errors in the data, we also conducted the tests after truncating extreme fitted residuals; see the supplementary material for details. Table 5 shows that the conclusions are consistent across different test statistics, and are robust to the use of truncation. Our results confirm the findings in Pardo-Fernández, Van Keilegom, and González-Manteiga (2007).

9. Discussion

We developed a partial permutation test for comparing across different groups the functional relationship between a response variable and some covariates, and studied its properties under null models (1) and (8) when the underlying function either is fixed or follows a Gaussian process. The key idea of the proposed tests is to keep invariant the projection of the response vector onto the space spanned by leading principle components of the kernel matrix, and permute the remaining (residual) part. Practically, we can also accommodate multiple kernels by conducting a partial permutation test that retains the projections of the response vector on the leading principle components of multiple kernel matrices. For example, if we use both the polynomial kernel of degree p and the Gaussian kernel, then the partial permutation test is exactly valid when the underlying function is polynomial up to degree p as implied by Theorem 2, and also has nice properties with flexible underlying functions as implied by Theorems 3, 4 and 5. Furthermore, based on the simulation study, we suggest to use test statistics based on a comparison between the null GPR model and its pseudo alternative as in (20), or a comparison between mean squared errors from pooled and group-specific kernel regressions. These test statistics are easy to calculate and have a superior power.

Our testing procedure is also related to Bayesian model checking, especially the conditional predictive p-value proposed by Bayarri and Berger (1997, 1999, 2000). The authors generated predictive samples from the model with parameters following the prior distribution, but only kept those samples that have the same value of a summary statistic U as the observed data. Then, they compared the test statistic of predictive samples with that of the observed samples. As pointed by Bayarri and Berger (1999), the intuition behind a suitable choice of U is that U should contain as much information about the unknown parameters as possible. In the extreme case where U is chosen to be the sufficient statistic for all parameters, the conditional predictive p-value is valid under the model where the parameters are fixed and unknown.

In our case, although we are considering a nonparametric model (1), the idea of conditional predictive p-value can still be applied. Suppose the variance of residuals is fixed and known and the underlying function follows a Gaussian process prior. We can perform the conditional predictive checking by choosing U to be $(\mathbf{X}, \mathcal{S}_y, \boldsymbol{Z})$, where $\mathcal{S}_y$ is from either the discrete or the continuous partial permutation test in Algorithm 1. Such choice of U contains information about the smooth components of the underlying function. However, it is generally computationally challenging to generate the predictive samples. From Theorem 4, under some regularity conditions on the Gaussian process prior, the predictive samples can be asymptotically equivalent to the ones from partial permutation, and the conditional predictive p-value can be approximated by the partial permutation p-value given the same choice of the test statistic.

In practice, we may face high-dimensional covariates, under which the comparison of functional relation among various groups becomes much more challenging. Generally, the permutation size of our partial permutation test decreases as the dimension of covariates increases, and will eventually lose power due to the lack of permutation size. This is intuitive due to the nature of the problem: with high-dimensional covariates, the underlying function can have a complex structure making it hard to distinguish whether multiple groups share the same functional relation or not, especially when there are limited sample size and limited overlaps of covariates from different groups. The issue may be mitigated by imposing additional structural assumptions, such as sparsity, and we leave it for future work.

Supplementary Material

The supplementary material contains computation details for maximizing the likelihood under GPR models, additional simulations for non-Gaussian or correlated noises and the choice of kernel parameters, and the proofs of all theorems, corollaries and propositions.

Acknowledgments

We thank the Associate Editor and two reviewers for constructive comments. The views expressed herein are the authors' alone and are not necessarily the views of Two Sigma Investments LP, or any of its affiliates.

Funding

This research is partly supported by the NSF grant DMS-1712714.

ORCID

Jun S. Liu http://orcid.org/0000-0002-4450-7239

References

Bayarri, M. J., and Berger, J. O. (1997), "Measures of Surprise in Bayesian Analysis," Duke University Institute of Statistics and Decision Sciences Working Paper, 97-46. [18]

——— (1999), "Quantifying Surprise in the Data and Model Verification," *Bayesian Statistics*, 6, 53–82. [18]

——— (2000), "P Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142. [18]

Behseta, S., and Kass, R. E. (2005), "Testing Equality of Two Functions Using Bars," *Statistics in Medicine*, 24, 3523–3534. [1]

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005), "Hierarchical Models for Assessing Variability Among Functions," *Biometrika*, 92, 419–434. [1]

Benavoli, A., and Mangili, F. (2015), "Gaussian Processes for Bayesian Hypothesis Tests on Regression Functions," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 74–82. [1]

Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019), "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs," *Journal of Statistical Planning and Inference*, 202, 14–30. [1]

Braun, M. L., Buhmann, J. M., and Müller, K.-R. (2008), "On Relevant Dimensions in Kernel Feature Spaces," *Journal of Machine Learning Research*, 9, 1875–1908. [4]

Cavazza, J., and Murino, V. (2016), "Active Regression With Adaptive Huber Loss," arXiv:1606.01568. [10]

Christmann, A., and Steinwart, I. (2007), "Consistency and Robustness of Kernel-Based Regression in Convex Risk Minimization," *Bernoulli*, 799–819. [6]

Degras, D., Xu, Z., Zhang, T., and Wu, W. B. (2012), "Testing for Parallelism Among Trends in Multiple Time Series," *IEEE Transactions on Signal Processing*, 60, 1087–1097. [15]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [10]

Freedman, D. A., and Peters, S. C. (1984), "Bootstrapping a Regression Equation: Some Empirical Results," *Journal of the American Statistical Association*, 79, 97–106. [3]

Hahn, J., Todd, P., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design," *Econometrica*, 69, 201–209. ISSN 00129682, 14680262. Available at *http://www.jstor.org/stable/2692190*. [1]

Hastie, T., and Zhu, J. (2006), "Comment," *Statistical Science*, 21, 352–357. [4,9]

Hinkley, D. V. (1988), "Bootstrap Methods," *Journal of the Royal Statistical Society*, Series B, 50, 321–337. [3]

Imbens, G. W., and Lemieux, T. (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635. [1]

Kühn, T. (1987), "Eigenvalues of Integral Operators Generated by Positive Definite Hölder Continuous Kernels on Metric Compacta," in *Indagationes Mathematicae (Proceedings)*, Vol. 90, 51–61. Amsterdam, Netherlands: Elsevier. [9]

Liu, M., and Cheng, G. (2018), "Early Stopping for Nonparametric Testing," in *Advances in Neural Information Processing Systems*, eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Vol. 31. Red Hook, NY: Curran Associates, Inc. [10]

Meng, X.-L. (1994), "Posterior Predictive p-Values," *The Annals of Statistics*, 22, 1142–1160. [2]

Micchelli, C. A., Xu, Y., and Zhang, H. (2006), "Universal Kernels," *The Journal of Machine Learning Research*, 7, 2651–2667. [6,9]

Neumeyer, N., and Dette, H. (2003), "Nonparametric Comparison of Regression Curves: An Empirical Process Approach," *The Annals of Statistics*, 31, 880–920. [1]

Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2007), "Testing for the Equality of k Regression Curves," *Statistica Sinica*, 17, 1115. [1,17,18]

Raskutti, G., Wainwright, M. J., and Yu, B. (2014), "Early Stopping and Non-Parametric Regression: An Optimal Data-Dependent Stopping Rule," *The Journal of Machine Learning Research*, 15, 335–366. [10]

Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press. [2,9]

Rischard, M., Branson, Z., Miratrix, L., and Bornn, L. (2018), "A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs: Do School Districts Affect NYC House Prices?" arXiv:1807.04516. [1]

Shang, Z., and Cheng, G. (2013), "Local and Global Asymptotic Inference in Smoothing Spline Models," *The Annals of Statistics*, 41, 2608–2638. [8,9]

Shi, J. Q., and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*. Boca Raton, FL: CRC Press. [2]

Thistlethwaite, D. L., and Campbell, D. T. (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317. [1]

Wang, L. (2005), *Support Vector Machines: Theory and Applications*, Vol. 177. Berlin, Germany: Springer Science & Business Media. [10]

Xing, X., Liu, M., Ma, P., and Zhong, W. (2020), "Minimax Nonparametric Parallelism Test," *Journal of Machine Learning Research*, 21, 1–47. [8,9,15,16]