Published in final edited form as:

Biometrics. 2019 September ; 75(3): 757–767. doi:10.1111/biom.13054.

A Hybrid Omnibus Test for Generalized Semiparametric Single-Index Models with High-Dimensional Covariate Sets[†]

Yangyi Xu¹, Inyoung Kim^{1,*}, Raymond J. Carroll²

¹ Department of Statistics, Virginia Tech., Blacksburg, Virginia, U.S.A.

² Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, and School of Mathematical Sciences, University of Technology, Sydney, Broadway NSW 2007

Summary

Numerous statistical methods have been developed for analyzing high-dimensional data. These methods often focus on variable selection approaches but are limited for the purpose of testing with high-dimensional data. They are often required to have explicit likelihood functions. In this paper, we propose a “hybrid omnibus test” for high-dimensional data testing purpose with much weaker requirements. Our hybrid omnibus test is developed under a semiparametric framework where a likelihood function is no longer necessary. Our test is a version of a frequentest-Bayesian hybrid score-type test for a generalized partially linear single index model, which has a link function being a functional of a set of variables through a generalized partially linear single index. We propose an efficient score based on estimating equations, define local tests, and then construct our hybrid omnibus test using local tests. We compare our approach with an empirical likelihood ratio test and Bayesian inference based on Bayes factors, using simulation studies. Our simulation results suggest that our approach outperforms the others, in terms of Type I error, power, and computational cost in both the low-and high-dimensional cases. The advantage of our approach is demonstrated by applying it to genetic pathway data for type II diabetes mellitus.

Keywords

Hybrid; Omnibus; Score function; Single index; Spline Approximation

1. Introduction

Variable selection and testing problems play important roles in genomics, data mining, image analysis, text mining, and other areas with high-dimensional data. Numerous statistical methods have been developed to analyze high-dimensional data (Tibshirani, 1996;

[†]This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1111/biom.13054]

*To whom correspondence should be addressed, inyoungk@vt.edu.

Supplementary Materials

Technical derivations, Tables, and Figures referenced in Section 4, the Figures referenced in Section 5, and the program code written as Matlab are also available with this paper at the Biometrics website on Wiley Online Library.

Zou & Hastie, 2005). These methods are mainly focused on variable selection and are based on various penalized regression models.

For testing purposes, statistical methods are limited in tests of high-dimensional data. Verzelen & Villers (2010) developed a method for testing high-dimensional goodness-of-fit but this method assumes linear models with Gaussian error. We develop our approach in a semiparametric framework so that the likelihood function would not be necessary. We propose a test statistic, called the “hybrid omnibus test” for high-dimensional data in single-index models.

Carroll et al. (1997) proposed a framework for estimating generalized partially linear single-index models. This framework allows for unknown single-index functions in low-dimensional cases. Carroll et al. (1997) focused on estimating, rather than testing. Although single-index models can be considered to be one of the most convenient ways of handling the “curse of dimensionality,” this problem was still difficult to handle with high-dimensional data. Radchenko (2015) thus derived an asymptotic theory for high-dimensional single-index models.

Upon launching the omnibus test for single-index models, we stabilized the procedure with a hybrid of the frequentist and Bayesian methods, which involved a redesigned efficient score without the requirement for an explicit likelihood. This type of test has been used (Hart, 2009; Ma et al., 2011) for goodness-of-fit tests in low-dimensional cases. Hart (2009) developed an omnibus goodness-of-fit tests using a hybrid of Bayesian and frequentist ideas, as well as a Laplace approximation. Ma et al. (2011) adapted this test to the general measurement-error framework and proposed both local and omnibus tests. Since omnibus test allows for alternative departures of X , this means all possible departures, including parametric or nonlinear departures, so this test was referred to as “omnibus” test.

The main idea of using these tests was in recognizing that for a given score type, similar tests can be constructed using semiparametric models even when the score itself cannot be calculated. In this paper, we propose a hybrid score-type test for single-index models with high dimensional sets of variables. Our hybrid omnibus test is useful for testing whether a set of variables, rather than an individual variable, is significantly associated with a response variable. We develop our hybrid omnibus test by starting from a generalized partially linear single-index model. An example of a case with such high-dimensional covariates is a genetic pathway, which is a set of genes that serve a particular cellular or physiological function. The genes within a particular pathway are expected to have a common function. The connection between clinical outcome and the genes for a given pathway is difficult to quantify using a parametric model. Hence, we consider the single-index model because this model is allowed to model a set of variables (i.e., all the genes within a pathway) together and connect nonparametrically between a response and a set of variables. Pathway-based analysis has the ability to detect subtle changes in a response variable that gene-based analysis might not identify (Mootha et al., 2003; Hosack, 2003; Rajagopalan and Agarwal, 2005). We use our approach to test the “overall” pathway effect rather than a single gene’s effect.

In this study, our goal is to develop a hybrid omnibus test that (a) is based on an estimating equation that cannot necessarily be obtained from an explicit likelihood function; (b) has high computational efficiency; (c) achieves reasonable power and a low type I error; (d) is robust to various estimating algorithms; (e) applies when either $p > n$ or $p \gg n$, where n is the sample size and p is the number of variables; and (f) is useful to test the overall effect, rather than individual variables. To the best of our knowledge, no previously developed test statistics have all of these features as we describe in an example in Section 2.

In Section 2, we describe our semiparametric model's framework. In Section 3, we propose a hybrid omnibus test; we describe how to obtain a score-type test in Section 3.1 and how to construct both local and hybrid omnibus tests in Section 3.2. In Section 4, we describe simulation studies that compare the performance of our hybrid omnibus test to that of alternative approaches, including the empirical-likelihood ratio test and Bayesian inference using Bayes' factor. In Section 5, we describe an application of our hybrid omnibus test to a genetic pathway analysis for type II diabetes mellitus. Finally, in Section 6, we provide concluding remarks.

2. Framework for the Semiparametric Single-Index Model

First, we illustrate the features of our hybrid omnibus test. Let Y be an $n \times 1$ binary response, X be an $n \times p$ matrix for high-dimensional predictors ($p > n$), and Z be an $n \times q$ low-dimensional predictor matrix ($n > q$). We want to test whether the effect of the predictor X is constant. Let $H(\cdot)$ be the logistic function and $f(\cdot)$ be a known parametric function of Z . Thus, the null model would be

$$\text{pr}(Y = 1 | X, Z) = H\{\kappa_0 + f(Z, \gamma)\}, \quad (1)$$

whereas, for a local test, an alternative model could allow for polynomial departures of X from constant κ_0 , as in

$$\text{pr}(Y = 1 | X, Z) = H\{\kappa_0 + h(X\alpha) + f(Z, \gamma)\}, \quad (2)$$

where $h(X\alpha)$ is a polynomial function with unknown parameters α . For an omnibus test, an alternative that allows for all departures of X from constant κ_0 would be

$$\text{pr}(Y = 1 | X, Z) = H\{\kappa_0 + g(X\alpha) + f(Z, \gamma)\} \quad (3)$$

for the unspecified global function $g(\cdot)$. Note that because omnibus testing allows for alternative departures of X from constant κ_0 , this test allows for all possible departures, including any parametric or nonlinear ones, so we refer this test to "omnibus" test.

Inferences can be easily drawn from Models (1) and (2) by using either sufficient scores or likelihood ratios. However, Model (3) is not trivial because $g(\cdot)$ is an unknown function and

because α is unknown and not identifiable in the null hypothesis. We propose using score-type tests that fit the model only under the null hypothesis so as to obtain estimates of γ , and then constructing a test statistic based on the estimating equations (Tsiatis & Ma, 2004; Ma & Carroll, 2006; Ma et al., 2011). Our test avoids parameter estimation for $g(\cdot)$, but we need a feasible solution for estimating the equation for α . Hence, our test statistic is only characterized under the null model. In this way, our approach is similar to the use of a score test. However, there is a crucial difference in that the ordinary score test is obtained from the likelihood, but our approach is not. We developed our approach using a semiparametric model, for which the explicit-likelihood function is not required. We also built our hybrid omnibus test on an efficient score, as that does not require the likelihood to be derived.

Because we did not specify the unknown function $g(\cdot)$, our approach do not use an exact likelihood. On the other hand, if we had approximated unknown function $g(\cdot)$ using wavelet-basis functions or splines, we could have obtained an approximate likelihood. For example, $g(\cdot)$ could be approximated with B -splines and J basis functions as

$$g(X\alpha) = \sum_{j=1}^J \beta_j b_j(X\alpha),$$

where β_j is an unknown coefficients of the B -spline basis functions and $b_j(\cdot)$ denotes the j th function in a cubic B -spline basis. However, this approach is not limited to B -splines.

For identifiability, $\|\alpha\| = 1$ is usually assumed. The benefit of single-index models is that they allow for convenient handling of high-dimensional predictors using $X\alpha$. For $\|\alpha\| = 1$, we use a polar-coordinate reparameterization. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$, $p \geq 3$. Each element of α can be represented as follows: $\alpha_1 = \sin(\phi_1)$, $\alpha_2 = \cos(\phi_1) \sin(\phi_2)$, ..., $\alpha_{p-1} = \cos(\phi_1) \dots \cos(\phi_{p-2}) \sin(\phi_{p-1})$, $\alpha_p = \cos(\phi_1) \dots \cos(\phi_{p-1})$, where $-\pi/2 \leq \phi_\ell \leq \pi/2$, $\ell = 1, \dots, p-1$. This parametrization provides an identifiable model via ϕ_ℓ which has a finite range for all values of ℓ .

The omnibus test that we constructed is based on the idea that a linear combination of sufficient many basis functions a smooth function can approximate arbitrarily well. Suppose that the J basis functions $h_j(\cdot)$, $j = 1, \dots, J$ express a linear form $g(X\alpha) = \sum_{j=1}^J \beta_j h_j(X\alpha)$, where the basis functions $h_1(X\alpha)$, $h_2(X\alpha)$, ..., $h_J(X\alpha)$ are arranged.

Let $P_{Y|X,Z}$ represent the model for Y given (X, Z) . The null model is thus

$$p_{Y|X,Z}(Y|X, Z) = p_{Y|X,Z}\{Y, \kappa_0 + f(Z, \gamma)\}.$$

For a local test in which $h_j(\cdot)$ can be a linear or polynomial function depending on testing, an alternative model that allows for linear and polynomial departures from the constant κ_0 is as follows:

$$p_{Y|X,Z}(Y|X, Z) = p_{Y|X,Z}\{Y, \kappa_0 + \beta_j h_j(X\alpha) + f(Z, \gamma)\}.$$

Hence, J local tests exist, depending on whether $\beta_j = 0$, for $j = 1, \dots, J$.

For an omnibus test, an alternative model that allows for any departure of X from constant κ_0 is

$$p_{Y|X,Z}(Y|X,Z) = p_{Y|X,Z}\{Y, \kappa_0 + g(X\alpha) + f(Z, \gamma)\},$$

again for the unspecified function $g(\cdot)$.

Under this setting, the null hypothesis (H_0) becomes whether $\beta = \mathbf{0}$ or $\alpha = \mathbf{0}$. However, $H_0: \beta = \mathbf{0}$ is more desirable because, unlike $\alpha = \mathbf{0}$, it does not suffer from the identifiability problem. Based on this framework, we propose a test procedure for the our hybrid omnibus test.

3. Hybrid Omnibus Test

Explicit likelihoods may be neither available nor necessary because the actual form of $g(\cdot)$ is unknown. Thus, we may be unable to derive an explicit likelihood. Hence, we first introduce a “score-type” test, which requires neither additional likelihood computation nor parameter utilization for the full model with the unspecified function $g(\cdot)$.

3.1 Score-Type Test

As we can build our omnibus test statistic using estimating equations, we can obtain estimating equations using either the log-likelihood function or the penalized log-likelihood function.

For the special logistic model, we can use either the following log-likelihood function:

$$\ell(\alpha, \beta, \gamma) = Y^T \left[\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + f(Z, \gamma) \right] - \mathbf{1}^T \log \left[\mathbf{1} + \exp \left[\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + f(Z, \gamma) \right] \right].$$

We then obtain a score function for $\beta = (\beta_1, \dots, \beta_J)$:

$$\Phi_{\beta}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = \frac{\partial \ell(\alpha, \beta, \gamma)}{\partial \beta}.$$

Or we can use the penalized log-likelihood function:

$$\begin{aligned} \ell_p(\alpha, \beta, \gamma) &= Y^T \left[\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + f(Z, \gamma) \right] - \mathbf{1}^T \log \left[\mathbf{1} + \exp \left[\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + f(Z, \gamma) \right] \right] \\ &\quad + \lambda \|\beta\|^2, \end{aligned}$$

where λ is a penalty parameter that can be obtained from cross-validation or from an information-based criterion. In this case, using $\ell_p(\alpha, \beta, \gamma)$ instead of $\ell(\alpha, \beta, \gamma)$, we can obtain a score function:

$$\Phi_{\beta}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = \frac{\partial \ell_p(\alpha, \beta, \gamma)}{\partial \beta}.$$

For the general model, a score-type function can be derived from estimating equations rather than strictly from a likelihood function whose characteristics are similar to those of the original score function. We can thus obtain a score-type function for $\beta = (\beta_1, \dots, \beta_J)$. Hence, we write the estimating equations as $\sum_{i=1}^n \Phi_{\beta}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = 0$,

$\sum_{i=1}^n \Phi_{\gamma}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = 0$, and $\sum_{i=1}^n \Phi_{\alpha}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = 0$, where the estimating functions $\Phi_{\beta}(\cdot)$, $\Phi_{\gamma}(\cdot)$, and $\Phi_{\alpha}(\cdot)$ have the same dimensions as β , γ , and α , respectively. These estimating functions cannot necessarily be derived from any profile likelihood, as that may not exist in our semiparametric framework.

Under the null hypothesis, the estimating equations are simply $\sum_{i=1}^n \Phi_{\gamma}(X_i, Z_i, Y_i, \alpha, 0, \gamma) = 0$ and $\sum_{i=1}^n \Phi_{\alpha}(X_i, Z_i, Y_i, \alpha, 0, \gamma) = 0$. The roots of these estimating equations are $\hat{\alpha}$ and $\hat{\gamma}$, respectively. However, we cannot use the estimating equation for $\hat{\alpha}$,

$\sum_{i=1}^n \Phi_{\alpha}(X_i, Z_i, Y_i, \alpha, 0, \gamma) = 0$ because of the identifiability between α and β . As our null hypothesis is based on β and not α , we could use any possible root of a feasible solutions of

$$\sum_{i=1}^n \Phi_{\alpha}(X_i, Z_i, Y_i, \alpha, \beta, \gamma) = 0.$$

Using Φ_{γ} , we obtain the root $\hat{\gamma}$ using the null model $\text{pr}(Y|Z) = P_{Y|X,Z}\{Y, \kappa_0 + f(Z, \gamma)\}$. Based on the score test, we propose the estimated score

$$\hat{U} = n^{-1/2} \sum_{i=1}^n \Phi_{\beta}(X_i, Z_i, Y_i, \hat{\alpha}, 0, \hat{\gamma}).$$

Analyzing $\hat{U}$ is not difficult. We first create the following definitions, with all expectations based on the null hypothesis: $A_1 = E\{\partial \Phi_{\gamma}(X, Z, Y, \alpha, 0, \gamma) / \partial \gamma^T\}$,

$$A_2 = E\{\partial \Phi_{\beta}(X, Z, Y, \alpha, 0, \gamma) / \partial \gamma^T\}, A_3 = E\{\partial \Phi_{\gamma}(X, Z, Y, \alpha, \beta, \gamma) / \partial \beta^T\},$$

$$A_4 = E\{\partial \Phi_{\beta}(X, Z, Y, \alpha, \beta, \gamma) / \partial \beta^T\}, B_{11} = E\{\Phi_{\gamma}(X, Z, Y, \alpha, 0, \gamma) \Phi_{\gamma}(X, Z, Y, \alpha, 0, \gamma)^T\},$$

$$B_{22} = \text{cov}\{\Phi_{\beta}(X, Z, Y, \alpha, 0, \gamma)\}, B_{12} = \text{cov}\{\Phi_{\gamma}(X, Z, Y, \alpha, 0, \gamma), \Phi_{\beta}^T(X, Z, Y, \alpha, 0, \gamma)\},$$

$V_{\gamma} = A_1^{-1} B_{11} (A_1^{-1})^T$, and $\Sigma_0 = \text{cov}\{\Phi_{\beta}(\cdot, \alpha, 0, \gamma) - A_2 A_1^{-1} \Phi_{\gamma}(\cdot, \alpha, 0, \gamma)\}$. We then further define the following matrices:

$$\mathcal{A} = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix}, \quad \mathcal{V} = \mathcal{A}^{-1} \mathcal{B} (\mathcal{A}^{-1})^T = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{pmatrix}.$$

All of these quantities can be estimated by replacing their expectations and covariance matrices with their sample versions. We denote the resulting sample estimate of Σ_0 is $\hat{\Sigma}_0$.

Based on the test statistic with nominal level α_0 , we propose rejecting the hypothesis if $\hat{T} = \hat{U}^T \hat{\Sigma}_0^{-1} \hat{U}$ exceeds the $(1 - \alpha_0)$ quantile of the χ^2 distribution of T with degrees of freedom p_T . $\hat{T}$ does not involve estimating β .

Theorem 1. Assume that p is fixed and $p < n$. Under the null hypothesis, as $n \rightarrow \infty$, $n^{1/2}(\hat{\gamma} - \gamma) \rightarrow N(0, V_\gamma)$ and $\hat{U} \rightarrow N(0, \Sigma_0)$. Hence, $\hat{T} = \hat{U}^T \hat{\Sigma}_0^{-1} \hat{U}$ is asymptotically χ^2 with degrees of freedom, p_T .

We use $\hat{T}$ as a score-type test to construct our hybrid omnibus test. The proof of Theorem 1 is shown in Appendix A of the Supplementary Materials.

The result of Theorem 1 also holds when $p > n$ based on the results of Radchenko (2015).

Lemma 1. Let $p_n = O\{\exp(n^\delta)\}$, where $0 < \delta < 1$. Then, there exists $\hat{\alpha}$ such that $\|\hat{\alpha} - \alpha\| = O\{n^{-\delta/2}(\log n)^{-1} \log \log(p_n)\}$.

We obtain the proof of Lemma 1 directly from Theorem 2 in Radchenko (2015). These assumptions and results can be applied to our study.

Theorem 2. Under the null hypothesis and Lemma 1, as $n \rightarrow \infty$, $n^{1/2}(\hat{\gamma} - \gamma) \rightarrow N(0, V_\gamma)$ and $\hat{U} \rightarrow N(0, \Sigma_0)$. Hence, $\hat{T} = \hat{U}^T \hat{\Sigma}_0^{-1} \hat{U}$ is asymptotically χ^2 with degrees of freedom, p_T .

Theorem 2 can be proven using an argument similar to that of Theorem 1. We use $\hat{T}$ as a score-type test when constructing our hybrid omnibus test. In the next section, we explain how the score-type test T plays an important role in this hybrid omnibus test. As T is an element of the hybrid omnibus test and is applicable for both the $p < n$ and $p > n$ cases, our test is also applicable to both these cases.

3.2 Construction of the Hybrid Omnibus Test

We consider J local tests based on the null hypothesis (H_0) and the j th alternative hypothesis (H_{j1}), $j=1 \dots J$:

$$\begin{aligned} H_0: & \quad p_{Y|X,Z}(Y|X, Z) = p_{Y|X,Z}\{Y, \kappa_0 + f(Z, \gamma)\}, \\ H_{j1}: & \quad p_{Y|X,Z}(Y|X, Z) = p_{Y|X,Z}\{Y, \kappa_0 + \beta_j h_j(X\alpha) + f(Z, \gamma)\}. \end{aligned}$$

This test is equivalent to $H_0: \beta_j = 0$ vs $H_{j1}: \beta_j \neq 0$. For the logistic model with $p_{Y|X,Z}(Y|X,Z) = H\{\kappa_0 + f(Z, \gamma)\}$ under H_0 and $H\{\kappa_0 + \beta_j h_j(X\alpha) + f(Z, \gamma)\}$ under H_{j1} , respectively, we apply a score-type test according to Theorem 1. For each $h_j(\cdot)$, we write the test statistic as $\hat{T}_j^2 = \hat{u}_j^2 / \hat{\sigma}_j^2$, where $\hat{u}_j$ and $\hat{\sigma}_j^2$ are one-dimensional versions of $\hat{U}_j$ and $\hat{\Sigma}_{0j}$, respectively. We further define A_{j2} and B_{j12} similarly for the one-dimensional versions of A_2 and B_{12} that are described in Section 3.1.

Let $\hat{T} = (\hat{T}_1, \dots, \hat{T}_J)^T$. We show that, asymptotically, $n^{-1/2}\hat{T} \sim N(\mathbf{0}, \Sigma)$ under the null hypothesis, for which the (j, k) th element of Σ is derived as

$$E\left(\frac{\hat{u}_j \hat{u}_k}{\hat{\sigma}_j \hat{\sigma}_k}\right) = \frac{1}{\sigma_j \sigma_k} \left[A_{j2} A_1^{-1} B_{11} A_1^{-1T} A_{k2}^T - A_{j2} A_1^{-1} B_{k12} - A_{k2} A_1^{-1} B_{j12} \right. \\ \left. + E\left\{ \Phi_{\beta_j}(\cdot; \kappa_0, 0) \Phi_{\beta_k}(\cdot; \kappa_0, 0) \right\} \right] \quad (4)$$

because $E(\hat{u}_j / \hat{\sigma}_j) = 0$ for any value of j . The marginal limit distribution is thus $\hat{u}_j / \hat{\sigma}_j \rightarrow N(0, 1)$ in distribution. The score-type function is thus

$$\hat{U} = n^{-1/2} \sum_{i=1}^n \Phi_{\beta}(X_i, Z_i, Y_i, \hat{\alpha}, 0, \hat{\gamma}).$$

We propose a hybrid omnibus test using local tests. We have combined the local test statistics to build an omnibus test such as the F test even though it is difficult to obtain either a closed-form or asymptotic distribution. Following the approach that we use to obtain a value of χ^2 from the sum of the square of the normal distributions, we designed a hybrid omnibus test statistic, which can be expressed as $\hat{\mathcal{T}} = \sum_{j=1}^J \omega_j \exp(\hat{T}_j^2/2)$, where ω_j is a weight. The purpose of this weight term is to assign larger weights to the bases of global features than to those of local features. We arrange basis functions from smallest to largest knot span using splines with J basis functions. We use cubic B -splines for convenience because B -spline functions enable the creation and management of complex shapes and surfaces using various numbers of knot points. Ruppert (2002) found that function approximations were not very sensitive to the number of knots beyond the minimum number and that having too many knots could worsen the mean squared error.

Assuming that each of the indices of j is proportional to the local test statistic that is associated with the corresponding basis, it is reasonable to specify that the prior of the corresponding local test statistic is $\pi_j = (1 + c^j)^{-1}$, $j = 1, \dots, J$, where $c > 1$ and the weight (ω_j) is $\pi_j / (1 - \pi_j)$. Using this weight, we rewrite the hybrid omnibus test statistic as $\hat{\mathcal{T}} = \sum_{j=1}^J \left\{ \pi_j / (1 - \pi_j) \right\} \exp(\hat{T}_j^2/2)$.

For $c = 2$, the result is $\omega_j = 1/j^2$. Thus the hybrid omnibus test statistic is

$$\widehat{\mathcal{T}} = \sum_{j=1}^J j^{-2} \exp(\widehat{T}_j^2/2). \quad (5)$$

The p value and the power of the omnibus test, as compared to a local test, can be obtained from the result that $\widehat{\mathcal{T}}$ has an asymptotically multivariate normal distribution. In practice, the p value and power can be approximated by generating samples from the $N(0, \cdot)$ distribution and then comparing them to observed values. We explain this procedure is explained in Appendix B of the Supplementary Materials.

4. Simulation

We conducted several simulation studies to understand the hybrid omnibus test's performance. We calculated the estimating equations using $\ell(\mathbf{a}, \mathbf{\beta}, \boldsymbol{\gamma})$, and $\ell_p(\mathbf{a}, \mathbf{\beta}, \boldsymbol{\gamma})$ as described in Section 3.1, and then conducted the hybrid omnibus tests, which denote as "HOT" and "HOT_p".

In this study, we consider the set of bases to be $h_j(X\mathbf{a}) = b_j(X\mathbf{a})$ cubic B -splines with $J = 22$ basis functions. We chose 22 basis functions for convenience, although we performed spot checks with as many as 42 basis functions, and the results for between 12 basis and 42 basis functions were similar. We consider HOT to be a special case of HOT_p for $\lambda = 0$. By comparing HOT to HOT_p, we understand how much, and under which situations, they differ. We chose λ in a grid using the range [0.01, 10000] with a Bayesian information criterion (BIC). The smallest value of BIC that we achieved was around $\lambda = 15$. The test results were similar performance for $\lambda \geq 5$.

We also investigated how much the initial value of $\hat{\mathbf{a}}$ affects the testing results, as we describe in Section D of the Supplementary Material. We found that the our type I error and power are not sensitive to the initial value of $\hat{\mathbf{a}}$.

We then compared our approaches two alternatives- the empirical likelihood ratio test (ELRT) and the Bayesian inference, based on the approximated Bayes factor (ABF) in terms of type I error and power. We considered any Bayes factor above the cutoff values ($BF_{cut} = 10$) to represent strong favor to H_1 . We explain the algorithms for ELRT and ABF in detail in Sections A and B of the Supplementary Material.

We considered two settings for nonlinear models: a "sine-bump" function and a polynomial function. For each setting, we considered two cases: a low-dimensional case with a large sample size (n) relative to the number of variables (p), as well as a high-dimensional case with a relatively small sample size compared to the number of variables. We applied the approaches for both the low- and high-dimensional cases, but ELRT is only applicable to low-dimensional cases because parameter estimators for high-dimensional cases are very unstable.

4.1 Simulation Setting

We applied nonlinear functions: the sine-bump and the 4th degree polynomial functions. The sine-bump function has significant nonlinearity. We chose the polynomial function to investigate the loss of efficiency when compared to a correct, parameterized likelihood approach. For each setting, we simulated data for the low-dimensional case and high-dimensional cases. We generated 1,000 simulated data sets in all.

4.1.1 Setting 1: Sine-Bump Function—The binary predictor Z takes the values 0 and 1 with 50% probability. We generated the continuous predictors X from normal and uniform distributions for the low- and high-dimensional cases, which we explain in the following.

- Case 1: Low-dimensional case

We set $n = 500$ and $p = 3$. We generated each value of X from the $N(0,1)$ distribution. We set the true parameters were set as $\alpha = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, $A = \sqrt{3}/2 - 1.645/\sqrt{12}$, $B = \sqrt{3}/2 + 1.645/\sqrt{12}$, and $\gamma = 0.3$.

- Case 2: High-dimensional case

We set $n = 100$ and $p = (100, 200, 300, 500, 800, 1,000)$. We generate each value of X from the $Uniform(0,1)$ distribution. We considered $\alpha = (0, \dots, 0, 1, \dots, 1)$ with half 0 and half 1 of p ; we then standardized α so that $\|\alpha\| = 1$. The settings for A , B , and γ were the same as in Case 1.

For each case, we generated the binary response Y using the sine-bump function

$$\text{pr}(Y = 1 | X, Z) = H[m \cdot \sin\{\pi(X\alpha - A)/(B - A)\} + \gamma Z],$$

where m is an amplifying multiplier in the range $[4, 6]$ with increments of 0.5. Because of this range, the nonlinear function highly affects the response variable.

4.1.2 Setting 2: Polynomial Function—We also conducted a simulation to compare our hybrid omnibus test with a traditional score test using a polynomial model in the high-dimensional case. The settings for n and p were the same as in Case 2 of Setting 1. The variable X was generated from the $Uniform(0,1)$ distribution as it was before. We set $\gamma = 0.3$ and $\alpha = (\mathbf{0}_{1,p/2}, \mathbf{1}_{1,p/2})$. We then standardized α so that $\|\alpha\| = 1$. We generate the binary response variable from the polynomial function

$$\text{pr}(Y = 1 | X, Z) = H[m\{-0.9619 + X\alpha + (X\alpha)^2 + (X\alpha)^3 + (X\alpha)^4\} + \gamma Z],$$

where m is an amplifying multiplier; we varied it in the range $[4, 6]$ with increments of 0.5. Because of this range, the polynomial function highly affects the response variable.

4.2 Simulation Results

We obtained the average values of the power and type I error for all of each setting's low and high p cases. We provided these results based on 1,000 simulated data sets. We set $\alpha_{p \times 1} = \mathbf{0}_{p,1}$ for type I errors. For power, we set $\alpha_{p \times 1}$, as described in each case for each setting. For

each method, we also summarized the overall means of type I error and power values for all of the low-dimensional and high-dimensional p cases, as shown in Table 1.

Type I error and power results for each low-dimensional p are also summarized in Table 2.

Type I error and power results for each high-dimensional p are summarized in Tables 3–4. The ABF's results are based on a 5000 MCMC chains after 2500 burn-in.

Regarding the overall means for the type I error and power values in both the low-and high-dimensional cases, as shown in Table 1, HOT_p provides the smallest average type I error and the largest average power in both the low-and high-dimensional cases. HOT and HOT_p have similar performance as m increases. In the low-dimensional case, HOT has an average type I error of 0.04 and an average power of 0.95. HOT_p has an average type I error of 0.02 and an average power of 0.99. In the high-dimensional case, HOT has an average type I error of 0.02 and an average power of 0.91. HOT_p has an average type I error of 0.01 and an average power of 0.98.

For the low-dimensional case, ELRT also has good performance because it uses known a likelihood functions. In the high-dimensional case, ELRT was very unstable, so we were unable to obtain the type I error or power. Each of our HOT and HOT_p methods outperformed the other methods. Our hybrid omnibus test approaches is comparable.

Under Case 1 of Setting 1, for which we used the low-dimensional case and the sine-bump function, as shown in Table 2, the average type I errors of the ELRT and the HOT methods were comparable under Case 1 of Setting 1. ELRT's type I error was between 0.04 and 0.10. The hybrid omnibus test's type I error was between 0.04 and 0.10. HOT_p 's type I error was between 0.01 and 0.05. ABF's type I errors are between 0.00 and 0.06. We considered any Bayes factor above the cutoff values ($BF_{cut}=10$) to represent strong favor to H_1 . HOT_p had the smallest type I error values because of λ .

As shown in Table 2, the average powers of ELRT increased as m increased in Case 1 of Setting 1, as expected. The powers of our HOT and HOT_p methods also increased as m increased. The average power of HOT_p was between 0.99 and 1, which was larger than those of ELRT and ABF. Therefore, both HOT and HOT_p were more powerful than either ELRT or ABF.

Under Case 2 of Setting 1, for which we used the high-dimensional case and sine-bump function, as shown in Table 3, the average type I error of the hybrid omnibus test was between 0.00 and 0.11 under Case 2 of Setting 1. The average power of the hybrid omnibus test was between 0.69 and 0.98. These values are summarized in Table 3. Therefore, our HOT method performed well in the high-dimensional case. As shown in Table 4, the average type I error of HOT_p under Case 2 of Setting 2 was between 0.00 and 0.10. The average value of type I error decreased as p and m increased. The average power of HOT_p was between 0.89 and 1, as shown in Table 4. Hence, HOT_p also performed well.

Therefore, our simulation results suggest that our hybrid omnibus tests, HOT and HOT_p , outperform the other tests, in terms of type I error, power, and computational cost in both the low-high-dimensional cases.

5. Application

We applied our hybrid omnibus tests to type II diabetes pathway data set (Pang et al., 2015, 2006; Mootha et al., 2003), resulting in 278 pathways: 128 KEGG pathways plus the 149 pathways that Mootha group curated. We then excluded 36 user-defined c_U133_probes user-defined pathways. Finally, we used the total number of pathways were 242. In addition, as Associate Editor suggested, we also applied the Bayesian inference based on ABF. For each pathway, we ran 100,000 MCMC chains and 5,000 burn-in and then collected pathways with the largest ABF values. We compared these pathways to the results from our hybrid omnibus test.

5.1 Significant Pathways from the Hybrid Omnibus Test

In our analysis, let Y be the binary response representing both normal samples and those samples with type II diabetes mellitus; let X be the $n \times p$ gene-expression levels within each pathway, where n is 35 (i.e., the number of subjects); let p be the number of genes in a specific pathway, which varies from 4 to 200 across these pathways; and let Z be the clinical predictors (e.g., BMI). There were no variables related to population stratification. We had an age variable, but we did not find that it had statistical significance, which may be due to the fact that the participants were between 61 and 69 years old. The BMI was significant, however, so we included BMI in our data analysis. Our goal was to identify the pathways that would distinguish between normal samples and those with type II diabetes mellitus after adjusting for the linear BMI effect. We used a set of B -spline basis functions and performed the hybrid omnibus test. We identified twenty nine pathways using a significance level 0.05, as summarized in Table 5. The p-values of all pathways are summarized in Figure 6 in the Supplementary Materials. The top pathways which have ABF values above 10 are also summarized in Table 14 in the Supplementary Materials. The eleven out of twenty nine pathways listed in Table 5 are also identified using Bayesian inference based on ABF.

The pathways identified in Table 5 include the MAPK signaling pathway as well as alanine and Aspartate metabolism and Oxidative phosphorylation pathways from (Mootha et al., 2003) analysis of the binary phenotype of interest. (Mootha et al., 2003) have found that oxidative phosphorylation expression is coordinately decreased in human diabetic muscle. PGC-1 alpha, a cold-inducible regulator of mitochondrial biogenesis, thermogenesis, and skeletal-muscle fiber-type switching has been hypothesized to introduce the oxidative phosphorylation pathway (Mootha et al., 2003). It is not surprising that ATP synthesis has also considered because it is a subset of “Oxidative phosphorylation”.

Several important pathways have been found to distinguish between normal samples and those with type II diabetes (Pang et al., 2006, 2015). One of these is the MAPK signaling pathway, which is a member of the MAPK family and which is activated by a variety of environmental stressors and inflammatory cytokines. As with other MAPK cascades, the membrane-proximal component is a MAPKKK, which is typically a MEKK or a mixed-

lineage kinase. The MAPKKK phosphorylate activates MKK3/6, which is a p38 MAPK kinase. ASK can also directly activate MKK3/6 as a result of apoptosis stimuli. A p38 MAPK is involved in the regulation of HSP27, MAPKAPK-2 (MK2), MAPKAPK-3 (MK3), and several transcription factors, including ATF-2, Stat1, the Max/Myc complex, MEF-2, Elk-1, and CREB (indirectly, via activation of MSK1).

Researchers have ranked the actions of “Nitric Oxide in the Heart” as one of the top pathways, that nitric oxide synthesis plays a role in the reduction of glucose uptake for individuals with the type II diabetes, as compared with individuals in control groups (Kingwell et al., 2002).

We also identified pathway 36, c17_U133_probes (Pang et al., 2006; Kim et al., 2012, 2013). The genes MEF2C, NR4A1, SOX1, and TPS1 are known to be related to glucose (Voisine et al., 2004; Zhang et al., 2006). The gene CAPI is related to human insulin signaling (Dahlquist et al., 2002). The genes MAP2K6, ARF6, and SGK are known to be related to human insulin signaling (Dahlquist et al., 2002). The gene ARF6 plays a role in the activating protein kinase and phospholipase under high-glucose conditions, researchers have hypothesized that that is important intracellular event linked to diabetic nephropathy (Padival et al., 2004). Researchers have shown that SGK haplotype is significantly more prevalent in individuals with type II diabetes than in healthy volunteers in the Romanian population (Schwab et al., 2008; Boini et al., 2006). In addition, salt intake decreases SGK-dependent glucose uptake in mice; thus, SGK plays a role in glucose intolerance in mice. We also found other pathways that no previous researchers has detected and need to be further biologically validated. The above findings can help scientists to identify potential biomarkers and drug targets, as well as generate further biological hypotheses for testing.

6. Discussion

In this paper, we have proposed a hybrid omnibus test for high-dimensional data. We have developed using a semiparametric framework in which no likelihood function is available. We thus propose using an efficient score, which serves as a local test statistic associated with estimating equations, to avoid likelihood derivations (when they are unavailable).

We compared our two approaches to an empirical likelihood ratio test and to ABF using a simulation study. The results suggest that our hybrid omnibus tests outperformed the other methods in both the low- and high-dimensional cases. ELRT performed well for the low-dimensional case, as expected, but our approaches were comparable to ELRT. However, we could not obtain ELRT for $p > n$. In addition, the algorithm require immense computational costs. However, for the high-dimensional case, our proposed hybrid omnibus tests performed well in terms of both type I error and power. However, ABFs did not provide good performance in the high-dimensional case.

Our hybrid omnibus tests have the following advantages: They (a) do not require a likelihood function; (b) are applicable even in high-dimension cases, where $p \gg n$; (c) do not rely on specified estimating equations; (d) are flexible to build using various basis functions such as the spline, Fourier, and wavelet functions; (e) do not depend on an estimating

algorithm; and (f) have high computational efficiency. To the best of our knowledge, our approach is novel because it provides all of these advantages.

We also conducted an additional simulation study to examine the performance of our omnibus tests with significance levels 0.001 and 0.01. These simulation results are summarized in Appendix F of the Supplementary Materials. Our omnibus tests perform reasonably well in terms of type I error and power for the case when $n = 200$ with a significance level of 0.01 and also when $n = 100$ with a significance level of 0.05. However, type I error was often smaller than the nominal error rate, even though the type I error approaches the nominal error rate as n increases. Further research is still needed to examine the theoretical properties of our omnibus tests to understand how the behaviors depend on (n , p , type I error). Deriving theoretical properties and distributions would be useful for studies of the theoretical bound of significance and distribution's degree of freedom.

We analyzed each pathway separately. However, pathways are not independent, as they share genes and interactions; this makes it difficult to adjust p -values for our testing procedure. In addition, our hybrid omnibus tests do not consider multiple comparisons. Developing such a multiple-comparison method using our omnibus hybrid test is an interesting and challenging problem because of the pathways' complex dependence structure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). We are also grateful to the associate editor and reviewers for their valuable suggestion and constructive input.

References

- Boini KM, Hennige AM, Huang DY, Friedrich B, Palmada M, Boehmer C, Grahammer F, Artunc F, Ullrich S, Avram D, et al. (2006). Serum-and glucocorticoid-inducible kinase 1 mediates salt sensitivity of glucose tolerance. *Diabetes*, 55(7), 2059–2066. [PubMed: 16804076]
- Carroll RJ, Fan J, Gijbels I, & Wand MR (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438), 477–489.
- Coleman TF & Li Y (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2), 418–445.
- Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, & Conklin BR (2002). Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31(1), 19–20. [PubMed: 11984561]
- Härdle W & Stoker TM (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408), 986–995.
- Hart JD (2009). Frequentist-Bayes lack-of-fit tests based on Laplace approximations. *Journal of Statistical Theory and Practice*, 3(3), 681–704.
- Hosack DA, Dennis G Jr., Sherman BT, clifford H, and Lempicki RA (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10), R70. [PubMed: 14519205]
- Ichimura H (1993). Semiparametric least squares (s1s) and weighted sis estimation of single-index models. *Journal of Econometrics*, 58(1), 71–120.

- Kingwell B, Formosa M, Muhlmann M, Bradley S, McConell G (2002). Nitric oxide synthase inhibition reduces glucose uptake during exercise in individuals with Type 2 diabetes more than in control subjects. *Diabetes*, 51, 2572–2580 [PubMed: 12145173]
- Kim I, Pang H, & Zhao H (2012). Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. *Statistics in Medicine*, 31(15), 1633–1651. [PubMed: 22438129]
- Kim I, Pang H, & Zhao H (2013). Statistical properties on semiparametric regression for evaluating pathway effects. *Journal of Statistical Planning and Inference*, 143(4), 745–763. [PubMed: 24014933]
- Ma Y & Carroll RJ (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, 101(416), 1465–1474.
- Ma Y, Hart JD, Janicki R, & Carroll RJ (2011). Local and omnibus goodness-of-fit tests in classical measurement error models. *Journal of the Royal Statistical Society, Series B*, 73(1), 81–98.
- Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JR, Golub TR, Tamayo R, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, & Groop LC (2003). Pgc-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3), 267–273. [PubMed: 12808457]
- Padival AK, Hawkins KS, & Huang C (2004). High glucose-induced membrane translocation of pkc β is associated with arf6 in glomerular mesangial cells. *Molecular and Cellular Biochemistry*, 258(1–2), 129–135. [PubMed: 15030177]
- Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MR, Floyd E, & Zhao H (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, 22(16), 2028–2036. [PubMed: 16809386]
- Pang H, Kim I, & Zhao H (2015). Random effects model for multiple pathway analysis with applications to Type II diabetes microarray data. *Statistics in Biosciences*, 7(2), 167–186. [PubMed: 26640601]
- Radchenko P (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139(16), 266–282.
- Rajagopalan DA and Agarwal P (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6), 788–793. [PubMed: 15509611]
- Ruppert D (2002). Selecting the number of knots for penalized splines. *Journal of Computational & Graphical Statistics*, 11(4), 735–757.
- Schwab M, Lupescu A, Mota M, Mota E, Frey A, Simon R, Mertens PR, Floege J, Luft F, Asante-Poku S, et al. (2008). Association of sgkl gene polymorphisms with type 2 diabetes. *Cellular Physiology and Biochemistry*, 21(1–3), 151–160. [PubMed: 18209482]
- Stoker TM (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6), 1461–1481.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 267–288.
- Tsiatis AA & Ma Y (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, 91(4), 835–848.
- Verzelen N & Villers F (2010). Goodness-of-fit tests for high-dimensional Gaussian linear models. *Annals of Statistics*, 38(2), 704–752.
- Voisine P, Ruel M, Khan TA, Bianchi C, Xu S-H, Kohane I, Libermann TA, Otu H, Saltiel AR, & Sellke FW (2004). Differences in gene expression profiles of diabetic and nondiabetic patients undergoing cardiopulmonary bypass and cardioplegic arrest. *Circulation*, 110(11 suppl 1), II–280.
- Weinberg MD (2012). Computing the Bayes factor from a Markov Chain Monte Carlo simulation of the posterior distribution. *Bayesian Analysis*, 7(3), 737–770.
- Yu Y & Ruppert D (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460), 1042–1054.
- Zhang D, Zhou Z, Li L, Weng J, Huang G, Jing P, Zhang C, Peng J, & Xiu L (2006). Islet autoimmunity and genetic mutations in Chinese subjects initially thought to have type 1b diabetes. *Diabetic Medicine*, 23(1), 67–71. [PubMed: 16409568]

Zou H & Hastie T (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301–320.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

The average values of type I error and power for four methods, ELRT=Empirical Likelihood Ratio Test; HOT=Hybrid Omnibus Test; HOT_p =HOT with penalized estimating equation; ABF=Approximated Bayes Factor that falls above the cutoff values ($BF_{cut} = 10$). This is under simulation setting 1, where $\text{pr}(Y = 1|X, Z) = H[m \sin\{\pi(\alpha_T X - A)/(B - A)\} + \gamma Z]$; The testing hypothesis is $H_0 : \text{pr}(Y = 1|Z) = H\{\kappa_0 + \gamma Z\}$ vs $H_1 : \text{pr}(Y = 1|X, Z) = H\{\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + \gamma Z\}$; $\alpha_{p \times 1} = \mathbf{0}_{p,1}$ for type I error; $\alpha = \{\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\}$ for power of low-dimensional case; $\alpha_{p \times 1} = \{\mathbf{0}_{1, p/2}, \mathbf{1}_{1, p/2}\}^T$ for power of high-dimensional case; These results are based on 1,000 simulated data sets; N/A for ELRT because of non-convergence in high-dimensional case

Method	Low dimensional case		High dimensional case	
	Average type I error	Average power	Average type I error	Average power
ELRT	0.069	0.873	N/A	N/A
HOT	0.042	0.945	0.021	0.916
HOT_p	0.021	0.999	0.012	0.988
ABF	0.025	0.954	0.001	0.332

Table 2:

The average value of type I error obtained from ELRT, HOT, HOT_p and ABF in simulation study under low-dimensional case; m =the amplifying multiplier; ELRT=Empirical Likelihood Ratio Test; HOT=Hybrid Omnibus Test; HOT_p=HOT with penalized estimating equation; ABF=Approximated Bayes Factor that falls above the cutoff values ($BF_{cut} = 10$); The data were generated from: $\text{pr}(Y = 1|X, Z) = H[m \sin\{\pi(\alpha_T X - A)/(B - A)\} + \gamma Z]$, with the following settings: $\alpha = \{0, 0, 0\}$ for type I error; $\alpha = \{\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\}$ for power; $A = \frac{\sqrt{3}}{2} - \frac{1.645}{\sqrt{12}}$, $B = \frac{\sqrt{3}}{2} + \frac{1.645}{\sqrt{12}}$, $\gamma = 0.3$, $H(\cdot)$ the inverse logistic link; The predictors are generated from the following setting: $X \sim N(0, 1)$, $Z = 0$ if observation is odd, $Z = 1$ if observation is even; The testing hypothesis is $H_0: \text{pr}(Y = 1|Z) = H\{\kappa_0 + \gamma Z\}$ vs $H_1: \text{pr}(Y = 1|X, Z) = H\{\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + \gamma Z\}$; These results are based on sample size $n=500$ and 1000 simulated data sets; Bayesian inference using MCMC chain=5000 after 2500 burn-in

	m	4.00	4.50	5.00	5.50	6.00
Type I error	ELRT	0.051	0.060	0.053	0.072	0.081
	HOT	0.086	0.103	0.104	0.043	0.014
	HOT _p	0.013	0.053	0.014	0.020	0.013
	ABF	0.011	0.001	0.010	0.042	0.030
Power	ELRT	0.873	0.841	0.870	0.882	0.913
	HOT	0.942	0.971	0.983	1.000	0.990
	HOT _p	0.988	0.988	0.999	1.000	1.000
	ABF	0.901	0.992	0.992	1.000	1.000

Table 3:

The average values of type I error vs the amplifying multiplier (m) in simulation study under high-dimensional case with HOT; HOT=Hybrid Omnibus Test; The data were generated from: $\text{pr}(Y=1|X,Z) =$

$H[m \cdot \sin\{\boldsymbol{\pi}(\boldsymbol{\alpha}^T X - A)/(B - A)\} + \gamma Z]$, with the following settings: $\boldsymbol{\alpha}_{p \times 1} = \mathbf{0}_{p,1}$ for type I error;

$\boldsymbol{\alpha}_{p \times 1} = \{\mathbf{0}_{1,p/2}, \mathbf{1}_{1,p/2}\}^T$ for power; $A = \frac{\sqrt{3}}{2} - \frac{1.645}{\sqrt{12}}$, $B = \frac{\sqrt{3}}{2} + \frac{1.645}{\sqrt{12}}$, $\gamma = 0.3$, $H(\cdot)$ = the inverse logistic link;

The predictors are generated from the following setting: $X \sim \text{Unif}(0,1)$, $Z = 0$ if observation is odd, $Z = 1$ if observation is even; The testing hypothesis is $H_0 : \text{pr}(Y=1|Z) = H\{\boldsymbol{\kappa}_0 + \gamma Z\}$ vs

$H_1 : \text{pr}(Y=1|X,Z) = H\{\boldsymbol{\kappa}_0 + \sum_{j=1}^J \beta_j h_j(X\boldsymbol{\alpha}) + \gamma Z\}$; These results are based on sample size $n=100$ and 1000 simulated data sets.

	$m \backslash \text{dim}(p)$	100	200	300	500	800	1000
Type I error	4.50	0.113	0.029	0.025	0.024	0.022	0.003
	5.00	0.029	0.029	0.023	0.011	0.019	0.012
	5.50	0.018	0.017	0.014	0.012	0.015	0.010
	6.00	0.005	0.001	0.001	0.001	0.001	0.001
Power	4.50	0.982	0.982	0.930	0.961	0.972	0.952
	5.00	0.970	0.973	0.931	0.890	0.973	0.960
	5.50	0.941	0.840	0.800	0.761	0.962	0.943
	6.00	0.820	0.801	0.710	0.689	0.812	0.920

Table 4:

The average value of type I error using HOT_p vs the the amplifying multiplier (m) and the dimension of predictor (p) in simulation study under sin-bump setting and high-dimension case; HOT_p = HOT with penalized estimating equation; The data were generated from: $\text{pr}(Y = 1|X, Z) = H[m \sin\{\pi(\alpha_T X - A)/(B - A)\} + \gamma Z]$, with the following settings: $\alpha_{p \times 1} = \mathbf{0}_{p,1}$ for type I error; $\alpha_{p \times 1} = \{\mathbf{0}_{1, p/2}, \mathbf{1}_{1, p/2}\}^T$ for power; $A = \frac{\sqrt{3}}{2} - \frac{1.645}{\sqrt{12}}$, $B = \frac{\sqrt{3}}{2} + \frac{1.645}{\sqrt{12}}$, $\gamma = 0.3$ $H(\cdot)$ = the inverse logistic link; The predictors are generated from the following setting: $X \sim \text{Unif}(0,1)$, $Z = 0$ if observation is odd, $Z = 1$ if observation is even; The testing hypothesis is $H_0 : \text{pr}(Y = 1|Z) = H\{\kappa_0 + \gamma Z\}$ vs $H_1 : \text{pr}(Y = 1|X, Z) = H\{\kappa_0 + \sum_{j=1}^J \beta_j h_j(X\alpha) + \gamma Z\}$; These results are based on sample size $n=100$ and 1000 simulated data sets.

	$m \backslash \text{dim}(p)$	100	200	300	500	800	1000
Type I error	4.50	0.108	0.093	0.051	0.039	0.032	0.030
	5.00	0.029	0.025	0.022	0.023	0.020	0.017
	5.50	0.012	0.009	0.011	0.010	0.003	0.009
	6.00	0.006	0.001	0.001	0.001	0.001	0.001
Power	4.50	0.998	0.996	0.965	0.951	0.962	0.990
	5.00	0.998	0.934	0.948	0.943	0.951	0.982
	5.50	0.997	0.988	0.934	0.940	0.920	0.980
	6.00	0.981	0.988	0.933	0.930	0.947	0.967

Table 5:

Significant pathways using our hybrid omnibus test approaches: HOT=Hybrid Omnibus Test; HOT_p =HOT with penalized estimating equation; Pathways identified by Bayesian inference based on ABF larger than 10 are marked as bold letter; Pathways with * are also identified by Mootha et al (2003)

Pathway Name	Num of Genes	P-value (HOT_p)
Tryptophan metabolism	60	0.007
Complement and coagulation cascades	47	0.010
MAP00380_Tryptophan_metabolism	60	0.011
MAPK signaling pathway	274	0.013
Starch and sucrose metabolism	54	0.015
Ubiquinone biosynthesis	46	0.016
MAP00071_Fatty_acid_metabolism	65	0.018
Oxidation Phosphorylation*	43	0.018
Taurine_and_hypotaurine_metabolism	8	0.02
MAP00710_Carbon_fixation	22	0.021
Propanoate metabolism	40	0.021
Pyrimidine metabolism	61	0.023
c17_U133_probes	116	0.023
<i>Alanine and aspartate metabolism*</i>	18	0.025
Apoptosis	92	0.025
Carbon fixation	25	0.025
Alzheimer's disease	50	0.027
Histidine metabolism	37	0.028
Glycerolipid metabolism	43	0.029
Parkinson's disease	40	0.029
Wnt signaling pathway	140	0.029
Nicotinate and nicotinamide metabolism	43	0.030
Ascorbate_and_aldarate_metabolism	10	0.033
Sulfur metabolism	9	0.037
Dichloroethane_degradation	10	0.038
Phosphatidylinositol signaling system	58	0.043
MAP00190_Oxidative_phosphorylation*	58	0.044
MAP00051_Fructose_and_mannose_metabolism	23	0.045
Circadian rhythm	20	0.048