Goodness-Of-Fit Test for Nonparametric Regression Models: Smoothing Spline ANOVA Models as Example

Sebastian J. Teran Hidalgo[a,*], Michael C. Wu[b], Stephanie M. Engel[c], Michael R. Kosorok[d]

[a]*Department of Biostatistics, Yale University, New Haven, Connecticut, U.S.A.*
[b]*Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.*
[c]*Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.*
[d]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.*

Abstract

Nonparametric regression models do not require the specification of the functional form between the outcome and the covariates. Despite their popularity, the amount of diagnostic statistics, in comparison to their parametric counterparts, is small. We propose a goodness-of-fit test for nonparametric regression models with linear smoother form. In particular, we apply this testing framework to smoothing spline ANOVA models. The test can consider two sources of lack-of-fit: whether covariates that are not currently in the model need to be included, and whether the current model fits the data well. The proposed method derives estimated residuals from the model. Then, statistical dependence is assessed between the estimated residuals and the covariates using the HSIC. If dependence exists, the model does not capture all the variability in the outcome associated with the covariates, otherwise the model fits the data well. The bootstrap is used to obtain p-values. Application of the method is demonstrated with a neonatal mental development data analysis. We demonstrate correct type I error as well as power performance through simulations.

Keywords: Goodness-of-fit; Interaction testing; Smoothing spline models;

*Corresponding author
Email address: sebastian.teranhidalgo@yale.edu (Sebastian J. Teran Hidalgo)

Bootstrap; Test of independence.

1. Introduction

Nonparametric regression models can provide a better fit when parametric assumptions are too restrictive (e.g., linearity of the mean). A popular nonparametric model from the machine learning literature is kernel ridge regression (KR) ([1, 2]). In KR regression, the input covariates are mapped to a high (possibly infinite) dimensional space through a kernel function, but which is only represent through a kernel matrix of the sample points. Another popular method is k-nearest neighbor (KNN) regression ([3]), in which for each observation, an average of the outcome is taken across all k-closest values with respect to the covariates. The fitted values KNN regression look jagged and are hard to interpret. Smoothing spline ANOVA models (SS-ANOVA) are also a popular nonparametric regression methodology this time arising from the statistical literature ([4, 5, 6, 7, 8]). SS-ANOVA models estimate the mean of an outcome as a smooth function with an ANOVA decomposition which partitions the variation of the outcome attributed to the covariates into main effects, two-way interactions, and all other higher-level interactions, but as a summation of functions, not constants, as with classical ANOVA. The KR, KNN, SS-ANOVA regressions as well as kernel smooth regression, local polynomial regression and others are *linear smoothers*, meaning that the fitted values are a linear function of the outcome.

Several goodness-of-fit tests exist for nonparametric regression models. The literature contains results on testing a parametric form under the null hypothesis against a nonparametric (semiparametric) one under the alternative hypothesis. Examples include tests of deviation from the parametric linear regression [9, 10]. Another method tests the goodness-of-fit of a linear model, which can potentially detect a nonparametric form under the alternative [11]. A test that allows a nonparametric form under the null exists [12], but it is only defined for a model with one covariate. A test that allows for multiple covariates in the

Nadaraya–Watson (NW) regression model [13] exists in the literature. Although similar to our method, the usage of the NW regression model is different from our current research for two fundamental reasons. First, the NW regression jointly models the mean of the outcome conditional on the covariates. Therefore, it is difficult to build examples where the conditional mean suffers from lack-of-fit and the test is essentially for heteroscedasticity of the residuals. The simulation scenarios are examples where the lack-of-fit comes only from the variance of the residuals being dependent on the covariates, and lack-of-fit coming from modelling the mean is never explored. Second, nonparametric methods like GAMs and SS-ANOVA models are not considered. Examples where the lack-of-fit comes from modelling the mean incorrectly could be explored by using these models. Given the popularity of these regression models, we believe it is important to develop a goodness-of-fit methodology for them.

Another article [14] generalizes a Tukey-type test of additivity proposed in[15]. This test can only detect two-way interactions that are a product of the main effects, which limits the type of departures of goodness-of-fit it can detect. Also, using generalized additive models (GAMs) a test for specific interaction terms exists [16]. That means that under the null the GAM has some main effects and interactions, and under the alternative the same model holds but with the addition of one more interaction. This is not a goodness-of-fit test since the user of this test has to specify the model under the alternative. A methodology exists to test whether an extra set of variables should be included in a nonparametric regression model [17].

Our literature review shows that a general test for nonparametric regression that detects lack-of-fit in modelling the mean does not exist. This paper resolves these issues by proposing a goodness-of-fit test for nonparametric regressions that are *linear smoothers* as defined in (2) (i.e., regression models that make use of a matrix to transform the outcome vector into a vector of fitted values). Our method is an innovation with respect to previous methods in that:

- Our goodness-of-fit methodology tests a nonparametric model under the

null against a general alternative, which can include parametric as well as nonparametric forms. Other methods require the model under the null to be parametric [18, 9, 11, 10], the regression to be univariate [12], require the form to be multiplicative, or have specific interactions[14], none of which are limitations in the current research.

- Other methods exist for testing the independence between residuals and covariates, as in the current research, but only in the context where the lack-of-fit comes from departures from the homoscedasticity assumption [19, 13]. The case where the lack-of-fit comes from incorrectly modelling the mean has not yet been analyzed. This is of importance given the fact that models like GAMs and SS-ANOVA have become highly popular as nonparametric models and they can suffer from lack-of-fit of the mean.

- Our method can incorporate testing of external variables, as [17]. Thus, we present a unified framework to test both for goodness-of-fit with respect to variables used to build the model or a set of external variables, an unification that was not attempted in the literature cited above.

- Our methodology provides a degree of freedom adjustment for the bootstrap null distribution, which was missing from the reviewed literature [11, 13].

We will use SS-ANOVA models throughout in examples, theory and simulations, but we emphasize that this methodology can be applied to any nonparametric linear smoother. The assessment of goodness-of-fit will be accomplished by fitting the model of interest and obtaining estimated residuals. The residuals contain the leftover information that remains unexplained by the model. Statistical dependence is then assessed between the estimated residuals and the covariates in the model, with the Hilbert-Schmidt independence criterion (HSIC). If dependence exists, the model does not capture all the variability in the outcome associated with the covariates. If no dependence exists, the model fits the data well. This process can also be used with covariates that are not

in the model, in order to assess whether their absence contributes to lack-of-fit. A test statistic is created from the HSIC between residuals and covariates to test for lack-of-fit. The bootstrap is used to derive p-values. The degrees of freedom of the model are calculated as the trace of the hat matrix and are used to adjusted the variance of the bootstrap distribution. The current article is an extension of the goodness-of-fit test for linear models proposed by Sen and Sen [11]. The major contributions we make to the literature include: identifying the need for assessing goodness-of-fit in a nonparametric regression, developing a test statistic, creating a variance adjustment to the bootstrap to improve the finite sample performance of the method, providing theoretical justification the use of the HSIC, and demonstrating correct type I error as well as power performance through numerical simulations.

This paper is organized as follows. In section 2 the method for goodness-of-fit for linear smoothers is introduced. Section 2 includes an introduction to linear smoothers, a formal definition of SS-ANOVA, a description of the evaluation of goodness-of-fit using the HSIC, the bootstrap for deriving p-values for the test statistic, and illustrative cases of lack-of-fit. In section 3, simulation results are presented, in section 4 application of the method is demonstrated with a neonatal mental development data analysis, and section 5 is a concluding discussion of the proposed method.

2. Goodness-Of-Fit Test For Nonparametric Regressions

This section describes linear smoothers, SS-ANOVA models, the HSIC, our proposed goodness-of-fit test based on residuals, and the bootstrap approximation to the null distribution. Then, theoretical results and illustrative cases are discussed.

2.1. Linear Smoothers

The current article will develop a goodness-of-fit test for nonparametric regression models which are linear smoothers. However, on the next section and

the rest of the paper the focus will be in one such type of linear smoothers: SS-ANOVA models. The reasons for this is that the ANOVA decomposition is very useful in creating examples where interactions and main effects are the source of lack-of-fit.

We assume the observed data consists of (Y, X), where Y is a dependent variable, $X \in [0,1]^p$ is a vector of covariates, and

$$Y = f_0(X) + \eta, \tag{1}$$

for an unknown function f_0 and random residual η, which is independent of X, with $E[\eta] = 0$. An i.i.d. sample $(\mathbf{X}_n, \mathbf{Y}_n) = \{(x_1, y_1),...,(x_n, y_n)\}$ is drawn from (1). In the current context, a linear smoother is a nonparametric regression that estimates f_0 in (1) such that

$$\hat{f} = \mathbf{A}\mathbf{Y}_n \tag{2}$$

that means that the vector of fitted values $\hat{\mathbf{Y}}_n$ can be written as a linear function of the outcome vector $\mathbf{Y}_n$. This is important because for a nonparametric regression that estimates $\hat{f}$ with the form of (2) the degrees of freedom of the model can be defined as

$$df(\hat{\mathbf{Y}}_n) = Tr(\mathbf{A}), \tag{3}$$

where $Tr(\mathbf{A})$ is the trace of the hat matrix $\mathbf{A}$. In the context of the current research, the degrees of freedom defined in (3) will be used to rescale the estimated residuals such that they have the correct variance. This is an essential step in generating the null distribution through the bootstrap for our proposed goodness-of-fit test. When the degrees of freedom are not accounted for, the variance of the null distribution is severely underestimated, something that was not address in the work by Sen and Sen [11].

2.2. SS-ANOVA

SS-ANOVA models are a special case of linear smoothers. In this section we describe how they work. A sample of size n denoted by $(x_1, y_1),...,(x_n, y_n)$ is drawn from (1). We assumed throughout that the response has been centered.

Estimation of f_0 can be done through minimization of the following penalized least squares with respect to f:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_n J(f). \tag{4}$$

In the case where $p = 1$, then $f(x)$ is just a univariate function and $J(f) = \int_0^1 f^{(k)}(x)^2 dx$, and $f^{(k)}$ is the k-th derivative of f. In the case where $p > 1$, then $f(X) = \sum_{j=1}^{p} f_j(X(j))$, where $X(j)$ is the j-th element of X, and $J(f) = \sum_{j=1}^{p} \theta_j^{-1} \int_0^1 f_j^{(k)}(x)^2 dx$. This corresponds to a semiparamtric additive model. In the general case

$$f(X) = \sum_j f_j(X(j)) + \sum_{j<k} f_{j,k}(X(j), X(k)) + \cdots \tag{5}$$

$$\text{and} \quad J(f) = \sum_{\alpha} \theta_{\alpha}^{-1} ||P_\alpha f||^2_{\mathscr{H}_\alpha} + \sum_{\alpha\beta} \theta_{\alpha\beta}^{-1} ||P_{\alpha\beta} f||^2_{\mathscr{H}_{\alpha\beta}} + \cdots,$$

where λ_n and $\boldsymbol{\theta}$ are tuning parameters which are selected through Generalized Cross Validation (GCV). We define the averaging operator as

$$\mathscr{E}_\alpha f = \int_0^1 f(t_1, \ldots, t_p) dt_\alpha. \tag{6}$$

Then, the main effects are defined as $f_\alpha = (I - \mathscr{E}_\alpha) \prod_{\beta \neq \alpha} \mathscr{E}_\beta f$, two-way interactions as $f_{\alpha,\beta} = (I - \mathscr{E}_\alpha)(I - \mathscr{E}_\beta) \prod_{\gamma \neq \alpha,\beta} \mathscr{E}_\gamma f$, and so forth. Because of this construction the terms of the decomposition satisfy side conditions of orthogonality and $\mathscr{E}_\alpha f_\alpha = 0$. Also, $P_\alpha f$ is the projection of f into the main effect space indexed by α after subtracting the first k-th polynomial terms depending on α, $P_{\alpha,\beta} f$ is the projection into the two-way interaction space indexed by α and β after subtracting the first k-th polynomial terms and their multiplicative interactions depending on α and β, each space with corresponding norms $||g||_{\mathscr{H}_\alpha} = \int_0^1 (g^{(k)}(t_\alpha))^2 dt_\alpha$ and $||g||_{\mathscr{H}_{\alpha\beta}} = \int_0^1 (g^{(k)}(t_{\alpha,\beta}))^2 dt_{\alpha,\beta}$, respectively. Higher order terms are defined similarly. For more details we refer the reader to [20].

2.3. Solution to Penalized Least Squares

For simplicity, this section assumes that the functional form of f is additive. Discussion of more complicated forms, i.e., which includes two-way

and other higher level interactions, can be found in [6]. The form in (4) will be minimized assuming that the data generating mechanism is (1) such that $f(x) = \sum_j f_j(x(j))$ or that $f \in \mathscr{H} = \oplus_{\beta=1}^{p} \mathscr{H}_\beta$. In this case, $J(f) = \sum_{j=1}^{p} \theta_j^{-1} \int_0^1 f_j^{(k)}(x(j))^2 dx_j$.

RKHS $\oplus_{j=1}^{p} \mathscr{H}_j$

To each $f_j \in \mathscr{H}_j$ corresponds a reproducing kernel. This happens because f_j can be decomposed by Taylor expansion at 0 as

$$f_j(x(j)) = \sum_{v=0}^{k-1} \frac{x^v(j)}{v!} f_j^{(v)}(0) + \int_0^1 \frac{(x(j)-u)_+^{k-1}}{(k-1)!} f_j^{(k)}(u) du.$$

Then $\mathscr{H}_j$ can be decomposed into a tensor sum $\mathscr{H}_j = \mathscr{H}_{j,0} \oplus \mathscr{H}_{j,1}$, where $\mathscr{H}_{j,1}$ is an RKHS with the following reproducing kernel

$$R_{j,1}(x(j), y(j)) = \int_0^1 \frac{(x(j)-u)_+^{k-1}}{(k-1)!} \frac{(y(j)-u)_+^{k-1}}{(k-1)!} du.$$

The space $\mathscr{H}_{j,0}$ has a polynomial basis of degree $k-1$ such that $\phi_j^k(x) = \{1, x(j), x^2(j), \ldots, x^{k-1}(j)\}$. Let ϕ^k be the polynomial basis of degree k of the tensor sum $\oplus_{j=1}^{p} \mathscr{H}_{j,0}$, such that $\phi^k(x) = \{\phi_1^k(x), \phi_2^k(x), ..., \phi_p^k(x)\} = \{1, x(1), x^2(1), \ldots, x^{k-1}(1), \ldots, x(p), x^2(p), \ldots, x^{k-1}(p)\}$. Moreover, the reproducing kernel of $\oplus_{j=1}^{p} \mathscr{H}_{j,1}$ will be $R_J(x,y) = \sum_{j=1} \theta_j R_{j,1}(x(j), y(j))$.

Solution as a Linear Smoother

With an i.i.d. sample $(\mathbf{X}_n, \mathbf{Y}_n) = \{(x_1, y_1), ..., (x_n, y_n)\}$ and form of $f(x) = \sum_{j=1}^{p} f_j(x)$, by the representer theorem ([8], [21]) the minimizer of (4) has the form

$$f(x) = \sum_{j=1}^{p} d_j \phi_j^k(x) + \sum_{i=1}^{n} c_i R_J(x_i, x),$$

where d_j is a vector of coefficients of length k. Then the estimation reduces to minimizing

$$(\mathbf{Y}_n - \mathbf{Sd} - \mathbf{Qc})^T (\mathbf{Y}_n - \mathbf{Sd} - \mathbf{Qc}) + n\lambda_n \mathbf{c^T Q c}.$$

with respect to the vectors $\mathbf{c}$ and $\mathbf{d}$ of length n and m, respectively, where m is the length of the basis $\phi^k(x)$. Here, $\mathbf{S}$ is $n \times m$ matrix, with the ith row

corresponding to $\phi^k(x_i)$, Q is $n \times n$ with the (i,j)th entry $R_J(x_i, x_j)$. Then by taking derivatives and setting equal to 0, we get that the solution of (4) with additive function is of the form

$$\hat{\mathbf{Y}}_n = \mathbf{A}(\lambda_n, \boldsymbol{\theta})\mathbf{Y}_n,$$

such that

$$\mathbf{A}(\lambda_n, \boldsymbol{\theta}) = \mathbf{I} - n\lambda_n(\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{M}^{-1}), \tag{7}$$

where $\mathbf{M} = \mathbf{Q} + n\lambda_n\mathbf{I}$. The fitted values $\hat{\mathbf{Y}} = \mathbf{A}(\lambda_n, \boldsymbol{\theta})\mathbf{Y}_n$ are in the form of linear smoothers as described in (2) and the degrees of freedom of the SS-ANOVA model are $Tr(\mathbf{A}(\lambda_n, \boldsymbol{\theta}))$.

The parameters λ_n and $\boldsymbol{\theta}$ are chosen by generalized cross-validation (GCV). The GCV statistic, as defined by [4], corresponds to

$$\text{GCV}(\lambda_n, \boldsymbol{\theta}) = \frac{n^{-1}||(\mathbf{I} - \mathbf{A}(\lambda_n, \boldsymbol{\theta}))\mathbf{y}||^2}{(n^{-1}Tr(\mathbf{I} - \mathbf{A}(\lambda_n, \boldsymbol{\theta})))^2},$$

where $\hat{\mathbf{Y}}_n = \mathbf{A}(\lambda_n, \boldsymbol{\theta})\mathbf{Y}_n$. λ_n and θ are chosen to minimize $\text{GCV}(\lambda_n, \boldsymbol{\theta})$. In the current research, the model used throughout will be the *Cubic* SS-ANOVA. This corresponds to the case where $k = 2$, or when the integral of the second derivative is being penalized, namely $\int_0^1 f_j''(x_j)^2 dx_j$.

After fitting a SS-ANOVA model (or any other nonparametric linear smoother), it is important to do some model diagnostics. Model diagnostics are statistics that check how well a model fits to the data. In the current research, the independence of the estimated residuals with respect to a set of covariates will be assessed using an independence statistic. The independence statistic that we will use is HSIC. A formal definition will be presented in the next subsection.

2.4. HSIC

Recent developments in tests of statistical independence are Brownian Distance Covariance ([22, 23, 24]) and HSIC ([25, 26]). Distance Covariance (DC) is defined as the weighted norm between the product of two random vectors' individual characteristic function and the joint characteristic function of these

two vectors. If this normed difference is 0 then these two vectors are statistically independent. The authors developed a sample version of DC that depends only on the Euclidean distances between the points. The HSIC is the Cross-Covariance Operator between two reproducing kernel Hilbert spaces (RKHSs). When this operator equals 0 for two vectors of random variables that are defined on the domain of two different RKHSs with universal kernels, then these two vectors are statistically independent. The sample version HSIC is exactly the same as the one for DC except that Euclidean distances are replaced by kernel distances.

The HSIC allows us to evaluate the statistical dependence between two random vectors of arbitrary dimensions. The goodness-of-fit statistic is based on the HSIC, because it can evaluate the statistical dependence between the estimated residuals and a set of covariates.

Let X and Y be vectors of random variables on the domain $\mathscr{X}$ and $\mathscr{Y}$, respectively, with $\mathscr{X} \subset \mathbb{R}^p$ and $\mathscr{Y} \subset \mathbb{R}^q$, and with joint probability measure P_{xy}. Let $\mathscr{F}$ and $\mathscr{G}$ be RKHSs on $\mathscr{X}$ and $\mathscr{Y}$ with reproducing universal kernel functions k and l. Gaussian kernels fulfill this requirement ([27]). The $HSIC(P_{xy}, X, Y)$ between X and Y is defined as

$$E_{\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}'}[k(X,Y')l(Y,Y')]] + E_{\mathbf{x},\mathbf{x}'}[k(X,X')]E_{\mathbf{y},\mathbf{y}'}[l(Y,Y')] \\ - 2E_{\mathbf{x},\mathbf{y}}[E_{\mathbf{x}'}[k(X,X')]E_{\mathbf{y}'}[l(Y,Y')]],$$

where the subscript under the operator E denotes which random variables we are taking the expectation with respect to. We will rely on the following theorem ([25]):

Theorem 1. *$HSIC(P_{xy}, X, Y) = 0$ if and only if X and Y are statistically independent, i.e., $P_{x,y} = P_x \times P_y$.*

With an i.i.d. sample $(\mathbf{X}_n, \mathbf{Y}_n) = \{(x_1, y_1),...,(x_n, y_n)\}$ from P_{xy}, $HSIC(P_{xy}, X, Y)$ can be estimated consistently with

$$T_n(\mathbf{X}_n, \mathbf{Y}_n) := n^{-2}Tr(KHLH),$$

where $H, K, L \in \mathbb{R}^{n \times n}$, $H_{i,j} := \delta_{i,j} - n^{-1}$, $K_{i,j} := k(x_i, x_j)$, $L_{i,j} := l(y_i, y_j)$, and $\delta_{i,j}$ is the kronecker delta. The statistic can be rewritten as

$$\frac{1}{n^2}\sum_{i,j}^{n} K_{ij}L_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} K_{ij}L_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} K_{ij}L_{iq}.$$

The kernels $k(x_i, x_j) = \exp(-||x_i - x_j||^2/\sigma^2)$ and $l(y_i, y_j) = \exp(-||y_i - y_j||^2/\sigma^2)$ are called Gaussian and satisfy the universal kernel conditions and will be the ones used throughout this paper with σ^2 held fix at 1 and $||\cdot||$ being the Euclidean norm. In the next subsection it will be shown how the ability of HSIC to discover arbitrary statistical dependencies can be used in conjunction with the estimated residuals from the SS-ANOVA model and a set of covariates to form a goodness-of-fit statistic.

2.5. Goodness-Of-Fit Test Based on Residuals

This subsection introduces the proposed goodness-of-fit test. After fitting a nonparametric regression with linear smoother form as in (2), the goodness-of-fit of the model can be evaluated by looking at the relationship between a set of covariates and the estimated residuals. If dependence exists, the model does not capture all the variability in the outcome associated with the covariates and further terms are needed. If no dependence is detected, all information in the covariates that is present in the response has been explained through the model. The test can consider two sources of lack-of-fit: whether the current model fits the data well, (i.e, whether the model captures all the variation in the outcome associated to the covariates,) and whether covariates that are not currently in the model need to be included.

The method will be developed in the context of SS-ANOVA models, meaning that theory, examples and simulations will be developed using SS-ANOVA models. This is so because SS-ANOVA models are particularly useful in illustrating lack-of-fit in a nonparametric setting (e.g., by non inclusion of ANOVA terms that should be included in the model). However, the test we developed can be applied to any nonparametric regression model with solution of the form (2).

We assume that the true data generating mechanism is as in (1). We specify a SS-ANOVA model between Y and X. Thus, our model is

$$Y = f(X) + \varepsilon, \tag{8}$$

and $f(X)$ has a ANOVA decomposition as in (5), but (most likely in applications) does not included all possible interactions. For example, a popular choice of model would be the main effects only model

$$f(X) = \sum_j f_j(X(j)), \tag{9}$$

or the model with all main effects and all two-way interactions

$$f(X) = \sum_j f_j(X(j)) + \sum_{j<k} f_{j,k}(X(j), X(k)). \tag{10}$$

We denote η by the true error and ε by the model error.

2.5.1. Lack-of-fit

To assess the goodness-of-fit of the SS-ANOVA model in (8) we define

$$\varepsilon := Y - f(X),$$

and test the following hypotheses:

$$\begin{aligned} H_0 &: HSIC(P_{x,\varepsilon}, X, \varepsilon) = 0 \\ H_A &: HSIC(P_{x,\varepsilon}, X, \varepsilon) > 0. \end{aligned} \tag{11}$$

If the null holds, the model error equals the true error ($\varepsilon = \eta$), ε is independent of X and $HSIC(P_{x,\varepsilon}, X, \varepsilon) = 0$. If the alternative holds, then $\varepsilon = \varepsilon(X) \neq \eta$, ε is dependent on X and $HSIC(P_{x,\varepsilon}, X, \varepsilon) > 0$. Dependence between ε and X would indicate lack-of-fit because the nonparametric model does not capture all the variation in the model with respect to X. It is important to emphasize that the alternative where ε is dependent on X can be capture a large class of lack-of-fit scenarios.

For example, if the model in (9) is taken in consideration, the alternative can hold because the assumption of main effects only model is incorrect, but for

example because the true model is (10), and in reality we have

$$\varepsilon = \sum_{j<k} f_{j,k}(X(j), X(k)) + \cdots + \eta,$$

which still depends on X and η is the true error term. Naturally, not all the terms of the decomposition have to exist under the alternative.

Let $(\mathbf{Y}_n, \mathbf{X}_n) = \{(y_1, x_1), ..., (y_n, x_n)\}$ be a random sample from the data generating mechanism described in (1), and we want to test the null and alternative hypotheses in (11). To accomplish this, we define

$$\hat{\varepsilon}_i := y_i - \hat{f}(x_i),$$

for $i = 1, \ldots, n$, where $\hat{f}$ is the solution to (4), with $f(X)$ having a ANOVA structure as in (5), and let $\hat{\boldsymbol{\varepsilon}}_n = \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$. Then, the statistic

$$nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n), \tag{12}$$

is used to test the hypotheses. This test procedure is intuitive, since $T_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ is an estimate of $HSIC(P_{x,\varepsilon}, X, \varepsilon)$, and later we show it is consistent both under the null and the alternative hypotheses.

2.5.2. Testing for Significance

We can also test whether covariates that are currently not in the model should be included. We assume a model for the data as in (8), meaning that $f(X)$ has a ANOVA decomposition as in (5). There exists another set of covariates, which is denoted by Z. To assess whether Z should be included in the model, in other words, if there is a lack-of-fit with respect to Z, we define

$$\varepsilon := Y - f(X),$$

and test the following hypotheses:

$$\begin{aligned} H_0 &: HSIC(P_{z,\varepsilon}, Z, \varepsilon) = 0 \\ H_A &: HSIC(P_{z,\varepsilon}, Z, \varepsilon) > 0. \end{aligned} \tag{13}$$

If the the null holds, the model error equals the true error ($\varepsilon = \eta$), ε is independent of Z, and $HSIC(P_{z,\varepsilon}, Z, \varepsilon) = 0$. This means that Y, after taking into

account the effect of X, is independent of Z. If the alternative holds, then Y depends on Z even after taking into account the effect of X, $HSIC(P_{z,\varepsilon}, Z, \varepsilon) > 0$, and there is a lack-of-fit with respect to Z (i.e., Z should be included in the model).

With an i.d.d. sample $(\mathbf{Y}_n, \mathbf{X}_n, \mathbf{Z}_n) = \{(y_1, x_1, z_1), ..., (y_n, x_n, z_n)\}$ from the data generating mechanism described in (1), we can test the null and alternative hypotheses in (11). To accomplish this, we define

$$\hat{\varepsilon}_i := y_i - \hat{f}(x_i),$$

for $i = 1, \ldots, n$, where $\hat{f}$ is the solution to (4) and let $\hat{\boldsymbol{\varepsilon}}_n = \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$. Then, the statistic

$$nT_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n). \tag{14}$$

is used to test the hypotheses. This makes sense since $T_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n)$ is an estimate of $HSIC(P_{z,\varepsilon}, Z, \varepsilon)$, and later we show it is consistent both under the null and the alternative hypothesis.

The test statistic in (12) and (14) is the proposed statistic to test the goodness-of-fit of the SS-ANOVA. The test statistic also works for any linear smoother of form (2), but by calculating the residuals in (12) or in (14) with their respective models (e.g., using the residuals of KNN, KR or any other linear smoother). The model fit can be easily assessed by first estimating the residuals and then calculating the test statistic in (12) or (14) to check for a lack-of-fit, with respect to the covariates used to create the model (denoted above by $\mathbf{X}_n$) or a set of external covariates not previously included in the model (denoted previously as $\mathbf{Z}_n$), respectively.

To perform the test, we need a valid distribution of (12) and (14) under the null hypothesis. An approximation to the null distribution is used. Details are shown in the next section.

2.6. Approximation to the Null Distribution of the Test Statistic with the Bootstrap

The difficulty in using (12) as a test statistic is that it is hard to derive analytically a distribution under the null hypothesis that will provide the critical values for a given significance level. One obvious first approach would be to randomly permute the vector $\hat{\boldsymbol{\varepsilon}}_n$ to obtain $\hat{\boldsymbol{\varepsilon}}_\pi$, calculate $nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_\pi)$ and repeat this process many times to obtain a distribution under the null. This approach happens to be flawed. When the vector $\hat{\boldsymbol{\varepsilon}}_n$ is permuted with respect to $\mathbf{X}_n$, complete independence between the two is created. Under the null, ε and X are independent. However, even under the null, $\hat{\boldsymbol{\varepsilon}}_n$ and $\mathbf{X}_n$ are not independent because of the simple fact that $\hat{\boldsymbol{\varepsilon}}_n$ is a statistic based on $\mathbf{X}_n$. Under the null, $\hat{\boldsymbol{\varepsilon}}_n$ is just a good approximation of ε. Therefore, a different procedure is needed.

A model based bootstrap, which needs to address the following issues: the bootstrap generating process must account for the fact that under the null X and ε are independent, and the bootstrap samples $\mathbf{X}_n^*$ and $\boldsymbol{\varepsilon}_n^*$ must be correlated in a similar way that $\mathbf{X}_n$ and $\hat{\boldsymbol{\varepsilon}}_n$ are correlated. A bootstrap that fulfills these requirements, and which will be used to derive a p-value for the test statistic, is described below.

Bootstrap Algorithm

Step 1

Calculate the estimated residuals $\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$ and create an empirical distribution P_{n,e^o} of the residuals with mass $1/n$ at each $e_i^o = \frac{\hat{\sigma}}{\hat{\sigma}'}(\hat{\varepsilon}_i - \bar{\varepsilon})$, where $\bar{\varepsilon} = \sum_{i=1}^n \frac{\hat{\varepsilon}_i}{n}$, $\hat{\sigma}'^2 = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\varepsilon})^2}{n}$ and $\hat{\sigma}^2 = \frac{||\mathbf{Y}-\mathbf{A}\mathbf{Y}||^2}{Tr(\mathbf{I}-\mathbf{A})}$. Below it will be explained why the term $\frac{\hat{\sigma}}{\hat{\sigma}'}$ is present in the empirical distribution P_{n,e^o}.

Step 2

Draw a bootstrap sample η^* from the empirical distribution P_{n,e^o} and draw a bootstrap sample $\mathbf{X}_n^*$ from the empirical distribution $P_{n,\mathbf{X}}$ of the $\mathbf{X}_n$'s independently of η^*. Then set y_i^* as

$$y_i^* = \hat{f}(x_i^*) + \eta_i^* \quad \text{for } i = 1, \ldots, n.$$

Step 3

We estimate $\hat{f}^*$ from $\mathbf{Y}_n^*$ and from $\mathbf{X}_n^*$, and create new bootstrap residuals as

$$\varepsilon_i^* = y_i^* - \hat{f}^*(x_i^*) \quad \text{for } i = 1, \ldots, n.$$

Step 4

Calculate the test statistic as $nT_n(\mathbf{X}_n^*, \boldsymbol{\varepsilon}_n^*)$.

Step 5

Repeat Step 1 through 4 B times, so as to create B bootstrapped test statistics $nT_n(\mathbf{X}_n^*, \boldsymbol{\varepsilon}^*)_b$, for $b = 1, \ldots, B$. This distribution approximates the distribution of $nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ under the null. The p-value is then calculated as

$$\text{p-value} = \frac{1}{B} \sum_{i=1}^{B} I(nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n) \leq nT_n(\mathbf{X}_n^*, \boldsymbol{\varepsilon}_n^*)_b).$$

Remark 1: If hypotheses in (13) need to be tested using test statistic in (14) the same bootstrap can be used with small changes. Details are shown in section A.1 of the appendix.

Remark 2: In **Step 1**, the matrix $\mathbf{A}$ corresponds to the hat matrix for linear smoothers as defined in (2). Thus, bootstrap procedure described above would work for any nonparametric linear smoother model. In the current context of SS-ANOVA models, it corresponds exactly to $\mathbf{A}(\lambda_n, \boldsymbol{\theta})$ as defined in (7).

The variance of a random draw from the empirical distribution of the estimated residuals $\hat{\varepsilon}_i$, $i = 1, ..., n$, in **Step 1**, is $\hat{\sigma}'^2$. Under the null-hypothesis model, $\hat{\sigma}'^2 \xrightarrow{p} \sigma^2$, the true error variance. However, $\hat{\sigma}'^2$ underestimates σ^2 whenever the degrees of freedom of the model $Tr(\mathbf{A})$ increase relative to n, in finite samples, as shown in figure 1. Hence, if we draw from the distribution of the estimated residuals, our sample will have lower variance than what we want. One simple solution is to use an estimator of σ^2 that takes into account p. The estimator we use is $\hat{\sigma}^2$ whose denominator takes into account the degrees of freedom of the model. Whenever we rescale the empirical distribution of the estimated residuals by $\frac{\hat{\sigma}}{\hat{\sigma}'}$ then a random draw from this empirical distribution will have variance equal to $\hat{\sigma}^2$, which does not underestimate σ^2. Asymptotically, there is no difference in rescaling or not because $\frac{\hat{\sigma}}{\hat{\sigma}'} \xrightarrow{p} 1$, but simulations show that it makes an important difference for small and moderate sample sizes in estimating the null-hypothesis appropriately even when $Tr(\mathbf{A})$ is only moderately big. This is an improvement over the bootstrap procedure in Sen and Sen [11], which was used in the goodness-of-fit setting too, but for linear models. This finite sample variance adjustment is a key contribution of our approach.

2.7. Large Sample Approximation of the Test Statistic and the Bootstrap Procedure

The rationale of using $T_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ is that it approximates $HSIC(X, \varepsilon)$. The following theorem helps to justify this choice. Assume the true data generating mechanism is as in (1). A function f is assumed for the relationship between Y and X. Estimated residuals are obtained by finding a solution to (4) and setting $\hat{\varepsilon}_i = y_i - \hat{f}(x_i)$ for $i = 1, \ldots, n$. Below we assume **A.1-A.3**, found in the appendix, hold.

Theorem 2. *Under H_0,*

$$T_n(\mathbf{X}_n, \hat{\varepsilon}_n) \xrightarrow{p} HSIC(X, \eta) = 0.$$

Under H_A,

$$T_n(\mathbf{X}_n, \hat{\varepsilon}_n) \xrightarrow{p} HSIC(X, \varepsilon(X)) > 0.$$

Under both H_0 and H_A,

$$T_n(\mathbf{X}_n^*, \varepsilon_n^*) \xrightarrow{p} 0.$$

The notation $\varepsilon(X)$ emphasizes the fact that under the alternative ε is dependent on X. The proof of this result can be found in section *A.1*. Under the null $\varepsilon = \eta$, in its turn η is independent of X, and hence $HSIC(X, \eta) = 0$. Thus under the null, $T_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ approximates 0. Under the alternative, ε depends on $\mathbf{X}$, and $HSIC(X, \varepsilon) > 0$. Thus under the alternative, $T_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ will be greater than 0. This is the behavior needed for the test statistic in (12) to work. Moreover, the bootstrapped version of the test statistic $T_n(\mathbf{X}_n^*, \boldsymbol{\varepsilon}_n^*)$ converges to 0 in probability under both the null and the alternative. This is what the behavior of the bootstrap needs to be, since it must reflect the situation where the correct model is being specified and there is no leftover information in the residuals.

Remark: The theorem also holds when $(\mathbf{X}_n^*, \mathbf{X}_n, X)$ is replaced by $(\mathbf{Z}_n^*, \mathbf{Z}_n, Z)$, where $(\mathbf{Z}_n^*, \mathbf{Z}_n, Z)$ is defined in the section on the bootstrap.

2.8. Illustrative Cases

The framework presented in the current article is a test for the Goodness-of-fit for nonparametric regression. SS-ANOVA models are particularly useful for creating examples of lack-of-fit in a nonparametric context because we can specify a model were one of the ANOVA components is missing that does exist in the true data generating mechanism. This subsection presents three examples of lack-of-fit in SS-ANOVA models. In all three cases, the null-hypothesis will correspond to the situation where the specified model equals the true model, and under the alternative hypothesis, the model is misspecificed by not containing one or more terms from the ANOVA decomposition that are present in

315 the true model. Constructing examples of lack-of-fit like these would have been difficult through KR or KNN regression models.

Case I: Missing Interactions Beyond the Main Effects

After fitting a main effects only model with p covariates, a goodness-of-fit test is run. The SS-ANOVA model specified is

$$Y = \sum_{j=1}^{p} f_j(X(j)) + \varepsilon,$$

which, under the null, equals the true model, but under the alternative hypothesis in (11) the true model is:

$$Y = \sum_{j=1}^{p} f_j(X(j)) + f_{1,...,p}(X(1), ..., X(p)) + \eta,$$

where $f_{1,...,p}(X_1, ..., X_p)$ is an unspecified function that could be in any functional space except for the main effects only space from the SS-ANOVA decomposition. Under the alternative assumption, the test will pick up any possible interactions that exist beyond the main effects. This case is relevant because in most situations it is hard to know which interactions to include among the combinations of main effects, but it is very possible that interactions exist even when they are hard to conceptualize.

Case II: Missing Interactions Beyond the Within Group Interactions

Two groups of variables indexed by the sets A and B exist. The sets A and B are disjoint and their union is equal to $\{1, ..., p\}$. An SS-ANOVA model is fit which includes all p main effects and all possible interactions between variables with indexes in set A and B, separately. The SS-ANOVA model specified is

$$Y = f_A(X(A)) + f_B(X(B)) + \varepsilon,$$

which, under the null, equals the true model, but under the alternative hypothesis in (11) the true model is:

$$Y = f_A(X(A)) + f_B(X(B)) + f_{A,B}(X(A \cup B)) + \eta.$$

Here, $f_A(X(A))$ includes main effects and all possible interactions among the variables indexed by the set A. The same holds for $f_B(X(B))$ but over the set B. The form $f_{A,B}(X(A \cup B))$ remains unspecified and includes any possible interactions between variables in group A and B. Under the alternative assumption, the test should detect any possible interactions between covariates in group A and covariates in group B not included in the model described in H_0. This case is relevant because it is possible to know two groups of covariates that are known to be interacting and hence all the interactions are included. However, some cross interactions could also happen.

Case III: Testing for Significance

We can test whether a model that includes covariates X needs also to include covariates Z. The SS-ANOVA model specified is

$$Y = f(X) + \varepsilon,$$

which, under the null, equals the true model, but under the alternative hypotheses in (13) is:

$$Y = f(X, Z) + \eta.$$

Here, $f(X)$ includes main effects and could also include interactions, among the elements of X, if they are believed to exist. The same definition holds for $f(X, Z)$, but over the set both X and Z. However, the form of $f(X, Z)$ remains unspecified, but covariates Z are specified. Under the alternative assumption, the test will detect any covariate Z that is present in $f(X, Z)$. This case is relevant because many situations arise where the interest comes in detecting a set of covariates which affect the outcome beyond a previously defined set of variables.

In all three cases shown above, in order to perform the test, the model under H_0 is fitted and a vector of estimated residuals $\hat{\varepsilon}$ is obtained. For the first two cases, $nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}})$ is calculated as the test statistic. For the third case the test statistic is $nT_n(\mathbf{X}_n(B), \hat{\boldsymbol{\varepsilon}})$. These three cases represent possible departures of

fitness, but they do not exhaust all possibilities. However, no matter what the departure is, the goodness-of-fit can always be assessed with respect to an $\mathbf{X}_n$ (either the matrix used to fit the model or a completely new set of covariates), by checking its independence from the estimated residuals.

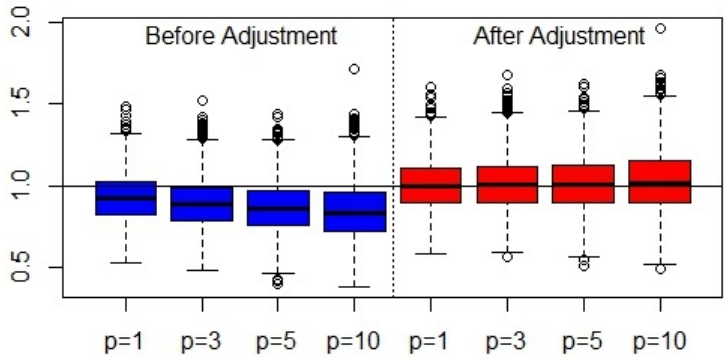

Figure 1: Variance Adjustment of the Distribution of the Estimated Residuals

Variance of the estimated residuals over 500 simulations. The variance is shown as the number of variables p in the model increases. The left panel shows the variance without adjustment, the right panel shows the variance with adjustment. The true variance is 1, which corresponds to the horizontal line.

3. Simulation Studies

This section will present simulation results comparing the variance of the empirical measure of the estimated residuals before and after the degrees of freedom of the model adjustment described in **Step 1** of the bootstrap algorithm. Also, it will present type I error and power results of the goodness-of-fit test under the three illustrative cases described above. It is important to reiterate that, for all three cases, a specific lack-of-fit has been specified under the alternative hypothesis, but that this is not known nor specified previously by

the researcher. The only objective of the test is to know if the current model under the null is sufficient. All simulations use SS-ANOVA models.

Variance Adjustment to the Bootstrap by the Degrees of Freedom of the Model
The left panel of figure 1 shows the box plots of $\hat{\sigma}'^2$ for 500 simulations of the null hypothesis for varying p and with a fixed sample size of 100. The right panel shows the same simulation scenarios but for $\hat{\sigma}^2$. The true variance for all the simulations is $\sigma^2 = 1$ denoted by the horizontal line. It can be seen that for moderate increments in dimension $\hat{\sigma}'^2$ underestimates the actual variance, whereas $\hat{\sigma}^2$ on average estimates σ^2 correctly. Thus, if the degrees of freedom of the model, as defined in (3), are not accounted for, the bootstrap procedure will not work properly for finite samples and will underestimate the actual variance of the residuals. Hence, our proposed adjustment is essential for the correct performance of the method.

In all cases shown below we simulated ε as $N(0,1)$ and all $X(j)$ as Uniform(0,1) independent of ε. Specific details of all simulations can be found in Appendix A of the supplementary materials.

Case I: Missing Interactions Beyond the Main Effects
Simulations were created where the null hypothesis only includes main effects. Therefore, we have $f(X) = \sum_{j=1}^{p} f_j(X(j))$, and under the alternative $f_{1,\ldots,p}(X(1), ..., X(p))$ is an interaction between covariates. The SS-ANOVA model specified is

$$Y = f(X) + \varepsilon,$$

whereas the models simulated under the null and alternative hypotheses of (11) are

$$\text{under } H_0,\ Y = f(X) + \eta, \text{ and}$$
$$\text{under } H_A,\ Y = f(X) + f_{1,\ldots,p}(X(1), ..., X(p)) + \eta.$$

When $p = 2$, the null model is $f(X) = 5\sin(\pi X(1)) + 2X(2)^2$ and the interaction added under the alternative is $f_{1,2}(X(1), X(2)) = 0.75\cos(\pi(X(1) - X(2)))$.

Type I Error				
Var.	Sig.	n=100	n=300	n=500
2	0.01	0.008	0.009	0.009
2	0.05	0.045	0.049	0.046
4	0.01	0.008	0.008	0.009
4	0.05	0.049	0.046	0.048
6	0.01	0.006	0.007	0.009
6	0.05	0.047	0.047	0.049
Power				
Var.	Sig.	n=100	n=300	n=500
2	0.01	0.019	0.152	0.512
2	0.05	0.106	0.493	0.886
4	0.01	0.008	0.0724	0.21
4	0.05	0.051	0.264	0.568
6	0.01	0.066	0.015	0.034
6	0.05	0.054	0.085	0.17

Var. corresponds to the number of variables used in the null model and Sig. corresponds to the significance level used in the test.

Table 1: Missing Interactions Beyond the Main Effects

When $p = 4, 6$ similar models were used.

A model of this nature would be hard to fit with a linear model given that the response depends sinusoidally on X_1 and depends quadratically on X_2, hence the handiness of SS-ANOVA. This simulation setting demonstrate how the ANOVA decomposition can be useful in picking up signals from interactions. After fitting a main effects only model, if the goodness-of-fit test is significant, then this would mean that main effects are insufficient and possibly some interactions exist. When the alternative is $f_{1,2}(X(1), X(2)) = 0.75\cos(\pi(X(1) - X(2)))$, the user of the test does not know between which covariates the interaction is happening. However, when the test is rejected, it is known that the main effects

only model is not sufficient and extra interactions might be needed. Thus, the test can be useful in finding interactions. From the simulation results in table (1) it can be seen that the method preserves the correct type I error both at the 0.01 and 0.05 significance levels. The size of the test gets sharper with increasing sample size. This happens because the bootstrap is a large sample method and will work best for larger sample size. For a given number of covariates, it can be seen that the power increases with larger sample size. Moreover, when more covariates are present in the main effects only model, the power decreases. This is due to the fact that the more main effects are included, the greater the number of possible interactions, and hence the alternative space becomes larger.

Case I: Departures from linearity through interactions

We created another set of simulations of Case I in order to compare our method to competitors. Under the null, the model is linear in the covariates, but under the alternative the covariates also have quadratic interactions. Our method tests the goodness-of-fit of a SS-ANOVA model with main effects only. Under the alternative the test statistic nT_n will detect the quadratic interactions. The test statistics $T^*_{AN,1}$ and S_n, described in [9] and [10] respectively, fit a linear model and detect departures from linearity. Although they are not testing the same null and alternative hypothesis, all three tests should have power to detect the alternative and should, in theory, approach the specified type I error under the null. In table 2 we show the results of the three tests, but below we only provide details of the hypotheses formulation for our method.

Simulations were created where the null hypothesis only includes linear main effects. Therefore, we have $f(X) = \sum_{j=1}^{p} X(j)\beta_j$, and under the alternative $f_{1,\ldots,p}(X(1), ..., X(p)) = \sum_{j<k}^{p} X^2(j)X^2(k)\alpha_{j,k}$ contains quadratic interactions between covariates. The SS-ANOVA model specified is

$$Y = f(X) + \varepsilon,$$

whereas the models simulated under the null and alternative hypotheses of (11)

Type I Error										
		nT_n			$T^*_{AN,1}$			S_n		
Var.	Sig.	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
2	0.01	**0.01**	**0.01**	0.008	0.009	0.01	**0.009**	0.011	0.010	0.009
2	0.05	**0.051**	**0.048**	0.047	0.044	0.0.45	**0.048**	0.053	0.060	0.054
3	0.01	**0.010**	**0.010**	**0.009**	0.014	0.014	0.012	0.012	0.016	0.011
3	0.05	**0.053**	**0.049**	**0.048**	0.070	0.068	0.045	0.063	0.062	0.053
Power										
		nT_n			$T^*_{AN,1}$			S_n		
Var.	Sig.	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
2	0.01	**0.024**	**0.13**	**0.398**	0.022	0.028	0.050	0.019	0.020	0.034
2	0.05	**0.123**	**0.40**	**0.758**	0.111	0.158	0.302	0.075	0.078	0.104
3	0.01	0.037	0.305	0.762	**0.203**	**0.487**	**0.782**	0.026	0.031	0.044
3	0.05	0.159	0.653	0.951	**0.582**	**0.925**	**0.99**	0.098	0.113	0.136

Var. corresponds to the number of variables used in the null model and Sig. corresponds to the significance level used in the test. Bolding describes which method performed better in terms of power and error.

Table 2: Departures from Linearity through Interactions

are

$$\text{under } H_0,\ Y = f(X) + \eta, \text{ and}$$

$$\text{under } H_A,\ Y = f(X) + f_{1,\ldots,p}(X(1), ..., X(p)) + \eta.$$

When $p = 2$, the null model is $f(X) = 2X(1)+2X(2)$ and the interaction added under the alternative is $f_{1,2}(X(1), X(2)) = 4X^2(1)X^2(2)$. When $p = 3$ similar models were used.

Our method fits a main effects only SS-ANOVA model and tests its goodness-of-fit. The two competing tests fit a linear model and then test if the form is non-linear. The simulation results in table (2) show that our methodology has correct type I error and good power performance. The type I error is sharp and power increases significantly with increased sample size. Across different n and p, our method has sharper type I error than the other two methods. For

$p = 2$, our method outperforms the competitors in terms of power. However, for $p = 4$ it underperforms with respect to $T^*_{AN,1}$. This makes sense since this latter test is specific to detecting deviations from linearity whereas our method is trying to detect any departures from the main effects only model, which is succeptible to the curse of dimensionality. Nevertheless, we have shown that our method performs similarly in this scenario when compared to other methods in the literature.

Case II: Missing Interactions Beyond the Within Group Interactions

Simulations were created where the null hypothesis includes two distinct groups of variables which contain all the main effects and all the interactions within each group, and the alternative adds interactions across the groups. Therefore, we have $f(X) = f_A(X(A))+f_B(X(B))$ and $f_{A,B}(X(A\cup B))$, where $f_{A,B}(X(A\cup B))$ contains interactions between variables in group A and B. The SS-ANOVA model specified is

$$Y = f(X) + \varepsilon,$$

whereas the models simulated under the null and alternative hypotheses of (11) are:

$$\text{under } H_0,\ Y = f(X) + \eta, \text{ and}$$
$$\text{under } H_A,\ Y = f(X) + f_{A,B}(X(A), X(B)) + \eta.$$

The first simulation has as null model $f_A(X(A)) = 5\sin(\pi X(1)) + 2X(2)^2$ and $f_B(X(B)) = 2\sin(\pi X(3)) + X(4)^2$, and the interaction between group A and B added under the alternative is $f_{A,B}(X(A \cup B)) = 0.75\cos(\pi(X(1) - X(3)))$. Under the alternative, there exists an interaction across group A and B between variables $X(1)$ and $X(3)$. In the second simulation setting, similar models were used. This setting is similar to Case I, but here under the null model, interactions have been included as well as main effects. The simulation results in table (3) show that the methodology has correct type I error and good power performance. In the first scenario, a model with 4 covariates with two sets of variables of size 2 each was fitted to the data. The two-way interaction between the 2 covariates in each group was included. The second scenario, denoted by

Type I Error				
Var.	Sig.	n=100	n=300	n=500
4	0.01	0.010	0.0120	0.0120
4	0.05	0.050	0.048	0.052
4^*	0.01	0.011	0.011	0.010
4^*	0.05	0.052	0.0531	0.051
Power				
Var.	Sig.	n=100	n=300	n=500
4	0.01	0.033	0.138	0.366
4	0.05	0.088	0.302	0.601
4^*	0.01	0.035	0.144	0.375
4^*	0.05	0.098	0.305	0.622

Var. corresponds to the number of variables used in the null model and Sig. corresponds to the significance level used in the test.

Table 3: Missing Interactions Beyond the Within Group Interactions

a 4^* on the table (3), corresponds to the same set-up defined previously of a model with 4 covariates and two groups of variables, but now in the first group there is only one covariate and in the second one there are 3 covariates. The model includes all the 3 two-way interactions and the 1 three-way interaction which corresponds to all the possible interactions between the 3 covariates in the second group. We see that power performance is comparable in either scenario. Moreover, comparing this to *Case I* we see that including extra interactions under the null-hypothesis increases power of the test. This is because it reduces the number of possible interactions under the alternative.

Case III: Testing for Significance

We compare our method with the test in [17] used to select variables for non-parametric regression. We denote this test statistic as D_n, which requires fitting

a nonparametric kernel regression under the null. The null and alternative hypotheses are the same for both D_n and our test. In the simulations the null hypothesis includes only main effects and the alternative adds covariates to the model. Therefore, we have $f(X) = \sum_{j=1}^{p} f_j(X(j))$ and $f_{p+1,...,p+q}(Z(1), ..., Z(q))$, where $f_{p+1,...,p+q}(Z(1), ..., Z(q))$ are variables leftover not included in $f(X)$. The SS-ANOVA model specified is

$$Y = f(X) + \varepsilon,$$

whereas the models simulated under the null and alternative hypotheses of (13) are:

$$\text{under } H_0,\ Y = f(X) + \eta, \text{ and}$$

$$\text{under } H_A,\ Y = f(X) + f_{p+1,...,p+q}(Z(1), ..., Z(q)) + \eta.$$

When $p = 2$, the null model is $f(X) = 5\sin(\pi X(1)) + 2X(2)^2$ and the covariate added under the alternative is $f_3(Z(1)) = \sin(\pi(Z(1)))$. When $p = 4$ similar models were used.

Under the alternative, the model fitted under the null is insufficient because $f_3(X(3)) = \sin(\pi(X(3)))$ also belongs in the model. However, this might not be known to the researcher or he/she might want to investigate this question precisely, i.e., through testing. This simulation setting shows that this can de done and that the goodness-of-fit test can be used as an omnibus test of the likes of [1], [28] and [17] where a set of covariates is tested to see if it is related to the outcome after a set of covariates have already being included in the model. From the simulation results in table 4, it can be seen that the proposed test has the appropriate size and power increases with sample size. In this simulation scenario, the setting with 4 covariates included in the model had more power compared to the setting with only 2 variables included in the model. This happened because under the former, 2 covariates were missing under the alternative whereas under the latter only 1 is missing. However, unlike Case I, in this scenario the number of variables included in the model under the null does not affect the power of the test, only the covariates added under the alternative affect the power. On the other hand, table 4 shows that the test statistic D_n

Type I Error							
		nT_n			D_n		
Var.	Sig.	n=100	n=300	n=500	n=100	n=300	n=500
2	0.01	**0.008**	**0.011**	**0.010**	0.005	0.036	0.013
2	0.05	**0.054**	**0.057**	**0.057**	0.084	0.093	0.063
4	0.01	**0.014**	**0.010**	**0.011**	0.03	0.096	0.096
4	0.05	**0.063**	**0.053**	**0.050**	0.16	0.256	0.27
Power							
		nT_n			D_n		
Var.	Sig.	n=100	n=300	n=500	n=100	n=300	n=500
2	0.01	0.014	0.034	0.072	**0.08**	**0.293**	**0.636**
2	0.05	0.077	0.154	0.321	**0.245**	**0.566**	**0.863**
4	0.01	**0.240**	**0.828**	**1**	0.07	0.14	0.236
4	0.05	**0.464**	**0.944**	**1**	0.213	0.37	0.46

Var. corresponds to the number of variables used in the null model and Sig. corresponds to the significance level used in the test. Bolding describes which method performed better in terms of power and error.

Table 4: Testing for Significance

has some problems. For the case $p = 2$, the type I error is moderatly larger than the specified level. However, as the sample size increases the actual type I error gets closer to its specified level, but never as close as our method. In terms of power, D_n outperforms our method, but this is expected given that it has larger type I error than our test. When $p = 4$, the test D_n performs quite badly. The type I error rate is wrong and the power does not increase with sample size as rapidly as our method does. It seems D_n does not perform properly when the dimension of p increases even moderatly. Thus, it seems our methodology is a better suited for larger p situations. We note that although D_n performed badly, there exist improved versions of this test adapted to high-dimensional situations [29]. We do not use those modifications as comparisons in our simulations since our focus is not in high-dimensional scenarios.

4. Application to Neonatal Psychomotor Development Data

The Mount Sinai Children's Environmental Health Cohort samples a prospective multiethnic cohort of primiparous women who presented for prenatal care with singleton pregnancies at the Mount Sinai prenatal clinic or two private practices ([30]). The target population was first-born infants with no underlying health conditions that might independently result in serious neurodevelopmental impairment. The continuous outcome of interest is the Psychomotor Development Index at age 2 (PDI), obtained by administration of the Bayley scales of infant development version 2. It is believed that PDI is affected by chemical exposures that can be assessed through urine and blood samples. Potentially, PDI could be affected by the mother's age (AGE), and certain chemical exposures, such as the amount of Bisphenol A (BPA), uM/L of di-2-ethylhexyl phthalate (DEHP) phthalate metabolites, and the amount of dialkylphosphate metabolites (DAP). Maternal exposure biomarkers were collected to asses the magnitude of exposure to the compounds. The data set consists of a sample of 237 maternal-child dyads. An SS-ANOVA model is built with PDI as the outcome and the four predictors variables as follows:

$$\text{PDI} = f_1(BPA) + f_2(DEHP) + f_3(DAP) + f_4(AGE) + \varepsilon. \tag{15}$$

The following paragraph provides an investigation into whether the model in (15) fits the data. A series of tests are conducted to evaluate the goodness-of-fit of this model, as well as possible alternative models. All p-values are shown in table (5). The first column of table (5) shows the null hypotheses that are tested, the second column shows the interaction terms that have been added to the basic model in (15), and the third column shows the p-value for each null hypothesis. Any p-value less than 0.05 is deemed as evidence of lack-of-fit.

Initially, we test the null hypothesis H_0, which corresponds to testing if the model in (15) fits the data well. The p-value of H_0 is 0.044, hence we detect a lack-of-fit. Since the model in (15) is not sufficient, it is possible that interactions need to be considered. Three models are possible extensions, and they only differ

Null	Added Interactions	p-value
H_0		0.044
$H_{1,2}$	$f_{1,2}$	0.158
$H_{1,3}$	$f_{1,3}$	0.077
$H_{2,3}$	$f_{2,3}$	0.029
$H_{1,2+1,3}$	$f_{1,2} + f_{1,2}$	0.114

P-values less than 0.05 are thought as evidence of lack-of-fit.

Table 5: Testing of Goodness-of-fit

from (15) with the addition of one of the following two-way interactions, respectively: $f_{1,2}(BPA, DEHP)$, $f_{1,3}(BPA, DAP)$, and $f_{2,3}(DEHP, DAP)$. These additions to each model correspond to interactions between the chemical exposures. It is theoretically unlikely that interactions exist between the exposures and the mother's age, or that there is a three-way interaction among the exposures; hence models that include such interactions are not considered. Testing null hypotheses $H_{1,2}$, $H_{1,3}$ and $H_{2,3}$ corresponds to testing the goodness-of-fit of these three models, which include three different two-way interactions between exposures. We detected lack-of-fit in the model with the interaction $f_{2,3}$, but we did not detect lack-of-fit for the models with the other two interactions: $f_{1,2}$ and $f_{1,3}$. We want to include all interactions that could potentially explain the outcome, so we include $f_{1,2}$ and $f_{1,3}$ in the model, since both models with those interactions do not show a lack-of-fit. As a last step, we check the goodness-of-fit of the model that includes the two relevant two-way interactions among exposures, and this corresponds to the null hypothesis $H_{1,2+1,3}$. The p-value of this hypothesis is 0.114. Thus, we do not have enough evidence for lack-of-fit of the model that includes both two-way interactions. The final form of our model

is:

$$\begin{aligned}\text{PDI} =& f_1(BPA) + f_2(DEHP) + f_3(DAP) + f_4(AGE) \\ &+ f_{1,2}(BPA, DEHP) + f_{1,3}(BPA, DAP) + \varepsilon.\end{aligned}$$

5. Discussion

In this article we have developed a general Goodness-of-fit statistic and test for nonparametric regression models which are linear smoothers. Particularly, examples, simulations, theory and the application were done in the context of the SS-ANOVA model with continuous outcome, one type of linear smoother. The method developed works by fitting a model currently of interest and tests for independence between the estimated residuals and the covariates used to fit the model, or covariates not yet in the model. A model based bootstrap is used to get critical values that preserve the correct type I error. In comparisons with competing methods, our tests performs favorably. The bootstrap method incorporates the degrees of freedom of the linear smoother. The test developed can deal with a useful variety of lack-of-fit settings. The major contributions we make to the literature include: identifying the need for assessing goodness-of-fit in nonparametric regression, developing a test statistic, creating a variance adjustment to the bootstrap to improve the finite sample performance of the method, providing theoretical justification of the use of the HSIC, and demonstrating correct type I error as well as power performance through numerical simulations. We note that our method can easily be extended to detect departures from homoscedasticity of variance, but we choose not to analyze it in the current research as we believe this was already done extensively in [13, 19].

Some caveats of the method are that when dimension increases and not many interactions have already been included in the model, the power decreases. This method might only be suitable for small models when the need is to detect any possible interactions among main effects. Once extra interactions are included, power increases and the problem becomes more manageable. On the other hand, when testing if extra variables not yet included in the model need to be included,

there is no such problem with the power. This is of importance because the test can be used as an omnibus or global test for testing significance of variables. One possible criticism of the method is that it rests on the assumption of homogeneity of variance. If this assumption is violated, then the test will pick up the lack-of-fit corresponding to the heterogeneous variance, and it will be more difficult to identify where the lack-of-fit is coming from. Another problem could arise if there exists a missing confounder correlated with a covariate in the current model. If the goodness-of-fit test were performed in this setting, with sufficient power it would reject the null, but it would be difficult to assess where the lack-of-fit is coming from, since the confounder is not available.

One of the possible extensions of this test would be to allow for heterogeneity of variance in the SS-ANOVA model, where the variance could be dependent on the covariates. In this way, whenever the homogeneity of variance is violated, the test would still have correct type I error and would be more powerful. Another aspect left unaddressed in the current research is how to choose the degree of the derivative being integrated in to the penalty term. We have used the second derivative in our examples. Other choices are possible too. Further research could extend this method to deal with dichotomous outcomes. Also, the Gaussian kernel in the HSIC has a parameter that has been fixed to 1 in the current report. However, further research could elucidate how to best choose this parameter following a suitable optimality criterion.

Acknowledgements

This work was funded, in part by NIH grants PO1 CA142538 (SJTH and MRK), U10 CA180819 (MCW), and in part by NIEHS/U.S. EPA Children's Center grants ES09584 and R827039, The New York Community Trust, and the Agency for Toxic Substances and Disease Registry/CDC/Association of Teachers of Preventive Medicine (SME).

Appendix

A.1. Details on the Bootstrap Algorithm

If hypotheses in (13) need to be tested using the test statistic in (14), the following bootstrap variation can be used:

Bootstrap Algorithm

Step 1

Calculate the estimated residuals $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$ and create an empirical distribution P_{n,e^o} of the residuals with mass $1/n$ at each $e_i^o = \frac{\hat{\sigma}}{\hat{\sigma}'}(\hat{\varepsilon}_i - \bar{\varepsilon})$, where $\bar{\varepsilon} = \sum_{i=1}^n \frac{\hat{\varepsilon}_i}{n}$, $\hat{\sigma}'^2 = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\varepsilon})^2}{n}$ and $\hat{\sigma}^2 = \frac{||\mathbf{Y}-\mathbf{AY}||^2}{Tr(\mathbf{I}-\mathbf{A})}$.

Step 2

Draw a bootstrap sample η^* from the empirical distribution P_{n,e^o} and draw a bootstrap sample $(\mathbf{X}_n^*, \mathbf{Z}_n^*)$ from the empirical distribution $P_{n,\mathbf{X},\mathbf{Z}}$ of the $(\mathbf{X}_n, \mathbf{Z}_n)$'s independently of η^*. Then set Y_i^* as

$$Y_i^* = \hat{f}(X_i^*) + \eta_i^* \quad \text{for } i = 1, \ldots, n.$$

Step 3

We estimate $\hat{f}^*$ from $\mathbf{Y}_n^*$ and from $\mathbf{X}_n^*$, and create new bootstrap residuals as

$$\varepsilon_i^* = Y_i^* - \hat{f}^*(X_i^*) \quad \text{for } i = 1, \ldots, n.$$

Step 4

Calculate the test statistic as $nT_n(\mathbf{Z}_n^*, \boldsymbol{\varepsilon}_n^*)$.

Step 5

Repeat Step 1 through 4 B times, so as to create B bootstrapped test statistics $nT_n(\mathbf{Z}_n^*, \boldsymbol{\varepsilon}^*)_b$, for $b = 1, \ldots, B$. This distribution approximates the distribution of $nT_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n)$ under the null. The p-value is then calculated as

$$\text{p-value} = \frac{1}{B} \sum_{i=1}^{B} I(nT_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n) \leq nT_n(\mathbf{Z}_n^*, \boldsymbol{\varepsilon}_n^*)_b).$$

A.2. Details on Simulation Studies

In this section, specific details on each simulation study of section 3 provided. We simulated η as $N(0,1)$, and all $X(j)$ and $Z(i)$ as Uniform(0,1) independent of each other. Three simulation cases are described. First, the general case is described, and then, for each subcase, the corresponding true model used under the null and under the alternative are shown. The form shown under the null is the model that was used both under the null and under the alternative. Hence, the alternative simulates lack-of-fit in the model.

Case I: Missing Interactions Beyond the Main Effects

This case corresponds to the simulation results shown in table 1. Simulations were created where the null hypothesis only includes main effects. We have $f(X) = \sum_j^p f_j(X(j))$ and $f_{1,...,p}(X(1), ..., X(p))$ is any interaction between covariates. The hypotheses then become

$$H_0 : Y = f(X) + \eta,$$
$$H_A : Y = f(X) + f_{1,...,p}(X(1), ..., X(p)) + \eta.$$

Below are shown all the instances of *Case I.*

Case I.1, p=2

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2,$$

$$f_{1,2}(X(1), X(2)) = 0.75\cos(\pi(X(1) - X(2))).$$

Case I.2, p=4

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2 + 2\sin(\pi X(3)) + X(4)^2,$$

$$f_{1,...,4}(X(1), ..., X(4)) = 0.5\cos(\pi(X(1) - X(2))) + 0.5\cos(\pi(X(3) - X(4))).$$

Case I.3, p=6

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2 + 2\sin(\pi X(3)) + X(4)^2 + 2\sin(\pi X(5)) + 3X(6)^3,$$

$$f_{1,...,6}(X(1),...,X(6)) = 0.75\cos(0.5\pi(X(1)-X(2))) + 0.5X(2)X(3) + 0.5\cos(\pi(X(4)-X(5)+2X(6))).$$

Case I: Departures from linearity through interactions

This case corresponds to the simulation results shown in table 2. Simulations were created where the null hypothesis only includes linear main effects. Therefore, we have $f(X) = \sum_{j=1}^{p} X(j)\beta_j$, and under the alternative $f_{1,...,p}(X(1),...,X(p)) = \sum_{j<k}^{p} X^2(j)X^2(k)\alpha_{j,k}$ contains quadratic interactions between covariates. The hypotheses then become

$$H_0 : Y = f(X) + \eta,$$
$$H_A : Y = f(X) + f_{1,...,p}(X(1),...,X(p)) + \eta.$$

Below are shown all the instances of table 2.

Case I.4, p=2

$$f(X) = 2X(1) + 2X(2),$$

$$f_{1,2}(X(1),X(2)) = 4X^2(1)X^2(2).$$

Case I.5, p=3

$$f(X) = 2X(1) + 2X(2) + 2X(3),$$

$$f_{1,2,3}(X(1),X(2),X(3)) = 4X^2(1)X^2(2) + 4X^2(2)X^2(3).$$

Case II: Missing Interactions Beyond the Within Group Interactions

This case corresponds to the simulation results shown in table 3. Simulations were created where two distinct groups of covariates, A and B, exist. Under the null hypothesis the model contains all main effects, and all the interactions within each group. Under the alternative, interactions across both groups also exist. Define $f(X) = f_A(X(A)) + f_B(X(B))$ and $f_{A,B}(X(A \cup X(B))$, where $f(X) = f_A(X(A)) + f_B(X(B))$ contains all main effects and all possible interactions within A and B, but not between A and B, and $f_{A,B}(X(A \cup B))$

are any interactions between covariates in group A and B. Our hypotheses then become

$$H_0 : Y = f(X) + \eta,$$
$$H_A : Y = f(X) + f_{A,B}(X(A \cup B) + \eta.$$

Below are shown all the instances of *Case II.*

Case II.1, p=4

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2 + 2\sin(\pi X(3)) + X(4)^2,$$

$$f_{A,B}(X(A \cup B)) = 0.75\cos(\pi(X(1) - X(3))),$$

with $A = \{X_1, X_2\}$ and $B = \{X_3, X_4\}$.

Case II.2, p=4

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2 + 2\sin(\pi X(3)) + X(4)^2 + 0.75\cos(\pi(X(2) - X(3))),$$

$$f_{A,B}(X(A \cup B)) = 0.75\cos(\pi(X(1) - X(4))),$$

with $A = \{X_1\}$ and $B = \{X_2, X_3, X_4\}$.

Case III: Testing for Significance

This case corresponds to the simulation results shown in table 4. Simulation were created where under the null hypothesis the model only includes main effects, and under the alternative, covariates are added to the model. Therefore, we have $f(X) = \sum_j^p f_j(X(j))$ and $f_{p+1,\ldots,p+q}(Z(1), \ldots, Z(q))$, where $f_{p+1,\ldots,p+q}(Z(1), \ldots, Z(q))$ are covariates not yet included in $f(X)$. The hypotheses then become

$$H_0 : Y = f(X) + \eta,$$
$$H_A : Y = f(X) + f_{p+1,\ldots,p+q}(Z(1), \ldots, Z(q)) + \eta.$$

Below are shown all the instances of *Case III.*

Case III.1, p=2

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2,$$

$$f_3(Z(1)) = \sin(\pi Z(1)).$$

Case III.2, p=4

$$f(X) = 5\sin(\pi X(1)) + 2X(2)^2 + 2\sin(\pi X(3)) + X(4)^2,$$

$$f_{5,6}(Z(1), Z(2)) = 0.5Z(1) + \sin(\pi Z(2)).$$

A.3. Theoretical Results

The main purpose of this section is to provide a justification for Theorem 2. This theorem shows that under the null and alternative $T_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n)$ and $T_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n)$ converge to the population $HSIC$, and that the bootstrap version $T_n(\mathbf{X}_n^*, \hat{\boldsymbol{\varepsilon}}_n^*)$ and $T_n(\mathbf{Z}_n^*, \hat{\boldsymbol{\varepsilon}}_n^*)$ converge to 0 under both the null and the alternative. To simplify the theoretical results, it is assumed that the alternative corresponds to the case were covariates are missing from the model, and goodness-of-fit is assessed with respect to $\mathbf{Z}_n$. Other cases where interactions are missing from the model, and where the goodness-of-fit is assessed with respect to $\mathbf{X}_n$ follow a similar proof and are omitted. Next, we present the setting and Lemmas needed for the proof of Theorem 2.

Set-Up

The theoretical results presented here are for the estimation of f_0 in (1) through the solution of the penalized least squares in (4). For simplicity, it will be assumed throughout that f_0 is additive. Let the metric $||\cdot||_n$ be defined by

$$||f||_n^2 = \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|^2.$$

Let $\mathscr{F}_j = \{f_j : [0,1] \to R, \int |f_j^{(m)}|^2 < M_j\}$, $M_j \geq 1$, and $\mathscr{F} = \bigoplus_{j=1}^{p} \mathscr{F}_j$. Let $(y_1, x_1), ..., (y_n, x_n) \in \mathbb{R} \times [0,1]^p$ be a sample from

$$Y_i = \sum_{j=1}^{p} f_{0,j}(X(j)) + \eta_i, \qquad i = 1, ..., n,$$

where $\sum_{j=1}^{p} f_{0,j}(X(j)) \in \mathscr{F}$. Moreover, we have

$$Z_{j,n} = \begin{bmatrix} 1 & x_1(j) & ... & x_1(j)^{m-1} \\ \vdots & \vdots & ... & \vdots \\ 1 & x_n(j) & ... & x_n(j)^{m-1} \end{bmatrix}.$$

Let $\Sigma_j = \lim_{n\to\infty} \frac{1}{n} Z_{j,n}^T Z_{j,n}$, where we assume this limit exists in probability, and let $\phi_{j,1}^2$ be the smallest eigenvalue of Σ_j.

Assumptions

A.1 We assume $\phi_{j,1}^2 > 0$ for all j.

A.2 Uniform subgaussianity of the residuals: there exist $\beta > 0$ and $\Gamma > 0$ such that

$$\sup_n \max_{1\le k\le n} E[\exp|\beta\eta_k|^2] \le \Gamma < \infty.$$

A.3 The tuning parameter is chosen such that $\lambda_n^{-1} = O_p(n^{2m/(2m+1)})$ and $\lambda_n \to 0$.

Lemma 3. *If we solve the penalized least squares model defined in (4) over the RKHS $\mathscr{F}$, and A.1-3 hold, then we have that*

$$||f_0 - \hat{f}||_n^2 = O_p(n^{-2m/(2m+1)}).$$

Proof:

Fix $\delta > 0$. We know that $\mathscr{N}_n(\frac{\delta}{p}; \sigma, \mathscr{F}_j) \le \exp\left(A(\frac{M_j}{\delta/p})^{(1/m)}\right)$, where $\mathscr{N}_n(\frac{\delta}{p}; \sigma, \mathscr{F}_j)$ is the smallest number of δ balls needed to cover the open ball $B(f_{0,j}, \sigma) = \{f_j \in \mathscr{F}_j : ||f_{0,j} - f_j|| \le \sigma\}$ with respect to the $||\cdot||_n$ norm, as defined in Lemma 2.1 in [31]. Let $f_0 = f_{0,1} + \cdots + f_{0,p} \in \mathscr{F}$ and let f_j be the function in the δ/p-covering such that $||f_{0,j} - f_j||_n < \delta/p$. Then,

$$\begin{aligned} &||f_0 - (f_1 + ... + f_p)||_n \\ \le &||f_{0,1} - f_1||_n + \cdots + ||f_{0,p} - f_p||_n \\ \le &\delta/p + \cdots + \delta/p = \delta. \end{aligned}$$

Hence, we have that

$$\begin{aligned}\mathscr{N}_n(\delta;\sigma,\mathscr{F}) &\leq \mathscr{N}_n(\delta/p;\sigma,\mathscr{F}_1)\cdots\mathscr{N}_n(\delta/p;\sigma,\mathscr{F}_p)\\ &\leq \exp\left(pA(\frac{Mp}{\delta})^{(1/m)}\right) \quad \text{with} \quad M=\max\{M_1,...,M_p\}.\end{aligned}$$

From here the proof of Theorem 6.2 in [31] follows for the additive model, hence proving the result. □

As stated before the proof shown here corresponds to the alternative where covariates are missing from the model.

Lemma 4. *We fit the following model using penalized least squares in (4):*

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1,...,n,$$

where $f(x_i) = \sum_{j=1}^{p} f_j(x_i(j))$. *Under the alternative, the model is misspecified by several additive terms, in other words* $\varepsilon_i = \sum_{j=p+1}^{p+q} f_j(z_i(j-p)) + \eta_i$. *Under this situation, we still have convergence of the properly specified terms* $f_j, j = 1,..,p$, *namely*

$$||\sum_{j=1}^{p} f_{0,j} - \sum_{j=1}^{p} \hat{f}_j||_n^2 = O_p(n^{-2m/(2m+1)}),$$

provided that ε *follows subgaussianity in A.2 and that* $(X(1),...,X(p)) \perp (Z(1),...,Z(q))$.

Proof:

If the ε_i are uniform subgaussian and $(X(1),...,X(p)) \perp (Z(1),...,Z(q))$ then Lemma 3 holds in this situation. □

Lemma 5. *We fit an additive model by minimizing the penalized least squares in (4). Under the assumptions A.1-3 we have that*

$$\sup_{x\in[0,1]^p} |\hat{f}(x) - f_0(x)| = O_p(n^{-m(2m-2)/(2m+1)(2m-1)}).$$

Proof:

By [32], we know that $||\hat{f} - f_0||_2^2 = O_p(n^{-2m/(2m+1)})$ where $||\cdot||_2$ is the L_2 norm. By applying Lemma 6 we conclude that

$$||\hat{f} - f_0||_\infty = O_p(n^{-m(2m-2)/(2m+1)(2m-1)}).$$

570

Lemma 6. *Let $\{f_n\}$ be a sequence of functions such that $f_n : [0,1] \to \mathbb{R}$, $||f_n||_2 \to 0$, and $\lim_{n\to\infty} \int_0^1 (f_n^{(k)}(u))^2 du < \infty$, then*

$$||f_n||_\infty = O(||f_n||_2^{(2k-2)/(2k-1)}).$$

Proof:

Let $f : [0,1] \to \mathbb{R}$ and $J_k^2(f) = \int_0^1 (f^{(k)}(u))^2 du < \infty$ for some integer $1 \le k < \infty$. Let $\Delta = 1/m$ for some integer $1 \le m < \infty$. Let $\tilde{f}$ be an approximation to f such that

$$\tilde{f}(x) = \sum_{j=1}^{m} \tilde{f}_j(x - (j-1)\Delta) 1\{(j-1)\Delta \le x < j\Delta\}$$

with $\tilde{f}_j(x) = f((j-1)\Delta) + f^{(1)}((j-1)\Delta)x + \cdots + \frac{f^{(k-1)}((k-1)\Delta)x^{k-1}}{(k-1)!}$ for $x \in [0, \Delta]$.

For $x \in [0, \Delta]$ we have that,

$$\begin{aligned} |\tilde{f}_j(x) - f(x + (j-1)\Delta)| &\le |\int_{(j-1)\Delta}^{(j-1)\Delta+x} \int_{(j-1)\Delta}^{(j-1)\Delta+u_{k-1}} \cdots \int_{(j-1)\Delta}^{(j-1)\Delta+u_1} f^{(k)}(w) dw du_1 \cdots du_{k-1}| \\ &= \frac{\Gamma(3/2)}{\Gamma(k+1/2)} x^{k-1/2} J_k(f). \end{aligned}$$

We also have that

$$\begin{aligned} ||f||_\infty &\le ||\tilde{f}||_\infty + ||f - \tilde{f}||_\infty \\ &\le ||\tilde{f}||_\infty + \frac{\Gamma(3/2)}{\Gamma(k+1/2)} \Delta^{k-1/2} J_k(f). \end{aligned}$$

For $x \in [0, \Delta]$,

$$||\tilde{f}_j||_{\Delta,\infty} = \sup_{x\in[0,\Delta]} |\tilde{f}_j(x)| \le \sup_{x\in[0,\Delta]} |\sum_{l=0}^{k-1} a_{j,l} x^l|,$$

where $a_{j,l} = \frac{f^{(l)}((j-1)\Delta)}{l!}$. Now,

$$\sup_{x\in[0,\Delta]} |\sum_{l=0}^{k-1} a_{j,l}x^l| = \sup_{u\in[0,1]} |\sum_{l=0}^{k-1} a_{j,l}\Delta^l u^l|$$
$$\leq (\sum_{l=0}^{k-1}(a_{j,l})^2\Delta^{2l})^{1/2}$$

where the last inequality follows from Cauchy-Schwarz inequality and setting $u = 1$. Now let,

$$M_k(u) = \begin{bmatrix} 1 & u & u^2 & ... & u^{k-1} \\ u & u^2 & u^3 & ... & u^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u^{k-1} & u^k & u^{k+1} & ... & u^{2k-2} \end{bmatrix},$$

where u is a uniform $[0,1]$ random variable. Now let C_k be the smallest eigenvalue of $E[M_k(u)]$ and if $C_k > 0$ then we have, by change of variables,

$$||\tilde{f}_j||^2_{\Delta,2} = \int_0^\Delta (\tilde{f}_j(x))^2 dx = \Delta \int_0^1 (\tilde{f}_j(\Delta u))^2 du$$
$$= \Delta E \begin{pmatrix} a_{j,0} \\ a_{j,1}\Delta \\ \vdots \\ a_{j,k-1}\Delta^{k-1} \end{pmatrix}^T M_k(u) \begin{pmatrix} a_{j,0} \\ a_{j,1}\Delta \\ \vdots \\ a_{j,k-1}\Delta^{k-1} \end{pmatrix}$$
$$\geq C_k\Delta \sum_{l=0}^{k-1} a_{j,l}^2\Delta^{2l}.$$

With this we can see that

$$||\tilde{f}_j||_{\Delta,\infty} \leq C_k^{-1/2}\Delta^{-1/2}||\tilde{f}_j||_{\Delta,2},$$
$$\text{and } ||\tilde{f}_j||_\infty \leq C_k^{-1/2}\Delta^{-1/2}(\max_j||\tilde{f}_j||_{\Delta,2})^{1/2}$$
$$\leq C_k^{-1/2}\Delta^{-1/2}||\tilde{f}||_2$$
$$\leq C_k^{-1/2}\Delta^{-1/2}(||f||_2 + \frac{\Gamma(3/2)}{\Gamma(k+1/2)}\Delta^{k-1/2}J_k(f)).$$

This implies that,

$$||f||_\infty \leq C_k^{-1/2}\Delta^{-1/2}||f||_2 + \frac{\Gamma(3/2)}{\Gamma(k+1/2)}C_k^{-1/2}\Delta^{k-1}J_k(f) + \frac{\Gamma(3/2)}{\Gamma(k+1/2)}\Delta^{k-1/2}J_k(f).$$

If we let $a = C_k^{-1/2}||f||_2$ and $b = \frac{\Gamma(3/2)}{\Gamma(k+1/2)} C_k^{-1/2} J_k(f)$ and we choose $\Delta = (\frac{a}{(2k-1)b})^{2/(2k-1)} \wedge 1$, then we have for some $0 < C^* < \infty$ that only depends on k that

$$||f||_\infty \leq C^*(||f||_2 \vee (||f||_2^{\frac{2k-2}{2k-1}} J_k^{\frac{2}{2k-1}}(f)) + J_k(f) \wedge (||f||_2^{\frac{2k-2}{2k-1}} J_k^{\frac{1}{2k-1}}(f)) + J_k(f) \wedge ||f||_2).$$

Hence, for a sequence of functions$\{f_n\}$ as defined in the assumptions of the Lemma we get

$$||f_n||_\infty = O(||f_n||_2^{\frac{2k-2}{2k-1}}).$$

Lemma 7. *Under the same assumptions as Lemma 4 and using the notation from Theorem 2 we have that*

$$\varepsilon_1^* - \eta_1^* \xrightarrow{p} 0 \quad \textit{and} \quad \hat{\varepsilon}_1 - \varepsilon_1 \xrightarrow{p} 0.$$

Proof:

Since, $\hat{f}^*$ is an estimator of $\hat{f}$, and $\hat{f}$ has the same properties as f_0, with probability going to 1 as n increases, we can apply Lemma 5 and conclude that $||\hat{f} - \hat{f}^*||_\infty \xrightarrow{p} 0$. Thus we have

$$\sup_{x \in [0,1]^q} |\hat{f}^*(x) - \hat{f}(x)| \geq \max_{X_i^* \in \mathbf{X}_n^*} |\hat{f}^*(X_i^*) - \hat{f}(X_i^*)| = \max_{X_i^* \in \mathbf{X}_n^*} |Y_i^* + \hat{f}^*(X_i^*) - Y_i^* - \hat{f}(X_i^*)|$$

$$= \max_{i \in \mathbb{N}_n} |\varepsilon_i^* - \eta_i^*|,$$

where $\max_{X_i^* \in \mathbf{X}_n^*}$ denotes that we are taking the maximum over a finite bootstrap sample $\mathbf{X}_n^*$. Hence, $\sup_{i \in \mathbb{N}_n} |\varepsilon_i^* - \eta_i^*| \xrightarrow{p} 0$ and we can also say that $\varepsilon_1^* - \eta_1^* \xrightarrow{p} 0$. The same argument follows for $\hat{\varepsilon}_1 - \varepsilon_1 \xrightarrow{p} 0$ by replacing $\hat{f}^*$ with $\hat{f}$ and $\hat{f}$ with f_0. □

Proof of Theorem 2:

We write $T_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n) = \frac{1}{n^2} \sum_{i,j}^n K_{ij} \hat{L}_{ij} + \frac{1}{n^4} \sum_{i,j,q,r}^n K_{ij} \hat{L}_{qr} - 2\frac{1}{n^3} \sum_{i,j,q}^n K_{ij} \hat{L}_{iq}$, where $\hat{L}_{ij} = \exp(-(\hat{\varepsilon}_i - \hat{\varepsilon}_j)^2)$, $L_{i,j} = \exp(-(\varepsilon_i - \varepsilon_j)^2)$ and

$K_{ij} = \exp(-||z_i - z_j||^2)$. Let

$$HSIC(X, \eta) = A_1 + A_2 - A_3$$

and

$$T_n(\mathbf{Z}_n, \hat{\varepsilon}_n) = \hat{A}_1 + \hat{A}_2 - \hat{A}_3 + O(n^{-1}),$$

where

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n^2} \sum_{i \neq j} K_{ij} \hat{L}_{ij}, \\
\hat{A}_2 &= \frac{1}{n^4} \sum_{i \neq j, q \neq r} K_{ij} \hat{L}_{qr} \\
\hat{A}_3 &= \frac{2}{n^4} \sum_{i \neq j \neq q} K_{ij} \hat{L}_{iq}, \\
A_1 &= E_{z_1, \varepsilon_1, z_2, \varepsilon_2}[K_{1,2} L_{1,2}], \\
A_2 &= E_{z_1, z_2}[K_{1,2}] E_{\varepsilon_1, \varepsilon_2}[L_{1,2}], \\
A_3 &= 2 E_{z_1, \varepsilon_1}[E_{z_2}[K_{1,2}] E_{\varepsilon_2}[L_{1,2}]].
\end{aligned}$$

First, it will be shown that

$$\hat{A}_1 - A_1 = \frac{1}{n^2} \sum_{i \neq j} K_{i,j} \hat{L}_{i,j} - E[K_{1,2} \hat{L}_{1,2}] \xrightarrow{p} 0.$$

By Markov's inequality we have that

$$\begin{aligned}
&Pr(|\frac{1}{n^2} \sum_{i \neq j} K_{i,j} \hat{L}_{i,j} - E[K_{1,2} \hat{L}_{1,2}]| > \epsilon) \\
&\quad \leq \frac{1}{n^2 \epsilon^2} \mathrm{Var}(K_{1,2} \hat{L}_{1,2} - E[K_{1,2} \hat{L}_{1,2}]) + \\
&\quad \frac{1}{n^4 \epsilon^2} \sum_{i \neq j} \sum_{p \neq q} \mathrm{Cov}(K_{i,j} \hat{L}_{i,j} - E[K_{i,j} \hat{L}_{i,j}], K_{p,q} \hat{L}_{p,q} - E[K_{p,q} \hat{L}_{p,q}]) \\
&= \frac{1}{n^2 \epsilon^2} O(1) + O(1) \mathrm{E}[(K_{1,2} \hat{L}_{1,2} - E[K_{1,2} \hat{L}_{1,2}])(K_{3,4} \hat{L}_{3,4} - E[K_{3,4} \hat{L}_{3,4}])].
\end{aligned}$$

The first $O(1)$ term comes from the fact that the variance is bounded because $|K_{i,j} \hat{L}_{i,j}|$ is bounded by 1. The second $O(1)$ term comes from the fact that the number of elements in the double summation compared to n^4 is of magnitude

$O(1)$. Under H_0, it holds that $\sum_{j=p+1}^{p+q} f_{0,j}(z_i(j)) = 0$ for all i. Then, it holds that $\varepsilon_i = \eta_i$ for all i. Now, by applying Lemma 7 we know that

$$(\hat{\varepsilon}_1, \hat{\varepsilon}_2, \hat{\varepsilon}_3, \hat{\varepsilon}_4) - (\eta_1, \eta_2, \eta_3, \eta_4) \xrightarrow{p} 0,$$

and by the continuous mapping theorem we have that

$$K_{1,2}\hat{L}_{1,2} - K_{1,2}L_{1,2} \xrightarrow{p} 0.$$

Since $K_{1,2}\hat{L}_{1,2}$ is bounded it is also uniformly integrable, thus

$$E[K_{1,2}\hat{L}_{1,2}] \to E[K_{1,2}L_{1,2}].$$

Moreover,

$$E[K_{1,2}\hat{L}_{1,2}K_{3,4}\hat{L}_{3,4}] \to E[K_{1,2}L_{1,2}K_{3,4}L_{3,4}] = E[K_{1,2}L_{1,2}]E[K_{3,4}L_{3,4}].$$

Hence, the covariance will go to 0 as $n \to \infty$. Then, we can conclude that

$$\tfrac{1}{n^2}\sum_{i\neq j} K_{i,j}\hat{L}_{i,j} - E[K_{1,2}\hat{L}_{1,2}] \xrightarrow{p} 0.$$

Above we have already shown that $E[K_{1,2}\hat{L}_{1,2}] \to E[K_{1,2}L_{1,2}]$ so we can conclude that

$$\tfrac{1}{n^2}\sum_{i\neq j} K_{i,j}\hat{L}_{i,j} - E[K_{1,2}L_{1,2}] \xrightarrow{p} 0.$$

Similar arguments follow for $A_2 - \hat{A}_2$ and $A_3 - \hat{A}_3$. Hence, we have that

$$T_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n) \xrightarrow{p} HSIC(Z, \eta) = 0.$$

Under H_A the same result holds by Lemma 4 and Lemma 7 except that η is replaced by ε and $HSIC(Z, \varepsilon) > 0$, since ε depends on Z. Hence, we have that

$$T_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n) \xrightarrow{p} HSIC(Z, \varepsilon) > 0.$$

Under H_0 and H_A, $HSIC(\mathbf{Z}_n^*, \boldsymbol{\eta}_n^*) = 0$ since $\mathbf{X}_n^*$ and $\boldsymbol{\eta}_n^*$ were sampled independently. From Lemma 7 we have that $\varepsilon_1^* - \eta_1^* \xrightarrow{p} 0$ and hence from the same arguments as above we have that

$$T_n(\mathbf{Z}_n^*, \boldsymbol{\varepsilon}_n^*) - HSIC(\mathbf{Z}_n^*, \boldsymbol{\eta}_n^*) \xrightarrow{p} 0.$$

□

References

[1] D. Liu, X. Lin, D. Ghosh, Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models, Biometrics 63 (4) (2007) 1079–1088.

[2] J. Shawe-Taylor, N. Cristianini, Kernel methods for pattern analysis, Cambridge university press, 2004.

[3] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, The Mathematical Intelligencer 27 (2) (2005) 83–85.

[4] P. Craven, G. Wahba, Smoothing noisy data with spline functions, Numerische Mathematik 31 (4) (1978) 377–403.

[5] G. H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (2) (1979) 215–223.

[6] C. Gu, Smoothing spline ANOVA models, Vol. 297, Springer Science & Business Media, 2013.

[7] G. Kimeldorf, G. Wahba, Some results on tchebycheffian spline functions, Journal of mathematical analysis and applications 33 (1) (1971) 82–95.

[8] G. Wahba, Spline models for observational data, Vol. 59, Siam, 1990.

[9] J. Fan, L.-S. Huang, Goodness-of-fit tests for parametric regression models, Journal of the American Statistical Association 96 (454) (2001) 640–652.

[10] M. Kuchibhatla, J. D. Hart, Smoothing-based lack-of-fit tests: variations on a theme, Journal of Nonparametric Statistics 7 (1) (1996) 1–22.

[11] A. Sen, B. Sen, Testing independence and goodness-of-fit in linear models, Biometrika 101 (4) (2014) 927–942.

[12] J. H. Einmahl, I. Van Keilegom, Tests for independence in nonparametric regression, Statistica Sinica (2008) 601–615.

[13] N. Neumeyer, Testing independence in nonparametric regression, Journal of Multivariate Analysis 100 (7) (2009) 1551–1566.

[14] R. Eubank, J. D. Hart, D. Simpson, L. A. Stefanski, et al., Testing for additivity in nonparametric regression, The Annals of Statistics 23 (6) (1995) 1896–1920.

[15] T. J. Hastie, R. J. Tibshirani, Generalized additive models, Vol. 43, CRC press, 1990.

[16] J. Roca-Pardiñas, C. Cadarso-Suárez, W. González-Manteiga, Testing for interactions in generalized additive models: application to so 2 pollution data, Statistics and Computing 15 (4) (2005) 289–299.

[17] M. A. Delgado, W. G. Manteiga, Significance testing in nonparametric regression based on the bootstrap, Annals of Statistics (2001) 1469–1507.

[18] J. Fan, Test of significance based on wavelet thresholding and neyman's truncation, Journal of the American Statistical Association 91 (434) (1996) 674–688.

[19] J. H. Einmahl, I. Van Keilegom, Specification tests in nonparametric regression, Journal of Econometrics 143 (1) (2008) 88–102.

[20] Y. Wang, G. Wahba, C. Gu, R. Klein, B. Klein, Using smoothing spline anova to examine the relation of risk factors to the incidence and progression of diabetic retinopathy, Statistics in Medicine 16 (12) (1997) 1357–1376.

[21] B. Schölkopf, A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press, 2002.

[22] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al., Measuring and testing dependence by correlation of distances, The Annals of Statistics 35 (6) (2007) 2769–2794.

[23] G. J. Székely, M. L. Rizzo, et al., Brownian distance covariance, The annals of applied statistics 3 (4) (2009) 1236–1265.

[24] G. J. Székely, M. L. Rizzo, The distance correlation t-test of independence in high dimension, Journal of Multivariate Analysis 117 (2013) 193–213.

[25] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with hilbert-schmidt norms, in: Algorithmic learning theory, Springer, 2005, pp. 63–77.

[26] L. Song, A. Smola, A. Gretton, J. Bedo, K. Borgwardt, Feature selection via dependence maximization, The Journal of Machine Learning Research 13 (1) (2012) 1393–1434.

[27] C. A. Micchelli, Y. Xu, H. Zhang, Universal kernels, The Journal of Machine Learning Research 7 (2006) 2651–2667.

[28] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, The American Journal of Human Genetics 89 (1) (2011) 82–93.

[29] J. C. Escanciano, J. C. Pardo-Fernández, I. Van Keilegom, et al., Asymptotic distribution-free tests for semiparametric regressions, Unpublished manuscript.

[30] S. M. Engel, J. Wetmur, J. Chen, C. Zhu, D. B. Barr, R. L. Canfield, M. S. Wolff, Prenatal exposure to organophosphates, paraoxonase 1, and cognitive development in childhood, Environmental health perspectives 119 (8) (2011) 1182.

[31] S. Van de Geer, Estimating a regression function, The Annals of Statistics (1990) 907–924.

[32] Y. Lin, Tensor product space anova models, Annals of Statistics (2000) 734–755.