APPLICATIONS OF INDEPENDENCE STATISTICS TO GOODNESS-OF-FIT, MULTIVARIATE CHANGE POINT ESTIMATION AND CLUSTERING OF VARIABLES

Sebastian Jose Teran Hidalgo

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Michael R. Kosorok

Michael C. Wu

Mengjie Chen

Kari North

Donglin Zeng

ABSTRACT

Sebastian Jose Teran Hidalgo: Applications of Independence Statistics to Goodness-of-Fit, Multivariate Change Point Estimation and Clustering of Variables
(Under the direction of Michael R. Kosorok and Michael C. Wu)

Independence statistics try to evaluate the statistical dependence between two random vectors of general dimension and type. Independence statistics do not assume a specific form of dependence, but they are sensitive to all forms of departures from independence. The current manuscript seeks to extend the use of independence statistics to three settings.

In the first part of the dissertation, we developed a goodness-of-fit test for smoothing spline ANOVA models, which are a nonparametric regression methodology with the useful property that the contribution of the covariates can be decomposed in a ANOVA fashion. The proposed method derives estimated residuals from the model. Then, statistical dependence is evaluated between the estimated residuals and the covariates using independence statistics. If no dependence exists, the model fits the data well. Application of the method is demonstrated with a neonatal mental development data analysis.

In the second part, we develop a method for the change point problem where two sets of random vectors are observed sequentially over a dimension, but at some unknown point, the relationship between these two vectors changes. We propose a methodology to estimate the unknown change point without assuming a model. This is accomplished by assessing, with an independence statistic, the strength of the association before and

after possible change points. A test for the hypothesis of existence of the change point is developed. We demonstrate its use with blood glucose and physical activity measurements on an individual with type 1 diabetes.

In the third part, we develop a method for hierarchical clustering of variables while controlling for type I error rate, which is not done in common clustering methods. We accomplish this by turning the decision of whether to join two clusters into a hypothesis testing problem. The strength of our method is shown by clustering genes from single cell data coming from different tumors.

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my advisor Professor Michael Kosorok for guiding me through this process. Prof. Kosorok has been an inspiration to me. Because of his example, I have pushed myself to learn more than I ever thought I could and I have aspired to be a better biostatistician every day. He has shared with me many of the experiences he had to go through to become a great scientist, from which I have learned greatly. I have also seen Prof. Kosorok display much patience and kindness not only to me, but also to others, while teaching statistics. I am really grateful for the time he spent advising me through this dissertation.

I would also like to thank my co-advisor Michael Wu. Mike has always been available to provide me with great advice about my research. Many of his advices have been extremely valuable and have improve my research significantly. He has displayed great patience with me while going over papers together. Even though he lives on a different city, he takes the time to read along, during a phone call, my research and provides detailed suggestions. At times when I might had been discouraged, he has helped me put things into perspective. He has also been a role model to me. I would like also to thank my committee members: Dr. Mengjie Chen, Dr. Kari North, and Dr. Donglin Zeng for their insightful knowledge, and overall support for my research projects. Dr. North and Dr. Zeng, you were great teachers when I took your courses. Mengjie, thank you for meeting with me several times and sharing interesting data applications.

I would like to give a special thanks to my family: Ruth Hidalgo, Alfonso Zambrano and Cesar Zambrano. This is a small family that has always been around for me. My mother has always been an inspiration of hard work, resilience and justice. I would also like to thank Ludmila Janda, who has been by my side during all of my dissertation adventure and has been very encouraging.

I would like to thank many of the friends that I have made through the Department of Biostatistics, and that have made this experience very enjoyable. Thank you my friends Roy, James, Siying, Yu, Poulami, Matt, Cynthia, Thomas, Habtamu, Sayan, Guanhua, Susan, PJ, Sean, Steve, Sujatro, Arkopal, Michael, Daniel, Jon, Arianna, Phoebe, Crystal, Dave, Rachel, Alison, Jennifer and Elizabeth.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Recent developments in tests of statistical independence are Brownian Distance Covariance (Székely et al. (2007), Székely et al. (2009), Székely and Rizzo (2013)) and HSIC (Gretton et al. (2005), Song et al. (2012)). 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 norm 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.

Both DC and HSIC are beautiful results from statistical and machine learning theory. They can both be standardized to be between 0 and 1, with 1 correspond to complete statistical dependence between the two vectors being analyzed. Thus, this dependence statistics allow us to tackle many statistical problems without the need to specify models in advance or to do fancy modeling. In this dissertation, we extend the uses of these methods to many statistical problems.

The dissertation is organized as follows. In Chapter 2, we review current literature on topics related to the three chapters that will follow. A focus will be on independence

statistics, change point problems and multiple testing adjustments. In Chapter 3, we develop a goodness-of-fit test for nonparametric models which we apply to the smoothing spline ANOVA models. In Chapter 4, we develop an estimator and test of existence for the change point in the relationship between two multivariate random vectors. In Chapter 5, we create a hierarchical clustering of variables algorithm that controls the family wise error rate of clustered variables that are otherwise unrelated. In chapter 6, we present some other applications where the use of independence statistics can be used.

CHAPTER 2: LITERATURE REVIEW

2.1 Smoothing Splines ANOVA Models

The mathematical foundation of smoothing spline ANOVA models (SS-ANOVA) is the Reproducing Kernel Hilbert Space (RKHS). The main reference for its mathematical theory is Aronszajn (1950). Grace Wahba is one of the major contributors to the development of SS-ANOVA models (Wahba (1969), Kimeldorf and Wahba (1971), Wahba (1985), Wahba (1990). Chong Gu, one of Wahba's students, generalized SS-ANOVA models to exponential families, survival models and distribution estimation (Gu (2013)).

2.1.1 SS-ANOVA

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

$$Y = f(\mathbf{X}) + \eta, \tag{2.1}$$

for an unknown function f and random residual η, which is independent of $\mathbf{X}$, with $E\eta = 0$. A sample $(\mathbf{X}_1, Y_1)$,...,$(\mathbf{X}_n, Y_n)$ is drawn from (2.1). Estimation of f can be done through minimization of the following penalized least squares:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2 + \lambda J(f). \tag{2.2}$$

In the case where $p = 1$, $f(X_i)$ 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 $p > 1$, $f(\mathbf{X}) = \sum_{j=1}^{p} f_j(X_j)$ and $J(f) = \sum_{j=1}^{p} \theta_j^{-1} \int_0^1 f_j^{(k)}(x_j)^2 dx_j$. This corresponds to an additive model. In the general case,

$$f(\mathbf{X}) = \sum_j f_j(X_j) + \sum_{j<k} f_{j,k}(X_j, X_k) + \cdots,$$

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

where λ and $\boldsymbol{\theta}$ are tuning parameters which are selected through Generalized Cross Validation (GCV).

2.1.2 Solution to Penalized Least Squares

The current section will assume that the functional form is additive. Discussion of more complicated forms, i.e., which include interactions, can be found in Gu, 2013. The form in (2.2) will be minimized assuming that the data generating mechanism is (2.1) such 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

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

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

Solution

By the representer theorem (Wahba (1990), Schölkopf and Smola (2002)) the minimizer of (2.2) has the form

$$f(x) = \sum_{v=1}^{k} d_v \phi^v(x) + \sum_{i=1}^{n} c_i R_J(x_i, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \boldsymbol{c},$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions, and $\mathbf{c}$ and $\mathbf{d}$ are vectors of real coefficients. Then the estimation reduces to minimizing

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

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is $n \times m$ with (i,v)th entry $\phi^v(x_i)$ and Q is $n \times n$ with the (i,j)th entry $R_J(x_i, x_j)$. Then by taking derivatives, the solution of (2.2) is of the form

$$(\mathbf{Q} + n\lambda \mathbf{I})\mathbf{c} + \mathbf{Sd} = \mathbf{Y},$$

$$\mathbf{S^T c} = 0.$$

From these equations the hat matrix $\mathbf{A}(\lambda, \boldsymbol{\theta})$ can be derived such that $\hat{\mathbf{y}} = \mathbf{A}(\lambda, \boldsymbol{\theta})\mathbf{y}$.

2.1.3 Generalized Cross Validation

The GCV statistic, as defined by Craven and Wahba (1978), corresponds to

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

where $\hat{\mathbf{y}} = \mathbf{A}(\lambda, \boldsymbol{\theta})\mathbf{y}$. λ and θ are chosen to minimize $\text{GCV}(\lambda, \boldsymbol{\theta})$. 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 an SS-ANOVA model, 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.

2.1.4 Goodness-of-fit Statistics for SS-ANOVA

Gu (Gu (2004), Gu (2013)) developed some goodness-of-fit statistics for SS-ANOVA based of the Kullback-Leiber distance. Suppose f in (2.1) is estimated by assuming that $f \in \mathscr{H}$ for some $\mathscr{H}$. Assume that in fact $f \in \mathscr{H}^* \subset \mathscr{H}$. Heuristic diagnostics based on the Kullback-Leibler (KL) distance can be used in this situation. Let $\hat{f}$ be the solution to (2.1) in $\mathscr{H}$, with the smoothing parameters selected through GCV. Let $\tilde{f}$ be the KL projection of $\hat{f}$ in $\mathscr{H}^*$, the minimizer of $KL(\hat{f}, f)$ for $f \in \mathscr{H}^*$. Let $f_c = C$ be the constant model for some constant C. From this, we can write

$$KL(\hat{f}, f_c) = KL(\hat{f}, \tilde{f}) + KL(\tilde{f}, f_c).$$

The ratio

$$\rho = \frac{KL(\tilde{f}, f_c)}{KL(\hat{f}, f_c)},$$

just like an R^2 statistic in the standard least square regression, this indicates how much of $\hat{f}$ actually sits in $\mathscr{H}^*$, and can be used to diagnose the feasibility of the null hypothesis $f \in \mathscr{H}^*$. The set-up in (2.1) can be generalized to an exponential family. In such a case, $KL(\hat{f}, f)$ can be written as

$$KL(\hat{f}, f) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(\mathbf{x}_i)[\theta(\hat{f}(\mathbf{x}_i)) - \theta(f(\mathbf{x}_i))] - [b(\theta(\hat{f}(\mathbf{x}_i))) - b(\theta(f(\mathbf{x}_i)))]\}, \quad (2.3)$$

where $\mu((x)) = (db/\theta)(\mathbf{x}) = \mathrm{E}[Y|\mathbf{x}]$. The minimization of (2.3) with respect to f can be accomplished through Newton-Raphson algorithm. The resulting $KL(\hat{f}, \tilde{f})$ depends on the tuning parameters $\boldsymbol{\theta}$. An outer loop of optimization needs to be performed to minimize $KL(\hat{f}, \tilde{f})$ with respect to $\boldsymbol{\theta}$. Different $\mathscr{H}^*$ can be chosen so as to evaluate goodness-of-fit. A $\mathscr{H}^*$ such that $\mathscr{H}^* \subset \mathscr{H}$ is chosen so that an interaction is missing or a covariate is missing.

2.2 Independence Statistics

This section will cover generalizations of correlation statistics to the multivariate setting. Moreover, these statistics can potentially detect any form of dependence.

2.2.1 Distance Covariance

Distance covariance was developed by Székely et al. (2007), and Székely et al. (2009). An extension to the high dimensional case also exists (Székely and Rizzo

(2013)). For random variables $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, let ϕ_x, ϕ_y and $\phi_{x,y}$ be the characteristic function of $\mathbf{X}$, $\mathbf{Y}$ and $(\mathbf{X}, \mathbf{Y})$, respectively. Distance covariance $\mathcal{V}$ can be used to measure the dependence between $\mathbf{X}$ and $\mathbf{Y}$ through the distance

$$||\phi_x(t)\phi_y(s) - \phi_{x,y}(t,s)||.$$

If $\mathbf{X} \not\perp \mathbf{Y}$ then this distance will be greater than 0. If $\mathbf{X} \perp \mathbf{Y}$ then this distance will be exactly 0. Then, using this distance the following hypotheses can be tested:

$$H_0 : \phi_{x,y} = \phi_x\phi_y \qquad vs. \qquad H_1 : \phi_{x,y} \neq \phi_x\phi_y.$$

Then, the measure of independence chosen to assess this hypotheses is

$$\begin{aligned}\mathcal{V}^2(\mathbf{X}, \mathbf{Y}; w) =& ||\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)||_w^2 \\ =& \int_{\mathbb{R}^{p+q}} |\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)|^2 w(t,s)dtds\end{aligned}$$

where $\mathcal{V}^2(\mathbf{X}, \mathbf{Y}; w) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. The weight function is chosen as

$$w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$$

with $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ and $|\cdot|_p$ being the Euclidean norm in $\mathbb{R}^p$. For finiteness of $||\phi_{x,y} - \phi_x\phi_y||^2$ it is sufficient that $E|\mathbf{X}|_p < \infty$ and $E|\mathbf{Y}|_q < \infty$. Distance variance can be defined as $\mathcal{V}^2(\mathbf{X}; w) = \mathcal{V}^2(\mathbf{X}, \mathbf{X}; w)$. The distance correlation between random vectors $\mathbf{X}$ and $\mathbf{Y}$ with finite first moments is the nonnegative number $DC(\mathbf{X}, \mathbf{Y})$ defined by

$$DC(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}^2(\mathbf{X})\mathcal{V}^2(\mathbf{Y})}}$$

if $\mathcal{V}^2(\mathbf{X})\mathcal{V}^2(\mathbf{Y}) > 0$ and equals 0 otherwise. The distance covariance statistics are defined as follows. For an observed random sample $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{X}_k, \mathbf{Y}_k) : k = 1, ..., n\}$ from the joint distribution of random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, define

$$a_{kl} = |X_k - X_l|_p, \qquad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \qquad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$
$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

for $k, l = 1, ..., n$. Similarly, define $b_k l = |\mathbf{Y}_k - \mathbf{Y}_l|_q$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ for $k, l = 1, ..., n$.

Definition

The empirical distance covariance $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ is the nonnegative number defined by

$$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

Similarly, $\mathcal{V}_n(\mathbf{X})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2.$$

Definition

The empirical distance correlation $DC_n(\mathbf{X}, \mathbf{Y})$ is defined as

$$DC_n^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}}$$

if $\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0$ and 0 otherwise. For a sample of size n let the empirical characteristic functions of X, Y and (X, Y) be

$$\phi_x^n(t) = \frac{1}{n}\sum_{k=1}^{n}\exp\{i\langle t, \mathbf{X}_k\rangle\}, \quad \phi_y^n(s) = \frac{1}{n}\sum_{k=1}^{n}\exp\{i\langle s, \mathbf{Y}_k\rangle\},$$
$$\text{and } \phi_{x,y}^n(s) = \frac{1}{n}\sum_{k=1}^{n}\exp\{i\langle t, \mathbf{X}_k\rangle + i\langle s, \mathbf{Y}_k\rangle\},$$

respectively. The following theorems and corollaries come from Székely et al. (2007) and Székely et al. (2009).

Theorem

If $(\mathbf{X}, \mathbf{Y})$ is a sample from joint distribution of $(\mathbf{X}, \mathbf{Y})$, then

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \|\phi_{x,y}^n(t, s) - \phi_x^n(t)\phi_y^n(s)\|^2.$$

Theorem

If $E|\mathbf{X}|_p < \infty$ and $E|\mathbf{Y}|_q < \infty$, then almost surely

$$\lim_{n\to\infty}\mathcal{V}_n(\mathbf{X}, \mathbf{Y}) = \mathcal{V}(\mathbf{X}, \mathbf{Y}).$$

Corollary

If $E(|\mathbf{X}|_p + |\mathbf{Y}|_q) < \infty$, then almost surely,

$$\lim_{n\to\infty} DC_n(\mathbf{X}, \mathbf{Y}) = DC(\mathbf{X}, \mathbf{Y}).$$

Corollary

If $E(|\mathbf{X}|_p + |\mathbf{Y}|_q) < \infty$, then:

i) If $\mathbf{X}$ and $\mathbf{Y}$ are independent, $n\mathcal{V}_n^2/S_2 \xrightarrow{\mathcal{L}} Q$ where $Q = \sum_{j=1}^{\infty}\lambda_j Z_j^2$, where Z_j are independent standard normal random variables, $\{\lambda_j\}$ are nonnegative constants that

dependent on the distribution of $(\mathbf{X}, \mathbf{Y})$, and $E[Q] = 1$.

ii) If $\mathbf{X}$ and $\mathbf{Y}$ are dependent, then $n\mathcal{V}_n^2/S_2 \xrightarrow{P} \infty$.

2.2.2 Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion is similar to Distance Covariance, it even has the same form, but with Euclidean norms replaced by kernel distances. However, the former is developed in an RKHS setting instead of in the context of distances of characteristic functions (Gretton et al. (2005)).

RKHS Theory

Consider a Hilbert space $\mathscr{F}$ of functions from $\mathscr{X}$ to $\mathbb{R}$. Then $\mathscr{F}$ is a reproducing kernel Hilbert space if for each $x \in \mathscr{X}$, the Dirac evaluation operator $\delta_x : \mathscr{F} \to \mathbb{R}$, which maps $f \in \mathscr{F}$ to $f(x) \in \mathbb{R}$, is a bounded linear functional. To each point $x \in \mathscr{X}$, there corresponds an element $\phi(x) \in \mathscr{F}$ such that $\langle \phi(x), \phi(x') \rangle_{\mathscr{F}} = k(x, x')$ where $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ a unique positive definite kernel. We require $\mathscr{F}$ to be separable. Define a second separable RKHS, $\mathscr{G}$, with kernel $k(\cdot, \cdot)$ and feature map ψ, on the separable space $\mathscr{Y}$.

Hilbert-Schmidt Norm

Let $C : \mathscr{G} \to \mathscr{F}$ be a linear operator. Then, provided the sum converges, the Hilbert-Schmidt (HS) norm of C is define as

$$\|C\|_{HS}^2 = \sum_{i,j} \langle C v_i, u_j \rangle_{\mathscr{F}}^2,$$

where u_i and v_j are orthonormal bases $\mathscr{F}$ and $\mathscr{G}$, respectively.

Hilbert-Schmidt Operator

A linear operator $C : \mathscr{G} \to \mathscr{F}$ is called a Hilbert-Schmidt operator if its HS norm exists. The set of Hilbert-Schmidt operators $HS(\mathscr{G}, \mathscr{F}) : \mathscr{G} \to \mathscr{F}$ is a separable Hilbert space

with inner product

$$\langle C, D\rangle_{HS} = \sum_{i,j} \langle Cv_i, u_j\rangle_{\mathscr{F}} \langle Dv_i, u_j\rangle_{\mathscr{F}}.$$

Tensor Product

Let $f \in \mathscr{F}$ and $g \in \mathscr{G}$. Then the tensor product operator $f \otimes g : \mathscr{G} \to \mathscr{F}$ is defined as

$$(f \otimes g)h : f\langle g, h\rangle_{\mathscr{G}} \text{ for all } f \in \mathscr{G}.$$

Moreover, by the definition of the HS norm, we can compute the HS norm of $f \otimes g$ via

$$\begin{aligned} \|f \otimes g\|^2_{HS} =& \langle f \otimes g, f \otimes g\rangle_{HS} = \langle f, (f \otimes g)g\rangle_{\mathscr{F}} \\ =& \langle f, f\rangle_{\mathscr{F}} \langle g, g\rangle_{\mathscr{G}}. \end{aligned}$$

The Cross-Covariance Operator

We assume that $(\mathscr{X}, \Gamma)$ and $(\mathscr{Y}, \Lambda)$ are furnished with probability measures P_x and P_y respectively. We may now define the mean elements with respect to these measures as those members of $\mathscr{F}$ and $\mathscr{G}$ respectively for which

$$\begin{aligned} \langle \mu_x, f\rangle_{\mathscr{F}} :=& E_x[\langle \phi(x), f\rangle_{\mathscr{F}}] = E_x[f(x)], \\ \langle \mu_y, g\rangle_{\mathscr{G}} :=& E_y[\langle \psi(y), g\rangle_{\mathscr{G}}] = E_y[g(y)], \end{aligned}$$

where ϕ is the feature map from $\mathscr{X}$ to the RKHS $\mathscr{F}$, and ψ maps from $\mathscr{Y}$ to $\mathscr{G}$. Finally, $\|\mu_x\|^2_{\mathscr{F}}$ can be computed by applying the expectation twice via

$$\|\mu_x\|^2_{\mathscr{F}} = E_{x,x'}[\langle \phi(x), \phi(x')\rangle_{\mathscr{F}}] = E_{x,x'}[k(x, x')].$$

Cross-Covariance Operator

The cross-covariance operator associated with the joint measure $P_{x,y}$ on $(\mathscr{X} \times \mathscr{Y}, \Gamma \times \Lambda)$

is a linear operator $C_{xy} : \mathscr{G} \to \mathscr{F}$ defined as

$$C_{xy} := E_{x,y}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] = E_{x,y}[\phi(x) \otimes \phi(y)] - \mu_x \otimes \mu_y.$$

Definition

Given separable RKHSs $\mathscr{F}$, $\mathscr{G}$ and a joint measure P_{xy} over $(\mathscr{X} \times \mathscr{Y}, \Gamma \times \Lambda)$, the Hilbert-Schmidt Independence Criterion (HSIC) is defined as the squared HS-norm of the associated cross-covariance operator C_{xy}:

$$HSIC(P_{xy}, \mathscr{F}, \mathscr{G}) = \|C_{xy}\|^2_{HS}. \tag{2.4}$$

The $HSIC(P_{xy}, \mathbf{X}, \mathbf{Y})$ between $\mathbf{X}$ and $\mathbf{Y}$ can be expressed in terms of kernels as:

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

HSIC allows us to evaluate the statistical dependence between two random vectors of arbitrary dimension.

Theorem

Assume $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ are universal kernels (Micchelli et al. (2006)). Then, $HSIC(P_{xy}, \mathbf{X}, \mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are statistically independent, i.e., $P_{x,y} = P_x \times P_y$.

Empirical Criterion

With an i.i.d ,sample $(\mathbf{X}_1, \mathbf{Y}_1)$,...,$(\mathbf{X}_n, \mathbf{Y}_n)$ from P_{xy}, $HSIC(\mathbf{X}, \mathbf{Y})$ can be estimated with

$$T_n(\mathbf{X}, \mathbf{Y}) = n^{-2}tr(KHLH), \tag{2.5}$$

where $H, K, L \in \mathbb{R}^{n \times n}$, $H_{i,j} := \delta_{i,j} - n^{-1}$, $K_{i,j} := k(\mathbf{x}_i, \mathbf{x}_j)$, and $L_{i,j} := l(\mathbf{y}_i, \mathbf{y}_j)$. 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 $k_{ij} = \exp(-\sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2/\sigma^2)$ and $l_{ij} = \exp(-\sum_{k=1}^{p}(y_{i,k} - y_{j,k})^2/\sigma^2)$ are called Gaussian and satisfy the universal kernel condition (Micchelli et al. (2006)) and will be the ones used throughout the current research with σ^2 held fix at 1.

2.3 Resampling-Based Multiple Testing Procedures

This section will introduce multiple testing procedures that will useful in the current research. Most important will be the concept of family wise error rate (FWER) and how to develop a valid null distribution of the test statistic or p-values in such a way that the FWER is preserved. Power will also be an important consideration. One of the earlier references on the subject of multiple testing is the book by Westfall and Young (1993). This book introduces the $maxT$ and $minP$ procedures for preserving FWER while at the same providing more power than the Bonferroni approach. The null distribution of the $maxT$ and $minP$ procedures is obtained through permutation (Westfall and Young (1993), Ge et al. (2003)).

Criticisms of obtaining the null distribution of the $maxT$ and $minP$ procedures through permutation can be found in Dudoit et al. (2004), Dudoit and Van Der Laan (2007), van der Laan et al. (2004), Pollard and van der Laan (2004) and Pollard et al. (2005). The main reason why the null distribution might not work is that permutation destroys some of the correlation among p-values, hence does not represent the state of nature correctly. A permutation based approach can be used instead where the test

statistic is shift and scaled transformed (Dudoit et al. (2004), Dudoit and Van Der Laan (2007), Pollard and van der Laan (2004)). This procedure can be used in conjunction with step-down procedures to control the FWER (van der Laan et al. (2004)).

2.3.1 Set-up

The notation and definitions used here will be similar to those presented in the book by Dudoit and Van Der Laan (2007). Let $\mathscr{X}_n = \{X_i : i = 1, ..., n\}$ denote a random sample of size n from a data generating mechanism P. Denote as P_n the empirical distribution based of $\mathscr{X}_n$. In the setting of multiple testing, there exist M pairs of null and alternative hypothesis. These pairs of hypotheses correspond to some property of the data generating mechanism P, i.e., means, correlations or other possible parameters of P. Each pair corresponds to a submodel of P denoted by $\mathscr{M}(m)$. Then, the mth pair of null and alternative hypotheses can be written as

$$H_0(m) : I(P \in \mathscr{M}(m)) \qquad \text{and} \qquad H_1(m) : I(P \notin \mathscr{M}(m)).$$

This means that $H_0(m)$ is true if P belongs to the submodel $\mathscr{M}(m)$, and $H_0(m)$ is false otherwise.

2.3.2 Type I Error Rates

Sets of true and false null hypotheses

Let

$$\mathscr{H}_0 = \mathscr{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathscr{M}(m)\}$$

denote the set of $h_0 \equiv |\mathscr{H}_0|$ true null hypotheses, where the longer notation $\mathscr{H}_0(P)$ emphasizes the dependence of this set on the data generating distribution P. Likewise,

let

$$\mathscr{H}_1 = \mathscr{H}_1(P) \equiv \{m : H_1(m) = 1\} = \{m : P \notin \mathscr{M}(m)\} = \mathscr{H}_0^c(P)$$

be the set of $h_1 \equiv |\mathscr{H}_1| = M - h_0$ false null hypotheses.

Complete null hypothesis

The complete null hypothesis H_0^C is defined as

$$H_0^C \equiv \prod_{m=1}^{M} H_0(m) = \prod_{m=1}^{M} I(P \in \mathscr{M}(m)) = I(P \in \cap \mathscr{M}(m)).$$

The complete null hypothesis is true if and only if all M individual null hypotheses $H_0(m)$ are true, i.e, if and only if the data generating distribution P belongs to the intersection $\cap_{m=1}^{M} \mathscr{M}(m)$ of the M submodels.

Type I and Type II errors

Each pair of null and alternative hypotheses corresponding to submodel $\mathscr{M}(m)$ has a test statistic $T_n(m)$. A given multiple testing procedure will have a rejection region denoted by $\mathscr{C}_n(m)$ for each submodel $\mathscr{M}(m)$. This rejection region will be chosen so that a certain definition of type I error rate will be preserved. $\mathscr{R}_n$ is the set of submodels $\mathscr{M}(m)$ such that their null hypothesis $H_0(m)$ is rejected. A type I error is committed by rejecting a true null hypothesis ($\mathscr{R}_n \cap \mathscr{H}_0$). A Type-II error is committed by failing to reject a false null hypothesis ($\mathscr{R}_n^c \cap \mathscr{H}_1$). The number of rejected null hypotheses is

$$R_n \equiv |\mathscr{R}_n| = \sum_{m=1}^{M} I(T_n(m) \in \mathscr{C}_n(m)),$$

the number of Type I errors or false positives is

$$V_n \equiv |\mathscr{R}_n \cap \mathscr{H}_0| = \sum_{m \in \mathscr{H}_0} I(T_n(m) \in \mathscr{C}_n(m)),$$

the number of Type II errors or false positives is

$$U_n \equiv |\mathscr{R}_n^c \cap \mathscr{H}_1| = \sum_{m \in \mathscr{H}_1} I(T_n(m) \notin \mathscr{C}_n(m)),$$

the number of true negatives is

$$S_n \equiv |\mathscr{R}_n \cap \mathscr{H}_1| = \sum_{m \in \mathscr{H}_1} I(T_n(m) \notin \mathscr{C}_n(m)).$$

Type I error rate: FWER

In the multiple testing problem they are many definitions of type I error rate of a test procedure. The current research focuses on one such definition: the family-wise error rate (FWER). FWER is the probability of at least one Type I error,

$$FWER \equiv Pr(V_n > 0) = 1 - F_{V_n}(0).$$

Adjusted p-values

Adjusted p-values will be used to provide easy to use decision rules on when to reject a given hypothesis in such a away as to preserve $FWER$. Consider any multiple testing procedures with rejection regions $\mathscr{C}_n(m;\alpha)$ for submodel $\mathscr{M}(m)$. Then, the M-vector of adjusted p-values, $\tilde{P}_{0n} = (\tilde{P}_{0n}(m) : 1, ..., M)$, is

$$\begin{aligned} \tilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_0(m) \text{ at nominal } FWER \text{ level } \alpha\} \\ &= \inf\{\alpha \in [0,1] : T_n(m) \in \mathscr{C}_n(m;\alpha)\}, \qquad m = 1, ..., M. \end{aligned}$$

The adjusted p-value $\tilde{P}_{0n}(m)$ for the mth test statistic, is the smallest nominal type I error level of the multiple hypothesis testing procedure, in this case $FWER$, at which one would reject $H_0(m)$, given T_n. Given this definition, we can provide an alternative

representation to the set of rejected hypotheses given α:

$$\mathscr{R}_n(\alpha) = \{m : T_n(m) \in \mathscr{C}_n(m;\alpha)\} = \{m : \tilde{P}_{0n}(m) \leq \alpha\}.$$

2.3.3 minP and maxT Procedures

There are several procedures that can control FWER. *Setp-down* procedures order the raw p-values or the test statistics starting with the most significant. One such procedure that will be used in the current research is $minP$, which starts by ordering the p-values from smallest to largest and $maxT$, which orders the test statistics from largest to smallest (Dudoit and Van Der Laan (2007), Ge et al. (2003), Westfall and Young (1993)). Let $O_n(m)$ denote the indices for the ordered unadjusted p-values $P_{0n}(O_n(m))$, so that $P_{0n}(O_n(1)) \leq ... \leq P_{0n}(O_n(M))$. Also, let $\bar{\mathscr{O}}_n(h) = \{O_n(h), ..., O_n(M)\}$. The step-down minP adjusted p-values are defined by

$$\tilde{p}_{0n}(o_n(m)) = \max_{h=1,...,m} \{\Pr(\min_{l \in \bar{\mathscr{O}}_n(h)} P_{0n}(l) \leq p_{0n}(o_n(h)))\}.$$

Let $O_n(m)$ denote the indices for the ordered test statistics $T_n(O_n(m))$, so that $T(O_n(1)) \geq ... \geq T(O_n(M))$. Also, let $\bar{\mathscr{O}}_n(h) = \{O_n(h), ..., O_n(M)\}$. The step-down maxT adjusted p-values are defined by

$$\tilde{p}_{0n}(o_n(m)) = \max_{h=1,...,m} \{\Pr\Big(\max_{l \in \bar{\mathscr{O}}_n(h)} T_n(l) \geq t_n(o_n(h))\Big)\}.$$

For minP and maxT adjusted p-values $\tilde{p}_{0n}(o_n(1)) \leq \tilde{p}_{0n}(o_n(2)) \leq ... \leq \tilde{p}_{0n}(o_n(M))$, reject all hypothesis corresponding to $\mathscr{M}(m)$ whose adjusted p-value is such that $\tilde{p}_{o_n(m)} \leq \alpha$. The strong control of the FWER for the $maxT$ and $minP$ procedures can be proven assuming the subset pivotality property (Ge et al. (2003), Westfall and

Young (1993)).

Usually, the marginals and the join distribution of the test statistics are unknown. The bootstrap or permutations can be used to estimate the adjusted p-values in the $minP$ and $maxT$ procedures. These p-values are harder to compute than the raw p-values because a double permutation algorithm needs to be used. Below is a description of the permutation algorithm that can be used for the computation of the step-down $maxT$ adjusted p-values (Ge et al. (2003)).

Permutation algorithm for the step-down maxT adjusted p-values.

Order the test statistics such that $t_n(o_n(1)) \geq t_n(o_n(2)) \geq ... \geq t_n(o_n(M))$. For the bth permutation, $b = 1, ..., B$:

1. Permute the n rows of the data matrix X.

2. Compute the test statistics $t_{n,b}(1), ..., t_{n,b}(M)$ for each hypothesis $\mathscr{M}(m)$.

3. Next, compute $u_{i,b} = \max_{l=i,...,M} |t_{n,b}(o_n(l))|$, the successive maxima of test statistics by

$$u_{M,b} = |t_{n,b}(o_n(M))|$$
$$u_i, b = \max(u_{i+1,b}, |t_{n,b}(o_n(i))|) \quad \text{for } i = M-1, ..., 1.$$

The above steps are repeated B times and the adjusted p-values are estimated by

$$\tilde{p}_{0n}(o_n(i)) = \frac{\#\{b : u_{i,b} \geq |t_n(o_n(1))|\}}{B} \quad \text{for } i = 1, ..., M,$$

with the monotonicity constraint enforced by setting

$$\tilde{p}_{0n}(o_n(1)) \leftarrow \tilde{p}_{0n}(o_n(1)), \qquad \tilde{p}_{0n}(o_n(i)) \leftarrow \max(\tilde{p}_{0n}(o_n(i-1)), \tilde{p}_{0n}(o_n(i)))$$

for $i = 2, \ldots, M$. There is one caveat with this permutation algorithm. If the test statistics are not identically distributed across hypotheses, the $maxT$ adjusted p-values may be different from the $minP$ adjusted ones, and may give different weights to different hypotheses. In such situations, it would be better to use the $minP$ procedure. Below is a description of the permutation algorithm that can be used to compute the step-down $minP$ adjusted p-values.

Permutation algorithm for the step-down minP adjusted p-values

0. Compute raw p-values for each hypothesis. Order the raw p-values such that $p_{0n}(o_n(1)) \leq p_{0n}(o_n(2)) \leq ... \leq p_{0n}(o_n(M))$.

Set $q_b(M+1) = 1$ for $b = 1, ..., B$.

Set $i = M$.

1. For hypothesis $H_0(i)$, compute the B permutation test statistics $t_{n,1}(i), t_{n,2}(i), ..t_{n,B}(i)$ and use the **raw p-values** algorithm described below to get the B raw p-values $p_{0n,1}(i), p_{0n,2}(i), ..., p_{0n,B}(i)$.

2. Update the successive minima $q_b(i)$:

$$q_b(i) \leftarrow \min(q_b(i+1), p_{0n,b}(i)), \quad b = 1, ..., B.$$

3. Compute the adjusted p-values for the hypothesis $H_0(i)$

$$\tilde{p}_{0n}(i) = \frac{\#\{b : q_b(i) \leq p_{0n}(i)\}}{B}.$$

4. Update $i \leftarrow i - 1$. If $i = 0$, go to step 5, otherwise, go to step 1.

5.- Enforce monotonicity of $\tilde{p}_{0n}(i)$ by

$$\tilde{p}_{0n}(1) \leftarrow \tilde{p}_{0n}(1), \qquad \tilde{p}_{0n}(i) \leftarrow \max(\tilde{p}_{0n}(i-1), \tilde{p}_{0n}(i)) \qquad \text{for } i = 2, ..., M.$$

Raw p-values algorithm

From the permutation distribution of $T_n(i)$, $t_{n,1}(i), t_{n,2}(i), ..., t_{n,B}(i)$, obtain $p_{0n,1}(i), p_{0n,2}(i), ..., p_{0n,B}(i)$, simultaneously from

$$p_{0n,b}(i) = \frac{\#\{b' : |t_{n,b'}(i)| \geq |t_{n,b}(i)|\}}{B}.$$

2.3.4 Subset Pivotality

Procedures based on the *maxT* and *minP* adjusted p-values control the FWER weakly under all conditions. Strong control of the FWER also holds under the assumption of subset pivotality (Dudoit and Van Der Laan (2007), Ge et al. (2003), Westfall and Young (1993)). The distribution of raw p-values $(P(1), ..., P(M))$ is said to have the subset pivotality property if for all subsets $\mathscr{K}$ of $\{1, ..., M\}$ the joint distributions of the sub-vector $\{P(i) : i \in \mathscr{K}\}$ are identical under the restrictions $H_0(\mathscr{K}) = \cap_{i \in \mathscr{K}} \{H_0(i) = 0\}$ and $H_0(\mathscr{M}) = \cap_{i=1}^{M} \{H_0(i) = 0\}$. This property is required to ensure that the procedure based on adjusted p-values computed under the complete null provide strong control of the FWER. A practical consequence of it is that resampling for computing the adjusted p-values may be done under the complete null $H_0(\mathscr{M})$ rather than the unknown partial null hypothesis $H_0(\mathscr{M}_0)$. This might be problematic under situations where subset pivotality does not hold.

2.3.5 Type I Error Rate Control and Choice of Null Distribution

This section describes how to create a null distribution of the test statistics through the bootstrap, instead of the permutation approach, such that the FWER error rate is preserved while using the $maxT$ or $minP$ procedures (Dudoit et al. (2004), van der Laan et al. (2004), Pollard et al. (2005)). In this section, the error rate of interest will be the FWER and it will be denoted by $FWER(V_n)$ to denote its dependence on the unknown number of falsely rejected null hypotheses V_n.

General Test Statistics Null Distribution

In a multiple testing setting, a rejection region is defined such that the type I error rate is controlled at a given level α, i.e., such that

$$FWER(V_n) \leq \alpha \tag{2.6}$$

$$\limsup_{n\to\infty} FWER(V_n) \leq \alpha. \tag{2.7}$$

Inequality (2.6) and (2.7) are called finite sample control and asymptotic control, respectively. The type I error rate $FWER(V_n)$ is defined under the true distribution $Q_n(P)$ of the test statistics T_n, which in turn are functions of the underlying data generating distribution P. However, P is usually unknown to the statistician or the researcher and therefore is replaced by an assumed null distribution Q_0, or an estimate of it, denoted by Q_{0n}. The null distribution Q_0 has be chosen in such a way that the type I error rate is controlled under the true distribution $Q_n(P)$ and not only Q_0. In order for this to happen, the error rate under the assumed null distribution Q_0 must be such that it dominates the rate under the true distribution $Q_n(P)$. In other words,

the following null domination condition must be satisfied:

$$FWER(V_n) \leq FWER(V_0),$$
$$\limsup_{n\to\infty} FWER(V_n) \leq FWER(V_0).$$

Here V_0 denotes the number of Type I errors under Q_0, i.e., for $T_n \sim Q_0$. Dudoit, Pollard and van der Laan, in the papers mentioned before, for controlling the type I error rate, propose a null distribution $Q_0(P)$ which is the asymptotic distribution of the M-vector Z_n of null value shifted and scaled test statistics

$$Z_n(m) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m)]}\right)}\Big(T_n(m) + \lambda_0(m) - E[T_n(m)]\Big), \qquad m = 1, ..., M.$$

For the test of single-parameter null hypotheses using the t-statistics, the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$. For testing the equality of K population means using the F-statistics, the null values are $\lambda_0(m) = 1$ and $\tau_0(m) = 2/(K-1)$, under the assumption of equal variances in the different populations. Stepwise procedures based on such a null distribution do indeed provide the desired asymptotic control of the Type I error rate $FWER(V_n)$, for general data generating distributions, null hypotheses, and test statistics.

Bootstrap-based multiple testing procedures

The test statistics null distribution $Q_0 = Q_0(P)$ depends on the true data generating distribution P and is therefore typically unknown. It can be estimated with the bootstrap as explained below.

Bootstrap estimation of the test statistics null distribution Q_0

Let P_n^* denote an estimator of the true data generating distribution P. For the non-parametric bootstrap, P_n^* is simply the empirical distribution P_n of the observed data

$\mathscr{X}_n = \{X_i : i = 1, ..., n\}$. For the model-based bootstrap, P_n^* belongs to a model $\mathscr{M}$ for the data generating distribution P, such as a family of multivariate Gaussian distributions. One then proceeds as follows to generate the bootstrap test statistics null distribution.

1.- Obtain the bth bootstrap dataset, $\mathscr{X}_n^b = \{X_i^b : i = 1, ..., n\}$, $b = 1, ..., B$, by generating n i.d.d. random variables X_i^b with distribution P_n^*.

2.- For each bootstrap dataset X_n^b, compute the M-vector of test statistics, $T_n(\cdot, b) = (T_n(m, b) : m = 1, ..., M)$, which can be arranged in an $M \times B$ matrix, $\mathbf{T}_n \equiv (T_n(m, b))$, with rows corresponding to the M null hypotheses and columns to the B bootstrap samples.

3.- For each null hypothesis $H_0(m)$, compute empirical means $E[T_n(m, \cdot)] \equiv \sum_b T_n(m, b)/B$ and variances $Var[T_n(m, \cdot)] \equiv \sum_b (T_n(m, b) - E[T_n(m, \cdot)])^2/B$ of the B bootstrap test statistics $T_n(m, b)$ (i.e., row means and variances of the matrix $\mathbf{T}_n$), to yield estimates of $E[T_n(m)]$ and $Var[T_n(m)]$, respectively, $m = 1, ...M$.

4.- Obtain an MXB matrix, $\mathbf{Z}_n \equiv (Z_n(m, b))$, of null value shifted and scaled bootstrap statistics $Z_n(m, b)$, by row-shifting and scaling the matrix $\mathbf{T}_n$ using the bootstrap estimates of $E[T_n(m)]$ and $Var(T_n(m))$ and the user-supplied null values $\lambda_0(m)$ and $\tau_0(m)$. That is,

$$Z_n(m, b) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m, \cdot)]}\right)}\Big(T_n(m, b) + \lambda_0(m) - E[T_n(m, \cdot)]\Big).$$

5.- The bootstrap estimate Q_{0n} of the null distribution Q_0 is the empirical distribution of the B columns $Z_n(\cdot, b)$ of the matrix $\mathbf{Z}_n$.

Bootstrap estimation of common cut-offs and adjusted p-values for the single-step maxT procedure:

0.- Apply the previous bootstrap to generate an $M \times B$ matrix, $\mathbf{Z}_n = (Z_n(m,b))$, of null value shifted and scaled bootstrap statistics $Z_n(m,b)$.

1.- Compute the maximum statistic, $\max_m Z_n(m,b)$, $b = 1,...,B$, for each bootstrap dataset $\mathscr{X}_n^b$, i.e., each column of the matrix $\mathbf{Z}_n$.

2.- For controlling the FWER at nominal level $\alpha \in [0,1]$, the bootstrap single-step $maxT$ common cut-off $c(Q_{0n},\alpha)$ is the $(1-\alpha)$-quantile of the empirical distribution of the B maxima $\{\max_m Z_n(m,b) : b = 1,...,B\}$.

3.- The bootstrap single-step $maxT$ adjusted p-value for null hypothesis $H_0(m)$ is the proportion of maxima $\{\max_n Z_n(m,b) : b = 1,...,B\}$ exceeding the corresponding observed test statistic $T_n(m)$,

$$\tilde{P}_{0n}(m) \equiv \frac{1}{B}\sum_{b=1}^{B} I(\max_m Z_n(m,b) \geq T_n(m)), \qquad m = 1,...,M.$$

2.3.6 Multiple Testing for Correlation Coefficients

In this section, we will describe two approaches on how to do the testing for correlation coefficients, the first approach is shown in Westfall and Young (1993). The second approach corresponds to Dudoit and Van Der Laan (2007), Pollard et al. (2005), and Van der Laan and Pollard (2003).

Setting

Let $X \sim P = N(\mathbf{0}_p, \sigma^2)$, with p-dimensional Gaussian distribution P and covariance

matrix $\sigma = (\sigma(j,j') : j,j' = 1,...,p)$ equal to the corresponding correlation matrix $\rho = (\rho(j,j') : j,j' = 1,...,p)$.

Let $\mathscr{X}_n =\equiv \{X_i : i = 1,...,n\}$ be an i.i.d. random sample from P. The hypotheses of interest concern the $M \equiv \binom{p}{2} = p(p-1)/2$ distinct entries $\phi = (\phi(m) : m = 1,...,M)$ of the $p \times p$ correlation matrix ρ. Consider a two-sided test of the $M = p(p-1)/2$ null hypotheses $H_0(m) = I(\psi(m) = \psi_0(m))$ vs. the alternative hypotheses $H_1(m) = I(\psi(m) \neq \psi_0(m))$, $m = 1,...,M$. We are interest in testing for zero correlation, namely $\psi_0(m) = 0$ for all m.

The M null hypotheses are tested based on the following t-statistics,

$$T_n(m) \equiv \sqrt{n-2}\frac{\psi_n(m)}{\sqrt{1-\psi_n^2(m)}}, \qquad m = 1,...,M,$$

where $\psi_n = (\psi(m) : m = 1,...,M)$ is the M-vector of empirical correlation coefficients. Specifically, the empirical correlation coefficient for the pair of random variables $(X(j), X(j'))$, corresponding to the mth null hypothesis, is defined as

$$\psi_n(m) = \rho_n(j,j') \equiv \frac{\sigma_n(j,j')}{\sqrt{\sigma_n(j,j)\sigma_n(j',j')}},$$

based on the empirical means $\bar{X}_n(j)$ and covariances $\sigma_n(j,j')$,

$$\bar{X}_n(j) \equiv \frac{1}{n}\sum_{i=1}^{n} X_i(j), \qquad \sigma_n(j,j') \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i(j) - \bar{X}_n(j))(X_i(j') - \bar{X}_n(j')).$$

Bootstrap Null Distribution

This construction is described in Westfall and Young (1993). It consists of resampling each component of $X_i(j)$ independently for each j to create bootstrap samples of $\mathscr{X}_n$ such that the columns are independent of each other. This forces the complete null

where all columns are not correlated with each other. The procedure is as follows for each bootstrap sample b:

1.- For each variable $X(j)$, $j = 1, .., p$, sample n j-specific entries $X_i^b(j)$, $i = 1, ..., n$, at random, with replacement, from the set of n j-specific observations $\{X_i(j) : i = 1, ..., n\}$. The ith bootstrap p-vectpr $X_i^b = (X_i^b(j) : j = 1, ..., p)$, $i = 1, ..., n$, is obtained by combining the p such independently sampled variables. Let $\mathscr{X}_n^b \equiv \{X_i^b : i = 1, ..., n\}$ denoted the resulting bootstrap dataset.

2.- Compute the M-vector $T_n(\cdot, n) = (T_n(m, b) : m = 1, ..., M)$ of bootstrap test statistics as in $T_n(m)$ above, but based on the bootstrap dataset $\mathscr{X}_i^b$.

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M-vectors $\{T_n(\cdot, b) : b = 1, ..., B\}$.

Bootstrap Null Distribution

This bootstrap procedure works by resampling entire p-vectors so as to maintain the correlation among the vectors, but recreates the null by shifting and scaling the test statistics for each bootstrap sample (Pollard et al., 2005). For each bootstrap sample $b = 1, ..., B$ we:

1.- Sample n p-vectors X_i^b at random, with replacement from the set of n observations $\mathscr{X}_n = \{X_i : i = 1, , , .n\}$. Let $\mathscr{X}_n^b \equiv \{X_i^b : i = 1, ..., n\}$ denote the resulting bootstrap data set.

2.- Compute an M-vector $T_n(\text{cot}, b) = (T_n(m, b) : m = 1, ..., M)$ of test statistics $T_n(m)$ but based on the bootstrap dataset $\mathscr{X}_n^b$.

3.- Compute an M-vector $Z_n(\cdot, b) = (Z_n(m, b) : m = 1, .., M)$ of bootstrap null value

shifted and scaled test statistics,

$$Z_n(m,b) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m,\cdot)]}\right)}\Big(T_n(m,b) - E[T_n(m,\cdot)]\Big),$$

where $E[T_n(\cdot,m)] \equiv \sum_b T_n(m,b)/B$ and
$Var[T_n(m,\cdot)] \equiv \sum_b (T_n(m,b) - E[T_n(m,\cdot)])^2/B$ denote, respectively, the empirical mean and variance of the B bootstrap test statistics $T_n(m,b)$ for the null hypothesis $H_0(m)$, $m = 1, ..., M$.

The test statistics null distribution is the empirical distribution Q_{0n} of the $B = 10,000$ M-vectors $\{Z_n(\cdot,b) : b = 1, .., B\}$.

2.4 Change Point Models and Estimation

This section will introduce some examples of estimating a change point for multivariate data nonparametrically. The examples shown look at the distribution of the data and look for changes over time of this distribution. Later, it will be shown that this problem can be made more specific and instead of looking at any changes in the distribution, partition the random variables into two groups and see if changes between the relationship between these two groups changes over time.

2.4.1 Nonparametric Change Point of Multivariate Data Distribution

Set-up

For random variables $X, Y \in \mathbb{R}^p$, let ϕ_x and ϕ_y denote their characteristic functions, respectively. A divergence measure between multivariate distributions may be defined

as

$$\int_{\mathbb{R}^p} |\phi_x(t) - \phi_y(t)|^2 w(t) dt,$$

where $w(t)$ denotes a positive weight function for which the above integral exists. In this setting, the $w(t)$ function used is

$$w(t;\alpha) = \left(\frac{2\pi^{p/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((p+\alpha)/2)} |t|^{p+\alpha} \right)^{-1},$$

for some fixed constant $\alpha \in (0,2)$. Then, if $E|X|^{\alpha} < \infty$ and $E|Y|^{\alpha} < \infty$, a characteristic function based divergence measure may be defined as

$$\mathscr{D}(X,Y;\alpha) = \int_{\mathbb{R}^p} |\phi_x(t) - \phi_y(t)|^2 \left(\frac{2\pi^{p/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((p+\alpha)/2)} |t|^{p+\alpha} \right)^{-1} dt.$$

An equivalent measure of divergence which can be used is

$$\mathscr{E}(X,Y;\alpha) = 2E|X-Y|^{\alpha} - E|X-X'| - E|Y-Y'|^{\alpha}.$$

This measure is equally useful because of the fact that $\mathscr{D}(X,Y;\alpha) = \mathscr{E}(X,Y;\alpha)$.

Lemma For any pair of independent random vectors $X, Y \in \mathbb{R}^p$ such that $E(|X|^{\alpha} + |Y|^{\alpha}) < \infty$, and for any $\alpha \in (0,2)$, then $\mathscr{E}(X,Y;\alpha) = 0$ if and only if X and Y are identically distributed.

Let $\mathbf{X}_n = \{X_i : i = 1,...,n\}$ and $\mathbf{Y}_n = \{Y_i : i = 1,...,m\}$ be independent i.d.d. samples from the distribution $X, Y \in \mathbb{R}^p$, respectively. An empirical analog of $\mathscr{E}(X,Y;\alpha)$ would be

$$\hat{\mathscr{E}}(X,Y;\alpha) = \frac{2}{mn} \sum_{i,j} |X_i - Y_j|^{\alpha} - \binom{n}{2}^{-1} \sum_{i<k} |X_i - X_k|^{\alpha} - \binom{m}{2}^{-1} \sum_{j<m} |Y_j - Y_k|^{\alpha}.$$

Estimating the Location of a Change Point

Let $Z_1, ..., Z_T \in \mathbb{R}^p$ be an independent sequence of observations let $1 \leq \tau \leq T$ be a constant. Now define the following sets: $X_\tau = \{Z_1, Z_2, ..., Z_\tau\}$ and $Y_\tau = \{Z_{\tau+1}, Z_{\tau+2}, ..., Z_T\}$. If there exists a point τ where there is a change in the distribution then this τ can be estimated as

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \frac{mn}{m+n} \hat{\mathscr{E}}(X, Y; \alpha).$$

2.4.2 Kullback-Leibler Importance Estimation Procedure

KLIEP (Liu et al. (2013), Nguyen et al. (2010), Sugiyama et al. (2008)) is a density-ratio estimation algorithm that is suitable for estimating the Kullback-Leibler (KL) divergence. Let $Y, Y' \in \mathbb{R}^p$ be two random variables with densities $p(Y)$ and $p'(Y)$, respectively. The density ratio

$$\frac{p(X)}{p'(Y)}$$

can be modeled using the following kernel model

$$g(\mathbf{Y}; \boldsymbol{\theta}) := \sum_{l=1}^{n} \theta_l K(\mathbf{Y}, \mathbf{Y}_l),$$

where $\boldsymbol{\theta} := (\theta_1, ..., \theta_n)^{\mathrm{T}}$ are parameters to be learned from data samples, and $K(\mathbf{Y}, \mathbf{Y}')$ is a kernel basis function. The Gaussian kernel,

$$K(\mathbf{Y}, \mathbf{Y}') = \exp\left(- \frac{\|\mathbf{Y} - \mathbf{Y}'\|^2}{\sigma^2} \right),$$

with σ^2 chosen by Cross-Validation, can be used.

KLIEP Algorithm

The parameters $\boldsymbol{\theta}$ in the model $g(\mathbf{Y}; \boldsymbol{\theta})$ are determined so that the KL divergence from

$p(\mathbf{Y})$ to $g(\mathbf{Y}l\boldsymbol{\theta})p'(\mathbf{Y})$ is minimized:

$$\begin{aligned} KL &= \int p(\mathbf{Y})\log\left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})g(\mathbf{Y};\boldsymbol{\theta})}\right)d\mathbf{Y} \\ &= \int p(\mathbf{Y})\log\left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})}\right)d\mathbf{Y} - \int p(\mathbf{Y})\log\big(g(\mathbf{Y};\boldsymbol{\theta})\big)d\mathbf{Y}. \end{aligned}$$

The first term is ignored because it does not depend on $\boldsymbol{\theta}$. Then the empirical criterion that optimizes KL is given by

$$\begin{aligned} &\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{l=1}^{n}\theta_l K(\mathbf{Y}_i,\mathbf{Y}_l)\right), \\ &\text{s.t.} \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\theta_l K(\mathbf{Y}'_j,\mathbf{Y}_l) = 1 \quad \text{and} \quad \theta_1,...,\theta_n \geq 0. \end{aligned}$$

The equality constraint assures that $g(\mathbf{Y};\boldsymbol{\theta})p'(\mathbf{Y})$ is a probability density function. The inequality constrain comes from the non-negativity of the density-ratio function. This problem is convex and can be solved by a gradient-projection iteration. Then, the density-ratio estimator is given as

$$\hat{g}(\mathbf{Y}) = \sum_{l=1}^{n}\hat{\theta}_l K(\mathbf{Y},\mathbf{Y}_l).$$

This procedure has optimal convergence rates (Nguyen et al. (2010)).

Change-Point Detection KLIEP

Once the $\hat{g}(\mathbf{Y})$ is obtained, an approximation of the KL divergence is given as

$$\widehat{KL} := \frac{1}{n}\sum_{i=1}^{n}\log\hat{g}(\mathbf{Y}_i).$$

As in the previous example, this KLIEP-based KL-divergence estimator has been applied to change-point detection (Kawahara and Sugiyama (2012)).

CHAPTER 3: GOODNESS-OF-FIT TEST FOR SS-ANOVA MODELS

3.1 Summary

Smoothing spline ANOVA models are a nonparametric regression methodology with the useful property that the contribution of the covariates can be decomposed into main effects, two-way interactions, and all other higher-level interactions. Despite the popularity of this methodology, little has been done to develop diagnostic statistics. In the current research, we propose a goodness-of-fit test for a smoothing spline ANOVA model with a continuous predictor. 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. If no dependence exists, the model fits the data well. This dependence statistic is the foundation for the proposed goodness-of-fit test, and the bootstrap is used to obtain p-values. Application of the method is demonstrated with a neonatal mental development data analysis. Our major contributions to the literature include: developing a goodness-of-fit test for the smoothing spline ANOVA model, creating a finite sample variance adjustment to the bootstrap, providing theoretical justification the use of the HSIC, and demonstrating correct type I error as well as power performance through simulations.

3.2 Introduction

Nonparametric regression models are an attractive alternative to parametric models; they provide greater flexibility, thus, can provide a better fit when parametric assumptions are too restrictive (e.g., linearity of the mean). Smoothing spline ANOVA models (SS-ANOVA) are a popular nonparametric regression alternative (Craven and Wahba (1978); Golub et al. (1979); Gu (2013); Kimeldorf and Wahba (1971); Wahba (1990)). SS-ANOVA models estimate the mean of an outcome Y as a smooth function f. Their ANOVA decomposition partitions the variation of the outcome attributed to the covariates into main effects, two-way interactions, and all other higher-level interactions, but as functions, not constants, as with classical ANOVA. Therefore, f is a multivariate function that can be written as a summation of many functions, each being either a main effect or an interaction of a given order. An element of this summation is a main effect if it is a univariate function, and if it is a k term multivariate function, then it is considered to be an interaction term of order k. The SS-ANOVA methodology estimates these mean functions by assuming that the integral of their derivatives of a certain degree to be finite. The function is estimated by minimizing a least squares term plus a penalty that controls the degree of smoothness of each function from the decomposition. Gu also extends the SS-ANOVA methodology to exponential families, density estimation, survival analysis, semiparametric models and mixed effects models.

SS-ANOVA models provide greater interpretability and structure than similar nonparametric models found in the machine learning literature, such as kernel ridge regression (Liu, Lin and Gosh, 2007; Shawe-Taylor and Cristianini, 2004). However, inference can be difficult with nonparametric models. After fitting an SS-ANOVA model, the researcher may want to investigate the quality of their model. Methodologies exist

for diagnostics of specific components of the SS-ANOVA model based on the Kullback-Leiber distance and cosine angles (Gu (1992), Gu (2004)). These methodologies are useful for running diagnostics on specific components of the ANOVA decomposition, but do not evaluate the overall goodness-of-fit of the model. Moreover, the methods offer rules of thumb, but do not provide a p-value to inform a decision regarding the goodness-of-fit. Moreover, the method suggests which terms that have been included may be unnecessary, but does not provide information on how good the overall fit is.

This paper resolves these issues by proposing a goodness-of-fit test for the SS-ANOVA model. 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 major contributions we make to the literature include: identifying the need for assessing goodness-of-fit in a smoothing spline ANOVA model, 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 chapter is organized as follows. In section 3.3 the method for goodness-of-fit in SS-ANOVA is introduced. Section 2 includes 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 3.4, simulation results are presented, in section 3.5 application of the method is demonstrated with a neonatal mental development data analysis, and section 3.6 is a concluding discussion of the proposed method.

3.3 Goodness-Of-Fit in SS-ANOVA

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

3.3.1 SS-ANOVA

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

$$Y = f(X) + \eta, \tag{3.1}$$

for an unknown function f and random residual η, which is independent of X, with $E\eta = 0$. A sample of size n denoted by (X_1, Y_1),...,(X_n, Y_n) is drawn from 3.1. Estimation of f can be done through minimization of the following penalized least squares:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda J(f). \tag{3.2}$$

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 an additive model. In the general case

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

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

where λ and θ are tuning parameters which are selected through Generalized Cross Validation (GCV). The GCV statistic is defined as

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

where $\hat{y} = \mathbf{A}(\lambda, \theta)y$. λ and θ are chosen to minimize $\text{GCV}(\lambda, \theta)$. 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 derivatives is being penalized, namely $\int_0^1 f_j''(x)^2 dx$.

After fitting an SS-ANOVA model, it is important to conduct some model diagnostics. Model diagnostics are statistics that assess how well a model fits 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 the HSIC.

3.3.2 HSIC

Recent developments in tests of statistical independence are Brownian Distance Covariance (Székely et al. (2007); Székely et al. (2009), Székely and Rizzo (2013)) and HSIC (Gretton et al. (2005); Song et al. (2012)). 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 (Micchelli et al. (2006)). The $HSIC(P_{xy}, X, Y)$ between X and Y is defined as

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

We will rely on the following theorem (Gretton et al. (2005)):

Theorem 3.3.1. *$HSIC(P_{xy}, X, Y) = 0$ if and only if X and Y are statistically independent, i.e., $P_{x,y} = P_x \times P_y$.*

With an i.d.d sample $(\mathbf{X}_n, \mathbf{X}_n) = \{(X_1, Y_1),...,(X_n, Y_n)\}$ 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 $H, K, L \in \mathbb{R}^{n \times n}$, $H_{i,j} := \delta_{i,j} - n^{-1}$, $K_{i,j} := k(X_i, X_j)$, and $L_{i,j} := l(Y_i, Y_j)$. 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.

3.3.3 Goodness-Of-Fit Test Based on Residuals

This subsection introduces the proposed goodness-of-fit test. After fitting an SS-ANOVA model, 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.

We assume that the same data generating mechanism as 3.1 holds. It is assumed that the model depends on main effects only such that

$$f(X) = \sum_j f_j(X(j)).$$

To assess the goodness-of-fit of the main effects only model we define

$$\varepsilon := Y - \sum_j f_j(X(j)),$$

and test the following hypotheses:

$$\begin{aligned} H_0 &: HSIC(P_{x,\varepsilon}, X, \varepsilon) = 0 \\ H_A &: HSIC(P_{x,\varepsilon}, X, \varepsilon) > 0. \end{aligned} \tag{3.3}$$

If the the null holds, we have that $\varepsilon = \eta$, the true model error. Hence, ε is independent of X and $HSIC(P_{x,\varepsilon}, X, \varepsilon) = 0$. If the alternative holds, then $\varepsilon \neq \eta$ and there is a lack-of-fit. Hence, ε is dependent on X and $HSIC(P_{x,\varepsilon}, X, \varepsilon) > 0$.

The alternative can hold because the assumption of main effects only model is incorrect, and in reality we have

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

which still depends on X. 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 3.1, we can test the null and alternative hypotheses in 3.3. To accomplish this, we define

$$\hat{\varepsilon}_i := Y_i - \sum_j \hat{f}_j(X_i(j)),$$

for $i = 1, \ldots, n$, where $\sum_j \hat{f}_j$ is the solution to 3.2 under the assumption of main effects only and let $\hat{\boldsymbol{\varepsilon}}_n = \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$. Then, the statistic

$$nT_n(\mathbf{X}_n, \hat{\boldsymbol{\varepsilon}}_n), \tag{3.4}$$

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 hypothesis.

We can also test whether covariates that are currently not in the model should be included. We assume that the data generating mechanism in 3.1 holds, and that f, the function that specifies the relationship between Y and X is correctly specified. 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:

$$\begin{aligned} H_0 &: HSIC(P_{z,\varepsilon}, Z, \varepsilon) = 0 \\ H_A &: HSIC(P_{z,\varepsilon}, Z, \varepsilon) > 0. \end{aligned} \tag{3.5}$$

If the the null holds, then $\varepsilon = \eta$, the true model error. Hence, ε is independent of Z, and $HSIC(P_{z,\varepsilon}, Z, \varepsilon) = 0$. This means the model has no terms that depend on Z. If the alternative holds, $\varepsilon \neq \eta$, and there is a lack-of-fit. Hence, ε is dependent on Z, $HSIC(P_{z,\varepsilon}, Z, \varepsilon) > 0$, and the model has terms that depend on Z.

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 3.1, we can test the null and alternative hypotheses in 3.3. 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 3.2 and let $\hat{\boldsymbol{\varepsilon}}_n = \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$. Then, the statistic

$$nT_n(\mathbf{Z}_n, \hat{\boldsymbol{\varepsilon}}_n). \tag{3.6}$$

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 3.4 and 3.6 is the proposed statistic to test the goodness-of-fit of the SS-ANOVA. The model fit can be easily assessed by first estimating the residuals and then calculating the test statistic in 3.4 or 3.6 to check for a lack-of-fit, with respect to a set of covariates $\mathbf{X}_n$ or $\mathbf{Z}_n$, respectively.

To perform the test, we need a valid distribution of 3.4 and 3.6 under the null

hypothesis. An approximation to the null distribution is used. Details are shown in the next section.

3.3.4 Approximation to the Null Distribution of the Test Statistic with the Bootstrap

The difficulty in using 3.4 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 η 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: If hypotheses in 3.5 need to be tested using test statistic in 3.6 the same bootstrap can be used with small changes. Details are shown in Appendix A.

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 p increases relative to n, in finite samples. 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 p. 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 p is only moderately big. This is an improvement over the bootstrap procedure in Sen and Sen (2014), 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.

3.3.5 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 data generating mechanism in 3.1. A function f is assumed for the relationship between Y and X. Estimated residuals are obtained by finding a solution to 3.2 and setting $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$ for $i = 1, \ldots, n$.

Theorem 3.3.2. *Under H_0,*

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

Under H_A,

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

Under both H_0 and H_A,

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

The proof of this result can be found in Appendix A. 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 3.4 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 $(\mathbf{Z}_n^*, \mathbf{Z}_n, Z)$.

3.3.6 Illustrative Cases

The framework presented here is a test for the Goodness-of-fit of the SS-ANOVA model. The test is very versatile and can detect any possible lack-of-fit. The versatility of the test comes from the fact that HSIC can detect any form of statistical dependence. However, since lack-of-fit can happen in many ways, the general case is not

particularly illuminating, and hence three cases of lack-of-fit will be used as illustrations for the method both in the theory and the simulation results. In all three cases, the null-hypothesis will correspond to the situation where the current model fits the data properly, and under the alternative hypothesis the model is misspecificed in some way.

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. In terms of the SS-ANOVA model, the hypotheses are

$$H_0 : Y = \sum_{j=1}^{p} f_j(X(j)) + \eta$$
$$H_A : Y = \sum_{j=1}^{p} f_j(X(j)) + f_{1,\ldots,p}(X(1), \ldots, X(p)) + \eta,$$

where $f_{1,\ldots,p}(X_1, \ldots, X_p)$ 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 $\{1, \ldots, p\}$. 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. In terms of the SS-ANOVA model the hypotheses are

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

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: Missing Covariates

We can test whether a model that includes covariates X needs also to include covariates Z. In terms of the SS-ANOVA model, the hypotheses are

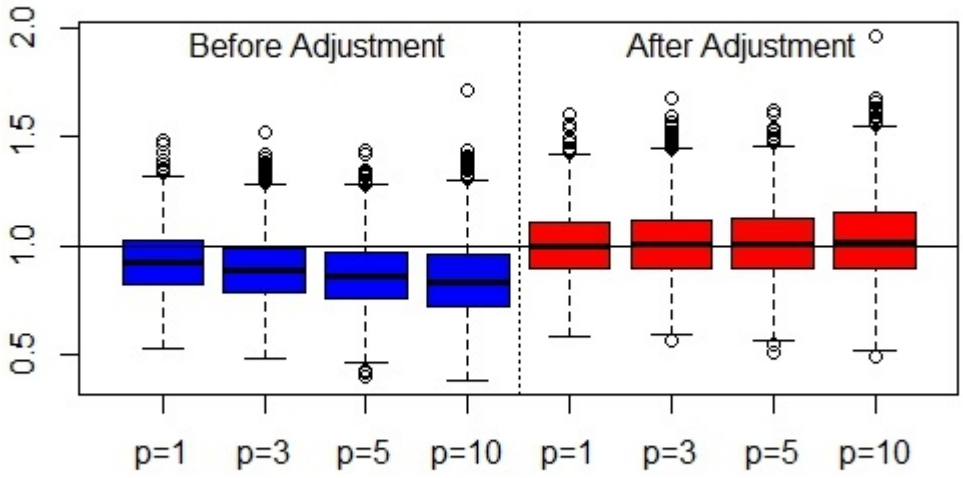
$$H_0 : Y = f(X) + \eta$$
$$H_A : Y = f(X, Z) + \eta.$$

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 3.1: 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.

3.4 Simulation Studies

This section will present simulation results comparing the variance of the empirical measure of the estimated residuals before and after the adjustment described in the previous section on the bootstrap. 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.

Variance Adjustment to the Bootstrap

The left panel of figure 3.1 shows the box plots of $\hat{\sigma}'^2$ for 2000 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.

In all cases shown below we simulated η as $N(0,1)$ and all $X(j)$ as Uniform(0,1) independent of η. 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 $f_{1,\ldots,p}(X(1),\ldots,X(p))$

is an interaction between covariates. Our hypotheses then become

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

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 $f_{1,2}(X(1), X(2)) = 0.75\cos(\pi(X(1) - X(2)))$. When $p = 4, 6$ similar models were used.

Table 3.1: Missing Interactions Beyond the Main Effects

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.

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 3.1 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 II: Missing Covariates

Simulations were created where 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 $f_{p+1,\ldots,p+q}(Z(1), \ldots, Z(q))$, where $f_{p+1,\ldots,p+q}(Z(1), \ldots, Z(q))$ are variables leftover not included in $f(X)$. Our hypotheses then becomes

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

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

Table 3.2: Missing Covariates

Type I Error				
Var.	Sig.	n=100	n=300	n=500
2	0.01	0.008	0.011	0.010
2	0.05	0.054	0.057	0.057
4	0.01	0.014	0.010	0.011
4	0.05	0.063	0.053	0.050
Power				
Var.	Sig.	n=100	n=300	n=500
2	0.01	0.014	0.034	0.072
2	0.05	0.077	0.154	0.321
4	0.01	0.240	0.828	1
4	0.05	0.464	0.944	1

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

$f_3(X(3)) = \sin(\pi(X(3)))$ 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 Liu et al. (2007), and Wu et al. (2011) 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 3.2, it can be seen that the 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.

Case III: 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 hypotheses then become

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

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 3.3 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 a 4^* on the table 3.3, 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*

Table 3.3: Missing Interactions Beyond the Within Group Interactions

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.

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.

3.5 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 (Engel et al. (2011)). 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:

$$\text{PDI} = f_1(BPA) + f_2(DEHP) + f_3(DAP) + f_4(AGE) + \varepsilon. \tag{3.7}$$

The following paragraph provides an investigation into whether the model in 3.7 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 3.4. The first column of table 3.4 shows the null hypotheses that are tested, the second column shows the interaction terms that have been added to the basic model in 3.7, 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.

Table 3.4: Testing of Goodness-of-fit

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.

Initially, we test the null hypothesis H_0, which corresponds to testing if the model in 3.7 fits the data well. The p-value of H_0 is 0.044, hence we detect a lack-of-fit. Since

the model in 3.7 is not sufficient, it is possible that interactions need to be considered. Three models are possible extensions, and they only differ from 3.7 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:

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

3.6 Discussion

In this article we have developed a general Goodness-of-fit statistic and test for nonparametric regression in the setting of the SS-ANOVA model with continuous outcome. 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. 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 a smoothing spline ANOVA model, 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.

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.

CHAPTER 4: NONPARAMETRIC MULTIVARIATE CHANGE POINT

4.1 Summary

Understanding the dynamics between physical activity and blood glucose in type I diabetes patients is an important yet difficult task. Physical activity is one of the major disruptors of blood glucose levels in type I individuals and could potentially drive them to a dangerous state of hypoglycemia. Understanding these dynamics is of great importance to the development of an artificial pancreas system. One of the challenges of modeling blood glucose and physical activity is that their relationship is not stable, this means that the relationship will vary depending on what time of the day it is, and even if the individual is doing a high intensity exercise. Blood glucose level and energy expenditure data were collected every 5 minutes on one individual over 23 hours in a metabolic chamber. A statistical method is proposed that estimates the change between blood glucose and energy expenditure nonparametrically. That is, no model or parametric form is assumed between the multivariate relationship of blood glucose and energy expenditure. The relationship is multivariate because we assume that several future values of blood glucose depend on many lagged values of energy expenditure. Two major change points are estimated from the metabolic chamber data, creating three time intervals for our data. We fit three models within each time interval and find that the relationship is in fact quite different. Changes in the relationship correspond to time intervals where the subject of study incur in high exercise activity.

4.2 Introduction

Individuals with type 1 diabetes (T1D) have to take into account diet and exercise when deciding to take extra insulin doses beyond the daily baseline amounts, in order to control their blood glucose (BG) level. If not controlled correctly, T1D individuals can easily suffer hypoglycemia, or low levels of glucose in the blood. Recently, progress has been made towards the creation of an artificial pancreas system that replicates the physiology that is lost in diabetes (Kowalski (2015)). Developing an artificial pancreas has many difficulties, one of which is to model the relationship between BG and physical activity (PA) (van Bon et al. (2011)). The objective of the current article is to evaluate, for individuals with T1D, how the relationship between PA and BG changes throughout a day, in order to better understand their dependence and dynamics. We have developed a change point estimation methodology to accomplish this. Our methodology estimates and tests for possible change points in the relationship between two sets of random variables, which in the current set up are PA and BG. With our method we wish to understand the changing relationship of PA and BG, while accounting for insulin use. We believe that this will be a contribution to the artificial pancreas, since it will increase understanding of the dynamics of PA and BG. If changes can be pinpointed, then this will ease the difficulty of modeling BG in terms of PA.

The reason for doing a change point analysis, is that, for T1D individuals, the relationship between PA and BG is subject to many changes depending on many circumstances. For instance, BG can fluctuate depending on the type, intensity and duration of PA. Light-to-moderate intensity areobic PA usually results in a fast drop in glucose in T1D subjects which could potentially results in hyplogycemia (Camacho et al. (2005), Tonoli et al. (2012), Kudva et al. (2014),Yardley et al. (2012), Yardley et al. (2013)). On the contrary, intense arerobic-anaerobic PA will frequently result

in hyperglycemia (Turner et al. (2015),Purdon et al. (1993)). Moreover, PA can also affect BG several hours after exercise (Maran et al. (2010), Iscoe and Riddell (2011)). In addition, the normal reduction in insulin secretion at the onset of exercise, either moderate or vigorous, cannot easily be emulated in T1D (Riddell et al. (2015)). In addition, exercise may increase insulin absorption rates from the subcutaneous depot, causing circulating levels to rise even if pump infusion rates remains constant or stop (Mallad et al. (2015)). There is some evidence that that the risk of hypoglycemia during and soon after exercise is not high when exercise is performed while plasma insulin levels are close to basal levels, particularly when exercise intensity is elevated (Shetty et al. (2016)). We have to our disposal data that will help us estimate change points in the relationship between PA and BG.

We collected data on energy expenditure (EE) and BG on one participant who stayed 23 hours in a metabolic chamber. The chamber recorded EE while the BG was collected through continuous glucose monitoring (CGM). Recordings were done every 5 minutes for a total of 271 observations. This setup is ideal to discover some of the dynamics between BG and EE. Specifically, our goal is to discover if the relationship between BG and EE changes over the course of these 23 hours. If changes do occur, we are interested in evaluating if these changes happen along with changes in the distribution of insulin on board (IOB) or PA, which is predicted by the literature previously mentioned. However, present and future values of BG can potentially dependent on present and past values of PA. Hence, the relationship is multivariate in nature. Also, it is not clear if the relationship will be at all linear or includes interactions between the lagged values. Our change point methodology is nonparametric, hence it does not require the specification of a model between the two multivariate vectors and can potentially detect changes in interactions. Our method will allow us to evaluate if the metabolic chamber data is consistent with the current knowledge in the literature and

also if we can discover new patterns that can be used in the development of an artificial pancreas.

Many approaches to the change point problems exist in either a nonparametric or multivariate context. For example, Zou et al. (2014) developed a nonparametric maximum likelihood approach to detecting multiple change points, with no assumptions on the distribution of the data. However, their method only works for the univariate case and thus only looks at changes in distribution. Similarly, Killick et al. (2012) create a method called PELT that detects multiple change points, but only in a univariate set up. Another approach was developed by Fryzlewicz et al. (2014) where a binary segmentation algorithm was used to select the change points, but it only works in the univariate case too. On the multivariate front, there are methodologies that detect change points in the joint distribution of a multivariate vector observed over time (or location), among these there are Song et al. (2012), Sugiyama et al. (2008). One such method that has received a lot of attention recently is the method of Matteson and James (2014) called energy change point (ECP) where a change point is selected by maximizing the difference between two sample characteristic functions. The main limitations of these methods for solving the question of interest in the current article are that existing methods do not look at changes in the relationship between two multivariate vectors, but instead look at changes in the distribution of a single multivariate vector. This methods will detect any changes in distribution, that might or might not be related to the change in relationship between BG and PA that we are interested in. Hence, these methods try to solve a problem that is more general and not fitted for our purpose.

For this reason, we developed a method to estimate nonparametrically the change point in the relationship between two multivariate vectors of general dimension. Our method works by evaluating nonparametrically the strength of the relationship between

the two multivariate vectors before and after a given time point, using a form of generalized correlation. If the correlation is highly different before and after this time point, then this would indicate that at this time point there has been a change in the strength of the relationship, i.e., a change point. If for all time points we inspect the difference in correlations before and after, then the point most likely to be a change point would be the one where the difference is the largest. This is exactly how our method will proceed. The estimator of the change point will be the time point where the largest difference in pre and post correlation occurs. The method fits well the current setup where it is believed that the relationship between BG and EE is dynamic but changes along time, depending on many factors. We analyze the relationship between BG and EE in two stages. First, using metabolic chamber data, we look at change points corresponding to when the relationship between EE and BG changes. This strategy will provide us with time intervals that, we hypothesize, correspond to periods of different PA intensity and type. At a second stage, we look at the distribution of EE, insulin and diet, and evaluate if they significantly change from interval to interval estimated, so that we can assess if the changes in the relationship between BG and EE are accompanied with changes in these distributions.

The remainder of the article is organized as follows. Section 4.3 proposes a general statistical method for estimating and testing a multivariate change point in the relationships between to vectors. Section 4.4 presents simulations results. Section 4.5 presents how the method will be applied to the metabolic chamber data. Section 4.6 concludes with some discussion.

4.3 Nonparametric Multivariate Change Point

4.3.1 Problem Set Up

As noted earlier, PA, and therefore EE, can affect BG after a certain amount of time, potentially an hour or more. The relationship can be very dynamic in nature and therefore we believe that, BG will depend on lagged values of EE. Moreover, we want to know how EE and its lagged values affect BG an hour into the future and its consecutive values after 5 and 10 minutes. Hence, we want to look at the relationship between a vector of lagged values of EE and a vector of future values of BG. Thus, our problem is multivariate in nature and we wish to estimate a change point in the relationship between these 2 vectors. We develop a method to solve this problem. The set up is described below.

We are interested in a sequence of a pair of multivariate vectors (Y_t, X_t) for $t = 1, \ldots, T$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. The pair (Y_t, X_t) are observed sequentially along a dimension denoted by the index t. In our context t can denote time, and Y_t can be several values of BG and X_t can be several values of EE, at time t. The change point problem that we are interested is characterized in the following data generating mechanism:

$$\begin{aligned} Y_t &= f_1(X_t, \eta_t) \quad \text{for} \quad t = 1, \ldots, \tau, \\ Y_t &= f_2(X_t, \eta_t) \quad \text{for} \quad t = \tau + 1, \ldots, T, \end{aligned} \tag{4.1}$$

where η_t and X_t are i.d.d for each t, and are independent of each other. Thus, Y_t depends on X_t through a function f_1. However, as t advances, a point τ is reached where Y_t depends on X_t through f_2. It is of interest to estimate τ and test whether

this change-point actually exists. The hypotheses that we are interested in are

$$\begin{aligned} H_0 &: f_1 = f_2, \\ H_A &: f_1 \neq f_2, \end{aligned} \tag{4.2}$$

and note that under the null hypothesis τ vanishes. Possible approaches in testing (4.2) would require estimating f_2 and f_2. It can be difficult to explicitly model the multivariate relationship between random X and Y parametrically; this becomes more difficult if Y is categorical, or if each element of Y is categorical of different type, i.e. ordinal, nominal or count; for estimating the change-point, a model is estimated at each (time) point and this can be computationally intensive; and there are approaches that model the change-point nonparametrically for the joint distribution of (X, Y), but not for the relationship between Y and X, and thus this approach is too general for our purpose.

In the current research, we avoid the complications of modeling f_1 and f_2. Instead, we will test (4.2) by creating a test statistic based on the distance covariance (DC). The DC statistic, denoted by $\mathcal{V}^2(X, Y)$ for the DC between X and Y, can evaluate how strong the statistical dependence between two random vectors is. The larger the value the stronger the dependence. A value of DC of exactly 0 is equivalent to statistical independence. In the current context, we will construct two DCs between two random vectors, one before and one after a given potential change point τ^*. If the difference between these two DCs is large for this specific τ^*, then this will provide evidence that τ^* constitutes a change-point. We will look at all such differences, times a rate, across all possible τ's and select the one position that maximizes this difference and denoted this $\hat{\tau}$; this will be our estimator of the change-point.

One caveat is that, while looking at changes of sample DC before and after a

change point τ^*, if the distribution of X changes while its relationship to Y remains intact, then this will generate slight differences in the sample DC that are not attributable to changes in the relationship between Y and X before and after τ^*. Hence, for the sake of stability, while evaluating if a given τ^* constitutes a true change point, the marginals of all the sample data of X and Y before τ^* will be transformed to be discrete uniform. The same will be done for all data marginals of X and Y after τ^*. Then, for this τ^*, the difference in DC will be assessed by using this uniformly transformed data instead of the original data. The uniform transformation will be performed anew for each potential change point and $\hat{\tau}$ will be chosen to be the one that maximizes the difference in DC. This will allow the estimator $\hat{\tau}$ to be less sensitive to changes in marginal distribution that are not related to changes in the relationship of X and Y. We will give more details of the transformation in Section 4.3.4.

Once an estimate $\hat{\tau}$ of the change point is attained, it is important to evaluate if it constitutes a true change point. The reason is that change point estimators always provide a change point regardless of whether one exists or not. Then, a test of the existence of the change point should be performed. One can test a similar set of hypotheses as (4.2) to accomplish this by testing instead

$$\begin{aligned} H_0 &: \mathcal{V}^2(X_1, Y_1) = \cdots = \mathcal{V}^2(X_T, Y_T), \\ H_A &: \cdots = \mathcal{V}^2(X_\tau, Y_\tau) \neq \mathcal{V}^2(X_{\tau+1}, Y_{\tau+1}) = \cdots, \end{aligned} \tag{4.3}$$

for some unknown τ. Our main assumption is that, if the change-points described in (4.1) exists, then $\mathcal{V}^2(X_\tau, Y_\tau) \neq \mathcal{V}^2(X_{\tau+1}, Y_{\tau+1})$, meaning that, if a change-point exist then, that change-point will be accompanied with a difference in DC before and after the change-point.

A test can be performed by permuting the order of the sequence (Y_t, X_t) for

$t = 1, \ldots, T$ and recalculating the test statistic many times to create a null distribution. Among the virtues of this estimation and testing procedure are that no form for f_1 and f_2 need ever be specified, that the test statistics is easy to compute and requires no regularization, and that it can accommodate any kind of data of any dimension.

4.3.2 Distance Covariance

Distance covariance was developed by Székely et al. (2007), Székely et al. (2009), and Székely and Rizzo (2013). For random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, let ϕ_x, ϕ_y and $\phi_{x,y}$ be the characteristic function of X, Y and (X, Y), respectively. Assume that $E|X|_p < \infty$ and $E|Y|_q < \infty$. Distance covariance ($\mathcal{V}$) can be used to measure the dependence between X and Y through the distance

$$\begin{aligned}\mathcal{V}^2(X,Y) =& ||\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)||^2 \\ =& \int_{\mathbb{R}^{p+q}} |\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)|^2 (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1} dtds\end{aligned}$$

with $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ and $|\cdot|_p$ is the Euclidean norm in $\mathbb{R}^p$. If $X \not\perp Y$ then $\mathcal{V}^2(X,Y)$ will be greater than 0. otherwise if $X \perp Y$ then it will be exactly 0. Distance variance can be defined as $\mathcal{V}^2(X) = \mathcal{V}^2(X,X)$. The distance correlation between X and Y is the nonnegative number $DC(X,Y)$ defined by

$$DC(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}$$

if $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$ and equals 0 otherwise. The distance covariance statistic are defined as follows. For an observed random sample $\{(X_k, Y_k) : k = 1, ..., n\}$ from the

joint distribution of random vectors $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$, define

$$a_{kl} = |X_k - X_l|_p, \qquad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \qquad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$
$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

for $k, l = 1, ..., n$. Similarly, define $b_{kl} = |Y_k - Y_l|_q$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ for $k, l = 1, ..., n$. The empirical distance covariance $\mathcal{V}_n(X, Y)$ and distance variance $\mathcal{V}_n(X)$ are the nonnegative numbers defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} \text{ and } \mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2.$$

The empirical distance correlation $DC_n(X, Y)$ is defined as

$$DC_n^2(X, Y) = \frac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}$$

if $\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0$ and 0 otherwise. Both $\mathcal{V}_n^2(X, Y)$ and $DC_n^2(X, Y)$ are a.s. consistent for $\mathcal{V}^2(X, Y)$ and $DC^2(X, Y)$, respectively.

4.3.3 Unbiased Distance Covariance

The estimator $\mathcal{V}_n^2(X, Y)$ is a V-statistic which is biased for $\mathcal{V}^2(X, Y)$, with bias disappearing asymptotically. Below we present an unbiased estimator of $\mathcal{V}^2(X, Y)$ in the form of a U-statistic.

Theorem 4.3.1 (U-Statistic)**.** *The statistic defined as*

$$\mathcal{U}_n^2(X, Y) = \frac{1}{n(n-3)}\Big[Tr(KL) + \frac{1^T K 1 1^T L 1}{(n-1)(n-2)} - \frac{2}{n-1}1^T KL1\Big],$$

where K and L are $n \times n$ matrices with $K_{i,j} = |X_i - X_j|_p$ and $L_{i,j} = |Y_i - Y_j|_q$, is unbiased

for $\mathscr{V}^2(X,Y)$ and is a U-statistic.

Proof: This follows immediately from Theorem 1 and 2 of Bounliphone et al. (2014), which follows Song et al. (2012) very closely, by replacing the Gaussian kernels with Euclidean distances instead.□.

We choose to use the estimator $\mathscr{U}_n^2(X,Y)$ because from our numerical results it performs better than $\mathscr{V}_n^2(X,Y)$ in the context of the change point problem.

4.3.4 Change Point Estimator

We will consider the data mechanism described in 4.1 where are pair of multivariate data (Y_t, X_t) is observed sequentially over an index $t = 1,\ldots,T$. For all $\tilde{\tau} \in \{1,\ldots,T\}$, let $X^*(\tilde{\tau}-) = \{X_1,...,X_{\tilde{\tau}}\}$, $Y^*(\tilde{\tau}-) = \{Y_1,...,Y_{\tilde{\tau}}\}$, $X^*(\tilde{\tau}+) = \{X_{\tilde{\tau}+1},\ldots,X_T\}$, and $Y^*(\tilde{\tau}+) = \{Y_{\tilde{\tau}+1},\ldots,Y_T\}$. Note that $X^*(\tilde{\tau}-), Y^*(\tilde{\tau}-), X^*(\tilde{\tau}+), Y^*(\tilde{\tau}+) \in \mathbb{R}_{\tilde{\tau}\times p}, \mathbb{R}_{\tilde{\tau}\times q}$, $\mathbb{R}_{T-\tilde{\tau}\times p}$ and $\mathbb{R}_{T-\tilde{\tau}\times q}$, respectively. The data $X^*(\tilde{\tau}-)$, $X^*(\tilde{\tau}+)$, $Y^*(\tilde{\tau}-)$, and $Y^*(\tilde{\tau}+)$ will be transformed so that all their marginals are discrete uniform. This is always possible to do whenever (Y_t, X_t) are continuous. This is described below.

Discrete Uniform Transformation

Let $X^*_{t,j}(\tilde{\tau}-)$ and $Y^*_{t,j}(\tilde{\tau}-)$ be the $t \times j$ entry of $X^*(\tilde{\tau}-)$ and $Y^*(\tilde{\tau}-)$, respectively. For a fixed $\tilde{\tau}$ and j, let $O(X^*_{t,j}(\tilde{\tau}-))$ be the rank of $X^*_{t,j}(\tilde{\tau}-)$ among all the $\tilde{\tau}$ observations in the j-th column vector of $X^*(\tilde{\tau}-)$. For a fixed $\tilde{\tau}$ and j, let $O(X^*_{t,j}(\tilde{\tau}+))$ be the rank of $X^*_{t,j}(\tilde{\tau}+)$ among all the $T-\tilde{\tau}$ observations in the j-th column vector of $X^*(\tilde{\tau}+)$. Similar definitions follow for $O(Y^*_{t,j}(\tilde{\tau}-))$ and $O(Y^*_{t,j}(\tilde{\tau}+))$. Then, we define the uniform transformed data of $X^*(\tilde{\tau}-)$ and $X^*(\tilde{\tau}+)$ as

$$X_{t,j}(\tilde{\tau}-) = \frac{O(X^*_{t,j}(\tilde{\tau}-))}{\tilde{\tau}} \qquad \text{and} \qquad X_{t,j}(\tilde{\tau}+) = \frac{O(X^*_{t,j}(\tilde{\tau}+))}{T-\tilde{\tau}},$$

for all $j = 1, \ldots, p$. Similarly, the uniform transformed data of $Y^*(\tilde{\tau}-)$ and $Y^*(\tilde{\tau}+)$ is

$$Y_{t,j}(\tilde{\tau}-) = \frac{O(Y^*_{t,j}(\tilde{\tau}-))}{\tilde{\tau}} \qquad \text{and} \qquad Y_{t,j}(\tilde{\tau}+) = \frac{O(Y^*_{t,j}(\tilde{\tau}+))}{T-\tilde{\tau}},$$

for all $j = 1, \ldots, q$. Following this transformation, we work exclusively with $X(\tilde{\tau}-)$, $X(\tilde{\tau}+)$, $Y(\tilde{\tau}-)$ and $Y(\tilde{\tau}+)$.

The reason to do the transformation is that we care about the dependence between Y_t and X_t, but not the marginal distributions and so we do not want to be sensitive to changes in them. One can see that the change point estimator (4.5) could be affected by changes in the marginal distribution of X_t or Y_t along t, were the uniform transformation is not applied, even though the relationship between Y_t and X_t remains intact. While using the transformation, changes in the marginals cannot perturb the estimate $\hat{\tau}$. This is the reason for the use of $X(\tilde{\tau}-)$, $X(\tilde{\tau}+)$, $Y(\tilde{\tau}-)$ and $Y(\tilde{\tau}+)$.

We define the absolute difference between two DCs, one using only data up to $\tilde{\tau}$ and the other one using only data after $\tilde{\tau}$ as

$$\Delta\mathscr{U}_n^2(X, Y; \tilde{\tau}) = |\mathscr{U}_n^2(X(\tilde{\tau}-), Y(\tilde{\tau}-)) - \mathscr{U}_n^2(X(\tilde{\tau}+), Y(\tilde{\tau}+))|. \tag{4.4}$$

Intuitively, the statistic $\Delta\mathscr{V}_n^2(X, Y; \tilde{\tau})$ will tend to be very small if the sample DC between Y_t and X_t changes little before and after $\tilde{\tau}$. This will tend to happen if the relationship between Y_t and X_t remains unchanged before and after $\tilde{\tau}$. On the other hand, $\Delta\mathscr{V}_n^2(X, Y; \tilde{\tau})$ will tend to be big if the sample DC between Y_t and X_t changes significantly before and after $\tilde{\tau}$, and this will tend to happen if the relationship between Y_t and X_t changes substantially before and after $\tilde{\tau}$. Therefore, large values of $\Delta\mathscr{U}_n^2(X, Y; \tilde{\tau})$ provide evidence that $\tilde{\tau}$ is a change point.

If a change point is believed to exist, it makes sense to estimate the change point

as the $\tilde{\tau}$ that maximizes $\Delta\mathscr{U}_n^2(X,Y;\tilde{\tau})$ (times a rate). In such a procedure, we would be selecting the $\tilde{\tau}$ that has the greatest evidence to be a change point. We will show in Theorem 4.3.2 that $a(\tilde{\tau})\cdot\Delta\mathscr{U}_n^2(X,Y;\tilde{\tau})$ will attain its maximum at the true change point τ, with probability tending to 1 under some conditions and as T increases, where $a(\tilde{\tau}) = \sqrt{\tilde{\tau}(T-\tilde{\tau})/T}$. Then, our estimator of the change point is

$$\hat{\tau} = \underset{\tau\in\{1,\ldots,T\}}{\arg\max}\ a(\tau)\cdot\Delta\mathscr{U}_n^2(X,Y;\tau). \tag{4.5}$$

Then, $\hat{\tau}$ is the point along the sequence that maximizes the difference in DC between Y_t and X_t. Hence, if a difference in the relationship between Y_t and X_t happens, it will translate into a difference in DC, and thus it will correspond to a maximum difference at $\hat{\tau}$.

The following theorem shows that as the number of observation along the sequence increases ($T\to\infty$), the estimator in (4.5) is consistent for the true change point.

Theorem 4.3.2 (Consistency)**.** *Assume the data mechanism describe in (4.1). Let* $DC_1 = DC(X_t,Y_t)$ *for* $t = 1,\ldots,\tau$, *and* $DC_2 = DC(X_t,Y_t)$ *for* $t = \tau+1,\ldots,T$. *Assume* $DC_1 \neq DC_2$. *Moreover, let* $\{\delta_T\}$ *be a sequence such that* $\delta_T \in [0,1]$, $\delta_T\to 0$ *and* $T\delta_T\to\infty$ *as* $T\to\infty$. *Then, as* $T\to\infty$ *and for all* $\epsilon > 0$,

$$Pr\Big(|\frac{\tau}{T} - \frac{\hat{\tau}}{T}| > \epsilon\Big) = 0.$$

Proof: Similar to the proof of theorem 1 of Matteson and James (2014) and it is given in Appendix B.

The assumption that $DC_1 \neq DC_2$ is our identifiability assumption, meaning that differences in f_1 and f_2 will translated into differences in $DC_1 \neq DC_2$. The case $DC_1 = DC_2$ would correspond to the null hypothesis in (4.3). In which case $\hat{\tau}$ is not consistent

because the true τ vanishes. This is why it is important to test (4.3).

4.3.5 Test Statistic and Null Distribution

As it is common in change point problems, we want to test whether the change point estimated $\hat{\tau}$ actually exists. This is so because, by using (4.5), we will always get an estimate of the change point, whether or not one exists. Testing for the existence of the change point is equivalent to testing (4.3). The test statistics to be used will be

$$\max_{\tau \in \{1,\ldots,T\}} a(\tau) \cdot \Delta \mathscr{U}_n^2(X, Y; \tau). \tag{4.6}$$

Under the alternative that there exist a change point, (4.6) will grow large as $T \to \infty$. To derive a null distribution for (4.6) a permutation strategy is used. Let π_b for $b = 1, \ldots, B$, denote random permutations of the ordered indices $\{1, \ldots, T\}$. Let $(Y_{\pi_b(t)}, X_{\pi_b(t)})$ be the pair of multivariate observations observed sequentially over the index t, but with the index permuted by π_b. Then, B such permuted statistics are calculated as

$$\max_{\tau \in \{1,\ldots,T\}} a(\tau) \cdot \Delta \mathscr{U}_n^2(X_{\pi_b}, Y_{\pi_b}; \tau), \text{for } b = 1, \ldots, B. \tag{4.7}$$

The B permuted statistics in (4.7) will form the null distribution for the test statistic in (4.6). If a change point τ exists, but we permute the data (Y_t, X_t) along the sequence t with π_b, we are effectively simulating the null hypothesis, because we are destroying the differences that exist before and after τ. By permuting by π_b any differences will be averaged out by the permutation, and the relationship between Y_t and X_t will be

on average the same across all t. The p-value of the test can be calculated as

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} I\Big(|\max_{\tau\in\{1,\ldots,T\}} a(\tau)\cdot\Delta\mathscr{U}_n^2(X,Y;\tau)| \le |\max_{\tau\in\{1,\ldots,T\}} a(\tau)\cdot\Delta\mathscr{U}_n^2(X_{\pi_b},Y_{\pi_b};\tau)\Big|\Big), \tag{4.8}$$

which corresponds to calculating the proportion of times the B permuted statistics in (4.7) were larger than the observed statistic in (4.6).

If the null hypothesis in (4.3) is rejected at α level with estimated change point $\hat{\tau}_1$, then we can test whether there exist further change points within the new 2 intervals created before and after $\hat{\tau}_1$, namely $\{1,\ldots,\hat{\tau}_1\}$ and $\{\hat{\tau}_1+1,\ldots,T\}$. Each test can be performed at the $\alpha/2$ level. This preserves the family-wise error rate. Further partitions can be created and change points can be estimated in this hierarchical fashion.

4.4 Simulation Results

4.4.1 Type I Error

To evaluate the type I error and the power of the test and the estimation error of the change point estimator, we carried out simulation studies under 3 major configurations. For all simulations scenarios, let $\eta_{T\times 2}$ be a matrix of dimension $T\times 2$, $X_{T\times 3}$ be a matrix of dimension $T\times 3$, $Y_{T\times 2}$ be a matrix of dimension $T\times 2$, where T denotes the number of samples observed in a sequence. The tth row of each matrix will be denoted by a subscript t as η_t, X_t and Y_t. Each (i,j)th entry of $\eta_{T\times 2}$ is an i.i.d sample from a standard normal distribution. T will take as values 100, 200 and 300. For the power calculations, τ will have as values $0.3T$, $0.5T$ and $0.7T$, and will be allowed to vary as T varies. Each scenario was simulated 10,000 times. We provide 3 simulation scenarios that represent a variety of possible applications, for both type I error and power. We

mainly focus on the mean absolute error (MAE) to assess the estimation error and on power to assess the test performance.

Linear Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard uniform distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} X_{t,1} + X_{t,2} + \eta_{t,1} & \text{for} \quad t = 1, ..., T, \end{cases}$$

$$Y_{t,2} = \begin{cases} X_{t,2} + X_{t,3} + \eta_{t,2} & \text{for} \quad t = 1, ..., T. \end{cases}$$

In this scenario, Y_t remains linear in X_t throughout and no change happens along $t = 1, \ldots T$. It can be seen that it would be difficult to evaluate this situation for a change point if it was not known in advance that the relationship is linear. From table (4.1) we see that the method performs well. The MAE decreases and the power increases as the sample size increases.

Table 4.1: Linear Association

Type I Error		
T	$\alpha = 0.01$	$\alpha = 0.05$
100	0.0127	0.0501
200	0.0108	0.0512
300	0.0102	0.0523
T is the sample size.		

Nonexistent Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard normal distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} \eta_{t,1} & \text{for} \quad t = 1, ..., T, \end{cases}$$

$$Y_{t,2} = \begin{cases} \\ \eta_{t,2} \end{cases} \quad \text{for} \quad t = 1, ..., T.$$

This scenario is important because it exemplifies many situations where a change point is believed to exist but the two sets of random variables are actually independent. This can happen in an applied set-up where many pairs of random variables are evaluate for change points, and many have no relationship at all. From table 4.2 we see that the method preserves the correct type I error within Monte Carlo error.

Table 4.2: Nonexistent Relationship

Type I Error		
T	$\alpha = 0.01$	$\alpha = 0.05$
100	0.0116	0.0527
200	0.0109	0.0520
300	0.0107	0.051
T is the sample size.		

Quadratic Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard normal distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} \\ X_{t,1} + X_{t,2} + 2X_{t,1}^2 + 2X_{t,2}^2 + \eta_{t,1} \end{cases} \quad \text{for} \quad t = 1, ..., T,$$

$$Y_{t,2} = \begin{cases} \\ X_{t,2} + X_{t,3} + 2X_{t,2}^2 + 2X_{t,3}^2 + \eta_{t,2} \end{cases} \quad \text{for} \quad t = 1, ..., T.$$

This scenario is relevant because it shows a situation where it would be difficult to model parametrically the relationship between Y and X without knowing a priori the their relationship. From table (4.3) we see that the method performs well.

Table 4.3: Quadratic Relationship

Type I Error		
T	$\alpha = 0.01$	$\alpha = 0.05$
100	0.0109	0.0521
200	0.0113	0.0478
300	0.0114	0.0491
T is the sample size.		

4.4.2 Power

From Linear to a Quadratic Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard uniform distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} X_{t,1} + X_{t,2} + \eta_{t,1} & \text{for} \quad t = 1, ..., \tau \\ X_{t,1} + X_{t,1}^2 + X_{t,2} + X_{t,2}^2 + X_{t,1}X_{t,2} + \eta_{t,1} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

$$Y_{t,2} = \begin{cases} X_{t,2} + X_{t,3} + \eta_{t,2} & \text{for} \quad t = 1, ..., \tau \\ X_{t,2} + X_{t,2}^2 + X_{t,3} + X_{t,3}^2 + X_{t,2}X_{t,3} + \eta_{t,2} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

for different τ. In this scenario, the relationship changes from being linear to quadratic plus an interaction. This represents a situation where the relationship starts off simply, but after a change point, it becomes more complex and stronger (in terms of distance correlation). This is a typical set-up for a change point problem. However, it can be seen that it would be difficult to evaluate this situation using a change point methodology that uses a linear regression to model the data. Even though before τ the relationship between the Y and X is linear, but after it has multiple quadratic terms plus an interaction. This would be difficult to model using linear regression when estimating

the change point if it is not known a priori that the change will take this form. This is even more problematic in the current scenario where what we are really interested in is a multivariate relationship. From table (4.4) we see that the method performs well. The MAE decreases and the power increases as the sample size increases.

Table 4.4: From Linear to Quadratic Relationship

T	τ	MAE	Power 0.01	Power 0.05
100	30	11.4%	0.29	0.541
100	50	7.3%	0.43	0.66
100	70	6.3%	0.34	0.53
200	60	10.3%	0.62	0.84
200	100	6.8%	0.76	0.92
200	140	5.1%	0.70	0.86
300	90	8.7%	0.84	0.95
300	150	4.9%	0.93	0.98
300	210	4.4%	0.88	0.96
T is the sample size, τ is the change point, and MAE is the mean absolute error.				

From a Nonexistent Relationship to a Linear Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard normal distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} \eta_{t,1} & \text{for} \quad t = 1, ..., \tau \\ 0.5X_{t,1} + 0.5X_{t,2} + \eta_{t,1} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

$$Y_{t,2} = \begin{cases} \eta_{t,2} & \text{for} \quad t = 1, ..., \tau \\ 0.5X_{t,2} + 0.5X_{t,3} + \eta_{t,2} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

for different τ's. This scenario is important because it exemplifies many situations where

two random variables are initially independent, but after a change point, a relationship now exists. In this scenario, the relationship after the τ is linear, but our method would be able to detect all other forms of deviations from statistical independence between Y and X. This makes our method ideal for applications where the a priori assumption is that there exists no dependence between the 2 vectors of interest. From table (4.5) we see that the method performs well. Moreover, it seems to have greater power and lower MAE than the other 2 scenarios presented here.

Table 4.5: From Nonexistent to Linear Relationship

T	τ	MAE	Power 0.01	Power 0.05
100	30	13.6%	0.44	0.64
100	50	8.1%	0.64	0.79
100	70	3.8%	0.60	0.72
200	60	12.9%	0.74	0.91
200	100	7.9%	0.89	0.97
200	140	3.3%	0.87	0.95
300	90	9.8%	0.84	0.93
300	150	7.5%	0.95	0.99
300	210	2.4%	93	0.98
T is the sample size, τ is the change point, and MAE is the mean absolute error.				

From a Quadratic to a Cubic Relationship The entries of $X_{T\times 3}$ are i.i.d. samples from a standard normal distribution. The rows of $Y_{T\times 2}$ are generated given the following mechanism:

$$Y_{t,1} = \begin{cases} X_{t,1} + X_{t,2} + 2X_{t,1}^2 + 2X_{t,2}^2 + \eta_{t,1} & \text{for} \quad t = 1, ..., \tau \\ 2X_{t,1}^3 + 2X_{t,2}^3 + \eta_{t,1} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

$$Y_{t,2} = \begin{cases} X_{t,2} + X_{t,3} + 2X_{t,2}^2 + 2X_{t,3}^2 + \eta_{t,2} & \text{for} \quad t = 1, ..., \tau \\ 2X_{t,2}^3 + 2X_{t,3}^3 + \eta_{t,2} & \text{for} \quad t = \tau + 1, ..., T, \end{cases}$$

for different τ's. This scenario is relevant because it shows a situation where it would be difficult to model parametrically the relationship between Y and X without knowing a priori the their relationship. Therefore, if one were to estimate a change point by creating a parametric model, all this information would be required before hand. The situation becomes more dire if the relationship happens to be multivariate as it is in this scenario. Even though in this case the relationship goes from quadratic to cubic, our method can potentially capture any form of change in relationship. From table (4.6) we see that the method performs well.

Table 4.6: From Quadratic to Cubic Relationship

T	τ	MAE	Power 0.01	Power 0.05
100	30	11.6%	0.49	0.73
100	50	7.0%	0.62	0.81
100	70	3.1%	0.52	0.71
200	60	10.6%	0.86	0.96
200	100	6.2%	0.93	0.98
200	140	3.5%	0.88	0.96
300	90	9.1%	0.97	0.99
300	150	4.8%	0.99	0.99
300	210	3.1%	0.99	0.99

T is the sample size, τ is the change point, and MAE is the mean absolute error.

General Comments on Power Simulations From all 3 simulation scenarios, we see that the test of existence of the change point is most powerful when the change point is located closest to the the middle of the sequence. In other words, whenever $\tau = 0.5/T$, the test is more powerful compared to the other cases (i.e., $\tau = 0.3T$ and $\tau = 0.7T$).

This makes intuitive sense, because whenever $\tau/T \approx 0.50$, there is the same amount of observations before and after τ. On the contrary, when one side has more data than the other, this affects the ability of the test statistic to estimate the changes in distance covariance.

Another important observation is that, the estimation error evaluated by MAE is smallest whenever $\tau/T \approx 0.7$. This is in part because of how our simulations were created. If we look at the distance correlation between Y and X before and after τ, distance correlation is always lower before the change point compared to after the change point. For instance, in the second scenario, the distance correlation between Y and X was 0 before the change point, whereas after the change point it was > 0. Therefore, MAE would be smallest whenever more data is available to estimate a weak relationship before the change point, and this happens whenever $\tau = 0.7T$ compared to our other simulation parameters. However, had we had done the reverse, meaning that the relationship between Y and X was stronger before τ but weaker after it, we would have also observed the reverse: a smaller MAE whenever $\tau = 0.3T$.

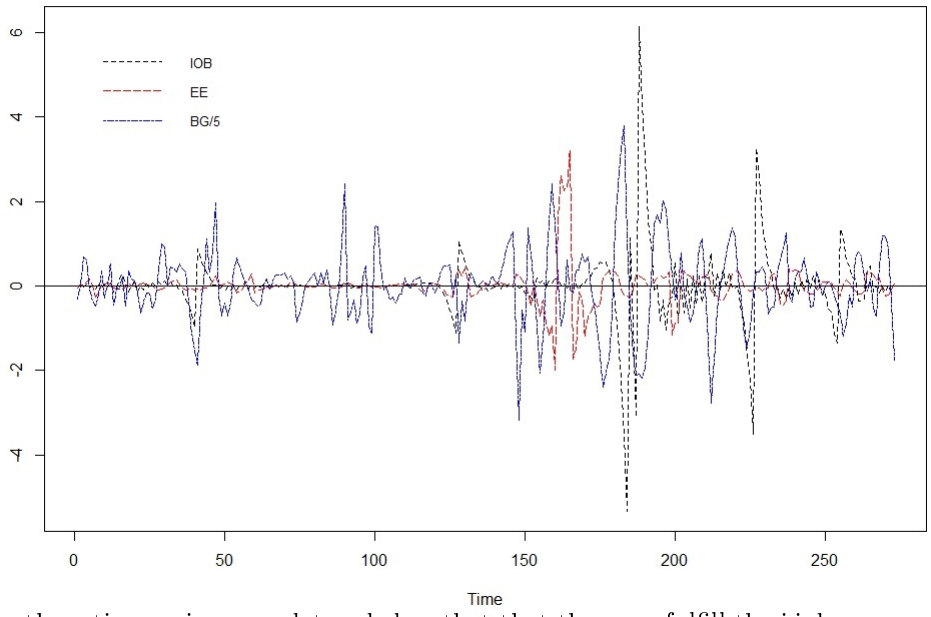
4.5 Data Analysis

4.5.1 Metabolic Chamber Data

Measurements of BG, insulin on board (IOB) and EE on a male subject with T1D were conducted in a metabolic chamber. The participant arrived at the metabolic chamber at night and measurements started at 8:33 pm on Day 1 and lasted until 7:13 pm the next day, denoted by Day 2. Measurements were taken at 5 minutes intervals for a total of 273 observations. A similar proof of concept study was developed in Maahs et al. (2012), but instead of EE, accelerometer data was collected. The data for

detrended BG, IOB and EE is displayed in figure 4.1. This is the data we will work with. We detrended the data so that the mean of the times series would be 0. This makes the data satisfy the i.i.d. assumption at each time point.

Figure 4.1: Detrended IOB, EE and BG



The three time series were detrended so that that they can fulfill the i.i.d. assumption.

The present value of BG can be affected by the present value of EE and IOB, but also by past or lagged values of EE and IOB. Moreover, it is known that there is a delayed effect of EE on BG, meaning EE really affects BG in the future and not immediately. Also, IOB has a lasting effect on BG. However, it can be difficult to know exactly how long this delayed effect is and at what time gap it is strongest, for both IOB and EE. In addition to these complicated time dynamics, the relationships between BG and EE, and between IOB and BG, are not necessarily linear and can include interactions among the lagged values. All these dynamics will be accounted for and dealt with by our method.

4.5.2 Changes in the Effect of EE and IOB on BG

We are going to use metabolic chamber data to establish if changes in the effect of EE and IOB on BG exist. Because we are uncertain on the time delay of the effect of both EE and IOB on BG, and the number of possible lagged values involved, we will incorporate into our analysis multiple values of EE, IOB and BG from different time points. Specifically, we will use the present values of EE, IOB and BG, as well as the values of EE and IOB every 20 minutes going back up to 1 hour into the past, and the values of BG every 20 minutes going forward up to 1 hour into the future. Our analysis will be performed jointly on EE and IOB and their lagged values, and on BG and its future values. This strategy allows us to evaluate the effect of EE and IOB on BG while capturing delayed effects up to 2 hours, lagged values, and interactions among lagged values, without specifying them explicitly. Moreover, this is accomplished nonparametrically, meaning that we never specify the relationship between EE and BG, or between IOB and BG. This is possible because our methodology makes use of DC, which can capture dependencies of any type between multivariate random vectors.

We denote by $BG_t(0)$, $BG_t(+20)$, $BG_t(+40)$ and $BG_t(+60)$, the value of BG at time t, and 20, 40, and 60 minutes into the future from t, respectively. We denote by $EE_t(0)$, $EE_t(-20)$, $EE_t(-40)$ and $EE_t(-60)$, the value of EE at time t, and 20, 40, and 60 minutes into the past from t, respectively. We ascribe the same meaning to $IOB_t(0)$, $IOB_t(-20)$, $IOB_t(-40)$ and $IOB_t(-60)$. We define three vectors as

$$\begin{aligned} BG_t &= \{BG_t(0), BG_t(+20), BG_t(+40), BG_t(+60)\}, \\ EE_t &= \{EE_t(0), EE_t(-20), EE_t(-40), EE_t(-60)\}, \\ IOB_t &= \{IOB_t(0), IOB_t(-20), IOB_t(-40), IOB_t(-60)\}. \end{aligned}$$

We are interested if any change points exist in the relationship between $\{EE_t, IOB_t\}$ and BG_t. The hypotheses that we will perform are as follow:

$$\begin{aligned} H_0 &: \mathscr{V}^2(\{EE_1, IOB_1\}, BG_1) = \cdots = \mathscr{V}^2(\{EE_T, IOB_T\}, BG_T), \\ H_A &: \cdots = \mathscr{V}^2(\{EE_\tau, IOB_\tau\}, BG_\tau) \neq \mathscr{V}^2(\{EE_{\tau+1}, IOB_{\tau+1}\}, BG_{\tau+1}) = \cdots, \end{aligned} \tag{4.9}$$

for some unknown change point τ. We also believe that there can be multiple change points, and we test for multiple points hierarchically. At a first stage, we test for the existence of any change point in the whole period of stay in the metabolic chamber. If the null is rejected, then at a second stage, we test if there exist a change point before and after the first change point discovered on the first stage. We continue until we are not able to reject the null anymore. This strategy will generate separate time intervals. Each interval will correspond to a section of time where the effect of EE and IOB on BG is different from the other time intervals. However, our goal is to hone in into the effect that EE has on BG, and how this effect changes.

Consequently, we will examine the relationship between EE and BG within each time interval discovered. For each time interval, a smoothing spline will be fitted with BG as outcome and EE as independent variable. This will be performed at different time gaps between EE and BG: concurrent values (no time gap), 1 hour time gap, and 2 hours time gap. If we see that the fitted smoothing spline is different within each time interval, this will help us visualize how the relationship between EE and BG changes. The result of this is shown in figures 4.2 to 4.4 in the next subsection. We will create a series of linear models within of the time intervals found at different time gaps where the outcome is BG and the variables are IOB and EE. Moreover, we expect IOB and EE to have different coefficients within each time intervals discovered by our change point methodology. If there is a difference in the coefficient of EE across time intervals, this will indicate that its effect on BG is in fact different.

Hereafter, we will evaluate if there exist changes in the distribution of EE or IOB across the time intervals found. Theory predicts that changes in IOB and intensity of EE can modify the relationship between BG and EE. We will evaluate changes in EE and IOB between time intervals using the Kolmogorov-Smirnov test. If changes exist, then it is possible that the change points occur because of this reason.

4.5.3 Testing for Change Points and Illustration of the Time Intervals

We start by testing if there exists a change point for the effect of $\{EE_t, IOB_t\}$ on EE_t. We applied the estimator (4.5) and detected a change point at 3:33 pm on day 2. This change point has a significant p-value as shown on table 4.7. We performed a test of existence of a change point in the time interval 8:33 pm on day 1 to 3:33 pm on day 2. We find a significant change point at 9:58 am on day 2. We did a further test between 8:33 pm on day 1 and 9:58 am on day 2. We estimate a change point at 4:13 am on day 2, but this change point is not statistically significant. Therefore, we found a total of 2 change points which provides us with 3 separate time intervals: from 8:33 pm on day 1 to 9:58 am on day 2, from 10:03 am to 3:33 pm on day 2 , and from 3:38 pm to 7:13 pm on day 2.

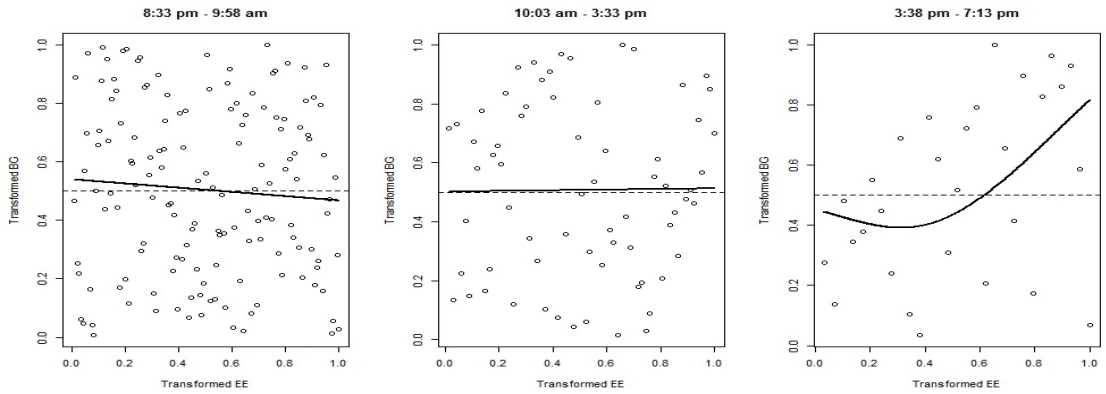
Table 4.7: Change Points

Stage	p-value	Sign.	Time($\hat{\tau}$)	Time Range
I	0.00001*	0.05	3:33 pm on day 2	8:33 pm on day 1 to 7:13 pm on day 2
II	0.006*	0.025	9:58 am on day 2	8:33 pm on day 1 to 3:33 pm on day 2
III	0.694	0.0125	4:13 am on day 2	8:33 pm on day 1 to 9:58 am on day 2

In figures 4.2 through 4.4, BG is represented on the y-axis and EE on the x-axis. Each figure has three plots corresponding to the three time intervals created by splitting the 23 hours of stay in the chamber where the change points are significant as given

in table 4.7. For each of the time intervals discovered, BG and EE are standardized to be discrete uniform. Thus within each interval, there are no differences in the marginal distributions of BG and EE. For each of the three time intervals, the results of a smoothing spline fit where the transformed BG and EE are used as dependent and independent variable are displayed. Figures 4.2 to 4.4correspond to different time gaps between BG and EE, concurrent values, 1 hour time gap and 2 hours time gap, respectively.

Figure 4.2: Concurrent BG and EE

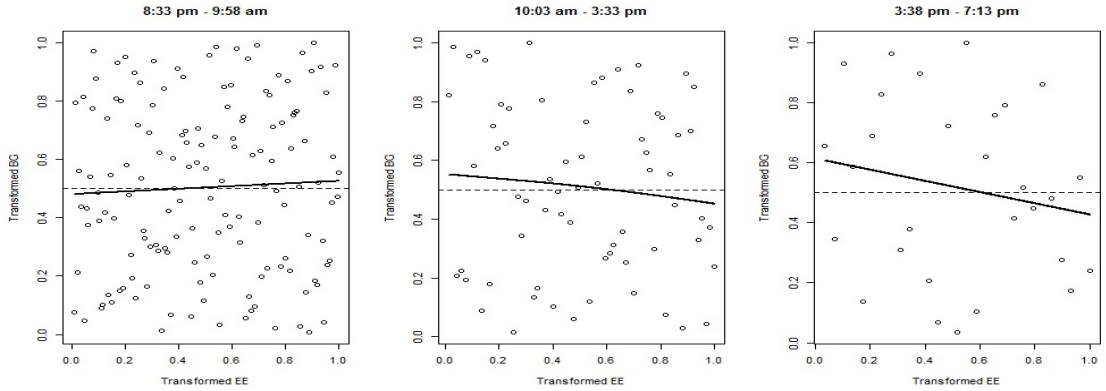


Each plot corresponds to a time interval found by the change point estimator. The lines are smoothing splines fit to the data which has been standardized to be uniform within each interval.

In figure 4.2, corresponding to concurrent values of BG and EE, we see that in the first time interval the relationship between BG and EE seems to be linear and negative, but very weak. However, on the next time interval the relationship almost disappears. On the last time interval, the relationship becomes overall positive, quadratic and strong. In figure 4.3, corresponding to the 1 hour gap between BG and EE, in the first time interval the relationship is linear, positive, but almost nonexistent. In the second time interval the relationship becomes negative and stronger compared to the first time interval. On the last time interval, the relationship remains negative but stronger compared to the two previous time intervals. Figure 4.4 corresponds to a time gap of 2 hours. In the first time interval, the relationship is overall negative, sinusoidal

and very strong. In the next two intervals the relationship becomes linear, but remains negative, with the relationship being stronger on the last time interval. From figure 4.2 to 4.4, we can derive some interesting observations. The relationship between BG and EE seem to be strongest in the last time interval across the three time gaps shown. Moreover, the relationship seems to be overall linear, even though there are a couple of instances, that seems to be quadratic and even sinusoidal. Moreover, it seems that the relationship is strongest between BG and EE at 2 hour gap, even though it seems also to be strong in the third time interval for concurrent values.

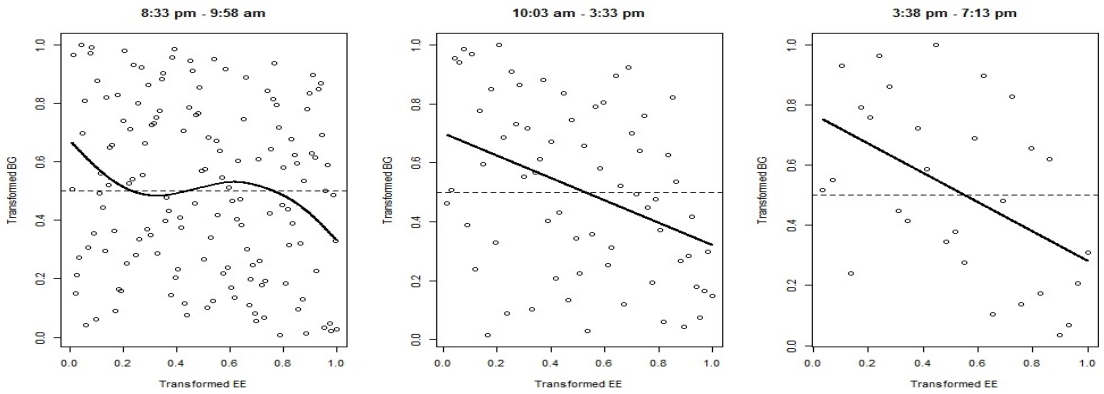
Figure 4.3: 1 Hour Gap Between BG and EE



Each plot corresponds to a time interval found by the change point estimator. The lines are smoothing splines fit to the data which has been standardized to be uniform within each interval.

Nevertheless, given that the use of IOB also affects BG and is correlated with EE, it is hard to arrive at a conclusion from these figures. Consequently, there is a need to adjust for IOB. We created three linear models within each of the three time intervals created by the change points estimated in table 4.7 with $BG_t(60)$ as outcome and as independent values all the lagged values of IOB and EE. We choose the farthest value in the future of BG because this seems to be when the relationship was strongest with EE by inspection of the smoothing spline plots in figures 4.2 to 4.4. The models are

Figure 4.4: 2 Hours Gap Between BG and EE



Each plot corresponds to a time interval found by the change point estimator. The lines are smoothing splines fit to the data which has been standardized to be uniform within each interval.

defined as

$$BG_t(60) = \beta_{0,i} + \beta_{1,i}EE_t(0) + \beta_{2,i}EE_t(-20) + \beta_{3,i}EE_t(-40) + \beta_{4,i}EE_t(-60) + \beta_{5,i}IOB_t(0) + \beta_{6,i}IOB_t(-20) + \beta_{7,i}IOB_t(-40) + \beta_{8,i}IOB_t(-60) + \varepsilon_t,$$

with $t \in J(i)$, $i = 1, 2, 3$, and each $J(i)$ represents the three time interval found by our change point estimator. The results are shown on table (4.8).

It can be seen that the coefficients related to EE_t change from interval to interval, even after adjusting for IOB_t. In terms of absolute value of their coefficients, all four lagged values of EE seem to be important. However, this importance changes from interval to interval. In figure 4.5, we displayed the coefficients of EE_t over the three intervals, the y-axis denotes the value of the coefficient, and each line corresponds to a sequence of coefficients associated with one of the lagged values of EE as it changes from time interval to the next. Larger boxes indicate more significant (smaller p-values) results. From Figure 4.5 some interesting observations can be made. Moving from the first time interval to the second one, half the coefficients decreased in value, while the other half increased in value. However, the coefficients that decreased in value were not

very significant and were close to 0 at the first time interval. More interesting, going from the second time interval to the third, all the coefficients associated with EE_t decreased in value. These results are interesting given that in the first time interval, not much activity was happening, in the second time interval activity rises and the coefficients change, and on the third time interval the coefficients drop once again after activity has receded.

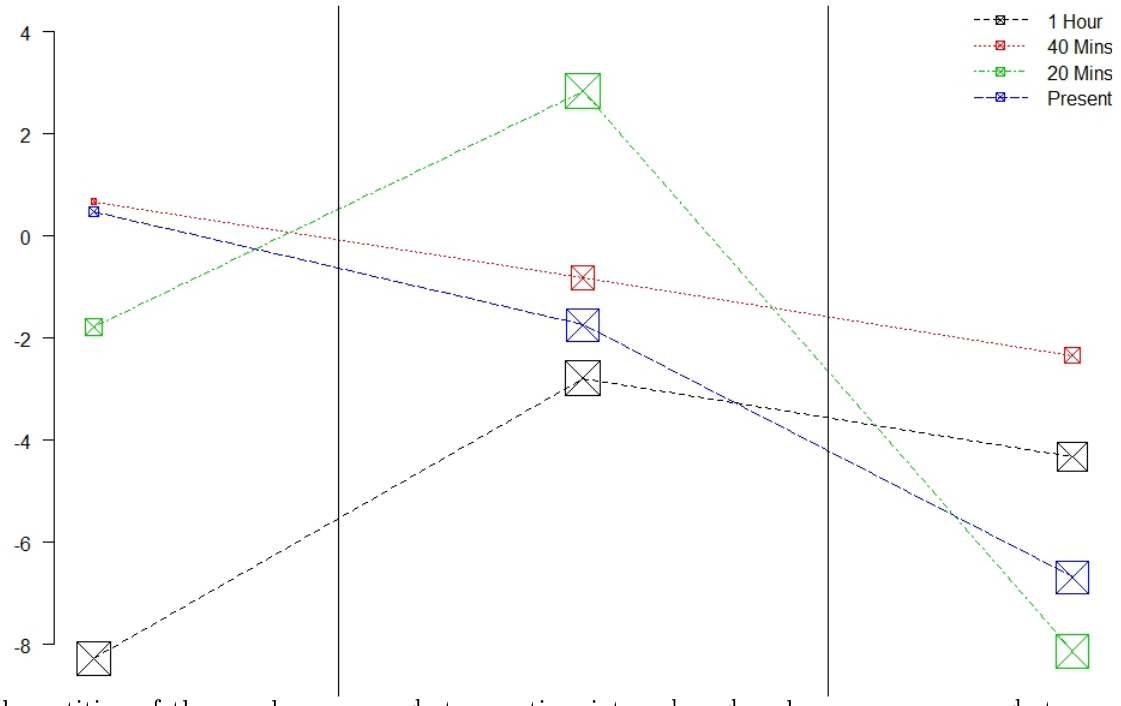
Table 4.8: Linear Model with $BG_t(60)$ as Outcome

	First Interval		Second Interval		Third Interval	
Variables	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	0.162	0.598	-0.018	0.977	-0.054	0.909
$EE_t(-60)$	-8.271	0.021*	-2.7827	0.001*	-4.321	0.157
$EE_t(-40)$	0.677	0.847	-0.8149	0.334	-2.336	0.538
$EE_t(-20)$	-1.774	0.513	2.8372	0.001*	-8.140	0.053
$EE_t(0)$	0.470	0.721	-1.739	0.056	-6.689	0.042*
$IOB_t(-60)$	-0.630	0.646	0.073	0.880	-0.821	0.079
$IOB_t(-40)$	0.734	0.600	-0.693	0.157	-0.827	0.098
$IOB_t(-20)$	-2.462	0.074	0.031	0.948	-0.354	0.444
$IOB_t(0)$	0.567	0.669	-0.264	0.553	0.671	0.368
P-values significant at the 0.05 level are denoted by an asterisk.						

4.5.4 Distribution of Insulin and EE

It is believed that the relationship between BG and EE is affected by the intensity of EE and the level of IOB in a T1D subject. While estimating change points in the effect of EE and IOB on BG, we have created three separated time intervals. We previously hypothesized that these change points discovered could potentially be due to changes in the intensity of EE and IOB, which would translate in changes in their distribution. Figure 4.6 shows the histograms of EE (left column) and IOB (right column) for the three time intervals (rows). There is a general pattern where the

Figure 4.5: Coefficients from the Linear Model



Each partition of the graph corresponds to one time interval, and each square corresponds to a coefficient shown on table 4.8. A line connecting several squares symbolizes the same coefficient as it changes from time interval to the next. Coefficients with smaller p-values are represented with larger symbols. The coefficients represented are those of the present value EE and its lagged values every 20 minutes going back an hour.

distribution of both EE and insulin in the second time interval (second row) have greater variance than the first or third time intervals (first and third rows). Hence, it seems that at the second time interval there was a surge of EE and IOB.

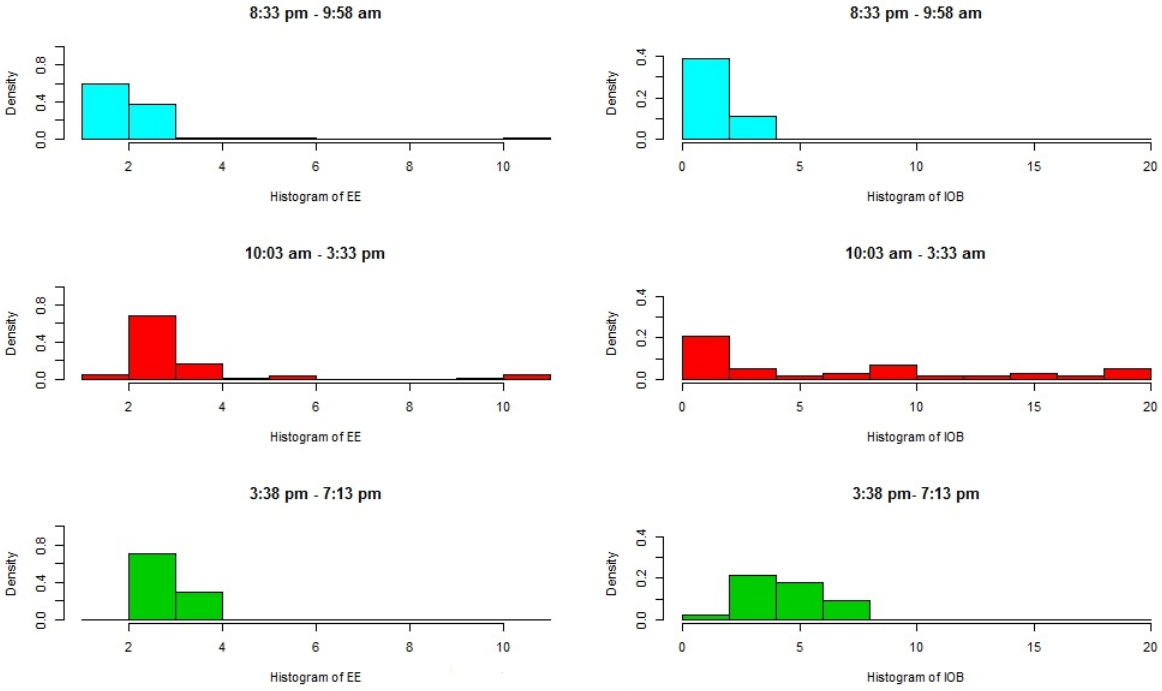
We performed a series of Kolmogorov-Smirnov tests between the distributions of first and second time interval, and the second and third time intervals of both EE and IOB. A significant Kolmogorv-Smirnov test would indicate that there exists a change in distribution between these time intervals. The results are shown in Table 4.9. We see that for both EE and insulin, their distribution had a significant change between the first and second time interval at 3:33 pm on the day 2. Moreover, the distributions in the second and third time interval were also statistically different, which can be seen

on the histograms where they appear quite different.

Table 4.9: Kolmogorov-Smirnov Tests

Test	p-value	Test	p-value
EEI vs EEII	0.0001*	IOBI vs IOBII	0.0001*
EEII vs EEIII	0.0001*	IOBII vs IOBIII	0.0001*

Figure 4.6: Distribution of EE and IOB by Interval



Each color represents a different interval of time. The intervals were created by separating time when a change point was statistically significant.

4.6 Conclusion and Discussion

We have developed a nonparametric change point methodology for the detection of the change in the relationship between two multivariate vectors and applied it to the analysis of metabolic chamber measurements on a T1D individual to gain insight into the dynamics between EE and BG, while adjusting for IOB. The results show

that changes in the effect EE has on BG do happen throughout the metabolic chamber time series. These changes happen whenever the intensity of EE along with IOB use increase, which happens at the same time in our data set. This supports the literature that says the relationship between PA and BG is variable for a T1D individual.

In addition to this, we tried to reproduce our findings with another set of data on the metabolic chamber, but with a shorter stay of only 8 hours. Our change point test did not find any significant change points. Nevertheless, we inspected the distributions of both IOB and EE and we find similar patterns where there is an increase in both IOB and EE after the change point estimated. Also, the distributions were significantly different with respect to the Kolmogorov-Smirnov test. Even though it was not significant, using the first change point estimated shows the same pattern found in the analysis of the longer metabolic chamber stay.

Our methodology is nonparametric in the sense that we did not have to specify explicitly the effect of EE and IOB on BG. In addition to this, we included a variety of lagged values of IOB and EE that could potentially affect several values of BG into the future. This allowed us to estimate change points without specifying the time gap at which BG is affected most by EE, given that this is hard to know a priori. We set up our estimation problem such that delayed effects of EE on BG of up to 2 hours can be detected. After detecting two change points in the metabolic chamber time series, creating three separated intervals, we perform an exploratory analysis where we look at the relationship between EE and BG within each of the three time intervals, and at different time gaps. It seems that the relationship between EE and BG is strongest whenever the time gap increases. We decided to inspect this further by creating a linear model, within each time interval, with future BG as outcome, and several lagged values EE and IOB as outcomes. We see that in fact, the coefficients associated with the lagged values of EE are quite different from interval to interval.

These results can be useful in the development of wearable technology and an artificial pancreas. In such settings, it can be hard to model PA and BG dynamics, but are incredibly important to the success of such devices. From the results of the current article we can arrive at several recommendations for such devices. Lagged values of PA are important, rigorously estimating the number needed would be beneficial. More importantly, it seems that the increase in PA is important in modifying the relationship of PA and BG. Thus, it is important to include an interaction between an indicator variable for PA past a certain threshold and the lagged values of PA. This would capture the changes in relationship between PA and BG. A similar approach was used in Colmegna et al. (2016) where their model incorporates many thresholds.

However, we want to emphasize that our method is very general and can be applied to the analysis of other data sets where there is a need to find a change point but it is difficult to model the relationship between the data explicitly. Hence, this methodology could be easily extended to genomics or environmental data sets where change point problems tend to arise naturally.

CHAPTER 5: NONPARAMETRIC CLUSTERING OF VARIABLES

5.1 Summary

In cancer research, it is important to correctly classify different types of cancer cells. A common practice for classifier creation, is to aggregate variables in clusters and summarize them using principal components. One popular way to create clusters of variables is to do hierarchical clustering and select a predefined number of subgroups. A notable drawback of existing hierarchical clustering of variables methods is that they do not control the type I error rate which can lead to falsely joining variables that are otherwise uncorrelated. More importantly, current methods have a bias towards creating large groups variables, whereas it might be possible that the true hierarchy has numerous clusters containing a small number of variables. We propose a statistical approach that can cluster variables while preserving a predefined family wise error rate. We accomplish this by turning the decision of whether joining clusters into a hypothesis testing problem. We use a generalized version of correlation to be able to test if two clusters are statistically independent or not. We demonstrate that the error rate is preserved through simulations. The strength of our method is shown by clustering gene expressions from single cell data coming from five primary glioblastoma tumors. In particular, our method confirms the variability in gene expression in different tumors, and principal components derived from our clusters classify single cells to their corresponding tumor with good accuracy.

5.2 Introduction

Hierarchical clustering is an extremely popular tool for detecting structure of data in both samples and variables. This method has been used extensively in the field of genomics, for classifying samples in subgroups but also detecting clusters of genes. In this setting, hierarchical clustering has been used to detect meaningful subgroups of genes within a cancer type that are associated with survival outcomes (Bhattacharjee et al. (2001), Sørlie et al. (2001), Shen et al. (2007)). Hierarchical clustering algorithms work in an agglomerative fashion meaning that, in the case of gene expressions, they join individual genes one by one until all genes belong to the hierarchy. Unfortunately, this type of analysis does not allow for genes to be uncorrelated with all other genes in the data set. Also, hierarchical clustering analysis is biased towards aggregating genes in very large groups, whereas the true structure could be small to medium size gene clusters. This is particularly troublesome given that the number of genes that are truly associated with survival may be in fact small. Another drawback is the measure used to create the clusters of genes, for example average linkage and others. These measures do not estimate a theoretical parameter, but are an *ad hoc* solution to the problem of creating distance measures between groups of variables. Thus, these measure do not really capture nonlinear dependency or interactions that could happen between groups of variables.

Traditional strategies for selecting subgroups or clusters of variables while using hierarchical clustering are to define before hand a number of groups to be selected. This requires providing the algorithm with knowledge not available in advance which is usually chosen based on convenience, which can create potential biases. Another approach is to use the elbow rule which consists of drawing average linkage as a function of the numbers of clusters and stop whenever an elbow shape is observed. Again, this

has no theoretical guarantees and is biased towards large clusters of variables. Another approach is to find a cut-off to determine where the hierarchy should stop, but knowledge of the appropriate cut-off is required too (Langfelder and Horvath (2008)). The bootstrap has been used to guide the decision of which clusters of variables to keep. It can be used to to establish the reproducibility of the clusters by fixing the cluster centers and to report for each gene the cluster-specific proportion of times it falls in that cluster out of many samples (Van Der Laan and Bryan (2001), Van der Laan and Pollard (2003)). A simpler approach is to create multiple replicates of the dendrogram by repeatedly applying the cluster analysis to the bootstrap samples and a probability value of a cluster is created as the frequency that it appears in the bootstrap replicates (Suzuki and Shimodaira (2006)). Options other than the bootstrap are possible. Another method improves the detection of outlying members of each cluster by identifying preliminary clusters as branches that satisfy a minimum number of variables, and variables too far from a cluster are excluded. Then, all previously unassigned genes are assigned to the nearest cluster (Langfelder et al. (2008)). Unfortunately, none of these methods deal with the problems discussed in the previous paragraph.

The current article aims to improve previous methodology on hierarchical clustering of variables. Our method deals with some of the issues with the current methodology using several strategies. First, the output of our method is not required to be a complete dendrogram. By complete dendrogram it is meant that all clusters join at the top of the tree to form one big cluster. Secondly, our method does not use heuristics to select cluster of variables, but instead relies on hypothesis testing in order to discover groups of coexpressed genes, while controlling for the type I error rate. Thirdly, our method does not have a bias towards large groups because it has control over the errors. Lastly, our method uses distance covariance which evaluates the statistical dependence

between clusters of groups, instead of relying on ad-hoc measures from the linkage family. We call our method the Nonparametric Hierarchical Clustering (NCH) algorithm. Our method then tackles the problems discussed previously by

- turning the decision of whether to create a cluster into a hypothesis testing problem,
- letting groups of variables come together into groups by discarding clusters that are not significant, and
- replacing average linkage by distance covariance which is a generalized version of correlation.

Moreover, this whole procedure will be performed by controlling the FWER. This will be accomplished by using the minP procedure with a modification on the permutation approach.

The rest of this article is organized as follow. In section 5.3, we describe our algorithm. In section 5.4, we present our simulation studies that evaluate if the NHC preserves the type I error rate. In section 5.5, we apply our method to a RNA-seq single cell data set. We end with a discussion in section 5.6.

5.3 Nonparametric Hierarchical Clustering Algorithm

5.3.1 Distance Covariance

Here we present the the distance covariance (DC) statistic and test. DC was developed by Székely et al. (2007), Székely et al. (2009), and Székely and Rizzo (2013). For random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, let ϕ_x, ϕ_y and $\phi_{x,y}$ be the characteristic function of X, Y and (X, Y), respectively. Assume that $E|X|_p < \infty$ and $E|Y|_q < \infty$.

Distance covariance ($\mathcal{V}^2$) can be used to measure the dependence between X and Y through the distance

$$\begin{aligned}\mathcal{V}^2(X,Y) =& ||\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)||^2 \\ =& \int_{\mathbb{R}^{p+q}} |\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)|^2 (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1} dtds\end{aligned}$$

with $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ and $|\cdot|_p$ is the Euclidean norm in $\mathbb{R}^p$. If $X \not\perp Y$ then $\mathcal{V}^2(X,Y)$ will be greater than 0. otherwise if $X \perp Y$ then it will be exactly 0. The DC statistic are defined as follows. For an observed random sample $\{(X_k, Y_k) : k = 1, ..., n\}$ from the joint distribution of random vectors $(X,Y) \in \mathbb{R}^p \times \mathbb{R}^q$, define

$$\begin{aligned}a_{kl} =& |X_k - X_l|_p, \qquad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \qquad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \\ \bar{a}_{\cdot\cdot} =& \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}\end{aligned}$$

for $k, l = 1, ..., n$. Similarly, define $b_{kl} = |Y_k - Y_l|_q$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ for $k, l = 1, ..., n$. The empirical distance covariance $\mathcal{V}_n^2(X,Y)$ is the nonnegative numbers defined by

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl},$$

which is a.s. consistent for $\mathcal{V}^2(X,Y)$. To test the hypotheses of independence between X and Y defined as

$$H_0 : \phi_x\phi_y = \phi_{x,y} \quad \text{and} \quad H_A : \phi_x\phi_y \neq \phi_{x,y}$$

the test statistic $n\mathcal{V}_n^2(X,Y)$ can be used with a null distribution calculated by permutation.

5.3.2 Sketch of the Algorithm

Here we present a brief description of the algorithm before giving extensive details on each of its components. The input of the algorithm will be a sample of size n of p random variables, $X_1^n, ..., X_p^n$. Common hierarchical agglomerative algorithms join clusters one by one. Even though NHC is also an agglomerative algorithm, it does not join clusters one by one, but instead proceeds by levels, and the collection of levels will be called a *tree*. Each level corresponds to a collection of clusters. At the bottom of the *tree*, level 0, each of the p variables are considered to be in p individual clusters. Next, at level 1, individual clusters start coming together into bigger clusters. Many clusters can come together at the same time to form a bigger cluster. This is the main difference from current methods, that instead of clusters joining one by one, here, in a given level, many clusters can come together as one. Individual clusters join into bigger clusters whenever they are statistically correlated with other members already present in the cluster, given a specified significance level. Next, at level 2, and in all subsequent levels, the same happens, any clusters formed in the previous level that are statistically correlated with other clusters are joined together to form a bigger cluster. This continues to happen until there are no more significant correlations or the algorithm has reached a maximum number of *levels* specified by the user. A sketch of the procedure is shown below.

Sketch of NHC: The input of the algorithm will be $\mathscr{X}_n = \{X_1^n, ..., X_p^n\}$, the maximum number of levels L, and the family wise error rate allowed at each level α_l, for $l = 1, ..., L$, satisfying $\alpha_l > 0$ and $\sum_{l=1}^{L} \alpha_l = \alpha$. Start with level 0 which includes p clusters each with one variable. The iteration and level index are denoted both by i. Also, DC here will be used for distance covariance test statistic defined in section 5.3.1 by $n\mathscr{V}_n^2(\mathbf{X}, \mathbf{Y})$. Moreover, $DC(X_i^n, X_j^n) \equiv n\mathscr{V}_n^2(X_i^n, X_j^n)$. DC or the plural DCs will be also used to denote $DC(X_i^n, X_j^n)$.

Initialize $i = 0$.

Step 1.- Increase the level index i by 1.

Step 2.- Calculate all pairwise DCs among all the clusters in level $i - 1$.

Step 3.- All pairs of clusters which have a DC that is statistically significant will be joined together into one cluster. This cluster now belongs to the collection of clusters at level i.

Step 4.- All other clusters will remain intact and be included as they were in level $i - 1$ in level i.

Step 5.- If $i = L$ stop the procedure. Otherwise, if no new clusters were created in the current iteration, in other words, if the ith level is exactly the same as level $i - 1$, stop the procedure and all the levels up to $i - 1$ will form the hierarchical tree. Otherwise go back to the **Step 1**.

At **Step 3**, p-values for corresponding DCs will be adjusted in such a way that the FWER can be preserved at the overall α level. Thus, it can be seen that variables are only joined together in clusters if they happen to be statistically dependent with each other, after adjusting for the error rate. The clusters generated by the algorithm account for the uncertainty of the random data and contain variables that are related to each other after a conservative error adjustment which is FWER. Moreover, since our algorithm proceeds by levels, we can easily visualize which variables join together in clusters at which level. At the first level we can visualize the variables that were statistically dependent to other variables. At later levels, we will start to see which groups of variables were dependent on other groups of variables. This will allow us to see associations between individual variables as as well as among group of variables.

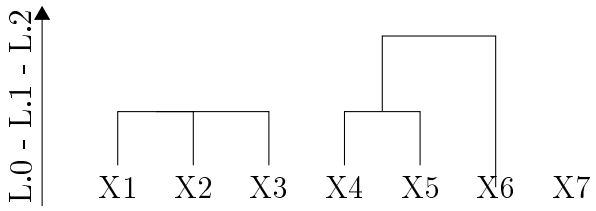
5.3.3 Formal Definition of NHC

Here we present a detailed description of our method. The output of the NHC algorithm is a tree structure with a hierarchy. The tree structure will be represented by a set denoted by $\mathscr{T}$. Each element of $\mathscr{T}$ is, in its turn, a collection of sets. Each element of $\mathscr{T}$ represents a *level* of the tree hierarchy which will be denoted by L_i, with $i = 0, ...K$ representing the levels. The ith level set, L_i, is a collection of sets, each representing a cluster, i.e., $L_i = \{C_i^1, C_i^2, ..., C_i^{K_i}\}$, K_i is the number of clusters in this level and C_i^k represents the kth cluster for the ith level. The lowest level of the tree, L_0, corresponds to the base of the tree, where there are p clusters with each including only one variable, i.e., $L_0 = \{C_0^1, C_0^2, ..., C_0^p\}$ and $C_0^k = \{X_k\}$. Clusters in higher levels are always either exactly equal to some cluster in the previous level or a union of clusters in the previous level. Hence, for level i and cluster $C_i^k \in L_i$, we have $C_i^k = \cup_{l \in \Delta} C_{i-1}^l$, for $C_{i-1}^l \in L_{i-1}$ for all $l \in \Delta$ and some Δ.

For any given level i of the tree, the union of all its member clusters makes the full set of variables, i.e, $\cup_{k=1}^{K_i} C_i^k = \{X\} = \{X_1, ..., X_p\}$, and all clusters are pairwise disjoint, i.e, $C_i^k \cap C_i^l = \{\varnothing\}$ for all $k \neq l$. Also, for any two levels i and i' such that $i < i'$, the cardinality of the higher level is smaller than the lower one, i.e. $|L_i| \geq |L_{i'}|$.

An example will help illustrate how this works. Let's say there is a sample of size n for 7 variables $X_1^n, ..., X_7^n$; a possible output of the NHC is described as follow. The bottom of the tree corresponds to $L_0 = \{C_0^1, C_0^2, ..., C_0^7\}$ where $C_0^k = \{X_k\}$ for $k = 1, ..., 7$. At the second level we could have $L_1 = \{C_1^1, C_1^2, C_1^3, C_1^4\}$, with $C_1^1 = \{X_1, X_2, X_3\}$, $C_1^2 = \{X_4, X_5\}$, $C_1^3 = \{X_6\}$ and $C_1^4 = \{X_7\}$. So we see that clusters at this level are unions of clusters in the previous level, namely $C_1^1 = C_0^1 \cup C_0^2 \cup C_0^3$, $C_1^2 = C_0^4 \cup C_0^5$, $C_1^3 = C_0^6$ and $C_1^4 = C_0^7$. At the third level we could have $L_2 = \{C_2^1, C_2^2, C_2^3\}$, with $C_2^1 = \{X_1, X_2, X_3\}$, $C_2^2 = \{X_4, X_5, X_6\}$ and $C_2^3 = \{X_7\}$. Again, the clusters at this level

are unions of clusters from the previous level, namely $C_2^1 = C_1^1$, $C_2^1 = C_2^2 \cup C_1^3$ and $C_2^3 = C_1^4$. After this level, no further clusters are selected, hence the final tree hierarchy corresponds to $\mathscr{T} = \{L_0, L_1, L_2\}$. The image depicting this tree is shown below:



An investigator can readily see from the display of the tree hierarchy that there is a natural arrangement of variables into groups and subgroups. They are three major groups $C_2^1 = \{X_1, X_2, X_3\}$, $C_2^2\{X_4, X_5, X_6\}$ and $C_2^3 = \{X_7\}$. The second group has two subgroups, namely $C_1^2 = \{X_4, X_5\}$ and $C_1^3 = \{X_6\}$. Hence, the investigator does not have to choose groups and subgroups based on some heuristic but they are available explicitly from the output of the NHC algorithm. Moreover, variables that do not belong to a group do not a enter cluster, i.e, X_7. This procedure can be accomplished by retaining a FWER less than a specified α.

5.3.4 Algorithm for NHC

The algorithm will go through several iterations, each indexed by i. Each iteration corresponds to a level of the tree hierarchy. At the initial level $i = 0$, $\mathscr{T} = \{\{L_0\}\}$, where $L_0 = \{C_0^1, C_0^2, ..., C_0^p\}$ and $C_0^k = \{X_k\}$ for each k. At the ith iteration, L_i is the ith level of the tree, with each element of L_i representing a cluster of variables at that level. Denote the number of clusters in the ith level by K_i and $K_i^{(2)} \equiv \binom{K_i}{2}$. Denoted by nT_i an ordered vector of DCs, the cardinality of nT_i is $K_i^{(2)}$ and its elements correspond to all the pairwise DCs between clusters in the previous level L_{i-1}. Denote by P_i^* the

ordered vector of raw p-values corresponding to each DC in nT_i. Denote by AdjP_i a vector of adjusted p-values corresponding to the adjusted versions of P_i^*.

Inputs: the number of permutations B, the maximum number of levels L, the error rates α_l such that $\sum_{l=1}^{L} \alpha_l = \alpha$, and the matrix $\mathscr{X}_n$.

Initialize $i = 1$.

Step 1.- Calculate all the pairwise DCs between all clusters in level L_{i-1}. This will make the vector nT_i of cardinality $K_{i-1}^{(2)}$. Also, calculate their corresponding raw p-values P_i^*.

Step 2.- For the current set of DCs in nT_i and corresponding set of raw p-values P_i^* each with cardinality $K_{i-1}^{(2)}$, calculate B permutations of all such $K_{i-1}^{(2)}$ DCs. This gives a matrix DC_i^{π} of dimension $K_{i-1}^{(2)} \times B$. Use the permutation procedure described in the section **Matrix of Permuted Statistics** DC_i^{π}.

Step 3.- Use the minP step-down algorithm to derive adjusted p-values for all $K_{i-1}^{(2)}$ test statistics from DC_i^{π}, nT_i and P_i^*. This makes a vector AdjP_i of length $K_{i-1}^{(2)}$ of adjusted p-values. The permutation procedure is described in **Step-down minP Adjusted p-values Algorithm**.

Step 4.- Each element $nT_i(j) \in nT_i$ corresponds to a DC between a pair of clusters in L_{i-1}, say $C_{j_1}^{i-1}$ and $C_{j_2}^{i-1}$. If adj$P_i(j) \leq \alpha_i$, then add the union to level L_i, so now $C_j^i \equiv C_{j_1}^{i-1} \cup C_{j_2}^{i-1} \in L_i$. If there exist a j' such that $C_{j_1}^{i-1} \cap C_{j'}^i \neq \varnothing$ or $C_{j_2}^{i-1} \cap C_{j'}^i \neq \varnothing$ and adj$P_i(j) \leq \alpha_i$, then first add $C_{j''}^i \equiv C_{j \cup j'}^i = C_{j_1}^{i-1} \cup C_{j_2}^{i-1} \cup C_{j'}^i$ to L_i and then remove $C_{j'}^i$ from L_i.

Step 5.- If no new clusters were created in the previous step, meaning $L_i = L_{i-1}$, then the last level of the tree is L_{i-1}, $\mathscr{T} = \{\{L_0\}, \{L_1\}, ..., \{L_{i-1}\}\}$. If $i = L$ then stop the procedure and $\mathscr{T} := \mathscr{T} \cup \{L_i\} = \{\{L_0\}, \{L_1\}, ..., \{L_{i-1}\}, \{L_i\}\}$. Otherwise,

$\mathscr{T} := \mathscr{T} \cup \{L_i\} = \{\{L_0\}, \{L_1\}, ..., \{L_{i-1}\}, \{L_i\}\}$, $i = i + 1$ and return to **Step 1**.

Remark: The second phrase in Step 4 says that, at the ith iteration, a cluster C_j^i currently existing in L_i will include new terms, or become larger, if there is a significant correlation between a cluster from the previous level L_{i-1} and any subset of C_j^i. This subset would have to correspond to some cluster in L_{i-1}.

5.3.5 Matrix of Permuted Statistics DC_i^π

In **Step 2** of the algorithm permutation was used. Typically done in permutation procedures for multiplicity adjustment, each test is permuted independently of one another. Our permutation procedure proceeds differently. Each DC evaluates the strength of the relationship of between two groups of variables. For each permutation iteration b, our procedure permutes one group of variables out of the two used in each of the DC tests. This preserves most of the correlation across tests, because one group of variables remains unpermuted and the other group is permuted similarly to all other tests. Below is a more detailed explanation of how the permutation proceeds.

For each permutation iteration indexed by b, there is a corresponding permutation denoted by π_b. Let $C_{n,j}^{i-1}$ denote the jth sample cluster of level L_{i-1}. The subscript n denotes that we are talking about a certain subset of the columns of the data matrix $\mathscr{X}_n$, i.e., $C_{n,j}^{i-1} = \{X_k^n : \text{for some } k \in \Delta\}$. Moreover, $\pi_b(C_{n,j}^{i-1})$ denote the permuted version of $C_{n,j}^{i-1}$ by π_b. Below is a matrix with each entry representing a distance covariance test. At the jth row of the matrix, all DC test statistics of the form $DC(C_{n,j}^{i-1}, C_{n,k}^{i-1})$ with

$j < k$ are represented. This matrix representation includes all the $K_{i-1}^{(2)}$ combinations.

$$\begin{pmatrix} DC(C_{n,1}^{i-1}, C_2^{i-1}) & DC(C_{n,1}^{i-1}, C_3^{i-1}) & DC(C_{n,1}^{i-1}, C_4^{i-1}) & \cdots & DC(C_{n,1}^{i-1}, C_{n,K_{i-1}}^{i-1}) \\ & DC(C_{n,2}^{i-1}, C_{n,3}^{i-1}) & DC(C_{n,2}^{i-1}, C_{n,4}^{i-1}) & \cdots & DC(C_{n,2}^{i-1}, C_{n,K_{i-1}}^{i-1}) \\ & & DC(C_{n,3}^{i-1}, C_{n,4}^{i-1}) & \cdots & DC(C_{n,3}^{i-1}, C_{n,K_{i-1}}^{i-1}) \\ & & & \ddots & \vdots \\ & & & & DC(C_{K_{i-1}-1}^{i-1}, C_{K_{i-1}}^{i-1}) \end{pmatrix}$$

For each bth permutation, a new set of $K_{i-1}^{(2)}$ statistics will be calculated from only one permutation π_b. For the jth row in the matrix above the cluster corresponding to that row, i.e, $C_{n,j}^{i-1}$, will be permuted by π_b for each entry of the jth row. The clusters corresponding to the columns remain unchanged. Hence, for each row j and all pairs $j < k$ we calculate the permuted statistic as $DC(\pi_b(C_{n,j}^{i-1}), C_{n,k}^{i-1})$. The same permutation π_b is used for all j rows. This situation is depicted in the matrix below

$$\begin{pmatrix} DC(\pi_b(C_{n,1}^{i-1}), C_2^{i-1}) & DC(\pi_b(C_{n,1}^{i-1}), C_3^{i-1}) & \cdots & DC(\pi_b(C_{n,1}^{i-1}), C_{n,K_{i-1}}^{i-1}) \\ & DC(\pi_b(C_{n,2}^{i-1}), C_{n,3}^{i-1}) & \cdots & DC(\pi_b(C_{n,2}^{i-1}), C_{n,K_{i-1}}^{i-1}) \\ & & \cdots & DC(\pi_b(C_{n,3}^{i-1}), C_{n,K_{i-1}}^{i-1}) \\ & & \ddots & \vdots \\ & & & DC(\pi_b(C_{n,K_{i-1}-1}^{i-1}), C_{n,K_{i-1}}^{i-1}) \end{pmatrix}.$$

These $K_{i-1}^{(2)}$ permuted statistics make one set of permutations. They will be a total of B sets. These permutations will constitute the entries in the matrix DC_i^{π} of dimension $K_{i-1}^{(2)} \times B$. To each column corresponds only one permutation π_b.

5.3.6 Step-down minP Adjusted p-values Algorithm

In **Step 3** of the algorithm, adjusted p-values are derived from the raw p-values and the permutation matrix. The raw p-values are denoted by p_j^*, and assume without loss of generality that $p_1^* \leq p_2^* \leq \cdots \leq p_{K_{i-1}^{(2)}}^*$. Otherwise, rearrange the order of the $K_{i-1}^{(2)}$ statistics such that it matches the order of the raw p-values. Next, three matrices will be needed to be defined. First, a matrix of permuted distance covariance test statistics

$$DC_i^\pi = \left(DC_i^\pi(j,b) \right),$$

a matrix of raw p-values

$$P = \left(p_{j,b} \right),$$

and a matrix of minima of raw p-values

$$Q = \left(q_{j,b} \right),$$

where $q_{j,b} = \min_{l=j,\ldots,q_i} p_{l,b}$ and the bth column of these matrices corresponds to each permutation π_b. With these definitions the minP algorithm is as follow:

0.- Compute raw p-values for each hypothesis. Assume $p_1^* \leq p_2^* \leq \cdots \leq p_{K_{i-1}^{(2)}}^*$ without loss of generality. Otherwise sort the $K_{i-1}^{(2)}$ test statistics and corresponding raw p-values according to the ordered p_j^*.

Initialize $q_{K_{i-1}^{(2)}+1,b} = 1$ for $b = 1, .., B$.

Initialize $j = K_{i-1}^{(2)}$.

1.- For the jth test statistic $nT_i(j)$ use the quick sort algorithm to get the B raw p-values $p_{j,1}, ..., p_{j,B}$ from the matrix of permuted statistics DC_i^π.

2.- Update the successive minima $q_{j,b}$

$$q_{j,b} = \min(q_{j+1,b}, p_{j,b}), \qquad b = 1, ..., B.$$

3.- Compute the adjusted p-values for jth test statistic $nT_i(j)$

$$\tilde{p}_j^* = \frac{\#\{b : q_{j,b} \leq p_j^*\}}{B}.$$

4.- Do $j \leftarrow j - 1$. If $j = 0$, go to step 5, otherwise go to step 1.

5.- Enforce monotonicity of p_j^*:

$$\tilde{p}_1^* \leftarrow \tilde{p}_1^*, \qquad \tilde{p}_j^* \leftarrow \max(\tilde{p}_{j-1}^*, \tilde{p}_j^*) \qquad \text{for } j = 2, ..., K_{i-1}^{(2)}.$$

6.- Set $AdjP_i = \{p_1^*, p_2^*, \cdots, p_{K_{i-1}^{(2)}}^*\}$.

Remember that the order of the $K_{i-1}^{(2)}$ test statistics has been made to match the order of the raw p-values.

5.4 Simulation Results

Simulation studies were performed to evaluate if our permutation strategy actually preserves the FWER. The simulation were performed in the following two scenarios with varying sample size. A matrix $X_{n\times p}$ was generated as n sample vectors of size p from a multivariate normal distribution, with mean vector $\mathbf{0}_p$, and covariance matrix $\Sigma_{p\times p}$. The covariance matrix is block diagonal, with each block corresponding to a subset of correlated variables out of the p variables. The block diagonal nature of the covariance

matrix creates clusters of variables corresponding to each block. The first scenario is for the case $p = 20$, and $\Sigma_{p\times p}$ is made from 2 blocks each of size 10×10. In this scenario there exist 2 groups of clusters of variables. The second scenario, is for the case where $p = 40$, and $\Sigma_{p\times p}$ is made from 3 blocks, one of size 20×20, and two of size 10×10. Thus, this scenario incorporates three clusters of variables with different cluster sizes. The sample size will take as values $n = 15, 30, 45$. Note that this set up incorporates both true nulls and true alternative hypotheses.

To assess if our proposed minP permutation procedure preserves the FWER, 10,000 simulations were created. The FWER was calculated as the proportion of times one or more null hypothesis were incorrectly rejected out of all the 10,000 simulations. Each simulation iteration will correspond to one *level* of the *tree*. If the procedure controls correctly the FWER for each *level*, then it should control it correctly for the whole *tree*. This is because we control the FWER of the *tree* by controlling it at each *level*. Each row of the table 5.1 represents a simulation scenario with different samples sizes. From table 5.1 it can be seen that for the case $p = 20$ the method seems

Table 5.1: Simulation Results

FWER at the 0.05 level							
n	p	q	Sign.	n	p	q	Sign.
15	20	190	0.030	15	40	780	0.040
30	20	190	0.031	30	40	780	0.043
45	20	190	0.035	45	40	780	0.051
n is sample size, p is the number of variables, and q is the number of tests.							

conservative with a calculated FWER of 0.030 when the sample size is 15. The method becomes less conservative as the sample size increases, reaching 0.035 when sample size is 45. For the case $p = 40$, the calculated FWER is closer to its prespecified value of 0.05 across all simulated sample sizes, compared to the case $p = 20$. Moreover, as the

sample size increases, the calculated FWER is very similar to the requested FWER of 0.05: in the case of $n = 45$ the calculated FWER is 0.051. The reason the calculated FWER is closer to the requested FWER, when the number of tests increases, is because our proposed procedure takes advantage of the correlation among tests by preserving the order of the samples among some of the tests while permuting them, as described in section 5.3.5. When $p = 40$, our simulation scenario generates a much larger number of pairs of variables that happen to be correlated. Hence, the amount of correlation that our procedure can use to become less conservative increases, and this makes the true FWER be closer to its specified value of 0.05.

Now, we wouldn't want to use our proposed permutation approach if it did not provide an improvement with respect to the permutation approach of Westfall and Young (1993) which permutes the data to create a null distribution of a given test, independent of other tests. Thus, we also created a simulation study where $p = 15$ and we vary the sample size as $n = 15, 30, 45, 100, 150$. In this numerical study the data generated is still multivariate normal and the covariance matrix is block diagonal with one block of size 10 and another one of size 5, which generates two clusters of dependent variables. We present in table 5.2 the results of this simulation study. The *Complete Null* left side of the table corresponds to the results using the method in Westfall and Young (1993) and the right side corresponds to our proposed permutation method. Across all sample sizes, our method is closer to the desired FWER of 0.05. Consequently, our method has more power than the permutation *minP* of Westfall and Young (1993) for correlation coefficients. The reason of this is that our method incorporates some of the correlation among tests.

Table 5.2: FWER at the 0.05 Level

Complete Null			Proposed Permutation		
n	p	Actual FWER	n	p	Actual FWER
15	15	0.014	15	15	0.032
30	15	0.024	30	15	0.033
45	15	0.026	45	15	0.035
100	15	0.035	100	15	0.039
150	15	0.033	150	15	0.042

5.5 Clustering RNA-seq Gene Expression of Glioblastoma Tumors

5.5.1 Tumor Heterogeneity and Glioblastomas Data Set

Tumor heterogeneity poses a big barrier to develop cancer treatments. This heterogeneity can manifest as variability between tumors, which is associated with distinct clinical outcomes. For example, patients with glioblastoma multiforme with a specific gene mutation had an increase in overall survival (Parsons et al. (2008)). Moreover, cells from the same tumor can have different mutations. In a study that used renal carcinomas it was shown that intratumor heterogeneity can present major challenges to personalized-medicine (Gerlinger et al. (2012)). For this reason, intratumoral heterogeneity plays a determinant role in treatment failure and disease recurrence (Bedard et al. (2013)).

Glioblastoma is a heterogeneous lethal brain cancer. Intratumor heterogeneity is the key to understanding treatment failure. Most patients display different glioblastoma subtypes within the same tumor, which affects treatment design (Sottoriva et al. (2013)). In order to examine the heterogeneity of glioblastoma tumors, single cell RNA-seq profiles of 430 cells from five primary glioblastomas were performed. It was found that these brain tumors are inherently diverse in their expression (Patel et al. (2014), Verhaak et al. (2010)). Moreover, it was shown that established glioblastoma subtypes

are variably expressed across individual cells within a tumor.

To illustrate our method we will use the data set of Patel et al. (2014) on single cell transcriptome analysis by RNA-seq (Ramsköld et al. (2012)). The authors isolated individual cells from five human glioblastoma tumors, which resulted in 6,000 genes in 430 cells. The five tumors analyzed consisted of heterogeneous mixtures of individual cells corresponding to different glioblastoma subtypes defined by the Cancer Genome Atlas Verhaak et al. (2010). They found that individual cells coming from the same tumor were more correlated to each other than cells from different tumors, but even within the same tumor was a large variation in correlations. This is consistent with the idea of intratumoral heterogeneity. Our method will be applied to the single cell data set to further analyze the heterogeneity across the five different tumors.

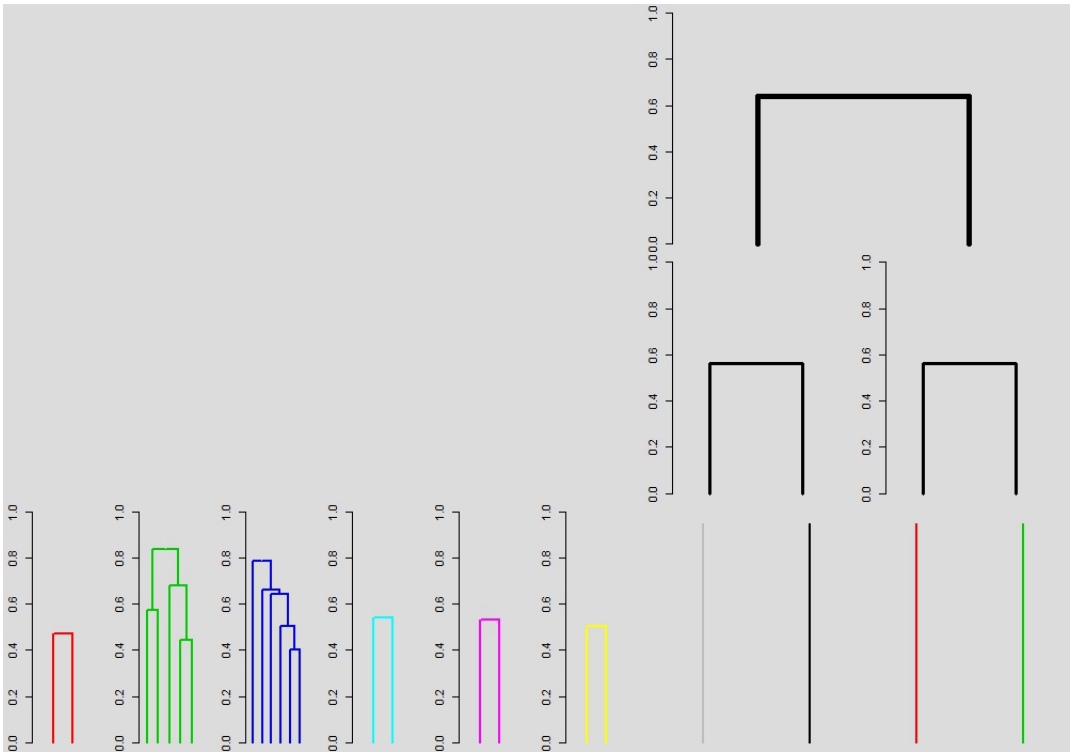
5.5.2 Clustering of Glioblastomas Genes and Prediction of Tumor Category

To assess the difference in RNA-seq profiles across three primary glioblastomas tumors, denoted by MGH 28, 29 and 31, we will apply our NHC algorithm to each of these tumors. The two other tumors, MGH 26 and 30, will not be analyzed initially, but will be used in later stage as comparisons on the classification rate. The data will be partitioned into training and testing set, with 70% and 30% of the data, respectively, and using all five glioblastoma tumors. The analysis will be performed on the training set, whereas the classification rate is calculated using the test set. We selected the 100 RNA-seq gene expressions with the greatest variation across all tumors from 6,000 genes. We will cluster RNA-seq expression of genes, for each of the three tumors separately and investigate if different clustering patterns occur. If there exist differences in clustering patterns, this will indicate across tumor variability in gene expression

profiles. We will perform the algorithm with three levels ($L = 3$). Also, we want to control the overall FWER to be no more than 0.05, thus we set each level to have an overall FWER of no more than 0.05/3, as defined in Section 5.3.4. After clustering the gene expressions for each tumor, coexpression modules will be created from groups of genes that cluster together. We will consider a module a set of genes, such that each gene belongs to the same cluster at level 3 ($L = 3$) of the algorithm. Modules will not be constructed from genes that did not cluster with any other genes at the end of the algorithm. Let $\mathscr{T}_{28}$, $\mathscr{T}_{29}$ and $\mathscr{T}_{31}$ be the *tree* constructed using the NHC algorithm from the training set of the tumors MGH 28, 29 and 31, respectively. Then, a module will be a set of genes that belong to a cluster at the third level, say C_j^3, such that $|C_j^3| > 1$. Many modules will be constructed out of the three glioblastoma tumors. Thus, modules across tumors might overlap on the gene expressions they contain.

Once modules are selected, the principal component (PC) with the largest eigenvalue on each module will be derived (Shen et al. (2007), Langfelder and Horvath (2007), Alter et al. (2000)). Thus, we will be able to represent each cluster of variables by using one variable that represents the direction of largest variance of that cluster. Given q modules of variables across all 3 glioblastoma tumors, we will denote the principal components obtained by $M_1, \ldots, M_q$. After the q PCs are obtained, we will calculate the distance correlation between all the q choose 2 combinations of PCs and the indicator variables for the five different tumors. Then, we rank the resulting combinations from lowest to largest distance correlation (with a larger number denoting greater dependence), and we pick the top 5 pairs that give the greatest association with the indicator variables of different tumors. This will allow us to select the most meaningful PCs for tumor classification. Moreover, since distance correlation can pick up nonlinearities and possible interactions, we believe this approach will be useful in

Figure 5.1: MGH 28

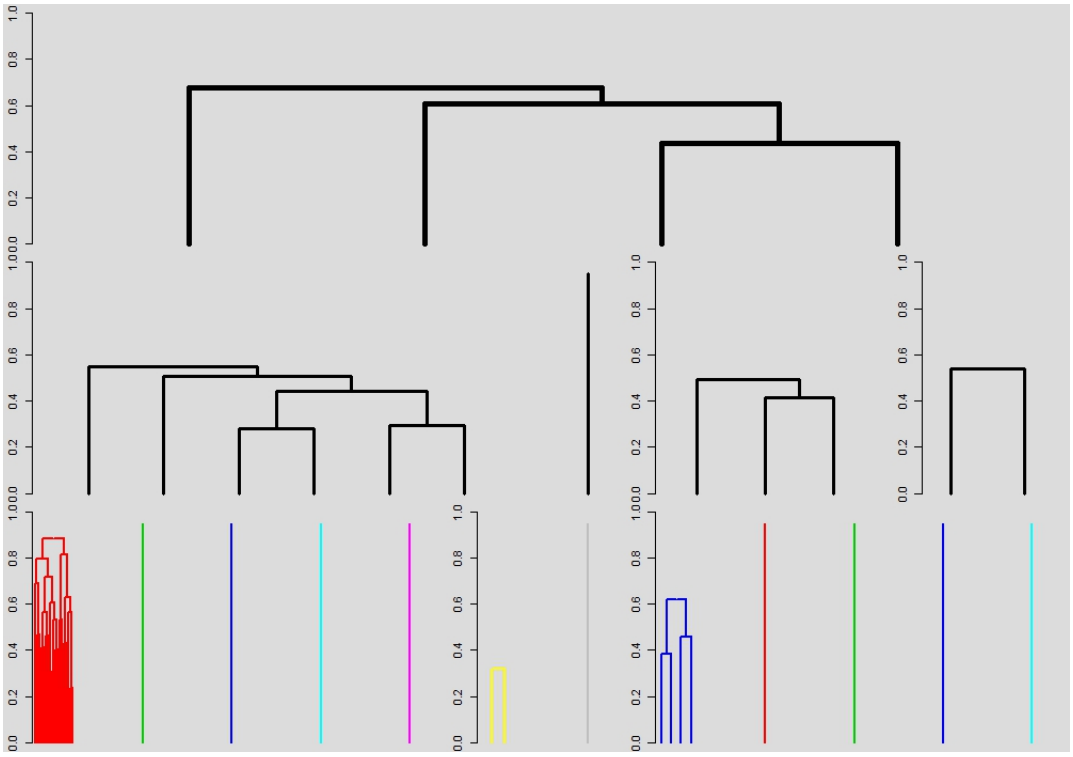


Application of NHC to 100 gene expressions of tumor MGH 28. Each row of the figure correspond to one of the three levels of the tree. Each cluster at each level is represented by a dendrogram created with average linkage and using distance correlation as the distance measure. Different colors denote different clusters at the first level. Clusters at higher levels are in black color.

detecting classification patterns that would be missed by linear and parametric methods. These pairs of PCs will be displayed in a series of figures, with each member of a pair in each axis. Each figure will help us visualize if the pair of PCs is highly associated with the indicator variables of tumor category. This strategy will allow us to investigate if our NHC method can derive a meaningful pattern from the RNA-seq single cell data. Moreover, it will help us visualize the intratumoral heterogeneity if cells of one tumor are present on the same space of cells from another tumor. This analysis will be performed only on the training set.

We will use all the PCs that had the strongest association as pairs with the tumor categories to create 5 separate logistic smoothing spline SNOVA models on the training

Figure 5.2: MGH 29



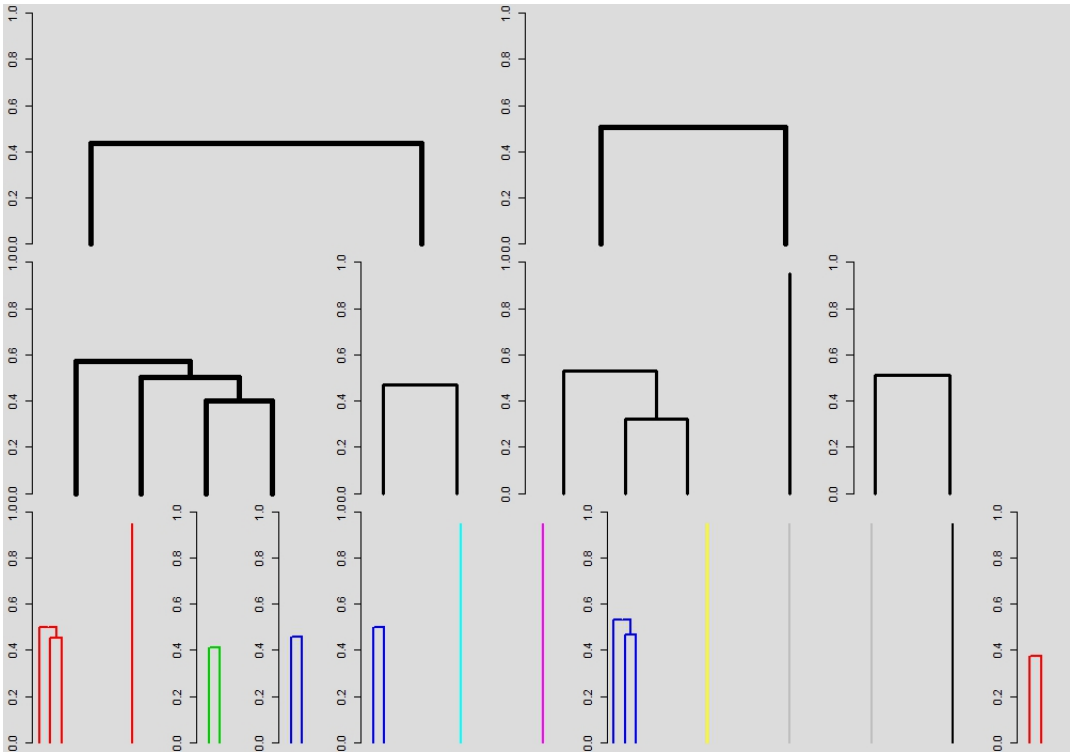
Application of NHC to 100 gene expressions of tumor MGH 29. Each row of the figure correspond to one of the three levels of the tree. Each cluster at each level is represented by a dendrogram created with average linkage and using distance correlation as the distance measure. Different colors denote different clusters at the first level. Clusters at higher levels are in black color.

set, each corresponding to an indicator variable of a tumor (i.e. MGH 26, 28, 29, 30 and 31) as outcome. Then, the test set will be use to evaluate the classification rate of the 5 models created. The overall classification rate will be calculated as well as the classification rate within each of the binary classes. If a relative high classification rate exists, then this will constitute evidence that the NHC algorithm is in fact capturing important structure that can differentiate one tumor cell from another.

5.5.3 Results

The results of the NHC algorithm are displayed in figures 5.1 to 5.3, each corresponding to a glioblastoma MGH 28, 29 and 31, respectively. The figures have 3 rows,

Figure 5.3: MGH 31

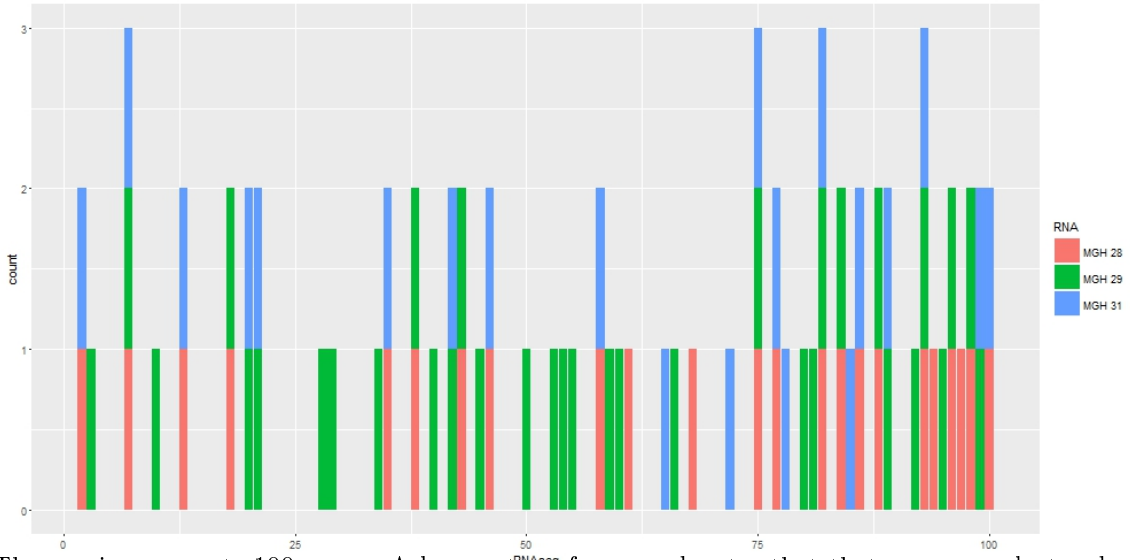


Application of NHC to 100 gene expressions of tumor MGH 31. Each row of the figure correspond to one of the three levels of the tree. Each cluster at each level is represented by a dendrogram created with average linkage and using distance correlation as the distance measure. Different colors denote different clusters at the first level. Clusters at higher levels are in black color.

each corresponding to a level of the NHC algorithm. We can see by a first look at each of the three figures that there exist a big difference in the outcome of the clustering algorithm across the three tumors. Figure 5.2 shows that they are 34 out of the 100 genes of MGH 29 that happen to be correlated with each other, all of which cluster together in one big group at the last step of the algorithm. From figure 5.3 we see that 21 genes were correlated with other genes, and that most of them happen to fall within 2 big clusters of genes, while there exist 2 other small clusters. This pattern of genes aggregating mostly in larger groups seems not to be followed by MHG 28. Figure 5.1 shows that the 21 genes that were correlated fall into 7 small groups. Hence, we have some evidence that the genes of MGH 28 are overall less correlated as a group than

MGH 29 and MGH 31, and form smaller groups of gene clusters.

Figure 5.4: Clusters by Tumor



The x-axis represents 100 genes. A bar on top of a gene denotes that that gene was clustered with some other gene (not shown) for that tumor. Bars of different color denote different tumors. Many genes were not clustered with one another, and only a few clustered for all three tumors.

Moreover, figure 5.4 shows which of the 100 gene expressions analyzed were correlated to other genes in the 3 tumors. We see that there is a great number of gene expressions that did not happen to be correlated to each other. Moreover, there are only 4 genes which happen to be dependent to other genes in all 3 tumors, there are 20 genes which happen to be dependent on others in 2 tumors out of 3, and there are 26 genes which are dependent in only 1 tumor. Also, MGH 29 has the largest amount of genes that are dependent which are not dependent in the other 2 tumors. The NHC algorithm shows that there exist a variability in the dependence structure of gene expressions in cells coming from different glioblastoma tumors.

Using our definition of modules, we constructed out of the 100 genes expressions, 7 modules for the MGH 28 tumor, 1 module for the MGH 29 tumor, and 4 modules for the MGH 31 tumor. From these modules representing groups of dependent gene

Figure 5.5: PCs of First Pair of Modules

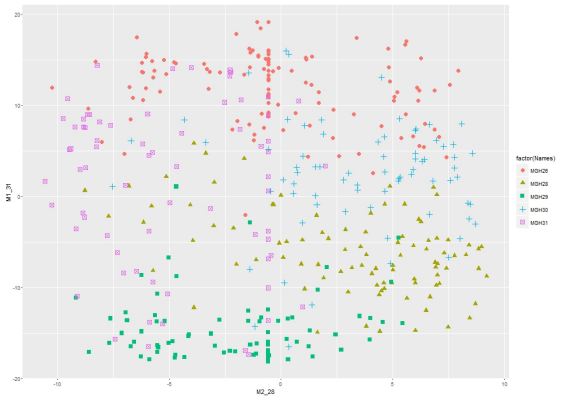


Figure 1 : Eigenvectors

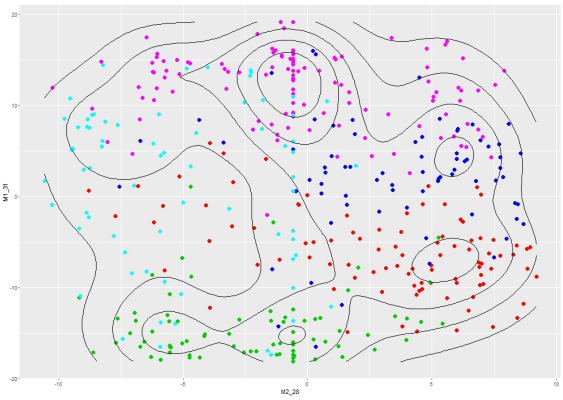


Figure 2 : Contour Plot

expressions, we derived PCs corresponding to the largest eigenvalue within each module. Even though the modules were created using only data for each tumor type, we created the PCs using all five tumor types in the training set. The reason for this is that we want to use these modules for predicting all of the 5 indicator variables of glioblastoma tumors, and thus we need to use all of the training data. From the 12 PCs we found, we chose the 5 pairs out of the possible 66 pairwise combinations, that were the most dependent on the tumor category. We display 3 scatter plots in figure 5.5 to 5.7 of some of the most illustrative pairs. From these figures, it is obvious that the PCs created separate quite well the 5 different glioblastoma tumors. Also, certain combinations of PCs are better at segregating different tumors from others. For example, in figure 5.5, the cells from the MGH 31 tumor are better separated from cells from MGH 30 than in the other 2 plots, where they share the same space.

The top 5 pairs strongly associated with tumor category have 6 unique PCs. We use these 6 PCs to train 5 different logistic smoothing spline ANOVA models, each corresponding to a tumor category. The main effects of the model are the 6 PCs. We train these models on the training data and test the classification rate on the test set.

Figure 5.6: PCs of Second Pair of Modules

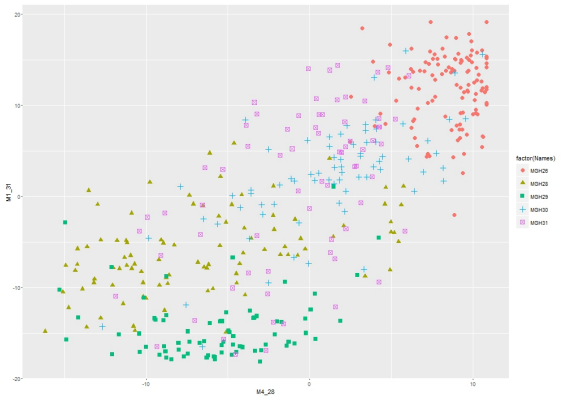


Figure 1 : Eigenvectors

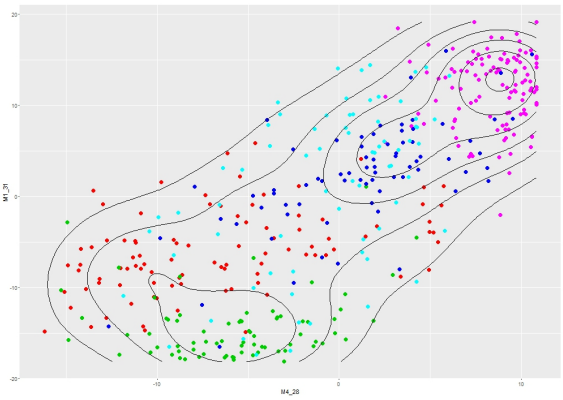


Figure 2 : Contour Plot

The results are shown on table 5.3. The gene expressions for the tumors MGH 26 and 30, were not clustered and PCs were not derived from their clusters of genes. However, in the case of MGH 26 we see that the three types of classification rate are really high. MGH 30 performed well with an overall rate of 88.3%, and with a class 1 rate of 91.3% which was larger than the same classification rate for tumors MGH 28, 29, and 31. From table 5.3 we can conclude that PCs generated from clusters of genes generated by the NHC algorithm can derive meaningful structure and classify tumors on the testing set very accurately.

Table 5.3: Classification Rate on the Testing Set

	MGH 26	MGH 30	MGH 28	MGH 29	MGH 31
Overall	95.3%	88.3%	86.8 %	94.5%	91.4%
Class 1	97.1%	91.3%	88.4 %	88%	85%
Class 0	94.6%	87.7%	86.4%	96.1%	92.6%

Figure 5.7: PCs of Third Pair of Modules

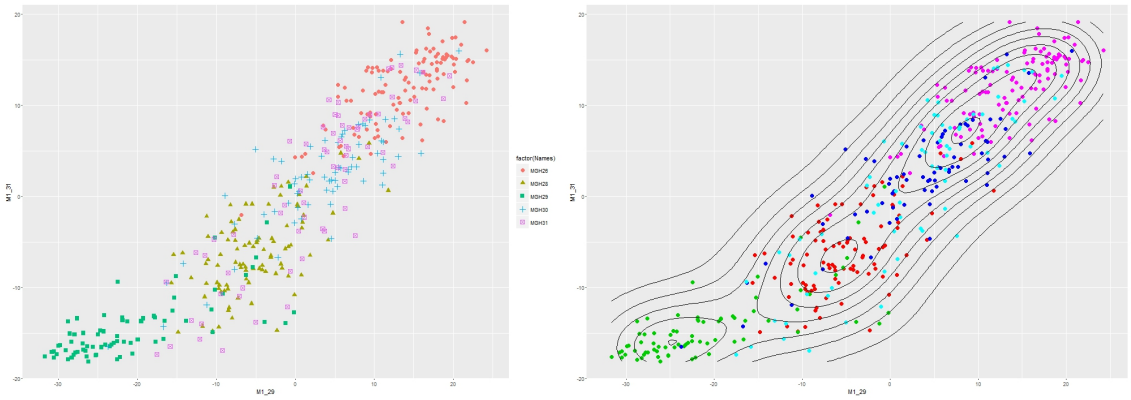


Figure 1 : Eigenvectors

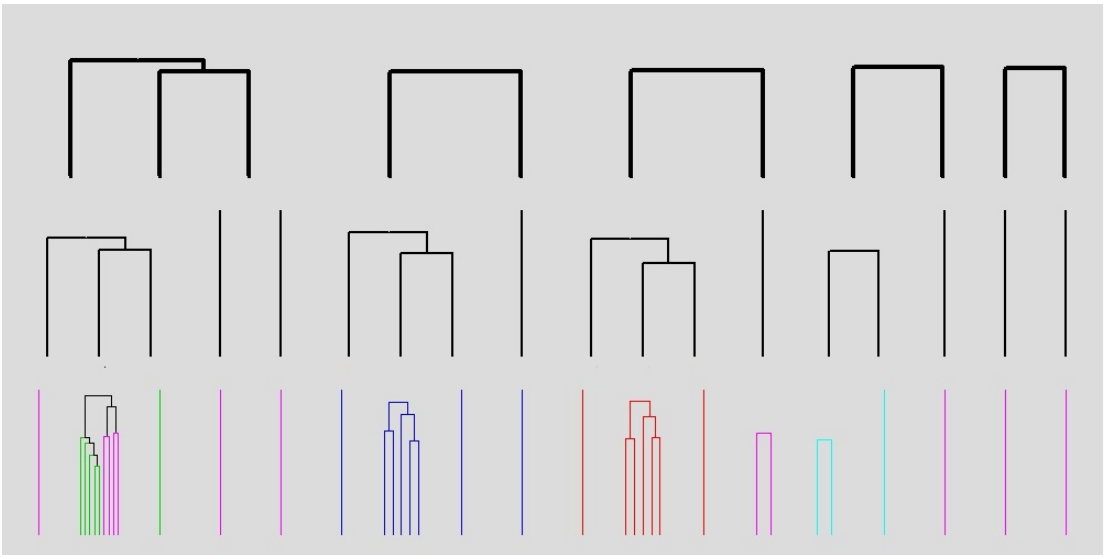
Figure 2 : Contour Plot

5.5.4 Clustering of Glioblastomas Samples

In the research by Patel et al. (2014), there was evidence of great intratumor heterogeneity. This heterogeneity was translated into glioblastomas being mostly correlated with samples of the same tumor type, but also of samples from different tumors. To evaluate this intratumoral heterogeneity, we apply the NHC algorithm on the samples. However, it is important to note that the basic assumption of our method is that the row of the data matrix are i.i.d. In the previous section we saw that many genes are correlated, so this assumption will be violated. Nevertheless, we proceed to see what type of results we obtain. We apply the NHC algorithm using the 200 genes with the most variance as rows, but now we use as columns 50 tumor cells sampled randomly from all the 430 five primary glioblastoma cells. The result of the clustering algorithm is displayed in figure 5.8.

Figure 5.8 shows that cells from tumor MGH 31, clustered together and were not dependent on cells from other tumors. Another cell type that remained almost completely independent of other types was MGH 28. Cells from this tumor were clustered

Figure 5.8: Clustering of Samples



Clustering of samples of five primary glioblastomas. Colors indicate different tumors, with MGH 26 magenta, MGH 28 red, MGH 29 green, MGH 30 light blue and MGH 31 dark blue.

to cells of the same type until level 2, after which some cells of MGH 26 were joined to the cluster. Also, all of MGH 29 and most of MGH 26 cells were clustered together in one group. There are two cells from MGH 26 that clustered alone with no other type. At level 2 of the algorithm, all of MGH 30 cells were clustered together, but at level 3 one MGH 26 cell joined the group. This indicates that there is great heterogeneity within tumor types. Some tumors, like MGH 31 are very homogeneous, whereas tumors MGH 26 and MGH 29 seem to be similar.

5.6 Discussion

In this paper, we introduced the NCH algorithm, a statistical framework and its implementation, to cluster variables while preserving a predefined FWER. Compared with existing methods, NHC uses a hypothesis testing framework to decide whether to join variables together to form larger clusters. NHC uses distance covariance, instead of average linkage or other forms of linkage, for clustering. This is an advantage because

distance covariance can detect non linearities and interactions among groups of variables, something that is impossible to detect using other forms of linkage. Our method detected clusters of genes within different tumors and confirmed the heterogeneity that exist between tumors. Moreover, our clusters of genes were used to generate principal components to predict tumor category which was done with significant accuracy.

A potential shortcoming of the NHC method is that, in its current form, it is computationally intensive. All the examples we have presented are limited in both samples size as well as the number of gene expressions. Therefore, NHC is applicable only if a screening procedure has been applied before hand. Although NHC works well in this set up, there is the open question of whether the NHC can be extend to a higher dimensional set up. We believed that this can be done by using other type I error rate adjustment methods, that while implementing them will make the method considerably faster, but also less powerful.

CHAPTER 6: FUTURE WORK

6.1 Extension of the Test for SS-ANOVA

One possible extension of the first chapter is to devise a test for importance of specific components of the SS-ANOVA model. For instance, if we start with two models that we are interested in comparing that look like this,

$$Y = f_1(X_1) + f_2(X_2) + \varepsilon,$$
$$Y = f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2) + \varepsilon,$$

where the first model has only main effects and the second has an interaction. Thus, we really want to know if $f_{1,2}(X_1, X_2) = 0$. One way to set this up is by using the HSIC, and testing the following hypotheses

$$H_0 : HSIC(Y, f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2)) \leq HSIC(Y, f_1(X_1) + f_2(X_2))$$
$$H_A : HSIC(Y, f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2)) > HSIC(Y, f_1(X_1) + f_2(X_2))$$

The test statistic would be $n(T_n(Y, \hat{f}_1(X_1) + \hat{f}_2(X_2) + \hat{f}_{1,2}(X_1, X_2)) - T_n(Y, \hat{f}_1(X_1) + \hat{f}_2(X_2)))$ and the null distribution would be derived also using the same bootstrap used in Chapter 3. Other terms of interest can be tested similarly.

6.2 Selection of the number of PCs for Nonparametric PCA Regression

Another possible extension of the bootstrap methodology is test for significant PCs to be included in a nonparametric PCA regression model for prediction (Hastie et al. (2005)). For example, if we denote $PC_1, \ldots, PC_p$ the PCs available from the variables data matrix, we can test, consecutively if more PCs are need in the regression. For example if we have two models like

$$Y = f_1(PC_1) + \varepsilon,$$
$$Y = f_1(PC_1) + f_2(PC_2) + \varepsilon,$$

and we want to decide which one is better, then we can test,

$$H_0 : HSIC(Y, \{PC_1, PC_2\}) \leq HSIC(Y, \{PC_1\}),$$
$$H_A : HSIC(Y, \{PC_1, PC_2\}) > HSIC(Y, \{PC_1\}).$$

The test statistic would be $n(T_n(Y, \{PC_1, PC_2\}) - T_n(Y, \{PC_1\}))$ and the null distribution would be derived also using the same bootstrap used in Chapter 3, but where f is replaced by PC_1. If this null hypothesis is rejected, then we can test if the third PC provides an increase in HSIC compared to only using the first two PCs, and so forth. Each test can be tested at some α level. Once a null is not rejected the sequence of tests is stopped. Thus, we keep rejecting the tests until one test is not significant anymore. This will preserve the family wise error rate. This is convenient because each of the p-values of the tests are compared at the α level. This means each test can have significant power.

6.3 Selection of the number of PCs for Spectral Clustering

Spectral clustering is an algorithm that uses graph representation of similarity matrices by converting them into a graph Laplacian and use k of its principal components to run k-means clustering on them (Von Luxburg (2007), Shi and Malik (2000), Ng et al. (2002)). Another way to decide the maximum number of clusters is to select only those that are not independent of each other. Non independence of PCs is useful while doing spectral clustering and if you have an upper bound on the number of PCs that are dependent, this is useful when deciding what k to choose in the the k-means clustering. One way you can go about it too is to do the following hypothesis testing

$$H_0 : HSIC(PC_1, PC_2) = 0,$$
$$H_A : HSIC(PC_1, PC_2) > 0,$$

where the PCs come from the graph Laplacian. The null hypothesis correspond to the case where the first two PCs are independent. If we reject this null, then we can move to test independence between $\{PC_1, PC_2\}$ and $\{PC_3\}$, and so forth. We can keep testing until a null hypothesis is not rejected anymore. Then, if q components are dependent, we wouldn't want to do k-means clustering with $k > q$. The test would be $nT_n(PC_1, PC_2)$ and we can derive the null by a modified version of the bootstrap approach shown in chapter 3.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 3

A.4 Details on the Bootstrap Algorithm

If hypotheses in 3.5 need to be tested using the test statistic in 3.6, 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 $\hat{\sigma}^2 = \frac{\|\mathbf{Y}-\mathbf{A}\mathbf{Y}\|^2}{Tr(\mathbf{I}-\mathbf{A})}$.

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.5 Details on Simulation Studies

In this section, specific details on each simulation study of section 3.4 of the main text are 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

This case corresponds to the simulation results shown in table 3.1 of the main text. Simulations were created where the null hypothesis only includes main effects. We have $f(X) = \sum_j^p f_j(X(j))$ and $f_{1,\ldots,p}(X(1), \ldots, X(p))$ is any interaction between covariates. The hypotheses then become

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

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

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

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

Case I.3, p=6

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

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

Case II

This case corresponds to the simulation results shown in fable 3.2 of the main text. 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 II.*

Case II.1, p=2

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

$$f_3(Z(3)) = \sin(\pi Z(3)).$$

Case II.2, p=4

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

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

Case III

This case corresponds to the simulation results shown in table 3.3 of the main text. 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 III.*

Case III.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 III.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 $B = \{X_2, X_3, X_4\}$.

A.6 Theoretical Results

The main purpose of this section is to provide a justification for Theorem 3.7. 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 3.7.

Set-Up

The theoretical results presented here are for the estimation of f in 3.1 through the solution of the penalized least squares in 3.2. For simplicity, it will be assumed

throughout that f 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 $\mathscr{F}_j = \{f_j : [0,1] \to R, \int |f^{(m)}|^2 < M_j\}$, $M_1 \geq 1$, 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

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

where $\sum_{j=1}^{p} f_j(X(j)) \in \mathscr{F}$. 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

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

Lemma A.6.1. *If we solve the penalized least squares model defined in 3.2 over the RKHS $\mathscr{F}$, and A.1 and A.2 hold, then we have that*

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

provided $n^{2m/(2m+1)}\lambda \geq 1$.

Proof:

Fix $\delta > 0$. We know that $\mathscr{N}_n(\frac{\delta}{p}; \sigma, \mathscr{F}_j) \leq \exp\left(A(\frac{M_j}{\delta/p})^{(1/m)}\right)$, where $\mathscr{N}_n(\frac{\delta}{p}; \sigma, \mathscr{F}_j)$ is the smallest number of δ balls needed to cover the open ball $B(f_{0,j}, \sigma) = \{f_j \in \mathscr{F}_j : \|f_{0,j} - f_j\| \leq \sigma\}$ with respect to the $\|\cdot\|_n$ norm, as defined in Lemma 2.1 in Van de Geer (1990). 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,

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

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 Van de Geer (1990) 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 A.6.2. *We fit the following model using penalized least squares in 3.2 in the main text:*

$$Y_i = f(X_i) + \varepsilon_i, \qquad i = 1, ..., n,$$

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*

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

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 A.6.1 holds in this situation. □

Lemma A.6.3. *We fit an additive model by minimizing the penalized least squares in 3.2. Under the assumptions A.1-2 we have that*

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

Proof:

By Lin (2000), we know that $\|\hat{f} - f\|_2^2 = O_p(n^{-2m/(2m+1)})$ where $\|\cdot\|_2$ is the L_2 norm. By applying Lemma A.6.4 we conclude that

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

Lemma A.6.4. *Let* $f : [0, 1] \to \mathbb{R}$ *such that* $\int_0^1 (f^{(k)}(u))^2 du < \infty$, *then*

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

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 $1 \leq k < \infty$. Let

$\Delta = 1/m$ for some integer $1 \leq m < \infty$. Let $\tilde{f}$ be an approximation to f such that

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

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

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

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

We also have that

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

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

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

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

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

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

Now let,

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

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,

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

With this we can see that

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

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 $\Delta = \left(\frac{a}{(2k-1)b}\right)^{2/(2k-1)} \wedge 1$, 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, if $\|f\|_2 \to 0$ and $J_k(f) = O(1)$, then

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

Lemma A.6.5. *Under the same assumptions as Lemma A.6.2 and using the notation from Theorem 3.7 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, with probability going to 1 as n increases, we can apply Lemma A.6.3 and conclude that $\|\hat{f} - \hat{f}^*\|_\infty \xrightarrow{p} 0$. Thus we have

$$\begin{aligned}\sup_{x\in[0,1]^q} |\hat{f}^*(x) - \hat{f}(x)| &\geq \max_{X_i^* \in \mathbf{X}_n^*} |\hat{f}^*(X_i^*) - \hat{f}(X_i^*)| = \max_{X_i^* \in \mathbf{X}_n^*} |Y_i^* + \hat{f}^*(X_i^*) - Y_i^* - \hat{f}(X_i^*)| \\ &= \max_{i \in \mathbb{N}_n} |\varepsilon_i^* - \eta_i^*|,\end{aligned}$$

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. □

Proof of Theorem 3.7:

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

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

where

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

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 $\sum_{j=p+1}^{p+q} f_j(Z_i(j)) = 0$ for all i. Then, it holds that $\varepsilon_i = \eta_i$ for all i. Now, by applying Lemma A.6.5 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

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

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

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

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 A.6.2 and Lemma A.6.5 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 A.6.5 we have that $\varepsilon_1^* - \eta_1^* \xrightarrow{p} 0$ and hence from the same arguments as above we have that

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

□

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 4

Before proving Theorem 4.3.2, we will prove Lemma B.0.6. In the results that follow, the change point τ will be assume to arise as the results of some true proportion γ such that $\tau = \lfloor \gamma T \rfloor$.

Lemma B.0.6. *Let $\{\delta_T\}$ be a sequence of positive numbers such that $\delta_T \to 0$ and that $T\delta_T \to \infty$. Let T be such that $\gamma \in [\delta_T, 1-\delta_T]$, then for any $\tilde{\gamma} \in [\delta_T, 1-\delta_T]$, let $X(\tilde{\gamma}-) = \{X_1, ..., X_{\lfloor \tilde{\gamma} T \rfloor}\}$, $Y(\tilde{\gamma}-) = \{Y_1, ..., Y_{\lfloor \tilde{\gamma} T \rfloor}\}$, $X(\tilde{\gamma}+) = \{X_{\lfloor \tilde{\gamma} T \rfloor + 1}, \ldots, X_T\}$, and $Y(\tilde{\gamma}+) = \{Y_{\lfloor \tilde{\gamma} T \rfloor + 1}, \ldots, Y_T\}$. Define $\tilde{r} = \lfloor \tilde{\gamma} T \rfloor$, $\tilde{s} = T - \lfloor \tilde{\gamma} T \rfloor$, $r = \lfloor \gamma T \rfloor$ and $s = T - \lfloor \gamma T \rfloor$. The U-statistic converges to*

$$|\mathscr{U}_n^2(X(\tilde{\gamma}-), Y(\tilde{\gamma}-)) - \mathscr{U}_n^2(X(\tilde{\gamma}+), Y(\tilde{\gamma}+))| \overset{a.s}{\to} Q(X, Y, \gamma; \tilde{\gamma}),$$

and

$$Q(X, Y, \gamma; \gamma) = |\mathscr{U}^2(X(\gamma-), Y(\gamma-)) - \mathscr{U}^2(X(\gamma+), Y(\gamma+))|,$$

where $Q(X, Y, \gamma; \tilde{\gamma})$ is defined in the proof.

Proof: The statistic $\mathscr{U}_n^2(X(\tilde{\gamma}-), Y(\tilde{\gamma}-))$ can be written as

$$\begin{aligned}
&\frac{1}{\tilde{r}(\tilde{r}-3)}\Bigg(2\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}} |X_k(\tilde{\gamma}) - X_j(\tilde{\gamma})||Y_k(\tilde{\gamma}) - Y_j(\tilde{\gamma})| \\
&+\frac{1}{(\tilde{r}-1)(\tilde{r}-2)}(2\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}} |X_j(\tilde{\gamma}) - X_k(\tilde{\gamma})|)(2\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}} |Y_j(\tilde{\gamma}) - Y_k(\tilde{\gamma})|) \\
&-\frac{2}{\tilde{r}-2}\sum_{l=1}^{\tilde{r}}(\sum_{j=1}^{\tilde{r}}|X_l(\tilde{\gamma}) - X_j(\tilde{\gamma})|)(\sum_{k=1}^{\tilde{r}}|Y_l(\tilde{\gamma}) - Y_k(\tilde{\gamma})|)\Bigg).
\end{aligned}$$

First assume that $\gamma < \tilde{\gamma}$. The first term converges by the strong law of large number

for U-statistics: Serfling (2009)

$$\begin{aligned}
&\frac{2}{\tilde{r}(\tilde{r}-3)}\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}}|X_k(\tilde{\gamma})-X_j(\tilde{\gamma})||Y_k(\tilde{\gamma})-Y_j(\tilde{\gamma})| \\
&\quad\overset{a.s}{\to}\frac{\gamma^2}{\tilde{\gamma}^2}E|X(\tau-)-X'(\tau-)|\cdot|Y(\tau-)-Y'(\tau-)| \\
&\quad+2\frac{\gamma(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^2}E|X(\tau-)-X(\tau+)|\cdot|Y(\tau-)-Y(\tau+)| \\
&\quad+\frac{(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^2}E|X(\tau+)-X'(\tau+)|\cdot|Y(\tau+)-Y'(\tau+)| \\
&\quad=A(\tilde{\gamma}-).
\end{aligned}$$

The second term converges to

$$\begin{aligned}
&\frac{2}{\tilde{r}(\tilde{r}-1)}\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}}|X_j(\tilde{\gamma})-X_k(\tilde{\gamma})|\frac{2}{(\tilde{r}-2)(\tilde{r}-3)}\sum_{j=1}^{\tilde{r}-1}\sum_{k=j+1}^{\tilde{r}}|Y_j(\tilde{\gamma})-Y_k(\tilde{\gamma})| \\
&\quad\overset{a.s}{\to}B(\tilde{\gamma}-)\cdot C(\tilde{\gamma}-) \\
&\quad=(B_1(\tilde{\gamma}-)+B_2(\tilde{\gamma}-)+B_3(\tilde{\gamma}-))(C_1(\tilde{\gamma}-)+C_2(\tilde{\gamma}-)+C_3(\tilde{\gamma}-)),
\end{aligned}$$

where we define

$$\begin{aligned}
B_1(\tilde{\gamma}-)&=\frac{\gamma^2}{\tilde{\gamma}^2}E|X(\tau-)-X'(\tau-)|, & C_1(\tilde{\gamma}-)&=\frac{\gamma^2}{\tilde{\gamma}^2}E|Y(\tau-)-Y'(\tau-)|, \\
B_2(\tilde{\gamma}-)&=2\frac{\gamma(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^2}E|X(\tau-)-X(\tau+)|, & C_2(\tilde{\gamma}-)&=2\frac{\gamma(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^2}E|Y(\tau-)-Y(\tau+)|, \\
B_3(\tilde{\gamma}-)&=\frac{(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^2}E|X(\tau+)-X'(\tau+)|, & C_3(\tilde{\gamma}-)&=\frac{(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^2}E|Y(\tau+)-Y'(\tau+)|.
\end{aligned}$$

The third term converges to

$$\begin{aligned}
&\frac{2}{\tilde{r}(\tilde{r}-2)(\tilde{r}-3)}\sum_{l=1}^{\tilde{r}}(\sum_{j=1}^{\tilde{r}}|X_l(\tilde{\gamma})-X_j(\tilde{\gamma})|)(\sum_{k=1}^{\tilde{r}}|Y_l(\tilde{\gamma})-Y_k(\tilde{\gamma})|)\\
&\qquad\overset{a.s.}{\to}\frac{2\gamma^3}{\tilde{\gamma}^3}E|X(\tau-)-X'(\tau-)|\cdot|Y(\tau-)-Y''(\tau-)|,\\
&\qquad+2\frac{(\tilde{\gamma}-\gamma)^3}{\tilde{\gamma}^3}E|X(\tau+)-X'(\tau+)|\cdot|Y(\tau+)-Y''(\tau+)|,\\
&\qquad+\frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3}E|X(\tau-)-X(\tau+)|\cdot|Y(\tau-)-Y'(\tau+)|\\
&\qquad+\frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3}E|X(\tau+)-X(\tau-)|\cdot|Y(\tau+)-Y'(\tau+)|\\
&\qquad+\frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3}E|X(\tau+)-X'(\tau+)|\cdot|Y(\tau+)-Y(\tau-)|\\
&\qquad+\frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3}E|X(\tau-)-X'(\tau-)|\cdot|Y(\tau-)-Y(\tau+)|\\
&\qquad+\frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3}E|X(\tau-)-X(\tau+)|\cdot|Y(\tau-)-Y'(\tau-)|\\
&\qquad+\frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3}E|X(\tau+)-X(\tau-)|\cdot|Y(\tau+)-Y'(\tau-)|.\\
&\qquad=D(\tilde{\gamma}-).
\end{aligned}$$

If $\tilde{\gamma}\le\gamma$, then we have that

$$\mathscr{U}_n^2(X(\tilde{\gamma}-),Y(\tilde{\gamma}-))\overset{a.s.}{\to}\mathscr{V}^2(X(\tau-),Y(\tau-)).$$

Now, the statistic $\mathscr{U}_n^2(X(\tilde{\gamma}+),Y(\tilde{\gamma}+))$ can be written as

$$\begin{aligned}
&\frac{1}{(T-\tilde{r})(T-\tilde{r}-3)}\Bigg(2\sum_{j=\tilde{r}+1}^{T-1}\sum_{k=j+1}^{T}|X_k(\tilde{\gamma})-X_j(\tilde{\gamma})||Y_k(\tilde{\gamma})-Y_j(\tilde{\gamma})|\\
&+\frac{1}{(T-\tilde{r}-1)(T-\tilde{r}-2)}(2\sum_{j=\tilde{r}+1}^{T-1}\sum_{k=j+1}^{T}|X_j(\tilde{\gamma})-X_k(\tilde{\gamma})|)(2\sum_{j=\tilde{r}+1}^{T-1}\sum_{k=j+1}^{T}|Y_j(\tilde{\gamma})-Y_k(\tilde{\gamma})|)\\
&-\frac{2}{(T-\tilde{r}-2)}\sum_{l=\tilde{r}+1}^{T}(\sum_{j=\tilde{r}+1}^{T}|X_l(\tilde{\gamma})-X_j(\tilde{\gamma})|)(\sum_{k=\tilde{r}+1}^{T}|Y_l(\tilde{\gamma})-Y_k(\tilde{\gamma})|)\Bigg).
\end{aligned}$$

The U-statistic will be decomposed in several terms. First, assume that $\tilde{\gamma} < \gamma$. The first term converges to

$$\begin{aligned}
&\frac{2}{(T-\tilde{r})(T-\tilde{r}-3)} \sum_{j=\tilde{r}+1}^{T-1} \sum_{k=j+1}^{T} |X_k(\tilde{\gamma}) - X_j(\tilde{\gamma})||Y_k(\tilde{\gamma}) - Y_j(\tilde{\gamma})| \\
&\qquad \overset{a.s}{\to} \frac{\gamma^2}{(1-\tilde{\gamma})^2} E|X(\tau-) - X'(\tau-)| \cdot |Y(\tau-) - Y'(\tau-)| \\
&\qquad + 2\frac{(1-\gamma)(\tilde{\gamma}-\gamma)}{(1-\tilde{\gamma})^2} E|X(\tau-) - X(\tau+)| \cdot |Y(\tau-) - Y(\tau+)| \\
&\qquad + \frac{(\tilde{\gamma}-\gamma)^2}{(1-\tilde{\gamma})^2} E|X(\tau+) - X'(\tau+)| \cdot |Y(\tau+) - Y'(\tau+)| \\
&\qquad = A(\tilde{\gamma}+).
\end{aligned}$$

The second term is

$$\begin{aligned}
&\frac{2}{(T-\tilde{r})(T-\tilde{r}-1)} \sum_{j=\tilde{r}+1}^{T-1} \sum_{k=j+1}^{T} |X_j(\tilde{\gamma}) - X_k(\tilde{\gamma})| \\
&\frac{2}{(T-\tilde{r}-2)(T-\tilde{r}-3)} \sum_{j=\tilde{r}+1}^{T-1} \sum_{k=j+1}^{T} |Y_j(\tilde{\gamma}) - Y_k(\tilde{\gamma})| \\
&\overset{a.s.}{\to} B(\tilde{\gamma}+) \cdot C(\tilde{\gamma}+) \\
&= (B_1(\tilde{\gamma}+) + B_2(\tilde{\gamma}+) + B_3(\tilde{\gamma}+))(C_1(\tilde{\gamma}+) + C_2(\tilde{\gamma}+) + C_3(\tilde{\gamma}+)),
\end{aligned}$$

with

$$
\begin{aligned}
B_1(\tilde{\gamma}+) =& \frac{(\gamma-\tilde{\gamma})^2}{(1-\tilde{\gamma})^2} E|X(\tau-) - X'(\tau-)|, \\
C_1(\tilde{\gamma}+) =& \frac{(\gamma-\tilde{\gamma})^2}{(1-\tilde{\gamma})^2} E|Y(\tau-) - Y'(\tau-)|, \\
B_2(\tilde{\gamma}+) =& 2\frac{(1-\gamma)(\gamma-\tilde{\gamma})}{(1-\tilde{\gamma})^2} E|X(\tau-) - X(\tau+)|, \\
C_2(\tilde{\gamma}+) =& 2\frac{(1-\gamma)(\gamma-\tilde{\gamma})}{(1-\tilde{\gamma})^2} E|Y(\tau-) - Y(\tau+)|, \\
B_3(\tilde{\gamma}+) =& \frac{(1-\gamma)^2}{(1-\tilde{\gamma})^2} E|X(\tau+) - X'(\tau+)|, \\
C_3(\tilde{\gamma}+) =& \frac{(1-\gamma)^2}{(1-\tilde{\gamma})^2} E|Y(\tau+) - Y'(\tau+)|.
\end{aligned}
$$

The third term converges to

$$
\begin{aligned}
&\frac{2}{(T-\tilde{r})(T-\tilde{r}-2)(T-\tilde{r}-3)} \sum_{l=\tilde{r}+1}^{T} \Big(\sum_{j=\tilde{r}+1}^{T} |X_l(\tilde{\gamma}) - X_j(\tilde{\gamma})| \Big) \Big(\sum_{k=\tilde{r}+1}^{T} |Y_l(\tilde{\gamma}) - Y_k(\tilde{\gamma})| \Big) \\
&\qquad \xrightarrow{a.s.} \frac{2\gamma^3}{\tilde{\gamma}^3} E|X(\tau-) - X'(\tau-)| \cdot |Y(\tau-) - Y''(\tau-)|, \\
&\qquad + 2\frac{(\tilde{\gamma}-\gamma)^3}{\tilde{\gamma}^3} E|X(\tau+) - X'(\tau+)| \cdot |Y(\tau+) - Y''(\tau+)|, \\
&\qquad + \frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3} E|X(\tau-) - X(\tau+)| \cdot |Y(\tau-) - Y'(\tau+)| \\
&\qquad + \frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3} E|X(\tau+) - X(\tau-)| \cdot |Y(\tau+) - Y'(\tau+)| \\
&\qquad + \frac{2\gamma(\tilde{\gamma}-\gamma)^2}{\tilde{\gamma}^3} E|X(\tau+) - X'(\tau+)| \cdot |Y(\tau+) - Y(\tau-)| \\
&\qquad + \frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3} E|X(\tau-) - X'(\tau-)| \cdot |Y(\tau-) - Y(\tau+)| \\
&\qquad + \frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3} E|X(\tau-) - X(\tau+)| \cdot |Y(\tau-) - Y'(\tau-)| \\
&\qquad + \frac{2\gamma^2(\tilde{\gamma}-\gamma)}{\tilde{\gamma}^3} E|X(\tau+) - X(\tau-)| \cdot |Y(\tau+) - Y'(\tau-)| \\
&\qquad = D(\tilde{\gamma}+).
\end{aligned}
$$

If $\gamma \leq \tilde{\gamma}$, then we have that

$$\mathscr{U}_n^2(X(\tilde{\gamma}+), Y(\tilde{\gamma}+)) \xrightarrow{a.s.} \mathscr{V}^2(X(\tilde{\gamma}+), Y(\tilde{\gamma}+)).$$

Define

$$\begin{aligned} Q(X, Y, \gamma; \tilde{\gamma}) = 1\{\gamma \leq \tilde{\gamma}\} & |A(\tilde{\gamma}-) + B(\tilde{\gamma}-)C(\tilde{\gamma}-) - A(\tilde{\gamma}-) - \mathscr{V}^2(X(\tau+), Y(\tau+))| \\ + 1\{\gamma > \tilde{\gamma}\} & |A(\tilde{\gamma}+) + B(\tilde{\gamma}+)C(\tilde{\gamma}+) - A(\tilde{\gamma}+) - \mathscr{V}^2(X(\tau-), Y(\tau-))|. \end{aligned}$$

It is clear that

$$|\mathscr{U}_n^2(X(\tilde{\gamma}-), Y(\tilde{\gamma}-)) - \mathscr{U}_n^2(X(\tilde{\gamma}+), Y(\tilde{\gamma}+))| \xrightarrow{a.s} Q(X, Y, \gamma; \tilde{\gamma}),$$

and that when $\tilde{\gamma} = \gamma$ we have that

$$Q(X, Y, \gamma; \gamma) = |\mathscr{V}^2(X(\gamma-), Y(\gamma-)) - \mathscr{V}^2(X(\gamma+), Y(\gamma+))|$$

□.

Theorem B.0.7. *Suppose the assumptions of the previous lemma hold. Let $\hat{\tau}$ denote the change point estimator. Then for T large enough, $\gamma \in [\delta_T, 1-\delta_T]$, and furthermore, for all $\epsilon > 0$*

$$\lim_{T\to\infty} P\left(\left|\gamma - \frac{\hat{\tau}}{T}\right| \geq \epsilon\right) = 0.$$

Proof: Let T be such that $\gamma \in [\delta_T, 1-\delta_T]$, then for any $\tilde{\gamma} \in [\delta_T, 1-\delta_T]$ we have that

$$\left|\mathscr{V}_n^2(X(\tilde{\gamma}-), Y(\tilde{\gamma}-)) - \mathscr{V}_n^2(X(\tilde{\gamma}+), Y(\tilde{\gamma}+))\right| \xrightarrow{a.s} Q(X, Y, \gamma; \tilde{\gamma})$$

as $T \to \infty$, because of Lemma B.0.6. The maximum of $\sqrt{\tilde{\gamma}(1-\tilde{\gamma})}Q(X, Y, \gamma; \tilde{\gamma})$ is

attained when $\tilde{\gamma} = \gamma$. Now, define

$$\hat{\tau}_T = \underset{\tau \in \lceil T\delta_T \rceil, \lceil T\delta_T \rceil + 1, \ldots, \lfloor T(1-\delta_T) \rfloor}{\operatorname{argmax}} a(\tau) \mathscr{V}_n^2(X, Y; \tau),$$

and in the interval $\hat{\Gamma}_T$ as

$$\hat{\tau}_T = \underset{\tilde{\gamma} \in [\delta_T, 1-\delta_T]}{\operatorname{argmax}} a(\tau) \mathscr{V}_n^2(X, Y; \tau),$$

with $\frac{\hat{\tau}_T}{T} \in \hat{\Gamma}$. Since $\hat{\tau}_T$ is the argmax, we have that

$$\frac{1}{\sqrt{T}} a(\hat{\tau}_T) \mathscr{V}_n^2(X, Y; \hat{\tau}_T) \geq \frac{1}{\sqrt{T}} a(\gamma T) \mathscr{V}_n^2(X, Y; \gamma T),$$

and thus we have

$$\frac{1}{\sqrt{T}} a(\hat{\tau}_T) \mathscr{V}_n^2(X, Y; \hat{\tau}_T) \geq \sqrt{\gamma(1-\gamma)} Q(X, Y; \gamma, \gamma) - o_p(1).$$

Now, let $\hat{\gamma} = \hat{\tau}/T$, we have that

$$\begin{aligned} 0 &\leq \sqrt{\gamma(1-\gamma)} Q(X, Y, \gamma; \gamma) - \sqrt{\hat{\gamma}(1-\hat{\gamma})} Q(X, Y, \gamma; \hat{\gamma}) \\ &\leq \frac{1}{\sqrt{T}} a(\hat{\gamma} T) \mathscr{V}_n^2(X, Y; \hat{\gamma} T) + o_p(1) - \sqrt{\hat{\gamma}(1-\hat{\gamma})} Q(X, Y, \gamma; \hat{\gamma}) \\ &\xrightarrow{p} 0. \end{aligned}$$

For every $\epsilon > 0$, there exists a η such that $|\tilde{\gamma} - \gamma| \geq \epsilon$ implies that

$$\sqrt{\tilde{\gamma}(1-\tilde{\gamma})} Q(X, Y, \gamma; \tilde{\gamma}) < \sqrt{\gamma(1-\gamma)} Q(X, Y, \gamma; \gamma) - \eta.$$

Thus, we have that

$$\lim_{T\to\infty} P\Big(\Big|\hat{\gamma}-\gamma\Big| \geq \epsilon\Big) \leq \lim_{T\to\infty} P\Big(\sqrt{\hat{\gamma}(1-\hat{\gamma})}Q(X,Y,\gamma;\hat{\gamma}) < \sqrt{\gamma(1-\gamma)}Q(X,Y,\gamma;\gamma) - \eta\Big)$$
$$\xrightarrow{p} 0$$

$\square$.

REFERENCES

Alter, O., Brown, P. O., and Botstein, D. (2000), "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, 97, 10101–10106.

Aronszajn, N. (1950), "Theory of reproducing kernels," *Transactions of the American mathematical society*, 68, 337–404.

Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. (2013), "Tumour heterogeneity in the clinic," *Nature*, 501, 355–364.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001), "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences*, 98, 13790–13795.

Bounliphone, W., Gretton, A., Tenenhaus, A., and Blaschko, M. (2014), "A low variance consistent test of relative dependency," *arXiv preprint arXiv:1406.3852*.

Camacho, R. C., Galassetti, P., Davis, S. N., and Wasserman, D. H. (2005), "Glucoregulation during and after exercise in health and insulin-dependent diabetes," *Exercise and sport sciences reviews*, 33, 17–23.

Colmegna, P. H., Sánchez-Peña, R. S., Gondhalekar, R., Dassau, E., and Doyle, F. J. (2016), "Reducing Glucose Variability Due to Meals and Postprandial Exercise in T1DM Using Switched LPV Control In Silico Studies," *Journal of diabetes science and technology*, 10, 744–753.

Craven, P. and Wahba, G. (1978), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377–403.

Dudoit, S. and Van Der Laan, M. J. (2007), *Multiple testing procedures with applications to genomics*, Springer Science & Business Media.

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004), "Multiple testing. Part I. Single-step procedures for control of general type I error rates," *Statistical Applications in Genetics and Molecular Biology*, 3, 1–69.

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

Fryzlewicz, P. et al. (2014), "Wild Binary Segmentation for multiple change-point detection," *The Annals of Statistics*, 42, 2243–2281.

Ge, Y., Dudoit, S., and Speed, T. P. (2003), "Resampling-based multiple testing for microarray data analysis," *Test*, 12, 1–77.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012), "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing," *New England Journal of Medicine*, 366, 883–892.

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

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005), "Measuring statistical dependence with Hilbert-Schmidt norms," in *Algorithmic learning theory*, Springer, pp. 63–77.

Gu, C. (1992), "Diagnostics for nonparametric regression models with additive terms," *Journal of the American Statistical Association*, 87, 1051–1058.

— (2004), "Model diagnostics for smoothing spline ANOVA models," *Canadian Journal of Statistics*, 32, 347–358.

— (2013), *Smoothing spline ANOVA models*, vol. 297, Springer Science & Business Media.

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

Iscoe, K. E. and Riddell, M. C. (2011), "Continuous moderate-intensity exercise with or without intermittent high-intensity work: effects on acute and late glycaemia in athletes with Type 1 diabetes mellitus," *Diabetic Medicine*, 28, 824–832.

Kawahara, Y. and Sugiyama, M. (2012), "Sequential change-point detection based on direct density-ratio estimation," *Statistical Analysis and Data Mining*, 5, 114–127.

Killick, R., Fearnhead, P., and Eckley, I. (2012), "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, 107, 1590–1598.

Kimeldorf, G. and Wahba, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of mathematical analysis and applications*, 33, 82–95.

Kowalski, A. (2015), "Pathway to artificial pancreas systems revisited: moving downstream," *Diabetes care*, 38, 1036–1043.

Kudva, Y. C., Carter, R. E., Cobelli, C., Basu, R., and Basu, A. (2014), "Closed-loop artificial pancreas systems: physiological input to enhance next-generation devices," *Diabetes care*, 37, 1184–1190.

Langfelder, P. and Horvath, S. (2007), "Eigengene networks for studying the relationships between co-expression modules," *BMC systems biology*, 1, 54.

— (2008), "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics*, 9, 1.

Langfelder, P., Zhang, B., and Horvath, S. (2008), "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," *Bioinformatics*, 24, 719–720.

Liu, D., Lin, X., and Ghosh, D. (2007), "Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models," *Biometrics*, 63, 1079–1088.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013), "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, 43, 72–83.

Maahs, D. M., Mayer-Davis, E., Bishop, F. K., Wang, L., Mangan, M., and McMurray, R. G. (2012), "Outpatient assessment of determinants of glucose excursions in adolescents with type 1 diabetes: proof of concept," *Diabetes technology & therapeutics*, 14, 658–664.

Mallad, A., Hinshaw, L., Schiavon, M., Dalla Man, C., Dadlani, V., Basu, R., Lingineni, R., Cobelli, C., Johnson, M. L., Carter, R., et al. (2015), "Exercise effects on postprandial glucose metabolism in type 1 diabetes: a triple-tracer approach," *American Journal of Physiology-Endocrinology and Metabolism*, 308, E1106–E1115.

Maran, A., Pavan, P., Bonsembiante, B., Brugin, E., Ermolao, A., Avogaro, A., and Zaccaria, M. (2010), "Continuous glucose monitoring reveals delayed nocturnal hypoglycemia after intermittent high-intensity exercise in nontrained patients with type 1 diabetes," *Diabetes technology & therapeutics*, 12, 763–768.

Matteson, D. S. and James, N. A. (2014), "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, 109, 334–345.

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

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002), "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, 2, 849–856.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010), "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, 56, 5847–5861.

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., et al. (2008), "An integrated genomic analysis of human glioblastoma multiforme," *Science*, 321, 1807–1812.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. (2014), "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, 344, 1396–1401.

Pollard, K. S., Birkner, M. D., Van Der Laan, M. J., and Dudoit, S. (2005), "Test statistics null distributions in multiple testing: Simulation studies and applications to genomics," *Journal de la société française de statistique*, 146, 77–115.

Pollard, K. S. and van der Laan, M. J. (2004), "Choice of a null distribution in resampling-based multiple testing," *Journal of Statistical Planning and Inference*, 125, 85–100.

Purdon, C., Brousson, M., Nyveen, S., Miles, P., Halter, J., Vranic, M., and Marliss, E. (1993), "The roles of insulin and catecholamines in the glucoregulatory response during intense exercise and early recovery in insulin-dependent diabetic and control subjects." *The Journal of Clinical Endocrinology & Metabolism*, 76, 566–573.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., et al. (2012), "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells," *Nature biotechnology*, 30, 777–782.

Riddell, M. C., Zaharieva, D. P., Yavelberg, L., Cinar, A., and Jamnik, V. K. (2015), "Exercise and the Development of the Artificial Pancreas One of the More Difficult Series of Hurdles," *Journal of diabetes science and technology*, 1932296815609370.

Schölkopf, B. and Smola, A. J. (2002), *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press.

Sen, A. and Sen, B. (2014), "Testing independence and goodness-of-fit in linear models," *Biometrika*, 101, 927–942.

Serfling, R. J. (2009), *Approximation theorems of mathematical statistics*, vol. 162, John Wiley & Sons.

Shen, L., Toyota, M., Kondo, Y., Lin, E., Zhang, L., Guo, Y., Hernandez, N. S., Chen, X., Ahmed, S., Konishi, K., et al. (2007), "Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer," *Proceedings of the National Academy of Sciences*, 104, 18654–18659.

Shetty, V. B., Fournier, P. A., Davey, R. J., Retterath, A. J., Paramalingam, N., Roby, H. C., Cooper, M. N., Davis, E. A., and Jones, T. W. (2016), "Effect of exercise intensity on glucose requirements to maintain euglycaemia during exercise in type 1 diabetes," *The Journal of Clinical Endocrinology & Metabolism*, jc–2015.

Shi, J. and Malik, J. (2000), "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22, 888–905.

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

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., et al. (2001), "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, 98, 10869–10874.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. (2013), "Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics," *Proceedings of the National Academy of Sciences*, 110, 4009–4014.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008), "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, pp. 1433–1440.

Suzuki, R. and Shimodaira, H. (2006), "Pvclust: an R package for assessing the uncertainty in hierarchical clustering," *Bioinformatics*, 22, 1540–1542.

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

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

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

Tonoli, C., Heyman, E., Roelands, B., Buyse, L., Cheung, S. S., Berthoin, S., and Meeusen, R. (2012), "Effects of different types of acute and chronic (training) exercise on glycaemic control in type 1 diabetes mellitus," *Sports medicine*, 42, 1059–1080.

Turner, D., Luzio, S., Gray, B., Bain, S., Hanley, S., Richards, A., Rhydderch, D., Martin, R., Campbell, M., Kilduff, L., et al. (2015), "Algorithm that delivers an individualized rapid-acting insulin dose after morning resistance exercise counters post-exercise hyperglycaemia in people with Type 1 diabetes," *Diabetic Medicine*.

van Bon, A. C., Verbitskiy, E., von Basum, G., Hoekstra, J. B., and DeVries, J. H. (2011), "Exercise in closed-loop control: a major hurdle," *Journal of diabetes science and technology*, 5, 1337–1341.

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

Van Der Laan, M. J. and Bryan, J. (2001), "Gene expression analysis with the parametric bootstrap," *Biostatistics*, 2, 445–461.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004), "Multiple testing. Part II. Step-down procedures for control of the family-wise error rate," *Statistical applications in genetics and molecular biology*, 3, 1–33.

Van der Laan, M. J. and Pollard, K. S. (2003), "A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap," *Journal of Statistical Planning and Inference*, 117, 275–303.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010), "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer cell*, 17, 98–110.

Von Luxburg, U. (2007), "A tutorial on spectral clustering," *Statistics and computing*, 17, 395–416.

Wahba, G. (1969), "On the numverical solution of Fredholm integral equations of the first kind." Tech. rep., DTIC Document.

— (1985), "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *The Annals of Statistics*, 1378–1402.

— (1990), *Spline models for observational data*, vol. 59, Siam.

Westfall, P. H. and Young, S. S. (1993), *Resampling-based multiple testing: Examples and methods for p-value adjustment*, vol. 279, John Wiley & Sons.

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

Yardley, J. E., Kenny, G. P., Perkins, B. A., Riddell, M. C., Balaa, N., Malcolm, J., Boulay, P., Khandwala, F., and Sigal, R. J. (2013), "Resistance Versus Aerobic Exercise Acute effects on glycemia in type 1 diabetes," *Diabetes Care*, 36, 537–542.

Yardley, J. E., Kenny, G. P., Perkins, B. A., Riddell, M. C., Malcolm, J., Boulay, P., Khandwala, F., and Sigal, R. J. (2012), "Effects of performing resistance exercise before versus after aerobic exercise on glycemia in type 1 diabetes," *Diabetes care*, 35, 669–675.

Zou, C., Yin, G., Feng, L., Wang, Z., et al. (2014), "Nonparametric maximum likelihood approach to multiple change-point problems," *The Annals of Statistics*, 42, 970–1002.