Nonparametric Tests to Detect Relationship between Variables in the Presence of Heteroscedastic Treatment Effects

by

SITI TOLOS

B.S., University of Wisconsin, 1991

M.S., University of Arkansas, 2000

M.S., University of Arkansas, 2002

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Abstract

Statistical tools to detect nonlinear relationship between variables are commonly needed in various practices. The first part of the dissertation presents a test of independence between a response variable, either discrete or continuous, and a continuous covariate after adjusting for heteroscedastic treatment effects. The method first involves augmenting each pair of the data for all treatments with a fixed number of nearest neighbors as pseudo-replicates. A test statistic is then constructed by taking the difference of two quadratic forms. Using such differences eliminate the need to estimate any nonlinear regression function, reducing the computational time. Although using a fixed number of nearest neighbors poses significant difficulty in the inference compared to when the number of nearest neighbors goes to infinity, the parametric standardizing rate is obtained for the asymptotic distribution of the proposed test statistics. Numerical studies show that the new test procedure maintains the intended type I error rate and has robust power to detect nonlinear dependency in the presence of outliers. The second part of the dissertation discusses the theory and numerical studies for testing the nonparametric effects of no covariate-treatment interaction and no main covariate based on the decomposition of the conditional mean of regression function that is potentially nonlinear. A similar test was discussed in Wang and Akritas (2006) for the effects defined through the decomposition of the conditional distribution function, but with the number of pseudo-replicates going to infinity. Consequently, their test statistics have slow convergence rates and computational speeds. Both test limitations are overcome using new model and tests. The last part of the dissertation develops theory and numerical studies to test for no covariate-treatment interaction, no simple covariate and no main covariate effects for cases when the number of factor levels and the number of covariate values are large.

Nonparametric Tests to Detect Relationship between Variables in the Presence of Heteroscedastic Treatment Effects

by

SITI TOLOS

B.S., University of Wisconsin, 1991

M.S., University of Arkansas, 2000

M.S., University of Arkansas, 2002

A DISSERTATION

submitted in partial fulfillment of the

requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics

College of Arts and Sciences

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2010

Approved by:

Major Professor
Haiyan Wang

Abstract

Statistical tools to detect nonlinear relationship between variables are commonly needed in various practices. The first part of the dissertation presents a test of independence between a response variable, either discrete or continuous, and a continuous covariate after adjusting for heteroscedastic treatment effects. The method first involves augmenting each pair of the data for all treatments with a fixed number of nearest neighbors as pseudo-replicates. A test statistic is then constructed by taking the difference of two quadratic forms. Using such differences eliminate the need to estimate any nonlinear regression function, reducing the computational time. Although using a fixed number of nearest neighbors poses significant difficulty in the inference compared to when the number of nearest neighbors goes to infinity, the parametric standardizing rate is obtained for the asymptotic distribution of the proposed test statistics. Numerical studies show that the new test procedure maintains the intended type I error rate and has robust power to detect nonlinear dependency in the presence of outliers. The second part of the dissertation discusses the theory and numerical studies for testing the nonparametric effects of no covariate-treatment interaction and no main covariate based on the decomposition of the conditional mean of regression function that is potentially nonlinear. A similar test was discussed in Wang and Akritas (2006) for the effects defined through the decomposition of the conditional distribution function, but with the number of pseudo-replicates going to infinity. Consequently, their test statistics have slow convergence rates and computational speeds. Both test limitations are overcome using new model and tests. The last part of the dissertation develops theory and numerical studies to test for no covariate-treatment interaction, no simple covariate and no main covariate effects for cases when the number of factor levels and the number of covariate values are large.

Table of Contents

List of Figures viii

List of Tables ix

List of Abbreviations xi

Acknowledgements xii

1 Introduction 1

2 Literature Review 5

2.1 Testing the Relationship of Two Variables Taking into Account the Existence of Treatment . . . 5

2.1.1 Likelihood Ratio Test . . . 5

2.1.2 Wald Type Test for Discrete Response Variable . . . 9

2.1.3 General Additive Models (GAM) . . . 10

2.1.4 Drop test . . . 11

2.2 Testing Relationship of Two Variables without Incorporating Treatment in the Model . . . 14

2.2.1 Goodness of Fit Tests . . . 14

2.2.2 Mutual Information (MI) . . . 15

2.2.3 Pearson's Correlation, Spearman's ρ and Kendal's τ . . . 16

2.3 Analysis of Covariance (ANCOVA) . . . 16

2.4 Nonparametric Methods for ANCOVA Model . . . 17

2.4.1 Fully Nonparametric (FNP) Model . . . 17

2.4.2 FNP Model in Higher-way ANCOVA with at Most Three Covariates 19

2.4.3 FNP Model in Testing No Main Covariate and No Main Covariate-treatment Interaction Effects. . . . 20

2.4.4 Further Comments on Wang and Akritas (2006)'s . . . 23

3 Method of Detecting Dependency of Two Variables in the Presence of Treatment Effect **24**

3.1 Introduction . . . 24
3.2 Main Results . . . 26
3.2.1 Construction of Test Statistics . . . 26
3.2.2 Results Under the Null Hypothesis . . . 29
3.3 Numerical Results . . . 41
3.3.1 Analysis of Ozone Concentration Data - Detection of Nonlinear Dependence . . . 41
3.3.2 Application to EFT Study - Resistance to Outliers . . . 42
3.3.3 Simulation Study . . . 45

4 New Nonparametric Tests when Treatment Level a is Small **51**

4.1 Nonparametric Tests of No Covariate-Treatment Interaction and No Main Covariate Effects . . . 51
4.2 Models and Hypotheses . . . 52
4.3 Test Statistics . . . 58
4.4 Asymptotic Distribution of the Test Statistics . . . 59
4.4.1 No Main Covariate Effect . . . 59
4.4.2 No Covariate-Treatment Interaction Effect . . . 66
4.5 Simulation studies . . . 66
4.5.1 Computational Time Comparison . . . 67
4.5.2 Linear Alternative . . . 68
4.5.3 Quadratic Alternative . . . 71
4.5.4 Bernoulli Responses . . . 75
4.6 Data Analysis . . . 76
4.6.1 Analysis of Ozone Concentration Data (continued from Chapter 3) . . . 76
4.6.2 Application to EFT Study (continued from Chapter 3) . . . 77

5 New Nonparametric Tests when Treatment Level a is Large **81**

5.1 Tests of No Covariate-Treatment Interaction, No Main Covariate and No Simple Covariate Effects when a is Large . . . 81
5.2 Asymptotic Distribution of Test Statistics Under the Null Hypotheses . . . 83
5.2.1 Test of No Covariate-treatment Interaction Effect. . . . 83
5.2.2 Test of No Main Covariate Effect . . . 89

5.2.3 Test for No Simple Covariate Effect . . . 92
5.3 Numerical studies . . . 93
5.3.1 Simulation Studies Setting . . . 93
5.3.2 Covariate-treatment Interaction . . . 95
5.3.3 When There is No Covariate-treatment Interaction : Test of No Simple Covariate effect and Test of No Main Covariate Effect . . . 99
5.4 Data Analysis . . . 114
5.4.1 Ozone Data Revisited . . . 114

6 Conclusion and Post-dissertation Research 117
6.1 Conclusion . . . 117
6.2 Post-dissertation Research . . . 119

Bibliography 123

Appendix 123

A R code 124
A.1 pNP Tests . . . 124
A.2 pNP Tests When a and N Are Large . . . 134
A.3 WA Tests . . . 135
A.4 Classical F Test (CF Test) . . . 139
A.5 Drop Test . . . 139
A.6 GAM Models (Spline and Loess) . . . 140
A.7 GAM Pspline . . . 141
A.8 Correlation Based Tests . . . 141
A.9 ACE . . . 141
A.10 Wald and Deviance Tests . . . 142
A.11 Comparing Computational Time . . . 142
A.12 Code for Simulation in Chapter 3 . . . 143
A.13 Data Generation in Chapter 4 . . . 146
A.14 Data Generation for Simulation Study in Chapter 5 . . . 148
A.15 Code for examples in Chapter 4 . . . 149

List of Figures

2.1 Curse of Dimensionality 23

3.1 Scatter plot of ozone vs wind or doy. 43
3.2 Scatter plot of Time vs EFT for each instruction group. 44
3.3 Empirical power at level 0.01 for data from a mixture of beta and lognormal 48

4.1 Scatter plots and graphs for the exponential with linear conditional mean and normal distribution with sinusoidal conditional mean 56
4.2 Computational Time Comparison 68

5.1 Plot of the conditional mean component and variance for the simulation $prop = 0.1$, and $\tau = 0.0625$. 96
5.2 Plot of the conditional mean component and variance for the simulation $b = 0.7$, $prop = 0.9$, and $\tau = 0.0625$. 106
5.3 Scatter plot for data generated under the null hypothesis of no covariate-treatment interaction with $\theta = \pi/4$, $b = 0.5$, $\tau = 0.25$ and $prop = 0.5$ following the model (5.3.2). 107
5.4 Scatter plot for data generated under the null hypothesis of no covariate-treatment interaction with $\theta = \pi/4$, $b = 0.5$, $\tau = 0.0625$ and $prop = 0.9$ following the model (5.3.2). 108
5.5 Scatter plot of data generated from equation (5.3.2) when $\tau = 0.25$ $prop = 0.33$ for power estimation of no covariate-treatment interaction test 109
5.6 Scatter plot of data generated from equation (5.3.2) when $\tau = 0.25$ $prop = 0.9$ for power estimation of no covariate-treatment interaction test 110
5.7 Scatter plot of ozone vs temp and box plot of ozone vs temp 115
5.8 Ozone Vs doy within temperature levels 116

List of Tables

3.1 $P-$values for test of no association before and after the outlier - adjusting for treatment effect . . . 44
3.2 $P-$values for test of no association before and after the outlier - correlation approach . . . 44
3.3 Proportion of rejections under H_0 in (3.2.1) at level 0.01 . . . 46
3.4 Proportion of rejections under H_0 (3.2.1) at level 0.01 - correlation based . . . 47
3.5 Empirical power at level 0.01 . . . 49
3.6 Empirical power at level 0.01 based on 2000 runs when $n_i = 20$ and 30. - correlation based tests . . . 50

4.1 Proportion of rejections for testing no covariate-treatment interaction effect using WA, pNP, GAM Spline and GAM Loess, Drop and CF tests - linear . . . 70
4.2 Proportion of rejections for testing no main covariate effect using WA, pNP, GAM Spline and GAM Loess, Drop and CF tests - linear . . . 72
4.3 Proportion of rejections for testing no covariate-treatment interaction effect using WA, pNP, Drop and CF tests - quadratic model . . . 73
4.4 Proportion of rejections for testing no main covariate effect for WA, pNP, GAM Spline and GAM Loess, Drop and CF tests - quadratic model . . . 74
4.5 Proportion of rejections for testing no covariate-treatment interaction effect for WA, pNP, GLM Wald GLM and Deviance tests . . . 78
4.6 Proportion of rejections for testing no main covariate effect for WA, pNP, GLM Wald GLM and Deviance tests . . . 79
4.7 P values for test of no doy-wind speed interaction effect of the ozone data . . . 79
4.8 P values for test of no main doy effect of the ozone data . . . 80
4.9 P values for test of no EFT-group interaction effect of the EFT data . . . 80
4.10 P values for test of no main EFT effect of the EFT data . . . 80

5.1 Proportion of rejections at level 0.01 under the null hypothesis of no covariate-treatment interaction effect for $\theta = \pi/4$ and $b = 0.5$ . . . 98

5.2 The values b's and θ's that generate the treatment levels combination for the simulation study for power performance of test of no covariate-treatment interaction effect. 99
5.3 Power performance at level 0.01 for testing of no covariate-treatment interaction 100
5.4 Power for no covariate-treatment interaction, no covariate simple and no main covariate effects for pNP test with additional values of *prop*. 101
5.5 Bootstrap Power performance for testing of no covariate-treatment interaction 102
5.6 Proportion of rejections at level 0.01 under the null hypothesis of no simple covariate effect 103
5.7 Power performance for testing no simple covariate effect 105
5.8 Bootstrap Power performance for testing of no simple covariate effect 111
5.9 Proportion of rejections at level 0.01 under the null hypothesis of no main covariate effect 112
5.10 (Power performance) Proportion of rejections at level 0.01 for testing of main covariate effect when $n_i = 20$ and $a = 20$ using the model 5.3.2 with $\theta = \pi/4$ and $b = 0.5$. 113
5.11 Table of Levels of temperature 114
5.12 P-values for test of no doy-temperature interaction, no simple doy and no main doy effects. 115

List of Abbreviations

ACE - Alternating Conditional Expectations

AIC - Akaike's Information Criterion

ANCOVA - Analysis of Covariance

ANOVA - Analysis of Variance

CDF - Cumulative Distribution Function

CF - Classical F

doy - day of year

EFT - Embedded Figure Test

FNP - Fully Nonparametric

GAM - Generalized Additive Model

GLM - General Linear Model

GLMz - Generalized Linear Model

iid - independent and identically distributed

LRT - Likelihood Ratio Test

LS - Least square

MI - Mutual Information

pdf - Probability Distribution Function

pNP - proposed Nonparametric

UMPI - Uniformly Most Powerful Invariant

WA - Wang and Akritas

WISC - Wechsler Intelligence Scale for Children

Acknowledgments

First, all Praise is for Allah (God) the Almighty for making the writing of this dissertation possible. With that I thank my major professor Dr Haiyan Wang for her persistent efforts in guiding and helping me with this dissertation. I also thank Dr. Soujin Wang from the Department of Statistics, Texas A& M University for his contribution. My sincere appreciation goes to Dr. Clive Fullagar who gave of his time from the Psychology Department to chair the committee, and that genuine appreciation extends as well to committee members Dr. James Higgins, Dr. Weixing Song and Dr. Susan Brown. I also include all the many faculty members of the Department of Statistics who have been very helpful throughout my studies here at K-State. Thank you also to all graduate students for their friendships. Last but not least, my sincere appreciation goes to all my family members; my mother, my late father, my late father-in-law and late mother-in-law, my husband, my sons, my daughters, my sisters and my brothers for all their support. Thank you.

Chapter 1

Introduction

Statistical tools to detect general relationships between variables are commonly needed in various research disciplines. The following examples describe some relationships between a response variable and a covariate in the presence of some discrete factors using a random sample (X_{ij}, Y_{ij}) observed from the i^{th} treatment, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$.

Example 1

(a) ANCOVA model: $Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \gamma_i X_{ij} + \epsilon_{ij}$, where X_{ij} and ϵ_{ij} are independent.

(b) Let $Y_{ij} = g_i(X_{ij} - \mu) + \epsilon_{ij}$ for some unknown functions $g_i(\cdot)$, where X_{ij} and ϵ_{ij} are independent. The dependence of Y_{ij} on X_{ij} can be through the main effect of the covariate or the covariate-treatment interaction on the conditional mean $E(Y_{ij}|X_{ij})$. Further, $E(Y_{ij}|X_{ij})$ may be arbitrarily linear or nonlinear functions of X_{ij}.

(c) Given X_{ij}, Y_{ij} is a Bernoulli random variable with success probability p_{ij}, such that the $logit(p_{ij}) = g_i(X_{ij})$ for some unknown functions $g_i(\cdot)$. Here Y_{ij} is binary and depends on X_{ij} in the success probability.

(d) Given X_{ij}, Y_{ij} is a Poisson random variable with mean $p_{ij} = g_i(X_{ij})$. In this example, Y_{ij} depends on X_{ij} through both the mean and variance function.

Example 1 (a) is the commonly used textbook model. The test for no main covariate effect and covariate-treatment interaction effect can be carried out using the likelihood

approach in classical linear models or the nonparametric approach by McKean and Schrader (1980) (see also the Drop test by Terpstra and Mckean 2005). Only linear relationship is captured in this model. In practice, the relationship may be nonlinear and the tests based on assumption of linearity may not detect such relationship. In fact, let Y_{ij} and X_{ij} be as specified in Example 1 (b) with $g_i(x)$ being symmetric functions around μ_i. Assume that X_{ij} are independent with mean μ_i and symmetric distribution around μ_i. Then X_{ij} and Y_{ij} are uncorrelated regardless of the distributions of X_{ij} and ϵ_{ij} as long as the above conditions are satisfied. This is easily verified by noting that $(x-\mu_i)\cdot g_i(x-\mu_i)$ is an odd function and X_{ij} has a symmetric distribution around μ_i. Example 1 (b) allows nonlinear relationship between the response and the covariate. When considering only the data for one treatment (i.e. $a=1$) and where the mean of Y is a continuous function of X, testing the independence of X and Y is accomplished by testing the hypothesis of constant regression against a general alternative. This is a special case of lack-of-fit or goodness-of-fit testing in regression, (cf Eubank and Hart 1992; Müller 1992; Hardle and Mammen 1993; Dette and Munk 1998; Dette 1999, to mention a few). When there is more than one treatment ($a>1$), one may carry out a test given in the aforementioned references for each treatment and the individual tests are combined to produce an overall measure of dependence between the response and covariate. If the conditional distribution of the response variable comes from the exponential family, then the nonlinear relationship may be studied through generalized additive models (GAM) using a smoother such as spline or loess (Hastie and Tibshirani 1990), or penalized smoothing spline (Wood, 2000, 2008). When the conditional distribution is beyond the exponential family, GAM may be very liberal as shown in the simulation studies presented in the dissertation.

To incorporate discrete observations (Example 1 (c)), mutual information (MI) is used as a dependency measure (D'haeseleer et al. 1998; Butte and Kohane 2000). However, one of the disadvantages of MI is the need to estimate the joint and marginal probability distribution functions for the response and covariate variables.

When significant dependency is found, a natural question is whether the response variable depends on the covariate in the same way for all treatments (covariate effect) or depends on

the covariate through its interaction with the treatment. Wang and Akritas (2006) proposed a test for no nonparametric main covariate effect and no treatment-covariate interaction effect adjusted for the effect of factors using the nonparametric ANCOVA model introduced by Akritas et al. (2000). The one-way ANCOVA is converted to an artificial two-way ANOVA design using pseudo-replicates. The asymptotic distribution of their test statistics have a standardizing rate $N^{1/2}k^{-1/2}$, where N is the total number of covariate values in all treatments and k is the number of pseudo-replicates per covariate value. In Wang and Akritas (2006), k is required to approach infinity at a rate faster than $\log(N)$ and typically at rate $N^{1/2}$. This rate falls in the range of a regular standardizing rate for nonparametric test statistics, N^{α}, where $0 < \alpha < 1/2$ (cf. Müller 1992). One disadvantage of their tests is that the test performance depends on the number of pseudo-replicates and the number of covariate values in each treatment group, denoted as n_i. For example, for $n_i = 30$, the estimated type I error at level 0.05 in one of their simulation studies is 0.089 when $k = 3$, and 0.042 when $k = 7$. For $n_i = 50$, the type I error estimate is 0.070 when $k = 5$, and 0.042 when $k = 9$ (see Table 2 of Wang and Akritas 2006). In addition, the tests in Wang and Akritas (2006) are very computationally extensive. For $n_i = 200$, it took 263 minutes to perform a single test for both the covariate and interaction effect. For $n_i = 500$, a single test of Wang and Akritas (2006) did not finish in 5 days. (The performance is based on a computer with Intel (R) Pentium M processor 1.86GHz, 1GB of RAM.) With such limitations, the tests in Wang and Akritas (2006) are not practicable.

This dissertation is divided into three parts. The first part consists of a computationally feasible nonparametric test to effectively detect general dependency between two variables after adjusting for the heteroscedastic treatment effects. A fixed number of nearest-neighbor pseudo-replicates augment each pair of treatment level and covariate value combinations. The test statistics are constructed as a difference between two quadratic forms, both of which are common estimates of linear combinations of the variances and conditional variances. The results are given under the null hypothesis. By using a fixed number of nearest-neighbors augmentation, the standardizing rate for the new test statistics achieved the rate for parametric analysis of $\sqrt{N}$.

Part 2 of the dissertation consists the tests for no main covariate effect and no covariate-treatment interaction effect under their corresponding null hypotheses. The same parametric standardizing rate $\sqrt{N}$ is achieved for all statistics. For exactly the same data that took 263 minutes for the Wang and Akritas (2006) test, the new tests finished all the hypotheses testing within 3.28 minutes; for the test that Wang and Akritas (2006) could not finish in 5 days, it only took 17 minutes for the new tests to perform all hypotheses considered in this dissertation. Therefore, comparing available literature on nonparametric hypotheses testing related to the effect of a covariate on the response, the contribution of this research is not only its parametric standardizing rate for the test statistics, but also on its efficient computational advantage.

The third part of the dissertation extends the nonparametric tests for no covariate-treatment interaction, no main covariate and no simple covariate effect to case when both the treatment level and covariate values in each treatment levels are large.

The rest of the dissertation is organized as follows. Chapter 2 will give a literature review of the available methods; Chapter 3 will be devoted to the theory and application of the new nonparametric test to detect general dependency between the response variable and covariate adjusted for heteroscedastic treatment effects; Chapter 4 presents the theory and numerical study for the new nonparametric of no main covariate and no covariate-treatment interaction effect for case when number of treatment level is small; Chapter 5 presents the theory and numerical study for the test of no main covariate and no covariate-treatment interaction effect for case when when both the treatment level and covariate values in each treatment levels are large. Chapter 6 presents a summary and post-dissertation research.

Chapter 2

Literature Review

In this chapter, reviews of available methods are given in conjunction with their relevance toward either testing for independence of covariate and response, covariate-treatment interaction or main covariate efffects.

2.1 Testing the Relationship of Two Variables Taking into Account the Existence of Treatment

2.1.1 Likelihood Ratio Test

The most intuitive way of testing for association of two variables is using a likelihood ratio test. A likelihood ratio test (LRT) is a general test procedure that is based on the ratio of likelihood functions. It is used to compare the fit of the two models for the data. Generally, the likelihood ratio test can also be performed under a general linear model (GLM) or a generalized linear model (GLMz). In this section we discuss two versions of LRT; one for the linear model, one for the generalized linear model. In both cases, the test for independence of response and covariate adjusted for treatment is implemented by testing for no covariate simple effect i.e. test for covariate plus treatment-covariate interaction effect equals zero.

a) **Likelihood Ratio Test for the Simple Effect of Covariate when Response Variable Is Continuous.**

Let (X_{ij}, Y_{ij}), $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, denote pairs of covariates and responses from

the j-th observation of the i-th treatment. Suppose a model

$$E(\mathbf{Y}|X) = \boldsymbol{\beta}^T\mathbf{X} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_1\mathbf{X}_2, \tag{2.1.1}$$

is fitted to the data where $\mathbf{X}_1$ is a dummy indicator variable for treatment effect, $\mathbf{X}_2$ is a continuous covariate variable and the matrix $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$.

To test any null hypothesis H_0 regarding the parameter $\boldsymbol{\beta}$, use the generalized likelihood ratio

$$\begin{aligned} LR &= -2\log\left(\frac{\text{Ł at } H_0}{\text{Ł at } H_a}\right) \\ &= 2(l_a - l_0), \end{aligned} \tag{2.1.2}$$

where Ł is a likelihood function, $l_a = \log$ Ł at H_a and $l_0 = \log$ Ł at H_0. When the sample size is large (2.1.2) has approximated χ^2 distribution with degree of freedom equals to the number of parameters being estimated (Harrel 2002).

The no simple effect of covariate of $\mathbf{Y}$ and $\mathbf{X}$ is specified by the hypothesis

$$H_0 : \beta_2 = \beta_3 = 0. \tag{2.1.3}$$

To test (2.1.3) using a likelihood ratio statistic, one can first fit two separate models M_0 and M_1. Define l_0 and l_1 as the log likelihoods under the models M_0 and M_1 respectively, where the model M_0 as $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_1\mathbf{X}_2$ and the model M_1 as $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X}_1$. The hypothesis (2.1.3) is tested by using statistic $LR_{stat} = -2(l_0 - l_1)$.

b) **Likelihood Ratio Test for the Effect of Covariate on Discrete Response Variables (Deviance Test)**

It is well known that when a response variable is nominal or ordinal, a traditional regression model will not apply. The generalized linear model has widely been used for the analysis of categorical responses. The proposed method will accommodate not only

continuous response variables but also categorical responses. In this section we will discuss the likelihood ratio test when the response variable is discrete. Let the distribution of random variables $Y_1, \cdots Y_N$ come from an exponential family, i.e.

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}$$

where θ_i and ϕ are parameters and $a_i(\phi), b(\theta_i)$ and $c(y_i, \phi)$ are known functions.

When Y_i comes from an exponential family, then

$$\begin{aligned} E(Y_i) &= \mu_i = b'(\theta_i) \\ \text{Var}(Y_i) &= \sigma_i^2 = b''(\theta_i)a_i(\phi) \end{aligned}$$

A generalized linear model uses a monotone link function g such that

$$g(\mu_i) = x_i^T\boldsymbol{\beta} = \eta_i, \tag{2.1.4}$$

where x_i is a p x 1 vector of explanatory variables and $\boldsymbol{\beta}$ is a p x 1 vector of parameters. The parameters $\boldsymbol{\beta}$ can be estimated by the maximum likelihood estimation method which then can be obtained by an iterative weighted least squares procedure.

The likelihood ratio test that is used to test the parameter $\boldsymbol{\beta}$ is called a deviance test. The deviance test is typically used to compare two nested models and therefore can be used to test for the significance of parameters. Define the likelihood ratio

$$\lambda = \frac{L(\mathbf{b}_{max}; \mathbf{y})}{L(\mathbf{b}; \mathbf{y})},$$

where $\mathbf{b}_{max}$ is the maximum likelihood estimator for $\boldsymbol{\beta}_{max}$ under a saturated model or full model, and $\mathbf{b}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$ under any other model. With the same assumed distribution and link function, the $L(\mathbf{b}_{max}; \mathbf{y})$ will be larger than any other likelihood function. The likelihood ratio λ can be used as a tool to test the goodness of fit for the model. However, the log likelihood of λ defined in (2.1.5) is more widely used in

practice.

$$\log \lambda = l(\mathbf{b}_{max}; y) - l(\mathbf{b}; \mathbf{y}). \tag{2.1.5}$$

In Dobson (2002), the expression 2 log λ was called a deviance. The deviance can be used in hypothesis testing for checking whether the alternative model M_1 fits better than the null model M_0. The hypotheses are written as; $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = [\beta_1 \cdots \beta_q]^T$ and $H_a : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = [\beta_1 \cdots \beta_p]^T$ where $q < p < N$. The difference of two deviances from the null model and the alternative is used to test H_0 against H_a, i.e

$$\begin{aligned} \Delta D &= D_0 - D_1 \\ &= 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})] - 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_1; \mathbf{y})] \\ &= 2\left[l(\mathbf{b}_1; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})\right]. \end{aligned} \tag{2.1.6}$$

McCullagh and Nelder (1993) discussed that the exact sampling distribution of (2.1.6) is not available except in the Normal-theory linear model and certain special cases including simple design from exponential and inverse Gaussian distribution. Dobson (2002) stated that if both models fit the data well, $D_0 \sim \chi^2(N - q)$ and $D_1 \sim \chi^2(N - p)$. Thus the estimated sampling distribution for ΔD is $\chi^2(p-q)$. A large value of ΔD indicates that the model under the alternative is preferred. However, McCullagh and Nelder (1993) commented further that the χ^2 approximation is not very good even when $n \to \infty$. In the case when the response is normally distributed or comes from other distributions with nuisance parameters that are not estimated, the data may not fully estimate the deviance. To eliminate the nuisance parameter, the test statistic

$$F = \frac{D_0 - D_1}{p - q} \Big/ \frac{D_1}{N - p} \tag{2.1.7}$$

is used instead of ΔD. When the null hypothesis is correct F will be approximated by a central $F(p - q, N - p)$ distribution. Thus, a large F value indicates that H_0 is not correct.

In R, the procedure is performed using the drop1 function after fitting the two generalized

linear models. Basically, the function drop1 will calculate the difference of the deviance from the null model and the alternative model. For instance, to test for no covariate treatment interaction, drop1 will calculate the difference of the deviance for models containing the interaction term with that of no interaction term. In addition to the p values calculated from the deviance test when each term is eliminated, the Akaike's information criterion (AIC) value is also provided. So, these AIC values are compared further to find which model is more appropriate.

To test the simple effect of covariate on the response when the response is discrete, proceed in a manner similar to part a) above except that the link function is modeled as a linear function of $\mathbf{X}$ instead of $E(\mathbf{Y}|\mathbf{X})$ using (2.1.7).

Although the likelihood ratio test for GLM or GLMz can be used to test the dependency of two variables, the performance of the LRTs is good only when (2.1.1) and (2.1.4) are satisfied. For continuous response variables the LRT test would not be able to powerfully detect any nonlinear contribution of X on the conditional mean of Y. A similar situation applies to the deviance test, i.e., the formulation is not general enough to detect the dependency of the conditional mean of $\mathbf{Y}$ on $\mathbf{X}$ if there is no linear relationship between $g(.)$ and $\mathbf{X}$.

2.1.2 Wald Type Test for Discrete Response Variable

In this subsection, we will briefly describe a Wald type test in the context of the generalized linear model in (2.1.4). The test will be used in the simulation studies to compare with the performance of our proposed test for Bernoulli response. A Wald statistic using the maximum likelihood estimate for the model in (2.1.4) is

$$(\mathbf{b}-\boldsymbol{\beta})^T \mathtt{I}(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta}), \tag{2.1.8}$$

where $\mathbf{b}$ is the maximum likelihood estimate for β and $\mathtt{I}$ is the information matrix. The statistics (2.1.8) can then be used to test hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = [\beta_1 \cdots \beta_p] = \mathbf{0}$ and $H_a : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. The asymptotic sampling distribution of (2.1.8) is $\chi^2(p)$. In the case where

the response is normally distributed, (2.1.8) is an exact result.

2.1.3 General Additive Models (GAM)

General additive models can be used as an alternative to GLM or GLMz. They allow a nonlinear dependency of the conditional mean of the response variable on the predictor variables. Hastie and Tibshirani (1990) discuss the theory and applications of GAM, while Venables and Ripley (1997) discuss how the GAM are implemented using S-Plus package called "gam" which is also available in R. In this section we discuss briefly the GAM to test the independence of response with explanatory variables for both cases of continuous and discrete response variables. Hastie and Tibshirani (1990) stated that the goal for the additive model is to generalize the GLM in the sense that the GAM will fit a general model not necessarily linear to the data. The idea is to let the data dictate the relationship of response variable and the explanatory variable. The general additive model (GAM)is defined as:

$$\mathbf{Y} = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{2.1.9}$$

where X_j's are independent of ϵ, $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The function f_j's are assumed to be smooth and could be estimated by a "scatter plot smoother" (Hastie and Tibshirani 1990). These $\sum_{j=1}^{p} f_j(X_j)$ are viewed as the estimates for p-variate response surfaces. There are many ways of estimating these f_j's. One of the ways is to estimate each function by an arbitrary smoother (Hastie and Tibshirani 1990). These smoothers include cubic smoothing splines, locally-weighted running-line, kernel and loess. However Venables and Ripley (1997) commented that these methods are very computer intensive.

The independence of response and covariate variables can be tested by an approach as in LRT except that the estimates of the parameters are obtained from GAM methods which may not involve the maximum likelihood estimation method. When the f_j's are linear, a least square method is used for estimation. In R the additive model is performed by the function "gam" which is included in the package "gam". The GAM can also be used for

discrete response variables.

Hastie and Tibshirani (1990) discussed how to fit the additive models to the data with the backfitting algorithm below;

Backfitting Algorithm

(i) Initialize: $\alpha = \text{ ave}(y_i), \quad f_j = f_j^0, \ j = 1, \ldots, p$

(ii) Cycle: $j = 1, \ldots, p, 1, \ldots, p, \ldots$

$f_j = S_j(\mathbf{y} - \alpha - \sum_{k \neq j} \mathbf{f}_k | x_j)$

(iii) Continue (ii) until the individual functions do not change.

The function $S_j(\mathbf{y}|\mathbf{x}_j)$ is the smooth function of $\mathbf{y}$ on $\mathbf{x}_j$. Basically, in the backfitting algorithm, we fit a smooth function to the residual $Y - \alpha - \sum_{k \neq j} \mathbf{f}_k | x_j)$ against x_j using scatterplot smoother. It is repeated until f_j does not change. Penalized smoothing spline (Wood 2000, 2008) can also be used to estimate the f_j.

Instead of equation (2.1.9), the alternating conditional expectation method (ACE) considers

$$\theta(Y) = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

where θ is an invertible smooth function. ACE based the choice of θ and f_j by maximizing the correlation between $\alpha + \sum_{j=1}^{p} f_j(X_j)$ and θ. The ACE algorithm was based on Breiman and Friedman (1985).

2.1.4 Drop test

From a nonparametric approach, robust nonparametric methods discussed by McKean and Schrader (1980) (see also Hettmansperger and McKean (1998) sec 3.6) can be used as an alternative to the traditional linear model. Therefore, the methods can accommodate the existence of treatment effects in the models. Terpstra and Mckean (2005) discussed a few rank-based linear model approaches and provided the R code for some of the techniques discussed.

These R codes can be downloaded from http://www.stat.wmich.edu/mckean/HMC/Rcode. These rank-based approaches are based on weighted Wilcoxon procedures. A robust ANCOVA model can be performed by a "*drop in dispersion test*" (Terpstra and Mckean 2005) which they also called a drop test.

Briefly discussed here is the procedure for the drop test as explained in Terpstra and Mckean (2005). The notation also follows from Terpstra and Mckean (2005). The general linear model is written as;

$$Y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + e_i \tag{2.1.10}$$

where Y_i is the ith response observation and $i = 1, \ldots, n$. $\mathbf{x}_i$ denote a p x 1 vector of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a p x 1 vector of regression parameter and β_0 is the intercept parameter. For the rank-based analyses of linear models, the ϵ_i are independent and identically distributed (iid) from a continuous distribution function F such that $F(0) = 1/2$ and the corresponding density function f such that $f(0) > 0$.

Terpstra and Mckean (2005) stated that when F deviates from normal distribution, the Wilcoxon procedure outperformed the least square procedures (LS). The Wilcoxon procedure basically estimates $\boldsymbol{\beta}$ by finding the solution that minimizes the dispersion function;

$$D_R(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[R[\epsilon_i(\boldsymbol{\beta})] - \frac{n+1}{2} \right] \epsilon_i(\boldsymbol{\beta}), \tag{2.1.11}$$

where $\epsilon_i(\boldsymbol{\beta}) = Y_i - \mathbf{X}_i^T \boldsymbol{\beta}$ and $R[\epsilon_i(\boldsymbol{\beta})]$ denotes the rank of $\epsilon_i(\boldsymbol{\beta})$ among $\{\epsilon_j(\boldsymbol{\beta})\}$.

According to Terpstra and Mckean (2005), instead of minimizing (2.1.11), an alternative objective function

$$D_{WR}(\boldsymbol{\beta}) = \sum_{1 \le i < j \le n} b_{ij} |\epsilon_j(\boldsymbol{\beta}) - \epsilon_i(\boldsymbol{\beta})| \tag{2.1.12}$$

is minimized to accommodate the possibility that the independent variable comes from an observational study and might be contaminated. This is the case since the Wilcoxon procedure is robust only in regard to a response variable. In (2.1.12), b_{ij} is the weight in

the i, j comparison. Hence the solution that minimized (2.1.12) is called a WW-estimate. Hettmansperger and McKean (1998) showed that when $b_{ij} = 1$ for $i \neq j$ and 0 otherwise, $D_{WR}(\boldsymbol{\beta}) = 2D_R(\boldsymbol{\beta})$.

According to Terpstra and Mckean (2005), the WW estimate is computed by using the L_1 regression routine by treating $b_{ij}(Y_j - Y_i)$ and $b_{ij}(\mathbf{X}_j - \mathbf{X}_i)$ as the response variables and design points respectively. The package **quantreg** written by Roger Koenker is used to calculate the WW estimate because the L_1 regression estimate is equivalent to quantile regression estimates implemented in the **quantreg** package. The quantiles regression was introduced by Koenker and Basset (1978).

Testing regarding the parameter $\boldsymbol{\beta}$ is done by the hypotheses;

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0} \text{ versus } H_0 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{0}, \tag{2.1.13}$$

where $\mathbf{A}$ is q x p matrix of full row rank. The drop in dispersion test statistic

$$SRD = \frac{\sqrt{12}}{n\hat{\tau}}[D_{WR}(\hat{\boldsymbol{\beta}}_r) - D_{WR}(\hat{\boldsymbol{\beta}}_f)] \tag{2.1.14}$$

is used to test (2.1.13). In (2.1.14), $\hat{\tau}$ is the consistent estimator of τ, where $\tau = \{\sqrt{12}E[f(\epsilon_1)]\}^{-1}$. In Terpstra and Mckean (2005), $\tau = 1/2$. The $\hat{\boldsymbol{\beta}}_r$ is the Wilcoxon (WIL) estimate for $\boldsymbol{\beta}$ in the reduced model and the $\hat{\boldsymbol{\beta}}_f$ is the Wilcoxon (WIL) estimate for $\boldsymbol{\beta}$ in the full model.

Then, $SRD \xrightarrow{d} \sum_{i=1}^{q} \lambda_i \chi_i^2$ where $\lambda_1, \lambda_2, \ldots, \lambda_q$ are q positive eigenvalues of $\mathbf{V}(C^{-1} - C^{+})$, where

$$C^{+} = \begin{bmatrix} C_r^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

and $C = lim_{n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is n x n matrix whose elements are

$$w_{ij} = \begin{cases} -\frac{1}{n}b_{ij} & ; i \neq j \\ \frac{1}{n}\sum_{k=1}^{n} b_{ik} & ; i = j. \end{cases}$$

Hettmansperger and McKean (1998) suggested bootstrapping or simulation to find the p-

value for test statistic (2.1.14). Because the drop test only considers Wilcoxon weight $b_{ij} = 1$, all the q eigenvalues are all equal to 1. Thus the limiting distribution for SRD is $\chi^2(q)$. However Terpstra and Mckean (2005) showed in the simulation studies that the test that rejects H_0 if $F_R = SRD/q > F_{1-\alpha}(q, n-p-1)$ is a better test for the testing hypotheses (2.1.13).

2.2 Testing Relationship of Two Variables without Incorporating Treatment in the Model

This section describes some of the methods that could be used to test the independence of two variables but without incorporating the treatment effect in the model, i.e. treatment level $a = 1$. These methods include parametric approaches and nonparametric approaches.

2.2.1 Goodness of Fit Tests

Goodness of fit tests could be performed to assess the relationship of two variables. Traditionally, in order to investigate whether there exists a relationship between two variables one would create a scatter plot, followed by the traditional parametric regression to analyze the data. A goodness of fit test then is used to test the fit of the postulated regression model. However, the existence of treatment in the model will be overlooked if the goodness of fit test for regression is to be used to see whether the two variables are related or not.

There is already much literature discussing goodness of fit; among them, Eubank and Hart (1992), Muller (1992), Azzalini and Bowman (1993), Hardle and Mammen (1993), Dette and Munk (1998), Dette (1999), Akritas and Papadatos (2004).

Consider the regression model used in Eubank and Hart (1992):

$$Y_j = g(x_j) + \epsilon_j, \qquad j = 1, \ldots, n, \tag{2.2.1}$$

where $(x_1, Y_1), \ldots, (x_n, Y_n)$ are the observed data, $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$ are the fixed design points, ϵ_j's are i.i.d random variables such that $E(\epsilon_j) = 0$ and $\mathrm{Var}(\epsilon_1) = \sigma^2$.

Assume that $g(x) = \sum_{j=1}^{p} \beta_j t_j(x)$ for all $x \in [0,1]$. Those goodness of fit tests make use of both parametric regression and nonparametric regression. The nonparametric regression approaches used in the above mentioned literature directly or indirectly deal with some smoothing parameter or bandwidth if using kernel density method, thus are generally computationally intensive. Testing if the conditional mean of Y depends on X in the model (2.2.1) is done by testing $H_0 : g(\cdot) = \beta$. Basically, the goodness of fit test is a special case of GAM where p=1. However, this dissertation does not compare the proposed test with the goodness of fit test.

2.2.2 Mutual Information (MI)

An approach that uses directly the concept that two variables are independent when the joint density function of the two variables is a product of the marginal density functions is a mutual information (MI) measure. D'haeseleer et al. (1998) and Butte and Kohane (2000) consider using MI as a dependency measure. This measure is used to detect a more general relationship between two variables including cases of discrete response variable.

For two continuous random variables X and Y, MI is defined as

$$I(X;Y) = \int_x \int_y f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy,$$

where $f(x,y)$ is the joint density of X and Y and $f(x)$ and $f(y)$ are the marginal densities of X and Y respectively. Note that X and Y are independent if and only if $I(X;Y) = 0$. Furthermore, the higher the MI the more closely the two variables X and Y are associated with one another. In order to estimate MI, the marginal density of X and Y and the joint density of X and Y will have to be estimated. Kraskov et al. (2004) discussed methods of estimating MI based on k^{th} nearest neighbor statistics. Another common method to estimate MI is to use the density functions by kernel density estimators discussed in Steuer et al. (2002). Because estimating MI entails estimating density functions, it is cost ineffective. In addition to estimating the density functions, there is no theory available for the distribution of the estimated MI. Furthermore, it is not clear how to extend the approach of MI measures

to include the presence of treatment effects in the model. Test procedures developed for the dissertation not only eliminate the need to estimate the probability density function (as in MI) but also derive the asymptotic distribution of the test statistics and take into account the existence of treatment effects in the model.

2.2.3 Pearson's Correlation, Spearman's ρ and Kendal's τ

A correlation based approach such as Pearson's correlation, Spearman's ρ and Kendal's τ could also be used to assess a monotone relationship of two variables. However, these methods do not incorporate any treatment effect in the models. Although the Spearman's ρ and the Kendal's τ do not assume a specific distribution assumption for the variables, the simulation study shows that these two tests do not perform well when the relationship of Y and X is quadratic.

2.3 Analysis of Covariance (ANCOVA)

A more common model that could be used as a basis of detecting the relationships of two variables in the existence of treatments in the model is the traditional analysis of covariance (ANCOVA). For a response variable Y and a covariate variable X, the traditional ANCOVA model is $y_{ij} = \mu + \mu_i + \beta_i\, x_{ij} + \epsilon_{ij}$, where $\epsilon_i \sim N(0, \sigma^2)$ and $i = 1, \ldots, a$ and $j = 1, \ldots, n_i$. In this model the slope in each treatment group is allowed to differ. In this dissertation, the aforementioned method is labeled as a CF test. Standard assumptions of ANCOVA model include homogeneity of variances and normality in the error term. The hypothesis $H_{0v} : \beta_1 = \cdots = \beta_a = 0$ could be used to test for no simple covariate effect. When H_{0v} is true, the conditional mean of the response variable Y does not vary linearly with the covariate. Clearly the ANCOVA model could be used as a test for the existence of linear relationships of two continuous variables but is not suitable for testing general association. In this dissertation, the test of no main covariate and no covariate-treatment interaction effects for the ANCOVA model is compared to that of the proposed test in Chapter 4 and Chapter 5. Because the ANCOVA model is a parametric approach, its inferences depend

on the satisfaction of assumptions such as constant variance of error. Often, the data do not satisfy these parametric assumptions. The proposed method should be free from any specific parametric distribution assumptions.

2.4 Nonparametric Methods for ANCOVA Model

In general, a test of association between two variables in the presence of treatment can be formulated by a one-way ANCOVA model. Because the proposed test is developed under a nonparametric model, which makes use of ANCOVA setting, this Section will discuss a few nonparametric methods and some factors that motivate the construction of the proposed nonparametric method.

2.4.1 Fully Nonparametric (FNP) Model

The proposed tests for general dependency between two variables in the presence of treatment effects could be formulated under the fully nonparametric model (FNP) initially used for nonlinear analysis of covariance introduced by Akritas et al. (2000). Let (X_{ij}, Y_{ij}), $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, denote the set of covariates and responses from the j-th observation of the i-th group. The FNP model assumes that the conditional distribution of Y_{ij} given $X_{ij} = x$ depends on the treatment group i and the covariate value x, i.e.,

$$Y_{ij}|X_{ij} = x \sim F_{ix}(y). \tag{2.4.1}$$

Akritas et al. (2000) defined the model (2.4.1) to be completely nonparametric because there is no specification how the $F_{ix}(y)$ changes for any i and x. This model can be used to test the independence of Y and X because if Y is independent of X the conditional distribution of Y given $X = x$ will not depend on X. Akritas et al. (2000) further discussed that for an arbitrary cumulative distribution of X, $G_X(x)$, sets

$$\overline{F}_{i\cdot}(y) = \int F_{ix}(y) dG_X(x), \text{ and } \overline{F}_{\cdot x}(y) = \frac{1}{a}\sum_{i=1}^{a} F_{ix}(y). \tag{2.4.2}$$

They suggested the following possible hypotheses of interest:

$$\text{No main treatment effect , or } \overline{\mathrm{F}}_{\mathrm{i}\cdot}(\mathrm{y}) \text{ do not depend on } i; \tag{2.4.3}$$

$$\text{No main covariate effect , or } \overline{\mathrm{F}}_{\cdot\mathrm{x}}(\mathrm{y}) \text{ do not depend on } x; \tag{2.4.4}$$

$$\text{No treatment covariate interaction effect , or } \mathrm{F_{ix}(y)} = \overline{\mathrm{F}}_{\mathrm{i}\cdot}(\mathrm{y}) + \mathrm{K_x(y)}; \tag{2.4.5}$$

$$\text{No simple treatment effect , or } \mathrm{F_{ix}(y)} \text{ do not depend on } i; \tag{2.4.6}$$

$$\text{No simple covariate effect , or } \mathrm{F_{ix}(y)} \text{ do not depend on } x, \tag{2.4.7}$$

where $K_x(y)$ in (2.4.5) is a function independent of i. The hypotheses (2.4.3) and (2.4.6) were considered by Wang and Akritas (2006). Akritas, Antoniou, and Wang (2003) considered testing (2.4.3) and (2.4.5), while Wang and Akritas (2006) tested (2.4.4) and (2.4.5).

Furthermore, Akritas et al. (2000) decomposed the conditional cumulative distribution function into

$$F_{ix}(y) = M(y) + A_i(y) + D_x(y) + C_{ix}(y). \tag{2.4.8}$$

where

$$M(y) = a^{-1}\sum_{i=1}^{a} \overline{F}_{i\cdot}(y), \qquad A_i(y) = \overline{F}_{i\cdot}(y) - M(y), \qquad D_x(y) = \overline{F}_{\cdot x}(y) - M(y),$$

$$\text{and} \qquad C_{ix}(y) = F_{ix}(y) - M(y) - A_i(y) - D_x(y).$$

Note that this notation is similar in Wang and Akritas (2006). In this decomposition, similar to the traditional ANCOVA model, $A_i(y)$ is the nonparametric covariate-adjusted main effect of treatment group i, $D_x(y)$ is the nonparametric main effect of the covariate value x, and $C_{ix}(y)$ is the nonparametric interaction effect between treatment group i and

covariate value x. From the above decomposition, the hypotheses

$$H_0(A) : A_i(y) = 0 \text{ for all x and all y} \tag{2.4.9}$$

$$H_0(D) : D_x(y) = 0 \text{ for all x and all y,} \tag{2.4.10}$$

$$H_0(C) : C_{ix}(y) = 0 \text{ for all x and all y,} \tag{2.4.11}$$

$$H_0(A + C) : A_i(y) + C_{ix}(y) = 0 \text{ for all x and all y,} \tag{2.4.12}$$

$$H_0(B) : D_x(y) + C_{ix}(y) = 0 \text{ for all x and all y,} \tag{2.4.13}$$

can also be used to test the hypotheses (2.4.3), (2.4.4), (2.4.5), (2.4.6) and (2.4.7) respectively.

Akritas et al. (2000) proposed a test for nonlinear higher-way ANCOVA under the FNP model. The hypotheses discussed were: no main treatment effect, no simple treatment effect and no interaction between treatments effects adjusted for covariate. The test developed in Akritas et al. (2000) can be used as an alternative to the classical ANCOVA because it allows nonlinear relationships between response and covariate and is completely nonparametric. Hence it does not require the assumptions needed for classical ANCOVA. However Akritas et al. (2000) did not discuss how this FNP model can be used in testing the independence of response and covariate adjusted for treatments, i.e. testing for no simple covariate effect adjusted for treatment.

2.4.2 FNP Model in Higher-way ANCOVA with at Most Three Covariates

The methodology of Akritas et al. (2000) is extended to include two and three covariates in Tsangari and Akritas (2004). Tsangari and Akritas (2004) also discussed the same hypotheses as in Akritas et al. (2000) with the addition of two and three covariates. Akritas et al. (2000) and Tsangari and Akritas (2004) approaches rely on consistent estimation of the conditional distribution function of the response given covariate values using a Nadaraya-Watson kernel estimator. In addition to the difficulty in determining the window bandwidth

k, they require the number of observations in each window to go to infinity. However, using such bandwidth also entails the number of observations per window to be of order $k(k^{-1/4})^p$, where p is the number of covariates in the model (Tsangari and Akritas 2004), see also Akritas, Antoniou, and Wang (2003). This term goes to infinity unless there are no more than three covariates in the model and therefore the method cannot be extended to handle cases with more than three covariates. In practice, many analyses require the use of multiple factors and multiple covariates.

2.4.3 FNP Model in Testing No Main Covariate and No Main Covariate-treatment Interaction Effects.

Wang and Akritas (2006) proposed a test for nonparametric no main covariate effect and no covariate-treatment interaction effect adjusted for treatment using the FNP for one-way ANCOVA model. The test in Wang and Akritas (2006) was developed based on treating a covariate as a factor with a large number of levels thus changing the setting of a one-way ANCOVA model into a two-way hypothetical ANOVA model. With this modification, the setting induces at most one observation in each cell, thus the need to introduce pseudo-replicates. The simple one-way ANCOVA design can now be treated as a two-way ANOVA design with one observation per cell. The pseudo-replicates are created within each cell in the two-way hypothetical ANOVA model. This method of creating pseudo-replicates is used both in Akritas et al. (2003) and Wang and Akritas (2006). First pool all the covariate values X_{ij} and put them in ascending order and relabel them as: $X_1, X_2, \ldots, X_N$, where $N = \sum_{i=1}^{a} n_i$. These ordered pooled covariate values act as a factor with levels $X_1, X_2, \ldots, X_N$ in the hypothetical two-way ANOVA design.

To create some replications in the analysis, a window or cell C_{ic} of size k centered at X_c is created, where $c = 1, 2, \cdots, N$. The window C_{ic} will consist of k paired observations (X_{ij}, Y_{ij}) whose covariate values X_{ij} are closest to X_c in ranks, among $X_{i1}, \cdots, X_{in_i}$ such

that Y_{ij} will be in the C_{ic} window if the corresponding X_{ij} satisfy

$$|\widehat{F}_{X,i}(X_{ij}) - \widehat{F}_{X,i}(X_c)| \leq \frac{k-1}{2n_i},$$

where $\widehat{F}_{X,i}(x)$ is the empirical cumulative distribution function of $\mathbf{X}$ in the ith group. If k is odd, then the window is symmetric about the center. If ($X_c \in$ group i) the window C_{ic} will include (X_c, Y_{ij}) whereas if ($X_c \notin$ group i), the window C_{ic} will be centered at X_{ij}, which is closest to X_c.

To differentiate between the original observation Y_{ij} and the observation in the hypothetical ANOVA, label the t-th observation in the (i, c) of the hypothetical ANOVA by U_{ict}. Note that after the augmentation, U_{ict}'s are not independent.

The test statistics in Wang and Akritas (2006) are constructed in a manner similar to the traditional ANOVA model. Define

$$MST_D(\mathbf{U}) = ak(N-1)^{-1} \sum_{c=1}^{N} \left(\overline{U}_{\cdot c \cdot} - \overline{U}_{\cdots}\right)^2 \tag{2.4.14}$$

$$MST_C(\mathbf{U}) = k(a-1)^{-1}(N-1)^{-1} \sum_{c=1}^{N} \sum_{i=1}^{a} \left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot} - \overline{U}_{\cdot c \cdot} + \overline{U}_{\cdots}\right)^2, \tag{2.4.15}$$

$$MSE(\mathbf{U}) = \{Na(k-1)\}^{-1} \sum_{i=1}^{a} \sum_{c=1}^{N} \sum_{t=1}^{k} \left(U_{ict} - \overline{U}_{ic\cdot}\right)^2. \tag{2.4.16}$$

Wang and Akritas (2006) studied the asymptotic distribution of

$N^{1/2}k^{-1/2}T_D = N^{1/2}k^{-1/2}(MST_D - MSE)$ and $N^{1/2}k^{-1/2}T_C = N^{1/2}k^{-1/2}(MST_C - MSE)$ which are used to test the hypotheses (2.4.10) and (2.4.11) respectively. In this dissertation, these tests are referred to as WA tests. The theory behind the result in Wang and Akritas (2006) assumes that the number of pseudo-replicates k, used in each cell (window) goes to infinity as the total number of pooled covariate values N becomes large. Specifically, their test statistics have a standardizing rate of $N^{1/2}k^{-1/2}$. Even though Wang and Akritas (2006) developed asymptotic distribution for testing no main covariate and no main covariate-treatment interaction effect, their simulation studies only show the observed type I error and power for testing no covariate-treatment interaction effect, not for test of no main

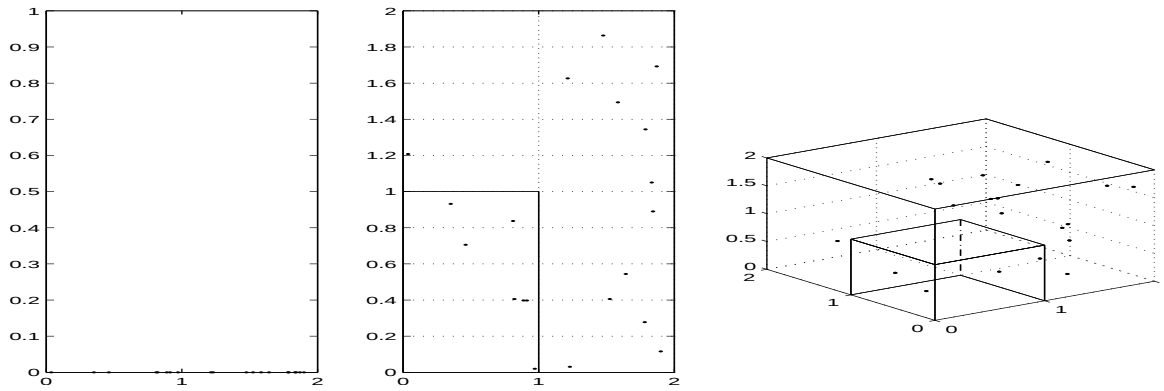
covariate effect.

Their simulation studies show that the performance of the WA test depends on the window size being used and the number of covariate values in each treatment group, denoted as n_i. From Table 2 of Wang and Akritas (2006) it is seen that the observed type I error is liberal for small k and as the window size increases the observed type I error rate seems to be conservative. For example, when $n_i = 30$ and $k = 3$, the observed type I error rate is 0.089 and when $k = 7$, the observed type I error rate is 0.042. Another drawback associated with the WA test is that the requirement of the number of pseudo-replicates increases with n_i, thus can escalate the computation time dramatically as shown in the simulation studies in Section 4.5.

Furthermore, with multiple covariates Wang and Akritas (2006) suggested to proceed the same way as in a single covariate case by treating the covariate factor as a whole. However, the covariate effect of interest to an applied researcher is often a specific covariate rather than the combined values of all the covariates. In addition, Wang and Akritas (2006) suggest constructing nearest neighbors windows in the same way as with the single covariate case, which augments the window with pseudo-replicates of size going to infinity. With high dimensional setting, as the number of covariates in the data set increases, the points spread out with additional dimensions that make the sample space sparse and this makes it impractical to use the tools that require window size to go to infinity (Parson et al. (2004)). This phenomenon is called the curse of dimensionality. Figure 2.1 illustrates this phenomenon where increasing dimensions cause a decrease in the number of data-points to be captured in a unit line, square or cube.

Hastie et al. (2001) discussed further the curse of dimensionality problem. Suppose a nearest neighbor approach was used to capture r percent of the observation from a uniformly distributed in a p-dimensional unit hypercube. Then, the expected edge length is $e_p(r) = r^{1/p}$. When p=10, $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.8$ with total range of each input is 1.0. This means that to capture 1% or 10% of the data to form a local average, 63% or 80% of each input variable must be used and such a neighborhood is not local anymore. Thus, high dimensional analysis is susceptible to the problem of curse of dimensionality.

Figure 2.1: Curse of Dimensionality



More importantly, besides the unsuitability of the WA test for high dimensional data, Wang and Akritas (2006) only derived tests for no main covariate and no main covariate-treatment interaction effects separately. They did not discuss a test for no simple covariate effect adjusted for treatment.

2.4.4 Further Comments on Wang and Akritas (2006)'s

Wang and Akritas (2006) stated that the performance of their test is generally good when the window size is $\sqrt{N}$ but should be of order $O(N^{3/5})$. However, the simulation studies performed in Wang and Akritas (2006) paper are limited to the use of smaller window sizes than the recommended value of $\sqrt{N}$. For example, in Table 2 of Wang and Akritas (2006), when $n_i = 30$ which corresponds to $N = 60$ for $a = 2$, the largest window size that is used in the simulation is 7, while the recommended window size is between 7.7 and 11.7. Because the assumption used for the theory is large k, we expect that the WA tests use the larger window sizes or at least the upper bound of the specified value. If Wang and Akritas (2006) included window sizes larger than 7 the power of the test would be smaller than 0.742 based on the pattern displayed in the table. Similarly with $n_i = 50$, $N = 100$ and $a = 2$, the largest window size used is 9, while the recommended window size is between 10 and 15.8. The power of the WA test seems to decrease when the window size is increased.

Chapter 3

Method of Detecting Dependency of Two Variables in the Presence of Treatment Effect

This chapter discusses the construction of the proposed nonparametric test of independence between two variables after adjusting for heteroscedastic treatment effects. Some applications and simulation studies are presented. The content of this chapter along with some additional materials are published in the Canadian Journal of Statistics, Volume 38, Issue 3, 2010, pages 408-433.

3.1 Introduction

Correlation based approaches such as Pearson's correlation, Spearman's ρ, or Kendall's τ evaluate linear relationship between two variables without accounting for the effect of factors. The method of alternating conditional expectations (ACE, Breiman and Friedman 1985) is an extended correlation approach that transforms both the response and covariate to achieve maximum correlation. Examples that allow hypothesis testing include likelihood methods from linear or generalized linear models, drop test (Terpstra and Mckean 2005), generalized additive models (GAM) using a smoother such as spline or loess (Hastie and Tibshirani 1990), or penalized smoothing spline (Wood 2000, 2008).

These approaches have provided flexible tools to discover the dependency between variables. However, there are often practical data that do not satisfy the assumptions required by these methods. For example, correlation based approaches typically are not sensitive enough to pick up nonlinear dependence; likelihood based methods are restrictive to the distributional assumptions; ACE assumes that conditional on the transformed covariates, the transformed response variable follows a normal distribution with constant variance; the GAM approaches are only applicable to exponential families and outliers can seriously distort the transformations leading to inaccurate inference. In a particular example (see the EFT study in subsection 3.3.2), all these methods except ACE found a significant relationship between the response and covariate with an outlier (influential observation) in the data and producing a contrary result when the outlier is replaced by a median response. A simulation study discovered that the type I error rate at level 0.01 in the presence of outliers produced from mixture distributions with a lognormal component is as high as 0.206 for the GAM methods, and 0.748 for the correlation based approaches. Robust methods valid for distributions beyond an exponential family that are resistent to outliers while maintaining high power to detect nonlinear dependence are developed here.

Whether two variables are independent or not is inherently defined through distribution functions. One may consider using mutual information (MI) as a dependency measure (D'haeseleer et al. 1998; Butte and Kohane 2000). The MI measures the expected distance (under the joint distribution) between the log of the joint probability density function (pdf) and the log of the product of the marginal density functions. It equals zero if and only if the variables are independent. Before the MI can be used, the joint and marginal pdfs need to be estimated from the same set of data. In addition, there is no MI theory available to determine the threshold for significance of the dependence. Other directions for testing mutual independence without estimating the pdfs are through combinations of asymptotically independent Cramér-von Mises statistics derived from a Möbius decomposition of the empirical copula process (Deheauvels 1981; Genest and Rémilland 2004, and the references therein), or based on a normalized estimated distance between the joint and the marginal characteristic functions. When there are heteroscedastic treatment effects, it is not clear

how to extend these tests to determine independence adjusted for treatment effects.

This chapter presents a nonparametric test that effectively detects general dependence between two variables after adjusting for the heteroscedastic treatment effects. A fixed number of nearest-neighbor pseudo-replicates will be used to augment each treatment level and covariate value combination. Test statistics are constructed by comparing two quadratic forms, both of which are common estimates of linear combinations of the variances and conditional variances. The asymptotic results are obtained under the null hypothesis. Note that the regular standardizing rate for a nonparametric test statistic is N^{α}, where $0 < \alpha < 1/2$. By using a fixed number of nearest-neighbors augmentation, the standardizing rate of the test statistics achieved the rate for parametric analysis $\sqrt{N}$ under the null hypotheses. The empirical studies show that the proposed test maintains intended type I error control while achieving competitive or better power compared to available methods when the data have a certain chance to have unusual observations from a skewed distribution such as a lognormal distribution.

3.2 Main Results

3.2.1 Construction of Test Statistics

The following notation and conditions will be used throughout this dissertation. Let (X_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, be a random sample from treatment i. Suppose $Y_{ij}|X_{ij} = x \sim F_i(y|x)$ for some unknown conditional distribution function $F_i(y|x)$.

Assume:

- The fourth conditional central moments of Y_{ij} given $X_{ij} = x$ are uniformly bounded for all i and x.

- Let $f_{X,i}(x)$ and $F_{X,i}(x)$ be the marginal density and distribution functions of X_{ij}. Assume $F_{X,i}(x)$ is differentiable at all x.

- Denote $\widehat{F}_{X,i}(x) = n_i^{-1} \sum_{j=1}^{n_i} I(X_{ij} \leq x)$ the empirical distribution of X_{ij}. Assume that $\min_{1 \leq i \leq a} n_i$ and $\max_{1 \leq i \leq a} n_i$ are of the same order. Denote $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{an_a})'$

to be the vector of all covariate values.

Independence between the two variables in all treatments is described in hypothesis:

$$H_0\text{: } F_i(y|x) \text{ does not depend on } x, \text{ for all } i, y. \tag{3.2.1}$$

Note that the difference between this problem and the testing of independence using a single sample from the same distribution is that the data in different treatment levels have different distributions. To effectively use these data, the samples from all treatment levels should contribute to the power of the test and therefore reduce the sample size requirement for each treatment. To achieve this, augment each treatment under the null hypothesis to have more observations using $k-$nearest neighbors. For convenience, take k to be an odd number. Specifically, treatment i and covariate value $X_{i_1 j_1}$ define a cell indexed by (i, c), where $c = \sum_{i'=1}^{i_1} n_{i'} - n_{i_1} + j_1$. In other words, for each i, there are $N = \sum_{i_1=1}^{a} n_{i_1}$ cells as i_1 goes from 1 to a and j_1 goes from 1 to n_{i_1}. The set of indices for the covariate values used in the augmented cell (i, c) is denoted by C_{ic}. Augments cell (i, c) using observations from treatment i as follows.

1. For $i_1 = i$, the cell (i, c) contains (X_{ij_1}, Y_{ij_1}). In addition, select $k-1$ pairs of other observations in treatment i whose covariate values are among the k-closest to X_{ij_1} in rank. That is, (X_{ij}, Y_{ij}) is selected for augmentation of cell (i, c) if and only if $n_i|\widehat{F}_{X,i}(X_{ij_1}) - \widehat{F}_{X,i}(X_{ij})| \leq \frac{k-1}{2}$.

2. For $i_1 \neq i$, i.e., $(X_{i_1 j_1}, Y_{i_1 j_1})$ is not in treatment i. First find the covariate value in treatment i that is closest to $X_{i_1 j_1}$ in absolute difference. Denote X_{ij} to be the closest. Then, select additional $k-1$ pairs of observations in treatment i such that their covariate values are among the k closest to X_{ij} in ranks centered at the rank of X_{ij}. Thus, $(X_{ij'}, Y_{ij'})$ is selected to augment cell (i, c) if $n_i|\widehat{F}_{X,i}(X_{ij'}) - \widehat{F}_{X,i}(X_{ij})| \leq \frac{k-1}{2}$.

The first part is similar to the idea used in the k-nearest-neighbor regression with a single identically distributed independent sample using a special weight function defined through

the ranks of the covariate values. The extra augmentation in the second part is aimed to capture possible dependence of the response variable on the covariate through its interactions with the factor. In both cases, the augmented response values in cell (i, c) are denoted as $U_{ict}, t = 1, \ldots, k$. Note that under the null hypothesis, the distribution of Y_{ij} does not depend on X_{ij}. The k-nearest neighbors are selected based on the covariate values. Therefore, the augmentation simply adds more observations under the null hypothesis. However, under the alternative, the conditional distribution of Y_{ij} depends on X_{ij}. Then such an augmentation will add some observations that increase the between-cell variations. The difference, $B_N - W_N$, between the average between- and within-cell variations for all treatments using the augmented observations is used as a test statistic, where B_N and W_N are defined below with $\overline{U}_{ic\cdot} = k^{-1}\sum_{t=1}^{k} U_{ict}$, $\overline{U}_{i\cdot\cdot} = N^{-1}\sum_{c=1}^{N}\overline{U}_{ic\cdot}$:

$$
\begin{aligned}
B_N &= ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot}\right)^2 \\
&= ka^{-1}(N-1)^{-1}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}\left[k^{-1}\sum_{j=1}^{n_i} Y_{ij} I\left(n_i|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij})| \le \frac{k-1}{2}\right) - \right. \\
&\qquad \left. (Nk)^{-1}\sum_{i_2=1}^{a}\sum_{j_2=1}^{n_{i_2}}\sum_{j=1}^{n_i} Y_{ij} I\left(n_i|\widehat{F}_{X,i}(X_{i_2 j_2}) - \widehat{F}_{X,i}(X_{ij})| \le \frac{k-1}{2}\right)\right]^2, \\
W_N &= \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t=1}^{k}\left(U_{ict} - \overline{U}_{ic\cdot}\right)^2 \\
&= \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}\sum_{j=1}^{n_i}\left[Y_{ij} I\left(n_i|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij})| \le \frac{k-1}{2}\right) - \right. \\
&\qquad \left. k^{-1}\sum_{j_2=1}^{n_i} Y_{ij_2} I\left(n_i|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij_2})| \le \frac{k-1}{2}\right)\right]^2.
\end{aligned}
$$

The idea seems straightforward. However, the technical difficulty is high as the augmented observations $\{U_{ict}, c = 1, \ldots, N, t = 1, \ldots, k\}$ are not independent since the observations are repeatedly used during the augmentation. If k is allowed to go to infinity with N, then techniques from nonparametric smoothing such as kernel regression is borrowed as the k here would play a similar role as the bandwidth for kernel regression. For a fixed finite

k, the inference basically relies on a combination of counting techniques, theory for spacings of order statistics, and some theory for quadratic forms.

3.2.2 Results Under the Null Hypothesis

To obtain the asymptotic distribution of $\sqrt{N}(B_N - W_N)$, first find a projection of it. The standardizing rate is $\sqrt{N}$ where k is finite. Note that even though U_{ict} are independent for different i, Hájek's projection cannot be applied because that projection will not simplify the problem where a is finite. Instead, by denoting $Z_{ict} = U_{ict} - E(U_{ict}|\mathbf{X})$, project B_N onto the space span by the functions of

$$\{\mathbf{Z}_c, c = 1, \ldots, N\} \text{ where } \mathbf{Z}_c = (Z_{1c1}, \cdots, Z_{ack})'. \tag{3.2.2}$$

Note that $\mathbf{Z}_c$, $c = 1, \ldots, N$, are not independent. Hence this projection is not implemented in a traditional sense. Meanwhile, B_N does not have to be centered before the projection as is required in Hájek's projection. Instead, B_N and W_N have the same expectation under the null hypothesis if the cell observations are true replicates. Proceed by partitioning the quadratic form B_N into a major summation over c and another summation over c and c', $c \neq c'$, i.e., under H_0,

$$B_N = P_B(\mathbf{Z}) + S_B(\mathbf{Z}), \text{ where } \mathbf{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_N')',$$

and

$$P_B(\mathbf{Z}) = \frac{k}{aN} \sum_{i=1}^{a} \sum_{c=1}^{N} \overline{Z}_{ic\cdot}^2, \quad S_B(\mathbf{Z}) = -\frac{k}{aN(N-1)} \sum_{i}^{a} \sum_{c \neq c'}^{N} \overline{Z}_{ic\cdot} \overline{Z}_{ic'\cdot}. \tag{3.2.3}$$

Then $P_B(\mathbf{Z})$ is a projection of B_N onto the space in (3.2.2) and $B_N - W_N = (P_B(\mathbf{Z}) - W_N) + S_B(\mathbf{Z}) = T_B + S_B(\mathbf{Z})$, where

$$
\begin{aligned}
T_B &= [a(k-1)N]^{-1} \sum_{i=1}^{a} \sum_{c=1}^{N} \sum_{t \neq t'}^{k} Z_{ict} Z_{ict'} && (3.2.4)\\
&= [a(k-1)N]^{-1} \sum_{i=1}^{a} \sum_{c=1}^{N} \sum_{t \neq t'}^{k} (U_{ict} - E(U_{ict}|\mathbf{X}))(U_{ict'} - E(U_{ict'}|\mathbf{X})) \\
&= [a(k-1)N]^{-1} \sum_{i=1}^{a} \sum_{j \neq j'}^{n_i} (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X})) \sum_{c=1}^{N} I(j \in C_{ic}) I(j' \in C_{ic}) \\
&= [a(k-1)N]^{-1} \sum_{i=1}^{a} \sum_{j \neq j'}^{n_i} (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X})) K_{ijj'}, && (3.2.5)
\end{aligned}
$$

where

$$
K_{ijj'} = \sum_{c=1}^{N} I(j \in C_{ic}) I(j' \in C_{ic}). \tag{3.2.6}
$$

Note that the term in (3.2.4) is closely related to the expected correlation between every pair of response values with a correlation induced by their dependence on $\mathbf{X}$. The $K_{ijj'}$ in (3.2.6) plays the role of a weight function which connects the response locally with the empirical distribution function of X_{ij}. The T_B term in (3.2.4) is more intuitive than $\sqrt{N}(B_N - W_N)$ to evaluate the effect of X_{ij} on Y_{ij}. However, T_B can not be calculated from the sample as $E(Y_{ij'}|\mathbf{X})$ is unknown. On the other hand, $\sqrt{N}(B_N - W_N)$ is directly obtained from the sample.

The following lemma shows that $\sqrt{N} S_B(\mathbf{Z})$ is asymptotically negligible. Derive the asymptotic distribution of $\sqrt{N} T_B$ by showing that it satisfies the conditions for the central limit theorem for clean quadratic forms by de Jong (1987). The result is stated in Theorem 3.2.2.

Lemma 3.2.1. *(Projection of B_N)*
Let $S_B(\mathbf{Z})$ be as defined in (3.2.3). If the assumptions in subsection 3.2.1 are satisfied, then as $N \to \infty$, $\sqrt{N} S_B(\mathbf{Z}) \to 0$ in probability.

Proof of Lemma 3.2.1

It is sufficient to show that $E(\sqrt{N}S_B(\mathbf{Z})) \to 0$ and $\text{Var}(\sqrt{N}S_B(\mathbf{Z})) \to 0$.

$$E(S_B(\mathbf{Z})) = -\frac{k}{aN(N-1)} \sum_{i=1}^{a} \sum_{c \neq c'}^{N} E\left\{E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})\right\}. \tag{3.2.7}$$

Because $E(Y_{ij}^2|X_{ij})$ is uniformly bounded for all i, j, there exists some finite $M_1 > 0$ such that

$$\begin{aligned} |E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})| &\leq \frac{1}{k^2}\sum_{t=1}^{k}\sum_{t'=1}^{k}|E(Z_{ict}Z_{ic't'}|\mathbf{X})| \\ &\leq \frac{1}{k^2}\sum_{t=1}^{k}\sum_{t'=1}^{k}[E(Z_{ict}^2|\mathbf{X})E(Z_{ict'}^2|\mathbf{X})]^{1/2} \leq M_1. \end{aligned} \tag{3.2.8}$$

When the observations in cell (i, c) and cell (i, c') do not have overlap, $E(Z_{ic\cdot}Z_{ic'\cdot}) = E(Z_{ic\cdot})E(Z_{ic'\cdot}) = 0$, giving the following result:

$$\sum_{i=1}^{a}\sum_{c\neq c'}^{N} E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X}) = O\left(\sum_{i=1}^{a}\sum_{c\neq c'}^{N} I(|c'-c| \leq k)E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})\right) = O(N),$$

implying that

$$E(\sqrt{N}S_B(\mathbf{Z})) = O(N^{-1/2}) \to 0 \text{ as } N \to \infty.$$

Next to be shown is that $\text{Var}(\sqrt{N}S_B(\mathbf{Z}))$ goes to 0 as $N \to \infty$. Because $E(\sqrt{N}S_B(\mathbf{Z}))$ goes to 0, it remains to show that $E(\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot})^2/N^3 \to 0$.

$$E(\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot})^2 \leq \sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N}|E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})|$$

$$\leq \sum_{c,c',c_1,c_1'}^{N}|E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})|(2I_1(c,c',c_1,c_1') + 3\ I_2(c,c',c_1,c_1') \tag{3.2.9}$$

$$+3\ I_3(c,c',c_1,c_1') + 4I_4(c,c',c_1,c_1')), \tag{3.2.10}$$

where $I_1(\cdot)$ in (3.2.9) is the indicator function for cases that either c, c', c_1, c_1' fall into three non-overlapping cells where two non-overlapping cells contain one of the $c's$ and one of the

cells contains two members of c, c', c_1, c_1'; $I_2(\cdot)$ in (3.2.9) is the indicator function for cases that c, c', c_1, c_1' are evenly divided into two non-overlapping cells; $I_3(\cdot)$ in (3.2.10) is the indicator function for cases that c, c', c_1, c_1' are in two non-overlapping cells, such that one cell contains three of the $c's$ and the other contains one of the $c's$. Finally $I_4(\cdot)$ in (3.2.10) is the indicator function for cases that c, c', c_1, c_1' are all in the same cell. The expectation in (3.2.9) is zero since the observations in non-overlapping cells are independent. Therefore,

$$\mathrm{Var}(\sqrt{N}S_B(\mathbf{Z})) = \frac{k^2}{a^2N(N-1)^2}E\left(\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right)^2 \leq \frac{k^2}{a^2N(N-1)^2}\{O(N^2)\} = O(N^{-1}),$$

and the proof is completed.

Theorem 3.2.2. *Under H_0 in (3.2.1) and the assumptions given in subsection 3.2.1,*

$$\sqrt{N}(B_N - W_N) \to N(0, \lim_{N\to\infty}\gamma_N^2),$$

where

$$\begin{aligned}\gamma_N^2 &= \frac{4}{Na^2(k-1)^2}\sum_{i}^{a}\sum_{j<j'}^{n_i} E\left\{\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})\left[B_{ijj'}^2 + B_{ijj'}\right.\right.\\ &\quad \left.\left. -2I(j_*' - j_* \leq (k-1)/2)\right]\right\} I(j_*' - j_* \leq k-1) + O(N^{-1}),\end{aligned}$$

with $B_{ijj'} = \sum_{i_1, i_1\neq i}^{a}\left(\frac{n_{i_1}}{n_i}d_{i_1i}(X_{ij}) + 1\right)[k-(j_*'-j_*)]I(j_*'-j_* \leq k-1)$, $d_{i_1i}(x) = f_{X,i_1}(x)/f_{X,i}(x)$ *and* $j_* < j_*'$, *where* j_*, j_*' *are the ranks of* X_{ij} *and* $X_{ij'}$ *among covariate values in treatment* i.

An estimator of $\sigma_i^2(X_{ij})$ is the sample variance $\widehat{\sigma}_i^2(X_{ij})$ calculated using the augmented observations for the cell determined by i and X_{ij}, i.e.,

$$\begin{aligned}\widehat{\sigma}_i^2(X_{ij}) &= \frac{k}{k-1}\left\{\frac{1}{k}\sum_{l=1}^{n_i}Y_{il}^2\ I\left[|\widehat{F}_{X,i}(X_{il}) - \widehat{F}_{X,i}(X_{ij})| \leq \frac{k-1}{2n_i}\right]\right.\\ &\quad \left. - \left(\frac{1}{k}\sum_{l=1}^{n_i}Y_{il}\ I\left[|\widehat{F}_{X,i}(X_{il}) - \widehat{F}_{X,i}(X_{ij})| \leq \frac{k-1}{2n_i}\right]\right)^2\right\}.\end{aligned}$$

The term $\frac{kn_{i_1}}{n_i}d_{i_1i}(X_{ij})$ is estimated by the number of times that (X_{ij}, Y_{ij}) is selected for

augmentation of cell determined by i and $X_{i_1 j_4}$ for all $j_4 = 1, \ldots, n_{i_1}$. That is,

$$\frac{n_{i_1}}{n_i}\widehat{d}_{i_1 i}(X_{ij}) = k^{-1}\sum_{j_4=1}^{n_{i_1}} I\left(|\widehat{F}_{X,i}(X_{ij}) - \widehat{F}_{X,i}(X_{i_1 j_4})| \leq \frac{k-1}{2n_i}\right). \tag{3.2.11}$$

This is because

$$\frac{d_{i_1 i}(X_{ij})}{n_i} = \frac{1}{k}\int I\left(|F_{X,i}(X_{ij} - F_{X,i}(x)| \leq \frac{k-1}{2n_i}\right) dF_{X,i_1}(x) + O_p(N^{-3/2}).$$

Sketch Proof of Theorem 3.2.2

By Lemma 3.2.1, $\sqrt{N}(B_N - W_N)$ has the same asymptotic distribution as $\sqrt{N}T_B$. The asymptotic variance of this statistic is obtained in Lemma 3.2.3. The asymptotic normality for the test statistic is shown here. Let $t^{(2)}_{iji'j'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{i'j'} - E(Y_{i'j'}|\mathbf{X}))K_{ijj'}$, where $K_{ijj'}$ is defined in (3.2.6), and write

$$\sqrt{N}T_B = \frac{\sqrt{N}}{Na(k-1)}\sum_{i,i',j,j'} t^{(2)}_{iji'j'} I(i=i')I(j \neq j') = \sum_{1\leq l_1 \leq N}\sum_{1 \leq l_2 \leq N} V_{l_1 l_2},$$

where $l_1 = l(i,j)$ and $l_2 = l(i,j')$ are defined through a one to one index mapping function

$$l(i,j) = \begin{cases} j & \text{for } i = 1 \\ \sum_{i_2=1}^{i-1} n_{i_2} + j & \text{for } i > 1, \end{cases} \tag{3.2.12}$$

and

$$V_{l_1 l_2} = \begin{cases} \frac{\sqrt{N}}{Na(k-1)}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))\ K_{l_1 l_2} & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.13}$$

Here $K_{l_1 l_2}$ is same as $K_{ijj'}$ but using index l_1, l_2:

$$K_{l_1 l_2} = \begin{cases} \sum_{i_1}^{a}\sum_{j_1}^{n_{i_1}} I(l_1 \in C_{iX_{i_1 j_1}})I(l_2 \in C_{iX_{i_1 j_1}}) & \text{for } i = 1 \\ \sum_{i_1}^{a}\sum_{j_1}^{n_{i_1}} I(\sum_{i_2=1}^{i-1} n_{i_2} + l_1 \in C_{iX_{i_1 j_1}})I(\sum_{i_2=1}^{i-1} n_{i_2} + l_2 \in C_{iX_{i_1 j_1}}) & \text{for } i > 1. \end{cases}$$

Note that $V_{l_1l_2} = V_{l_2l_1}$. Therefore

$$\sqrt{N}T_B = 2 \sum_{1 \le l_1 < l_2 \le N} V_{l_1l_2} \tag{3.2.14}$$

is a clean quadratic form as in de Jong (1987). To show that $\text{Var}(\sqrt{N}T_B)^{-1/2}\sqrt{N}T_B \xrightarrow{\mathcal{L}} N(0,1)$, show that Proposition 3.2 in de Jong (1987) can be applied, i.e., show that G_1, G_2 and G_3 (defined below) are of smaller order than that of $[\text{Var}(\sqrt{N}T_B)]^4 = O(1)$. Let $l_3 = l(i, j_3)$, and $l_4 = l(i, j_4)$. Define $G_1 = \sum_{1 \le l_1 < l_2 \le N} E(V_{l_1l_2}^4)$,

$G_2 = \sum_{1 \le l_1 < l_2 < l_3 \le N} \{E(V_{l_1l_2}^2 V_{l_1l_3}^2) + E(V_{l_2l_1}^2 V_{l_2l_3}^2) + E(V_{l_3l_1}^2 V_{l_3l_2}^2)\}$, and

$G_3 = \sum_{1 \le l_1 < l_2 < l_3 < l_4 \le N} \{E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) + E(V_{l_1l_2}V_{l_1l_4}V_{l_3l_2}V_{l_3l_4}) + E(V_{l_1l_3}V_{l_1l_4}V_{l_2l_3}V_{l_2l_4})\}$.

First, show that the order of G_1 is $o(1)$. It suffices to consider only the case that $V_{l_1l_2} \neq 0$. When the response has finite conditional fourth moment, there exists some finite $M_0 > 0$, such that

$$\begin{aligned} E(V_{l_1l_2}^4 I(V_{l_1l_2} \neq 0)) &= \frac{16}{N^2a^4(k-1)^4} E\{E[((Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))|\mathbf{X}) (K_{l_1l_2})]^4\} \\ &= \frac{16}{N^2a^4(k-1)^4} E\{E[((Y_{l_1} - E(Y_{l_1}|\mathbf{X}))^4 E(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))^4|\mathbf{X}) K_{l_1l_2}^4]\} \\ &\le \frac{M_0}{N^2a^4(k-1)^4} E(K_{l_1l_2}^4). \end{aligned}$$

Thus

$$E(K_{l_1l_2}^4) = E(K_{ijj'}^4) = E\left\{ E\left[\sum_{c=1}^{N} I(j \in C_{ic}) I(j' \in C_{ic})\right]^4 \middle| X_{ij}, X_{ij'} \right\} = E(D_1 + D_2 + D_3 + D_4),$$

where

$$
\begin{aligned}
D_1 &= E\left(\sum_{c=1}^{N} I(j \in C_{ic})I(j' \in C_{ic}) \middle| X_{ij}, X_{ij'}\right), \\
D_2 &= E\left[\sum_{c_1 \neq c_2} I(j \in C_{ic_1})I(j' \in C_{ic_1})I(j \in C_{ic_2})I(j' \in C_{ic_2}) | X_{ij}, X_{ij'}\right] I(c_1 \neq c_2), \\
D_3 &= E\left\{E\left[\sum_{c_1 \neq c_2 \neq c_3} I(j \in C_{ic_1})I(j' \in C_{ic_1})I(j \in C_{ic_2})I(j' \in C_{ic_2})\right.\right. \\
&\quad \left.\left. I(j \in C_{ic_3})I(j' \in C_{ic_3}) | \, X_{ij}, X_{ij'}\right]\right\}, \\
D_4 &= E\{E[\sum_{c_1}\sum_{c_2}\sum_{c_3}\sum_{c_4} I(j \in C_{ic_1})I(j' \in C_{ic_1})I(j \in C_{ic_2})I(j' \in C_{ic_2})I(j \in C_{ic_3}) \\
&\quad I(j' \in C_{ic_3})]I(j \in C_{ic_4})I(j' \in C_{ic_4})|X_{ij}, X_{ij'}\}I(c_1 \neq c_2 \neq c_3 \neq c_4).
\end{aligned}
$$

It can be shown that the D_m, $m = 1, 2, 3, 4$, are of $O_p(1)$ and thus $E(K_{l_1l_2}^4) = O(1)$. In fact, $K_{ijj'}$ are bounded counts, so that

$$D_1 = E(K_{ijj'}|X_{ij}, X_{ij'}) = O_p(1)I(j'_* - j_* \leq (k-1)). \tag{3.2.15}$$

The result in (3.2.15) is obtained from (3.2.22) in the Appendix. Next $D_2 \leq E^2(K_{ijj'}|X_{ij}, X_{ij'}) = O(1)I(j'_* - j_* \leq (k-1))$. Similarly $D_3 \leq E^3(K_{ijj'}|X_{ij}, X_{ij'}) = O(1)I(j'_* - j_* \leq (k-1))$. Lastly, $D_4 \leq E^4(K_{ijj'}|X_{ij}, X_{ij'}) = O(1)I(j'_* - j_* \leq (k-1))$. Therefore, $E(K_{l_1l_2})^4 = O(1)I(l_{2*} - l_{1*} \leq (k-1))$, and $E(V_{l_1l_2}^4) = O(N^{-2})I(l_{2*} - l_{1*} \leq k-1)$, where $l_{1*} = l(i, j*)$, $l_{2*} = l(i, j'*)$. Thus, $G_1 = O(N^{-1}) = o(1)$.

Next is to show that the order of G_2 is $o(1)$ when $l_1 < l_2 < l_3$, that is, $i = i'$, $j < j'$ and $j < j_3$. First show that $E(K_{l_1l_2}^2 K_{l_1l_3}^2)$ is bounded and $E(V_{l_1l_2}^2 V_{l_1l_3}^2)$ is of order $O(N^{-2})$.

By Equation (3.2.13),

$$\begin{aligned} E(V_{l_1l_2}^2 V_{l_1l_3}^2) &= E\{E[16(N^2a^4k^4)^{-1}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))^2(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))^2 K_{l_1l_2}^2 \\ &\quad (Y_{l_1} - E(Y_{l_1}|\mathbf{X}))^2(Y_{l_3} - E(Y_{l_3}|\mathbf{X}))^2 K_{l_1l_3}^2|\mathbf{X}]\} \\ &= E\{16(N^2a^4k^4)^{-1}E[(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))^4|\mathbf{X}]E[(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))^2|\mathbf{X}]K_{l_1l_2}^2 \\ &\quad E[(Y_{l_3} - E(Y_{l_3}|\mathbf{X}))^2|\mathbf{X}]K_{l_1l_3}^2\} \\ &\leq \frac{M_2}{N^2a^4k^4}E(K_{l_1l_2}^2 K_{l_1l_3}^2) \text{ for some finite } M_2 > 0. \end{aligned}$$

Applying the Cauchy-Schwartz inequality obtains

$$E(K_{l_1l_2}^2 K_{l_1l_3}^2) \leq [E(K_{l_1l_2})^4 E(K_{l_1l_3})^4]^{1/2} = O(1)I(l_2^* - l_1^* \leq k-1)I(l_3^* - l_1^* \leq k-1) \quad (3.2.16)$$

The last equation in (3.2.16) follows from the previous result that $E(K_{l_1l_2})^4 = O(1)$. Therefore

$$E(V_{l_1l_2}^2 V_{l_1l_3}^2) = O(N^{-2})I(l_2^* - l_1^* \leq k-1)I(l_3^* - l_1^* \leq k-1).$$

It can be shown similarly that the order for $E(V_{l_2l_1}^2 V_{l_2l_3}^2)$ and $E(V_{l_3l_1}^2 V_{l_3l_2}^2)$ is $O(N^{-2})$. Therefore, $G_2 = O(N^{-1})I(l_2^* - l_1^* \leq k-1)I(l_3^* - l_1^* \leq k-1) = o(1)$.

Next, to establish that $G_3 = o(1)$ for the case $i = i', j < j' < j_3 < j_4$, i.e., $l_1 < l_2 < l_3 < l_4$, requires first showing that $E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3})$ is bounded and $E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) =$

$O(N^{-2})$. Consider

$$
\begin{aligned}
& E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) \\
= \; & E\{E[\frac{1}{N^2a^4k^4}(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))K_{l_1l_2}(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))(Y_{l_3}-E(Y_{l_3}|\mathbf{X}))K_{l_1l_3} \\
& (Y_{l_4}-E(Y_{l_4}|\mathbf{X}))(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))K_{l_4l_2}(Y_{l_4}-E(Y_{l_4}|\mathbf{X}))(Y_{l_3}-E(Y_{l_3}|\mathbf{X}))K_{l_4l_3}|\mathbf{X}]\} \\
= \; & E\{E[\frac{1}{N^2a^4k^4}(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))^2(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))^2(Y_{l_3}-E(Y_{l_3}|\mathbf{X}))^2(Y_{l_4}-E(Y_{l_4}|\mathbf{X}))^2] \\
& K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}\} \\
\leq \; & \frac{M_3}{N^2a^4k^4}E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}).
\end{aligned}
$$

This leads to

$$
\begin{aligned}
E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}) &\leq [E(K_{l_1l_2}K_{l_1l_3})^2E(K_{l_4l_2}K_{l_4l_3})^2]^{1/2} \\
&= O(1)I(l_2^*-l_1^*\leq k-1)I(l_3^*-l_1^*\leq k-1) \\
&\quad I(l_2^*-l_4^*\leq k-1)I(l_3^*-l_4^*\leq k-1).
\end{aligned}
$$

It is shown similarly that $E(K_{l_1l_2}K_{l_1l_4}K_{l_3l_2}K_{l_3l_4})$ and $E(K_{l_1l_3}K_{l_1l_4}K_{l_2l_3}K_{l_2l_4})$ are also of $O(1)$. Therefore, $E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) = O(N^{-2})I(l_2^*-l_1^*\leq k-1)I(l_3^*-l_1^*\leq k-1)I(l_2^*-l_4^*\leq k-1)I(l_3^*-l_4^*\leq k-1)$. So, $G_3 = O(N^{-1}) = o(1)$.

Lemma 3.2.3. *Under the conditions of Theorem 3.2.2, as $N\to\infty$, $Var(\sqrt{N}T_B)-\gamma_N^2\to 0$.*

Proof of Lemma 3.2.3. Write $\mathrm{Var}(\sqrt{N}T_B) = E(\mathrm{Var}(\sqrt{N}T_B|\mathbf{X})) + \mathrm{Var}(\sqrt{N}E(T_B|\mathbf{X}))$. Show that $\mathrm{Var}(\sqrt{N}E(T_B|\mathbf{X}))=0$ and $E(\mathrm{Var}(\sqrt{N}T_B|\mathbf{X}))-\gamma_N^2\to 0$.

It is clear that $\mathrm{Var}(\sqrt{N}E(T_B|\mathbf{X}))=0$ because by the definition of T_B in (3.2.4),

$$
\begin{aligned}
E(\sqrt{N}T_B|\mathbf{X}) &= E\left(\frac{N^{-1/2}}{a(k-1)}\sum_{i=1}^{a}\sum_{j\neq j'}(Y_{ij}-E(Y_{ij}|\mathbf{X}))(Y_{ij'}-E(Y_{ij'}|\mathbf{X}))\middle|\mathbf{X}\right)K_{ijj'} \\
&= 0 \quad a.s. \qquad (3.2.17)
\end{aligned}
$$

Next, show that $E(\mathrm{Var}(\sqrt{N}T_B|\mathbf{X})) - \gamma^2 \to 0$. Let

$$t_{ijj'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))K_{ijj'}.$$

Then

$$\begin{aligned}
& E(\mathrm{Var}(\sqrt{N}T_B|\mathbf{X})) \\
= \; & E\left[\frac{1}{Na^2(k-1)^2}\mathrm{Var}\left(\sum_{i=1}^{a}\sum_{j\neq j'} t_{ijj'}|\mathbf{X}\right)\right] \\
= \; & \frac{2}{Na^2(k-1)^2}E\left(\sum_{i}^{a}\sum_{j\neq j'} E(t_{ijj'}^2|\mathbf{X})\right) \\
= \; & \frac{2}{Na^2(k-1)^2}E\left[\sum_{i}^{a}\sum_{j\neq j'} E\left((Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))\; K_{ijj'}|\,\mathbf{X}\right)^2\right] \\
= \; & \frac{2}{Na^2(k-1)^2}E\left[\sum_{i}^{a}\sum_{j\neq j'} \sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})K_{ijj'}^2\right] \\
= \; & \frac{4}{Na^2(k-1)^2}E\left\{\sum_{i}^{a}\sum_{j<j'} \sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})E\left[K_{ijj'}^2\big|\, X_{ij}, X_{ij'}\right]\right\} \\
= \; & \frac{4}{Na^2(k-1)^2}E\left\{\sum_{i}^{a}\sum_{j<j'} \sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})[E^2(K_{ijj'}|X_{ij},X_{ij'}) + \mathrm{Var}(K_{ijj'}|X_{ij},X_{ij'})]\right\} \quad (3.2.18)
\end{aligned}$$

Let $X_{i(j_*)}$ be the order statistic for X_{ij} within group i. Without loss of generality, assume that $X_{ij} < X_{ij'}$ so that $j_* < j'_*$. The conditional expectation is obtained by considering whether a covariate value X_c is in group i or not. Denote ($X_c \in$ group i_1) as $X_{i_1 j_1}$. Then, if $i_1 \neq i$,

$$\begin{aligned}
\Lambda_{ijj'i_1} &= E(j \in C_{ic_1}, j' \in C_{ic_1})|X_{ij}, X_{ij'}) \\
&= P(X_{ij} \in C_{ic_1}, X_{ij'} \in C_{ic_1}|(X_{ij}, X_{ij'})) = \int_{X_{ij}-L_{ij}}^{X_{ij}+D_{ij}} f_{X,i_1}(x)dx I(j'_* - j_* \leq k-1),
\end{aligned}$$

where D_{ij} = the upper $k/2$ spacing and L_{ij} = the lower $(k/2 - (j'_* - j_*))$ spacing from X_{ij}.

Applying Taylor's expansion twice, write

$$\Lambda_{ijj'i_1} = \left[f_{X,i_1}(X_{ij}) \frac{F_{X,i}(X_{ij} + D_{ij}) - F_{X,i}(X_{ij} - L_{ij})}{f_{X,i}(X_{ij})} + O_p(N^{-2}) \right] I(j'_* - j_* \leq k-1).$$

From properties of spacings in Pyke (1965)

$$E(F_{X,i}(X_{ij} + D_{ij}) - F_{X,i}(X_{ij} - L_{ij}) | X_{ij}, X_{ij'}) = \frac{[k - (j'_* - j_*)]}{n_i + 1} I(j'_* - j_* \leq k-1).$$

Therefore, for $X_c \in$ group $i_1 \neq i$,

$$E(\Lambda_{ijj'i_1} | X_{ij}, X_{ij'}) = \left[\frac{f_{X,i_1}(X_{ij})}{f_{X,i}(X_{ij})} \frac{k - (j'_* - j_*)}{n_i + 1} + O_p(N^{-2}) \right] I(j'_* - j_* \leq k-1). \quad (3.2.19)$$

If $i_1 = i$ and $X_{ij_1} \neq X_{ij}$ and $X_{ij_1} \neq X_{ij'}$, detailed inspection yields

$$E(\Lambda_{ijj'i} | X_{ij}, X_{ij'}) = \left[\frac{k - (j'_* - j_*) - 2I(j'_* - j_* \leq (k-1)/2)}{n_i + 1} + O_p(N^{-2}) \right] I(j'_* - j_* \leq k-1) \quad (3.2.20)$$

if $i_1 = i$ and $X_{ij_1} = X_{ij}$ (or symmetrically $X_{ij_1} = X_{ij'}$), then

$$\Lambda_{ijj'i} = I(j'_* \in C_{iX_{i(j_*)}}) = I\left(j'_* - j_* \leq (k-1)/2\right). \quad (3.2.21)$$

Collecting terms from (3.2.19), (3.2.20), and (3.2.21), with $B_{ijj'}$ defined in Theorem 3.2.2 giving

$$E\left(K_{ijj'} | X_{ij}, X_{ij'}\right) = \left[B_{ijj'} + O_p(N^{-2})\right] I(j'_* - j_* \leq k-1). \quad (3.2.22)$$

Now consider the conditional variance. Note that when $X_c \in \{X_{ij}, X_{ij'}\}$, the term in

$K_{ijj'}$ is a constant. Therefore,

$$
\begin{aligned}
&\mathrm{Var}\left(K_{ijj'}|X_{ij}, X_{ij'}\right) \\
&= \mathrm{Var}\left(\sum_{c=1}^{N} I(j \in C_{ic}) I(j' \in C_{ic}) I(X_c \notin \{X_{ij}, X_{ij'}\}) \middle| X_{ij}, X_{ij'} \right) \\
&= \sum_{c_1}^{N} \sum_{c_2}^{N} \{ E[I(j \in C_{ic_1}) I(j' \in C_{ic_1}) I(j \in C_{ic_2}) I(j' \in C_{ic_2}) | X_{ij}, X_{ij'}] \\
&\quad - E[I(j \in C_{ic_1}) I(j' \in C_{ic_1}) | X_{ij}, X_{ij'}] E[I(j \in C_{ic_2}) I(j' \in C_{ic_2}) | X_{ij}, X_{ij'}] \} \\
&\quad \times I(X_{c_1} \notin \{X_{ij}, X_{ij'}\}) I(X_{c_2} \notin \{X_{ij}, X_{ij'}\}) \\
&= \sum_{c}^{N} E\left[I(j \in C_{ic}) I(j' \in C_{ic}) I(X_c \notin \{X_{ij}, X_{ij'}\}) | \, X_{ij}, X_{ij'} \right] \qquad (3.2.23) \\
&\quad - \sum_{c=1}^{N} [E(I(j \in C_{ic}) I(j' \in C_{ic}) | X_{ij}, X_{ij'})]^2 I(X_c \notin \{X_{ij}, X_{ij'}\}),
\end{aligned}
$$

where the last equality is due to the fact that the indicator functions involving c_1 and c_2 are conditionally independent when $c_1 \neq c_2$ and neither c_1, c_2 is X_{ij} or $X_{ij'}$. Plugging (3.2.19) through (3.2.22) into (3.2.23), obtains

$$
\begin{aligned}
\mathrm{Var}\left(K_{ijj'}|X_{ij}, X_{ij'}\right) &= \left[\left(\sum_{i_1, i_1 \neq i}^{a} \frac{n_{i_1}}{n_i} d_{i_1 i}(X_{ij}) + 1 \right) [k - (j'_* - j_*)] \right. \qquad (3.2.24) \\
&\quad \left. - 2I(j'_* - j_* \leq (k-1)/2) + O_p(N^{-1}) \right] I(j'_* - j_* \leq k-1).
\end{aligned}
$$

Putting (3.2.22) and (3.2.24) into (3.2.18),

$$
\begin{aligned}
E(\mathrm{Var}(\sqrt{N} T_B | \mathbf{X})) &= \frac{4}{N a^2 (k-1)^2} \sum_{i}^{a} \sum_{j<j'}^{n_i} E\left\{ \sigma_i^2(X_{ij}) \sigma_i^2(X_{ij'}) \left[B_{ijj'}^2 + B_{ijj'} \right. \right. \qquad (3.2.25) \\
&\quad \left. \left. - 2I(j'_* - j_* \leq (k-1)/2) \right] \right\} I(j'_* - j_* \leq k-1) + O(N^{-1}) = \gamma_N^2,
\end{aligned}
$$

where γ_N^2 is defined in Theorem 3.2.2.

3.3 Numerical Results

The following tests will be considered for comparison with the proposed test (pNP) in this section: the score test from GAM using spline (GAM Spline) or loess smoothing (GAM Loess) with quasilikelihood, drop test, likelihood ratio test from GAM using penalized splines (GAM Pspline), likelihood ratio test from linear models (LRT), test of association based on Pearson's correlation, Spearman's ρ and Kendall's τ. All the computation is carried out in R 2.8.1. Package *gam* is used for GAM Spline or loess smoothing; package *mgcv* is used for GAM Pspline; package *acepack* is used for ACE test. The drop test is obtained from http://www.stat.wmich.edu/mckean/HMC/Rcode/ and command cor.test is used for the three correlation based tests. Except for the proposed test and three correlation based tests, the significance of dependence on the covariate for the rest of the tests is obtained through comparing the log-likelihood or residual deviance from two models using an F test (see Chap. 12 of Faraway (2006)), one model includes the covariate, treatment, and their interaction effects, and the other model includes only the treatment effect. Comparison with ACE is given only in subsection 3.3.3 and has been removed from other comparisons because this test consistently produces highly inflated type I error rates.

For the proposed test, trials were conducted with the number of nearest neighbors $k = 3, 5, 7$ when $n_i = 30$ and $n_i = 50$ for a few data generation settings (linear alternative, quadratic alternative, binary data with log-odds to be the cosine function of the covariate). A slight reduction in the type I error and slight increase in the power was observed as k increases. However, the difference was too small to discriminate among the different k values. Therefore, the rest of the simulation and data analysis for this chapter, solely provides results for $k = 3$.

3.3.1 Analysis of Ozone Concentration Data - Detection of Non-linear Dependence

The ozone data in R faraway package contains daily measurements of ozone concentrations (O_3) and 9 meteorological variables in the Los Angeles basin for 330 days of 1976. The

relationship of ozone concentration with two other variables, day of year (doy) and wind speed is considered here for illustration. Wind has only 11 integer values and is split into 4 intervals. The intervals are low for values $0, 1, 2$, medium for values $3, 4, 5$, medium high for values $6, 7, 8$, and high for values $9, 10, 11$.

The scatter plot of the data in Figure 3.1 suggests that the variable doy is related to the O_3 in a quadratic relationship. However, this relationship is not evident due to large variations of O_3. The variation of O3 is low at small or large values of doy and increases as O_3 value approaches the peak concentration. A similar variation patten is observed for O_3 versus wind. This suggests strong heteroscedasticity for wind levels and that the conditional variance of O_3 given doy changes with doy. Regression based methods typically only evaluate if the mean regression function depends on the covariate regardless of whether the conditional variance depends on the covariate or not. In this example, even if the quadratic relationship of O_3 on doy can be attributed to its dependence on the wind level, the dependence of O_3 on doy through variances is still apparent. When applying all the tests mentioned in the beginning of this section, a significant doy effect on O_3 was detected by the new test (p-value $= 0$), GAM with spline ($p = 9.6 \times 10^{-34}$), GAM with loess smoothing ($p = 1.9 \times 10^{-33}$), GAM with penalized spline ($p = 6.9 \times 10^{-36}$). None of the other tests was significant (p-values are 0.390 for the drop test, 0.214 for the likelihood ratio test, 0.186 for Kendall's correlation test, 0.335 for Spearman's correlation test, and 0.220 for Pearson's correlation test). This is reasonable because this group of tests only access monotone relationships.

3.3.2 Application to EFT Study - Resistance to Outliers

In this subsection, the new test was applied to a data set in Aitkin et al. (1989, p. 70) containing a sample of 24 children randomly selected from fifth-grade students attending a state primary school in a Sydney suburb. Each student was assigned to one of two experimental groups given different instructions: Corner group and Row group. The total time in seconds to conduct a test of Wechsler Intelligence Scale for Children (WISC) was

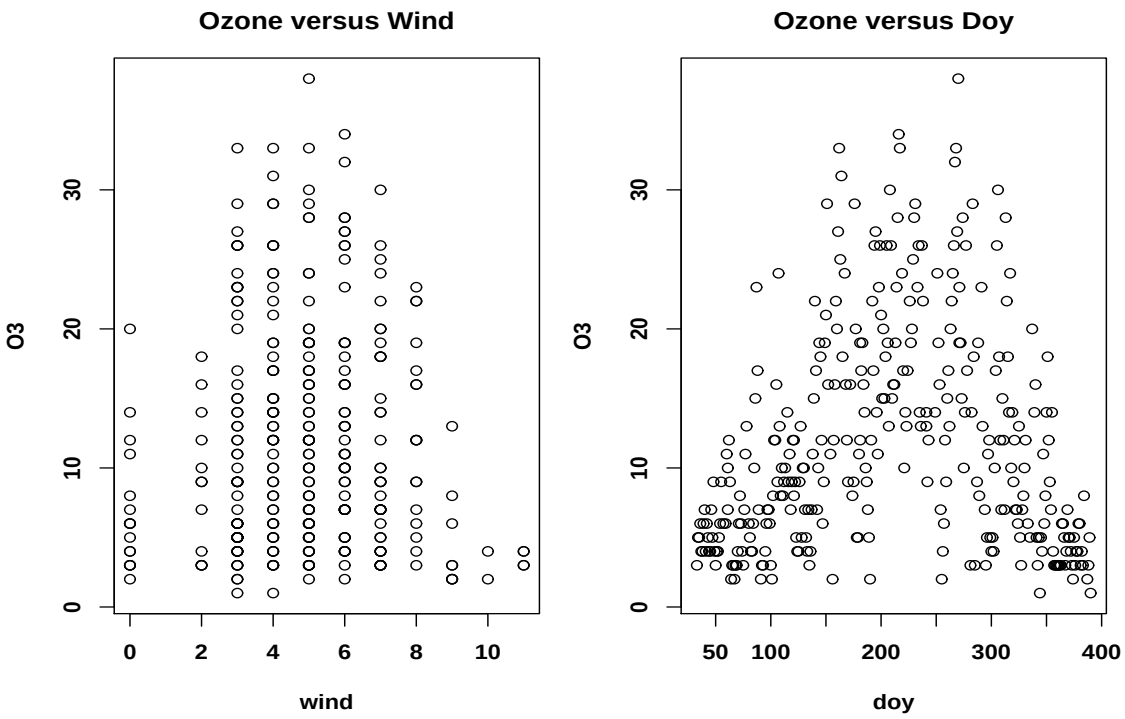


Figure 3.1: Scatter plot of ozone vs wind or doy.

recorded for each child. Each child was also tested for "field dependence" using an Embedded Figures Test (EFT). The objective of the study was to evaluate if the time to complete a WISC test was affected by field dependence. Figure 3.2 gives a scatter plot of Time vs EFT for each group. The observation at the upper right corner $(139, 739)$ is an unusual observation that has large influence for linear or nonlinear regression fit.

Five different linear models were considered in Aitkin et al. (1989, p.83 - p.104) with extensive discussions. They advised the readers to be cautious with small sample sizes because some of the fitted models produced conflicting interpretations. All the tests considered in this section are applied to this data set. The $p-$value for the test is given in the top row of Table 3.1 and Table 3.2. The new test is the only one that yielded a nonsignificant result. All other tests are significant at 0.05 level though some are not significant at 0.01 level. The second row in Table 3.1 and Table 3.2 gives the $p-$values of the tests when the outlier $(139, 739)$ is replaced by the median time in the Row group. With this single change, the new test produced consistent results but all the other tests change their p-values

dramatically yielding non-significant results at 0.05 level.

Figure 3.2: Scatter plot of Time vs EFT for each instruction group.

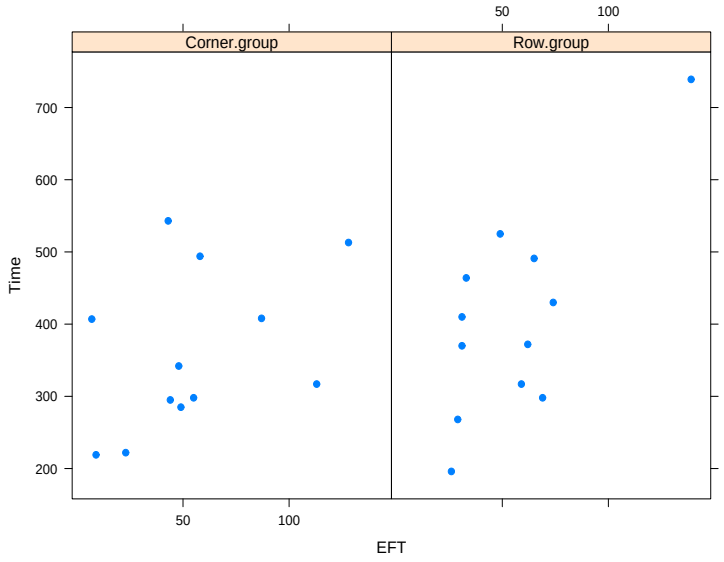


Table 3.1: $P-$values for test of no association before and after the outlier is replaced by the median time in the Row group- adjusting for treatment effect

	pNP	GAM Spline	GAM Loess	GAM Pspline	Drop test	LRT
Original data	0.729	0.033	0.041	0.013	0.035	0.002
Outlier replaced	0.369	0.385	0.347	0.145	0.305	0.306

Table 3.2: $P-$values for test of no association before and after the outlier is replaced by the median time in the Row group-correlation approach

	Kendall	Spearman	Pearson
Original data	0.017	0.021	0.006
Outlier replaced	0.059	0.067	0.178

The ground truth of whether time is associated with EFT or not is not known, but Aitkin et al. (1989, p. 76) did give a comment: "It is worth stressing that none of the models is a true representation of the population. If we could take a complete census of fifth grade children in the school, and administer the EFT and WISC tests to all of them, we would find that the mean completion time for children with each EFT score in each experimental

group did not lie on a straight line." In the next subsection, the performance of these tests is explored using simulation studies.

3.3.3 Simulation Study

This subsection reports a simulation study conducted to investigate type I error and power performance for the tests applied to the previous two applications. The type I error estimates are obtained for data having various probabilities of containing outliers. The power is presented for one setting. For group 1, the data were generated following:

$$X_{1j} \sim Unif(7, 128) \text{ and } Y_{1j} \sim Unif(219, 543), \tag{3.3.1}$$

where 7 and 128 are the minimum and maximum values of EFT in the corner group and 219 and 543 are the minimum and maximum values of Time in the corner group. The response and covariate for the other group were generated from a mixture of a Beta and a lognormal distribution as follows

$$\begin{cases} (r_2 - r_1)Z_{2j}, \text{ where } Z_{2j} \sim Beta(1.2, 3) & \text{with probability } p_0 \\ 10Q_{2j}, \text{ where } Q_{2j} \sim lognormal(1.2, 2) & \text{with probability } 1 - p_0, \end{cases} \tag{3.3.2}$$

where r_1 and r_2 are the lower and upper bound of the observed real data. That is, $r_1 = 26, r_2 = 74$ for EFT were used to generate X_{2j}, and $r_1 = 196, r_2 = 525$ for Time were used to generate Y_{2j}.

The type I error estimates at level 0.01 based on 2000 runs for different values of p_0 and n_i are given in Table 3.3. The new test is the only test having an acceptable type I error estimate under all mixing proportions. Smaller p_0 corresponds to bigger mixing percentage for the lognormal observations which leads to a higher chance of outliers. The type I error rates for the GAM tests increase as the chance of outliers increases. The drop test has a similar pattern as the GAM Loess test even though it has smaller type I errors. An opposite pattern was observed for the three correlation based tests. The type I error for the LRT test is inflated but does not change as dramatically as the other available tests. The ACE

test has consistently high type I errors (at least 0.22) for all cases. Therefore, ACE was eliminated from further comparisons.

Table 3.3: Proportion of rejections under H_0 in (3.2.1) at level 0.01 following the model (3.3.1) and (3.3.2).

	Mixture proportion		Estimated Type I error at 0.01 level						
n_i	p_{0X}	p_{0Y}	pNP	ACE	GAM Spline	GAM Loess	GAM Pspline	Drop Test	LRT
12	0.1	0.6	0.012	0.307	0.139	0.110	0.183	0.129	0.077
	0.1	0.1	0.013	0.351	0.121	0.097	0.149	0.056	0.074
	0.2	0.2	0.008	0.340	0.094	0.084	0.134	0.050	0.061
	0.4	0.4	0.006	0.309	0.085	0.070	0.104	0.047	0.064
	0.5	0.5	0.007	0.289	0.073	0.071	0.088	0.035	0.054
	0.6	0.6	0.010	0.282	0.044	0.039	0.066	0.034	0.043
20	0.1	0.6	0.013	0.330	0.101	0.151	0.202	0.147	0.056
	0.1	0.1	0.005	0.366	0.134	0.090	0.163	0.058	0.060
	0.2	0.2	0.005	0.342	0.130	0.081	0.156	0.052	0.058
	0.4	0.4	0.004	0.311	0.092	0.065	0.126	0.048	0.055
	0.5	0.5	0.008	0.298	0.093	0.068	0.096	0.037	0.045
	0.6	0.6	0.009	0.284	0.070	0.059	0.072	0.034	0.037
30	0.1	0.6	0.006	0.270	0.082	0.150	0.206	0.176	0.047
	0.1	0.1	0.008	0.336	0.143	0.082	0.185	0.063	0.048
	0.2	0.2	0.005	0.306	0.146	0.072	0.172	0.054	0.048
	0.4	0.4	0.004	0.260	0.111	0.062	0.126	0.046	0.049
	0.5	0.5	0.004	0.251	0.101	0.066	0.107	0.044	0.042
	0.6	0.6	0.005	0.228	0.077	0.052	0.084	0.036	0.042

For power comparisons, the departures from the null hypothesis in a quadratic relationship is considered where for the variables in one group, the data were generated from

$$X_{1j} \sim Unif(7, 128) \text{ and } Y_{1j} = \tau(X_{1j} - E(X_{1j}))^2 + \epsilon_{1j}, \text{ where } \epsilon_{1j} \sim Unif(-5, 15). \quad (3.3.3)$$

For the other group, X_{2j} were generated from the mixture distribution in (3.3.2) with $p_0 = 0.1$; Y_{2j} were independently generated from the mixture distribution in (3.3.2) with $p_0 = 0.6$ and were independent from X_{2j}.

The proportion of rejections at level 0.01 when $n_i = 12$ are presented in Figure 3.3 as τ increases from 0 to 2.5. The plot is busy for smaller values of τ so these values are presented also in Table 3.5 and Table 3.6. The power estimates were also obtained for some additional values of τ between 2.5 and 10. But the power stays at the plateau so they are not presented.

Table 3.4: Proportion of rejections under H_0 (3.2.1) following model (3.3.1) and (3.3.2) at level 0.01 - correlation based tests

	Mixture		Estimated Type I error at 0.01 level		
n_i	p_{0X}	p_{0Y}	Kendall	Spearman	Pearson
12	0.1	0.6	0.041	0.049	0.080
	0.1	0.1	0.037	0.037	0.048
	0.2	0.2	0.071	0.075	0.050
	0.4	0.4	0.134	0.142	0.075
	0.5	0.5	0.191	0.213	0.118
	0.6	0.6	0.256	0.292	0.153
20	0.1	0.6	0.068	0.071	0.072
	0.1	0.1	0.069	0.069	0.043
	0.2	0.2	0.119	0.128	0.048
	0.4	0.4	0.236	0.246	0.061
	0.5	0.5	0.374	0.403	0.090
	0.6	0.6	0.508	0.547	0.134
30	0.1	0.6	0.117	0.120	0.070
	0.1	0.1	0.094	0.095	0.040
	0.2	0.2	0.159	0.169	0.040
	0.4	0.4	0.416	0.426	0.050
	0.5	0.5	0.560	0.575	0.063
	0.6	0.6	0.717	0.749	0.114

The value $\tau = 0$ corresponds to the null hypothesis. The GAM Loess and GAM Pspline have similar power to the proposed test but they have inflated type I error rates. The GAM Spline has lower power than the other two GAM tests. The three correlation based tests have inflated type I error under H_0 due to outliers and the proportion of rejections reduces to the true level as τ increases. This is because X_{ij} and Y_{ij} are uncorrelated although Y_{1j} is not independent of X_{1j} and the signal to noise ratio increases as τ increases. The power of the drop test and LRT test lies in between the GAM Spline and the three correlation based tests. For $n_i = 20$ or 30, the proportion of rejections for all tests are reported in Table 3.5 and Table 3.6. In this simulation setting, the proposed test outperforms all other tests in terms of both the estimated type I error and power. The GAM tests were developed for the exponentially family and the mixture component log-normal distribution is not a member of the exponential family. This explains the observed lower power for the GAM tests.

In summary, the simulation study suggests that the proposed test not only offers reliable type I error estimates for our simulated data in the presence of outliers which lead to inflated

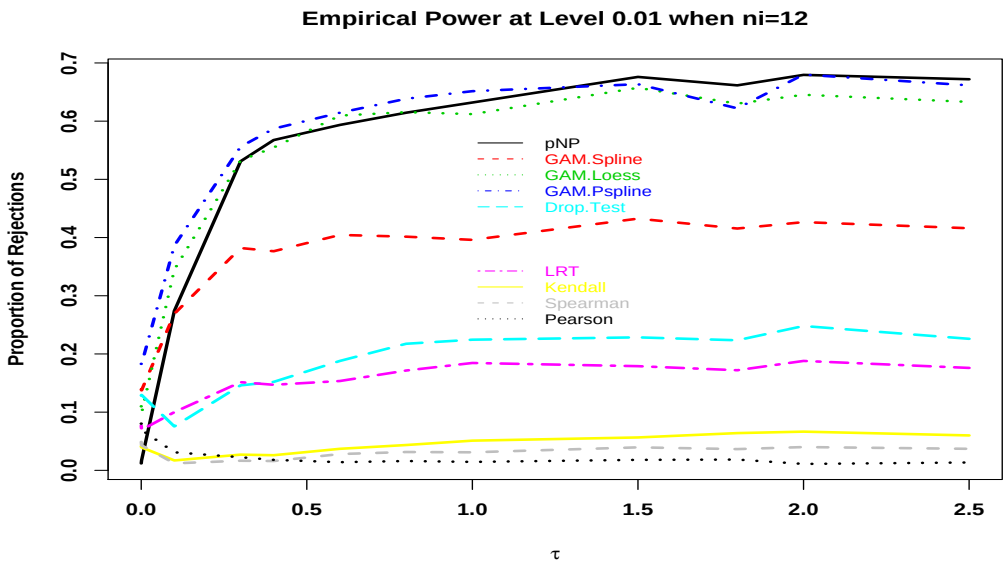


Figure 3.3: Empirical power at level 0.01 based on 2000 runs when the data from one group were from a mixture of beta and lognormal distribution with $n_i = 12$. and the other group were from equation (3.3.3).

type I error estimated for the GAM and other tests, but also maintains high power to detect nonlinear dependence.

Table 3.5: Empirical power at level 0.01 based on 2000 runs when $n_i = 20$ and 30 where the data from one group was from a mixture of beta and lognormal distribution with $n_i = 12$. and the other group was from equation (3.3.3).

n_i	τ	Estimated power at 0.01 level					
		pNP	GAM Loess	GAM Spline	GAM Pspline	Drop	LRT
12	0	0.012	0.110	0.139	0.183	0.129	0.077
	0.1	0.274	0.345	0.269	0.386	0.075	0.100
	0.3	0.531	0.532	0.382	0.556	0.146	0.151
	0.4	0.568	0.555	0.376	0.587	0.152	0.147
20	0	0.013	0.101	0.151	0.202	0.147	0.056
	0.1	0.485	0.423	0.273	0.435	0.085	0.080
	0.2	0.692	0.599	0.336	0.556	0.133	0.106
	0.3	0.775	0.651	0.364	0.608	0.160	0.116
	0.5	0.853	0.768	0.430	0.666	0.215	0.117
	1.0	0.927	0.831	0.440	0.736	0.248	0.128
	1.5	0.959	0.869	0.460	0.730	0.246	0.121
	2.0	0.964	0.875	0.469	0.730	0.245	0.126
	2.5	0.976	0.883	0.478	0.751	0.260	0.126
	3.0	0.973	0.883	0.455	0.740	0.254	0.122
30	0	0.006	0.082	0.150	0.206	0.176	0.047
	0.1	0.438	0.458	0.263	0.453	0.098	0.061
	0.2	0.682	0.666	0.340	0.536	0.127	0.077
	0.3	0.778	0.750	0.372	0.580	0.176	0.086
	0.5	0.879	0.830	0.415	0.643	0.232	0.102
	1.0	0.955	0.895	0.446	0.688	0.270	0.105
	1.5	0.972	0.909	0.452	0.701	0.283	0.110
	2.0	0.989	0.935	0.459	0.708	0.275	0.104
	2.5	0.991	0.936	0.457	0.704	0.299	0.123
	3.0	0.989	0.936	0.464	0.694	0.275	0.103

Table 3.6: Empirical power at level 0.01 based on 2000 runs when $n_i = 20$ and 30 where the data from one group was from a mixture of beta and lognormal distribution with $n_i = 12$. and the other group was from equation (3.3.3) - correlation based tests.

		Estimated power at 0.01 level		
n_i	τ	Kendall	Spearman	Pearson
12	0	0.041	0.049	0.080
	0.1	0.017	0.012	0.031
	0.3	0.027	0.016	0.022
	0.4	0.026	0.016	0.018
20	0	0.068	0.071	0.072
	0.1	0.024	0.021	0.031
	0.2	0.029	0.021	0.019
	0.3	0.032	0.025	0.012
	0.5	0.041	0.024	0.008
	1.0	0.064	0.042	0.010
	1.5	0.074	0.053	0.007
	2.0	0.102	0.074	0.007
	2.5	0.084	0.058	0.006
	3.0	0.073	0.048	0.004
30	0	0.117	0.120	0.070
	0.1	0.022	0.016	0.047
	0.2	0.032	0.020	0.027
	0.3	0.040	0.026	0.018
	0.5	0.056	0.034	0.005
	1.0	0.090	0.064	0.007
	1.5	0.093	0.066	0.004
	2.0	0.107	0.076	0.002
	2.5	0.116	0.084	0.002
	3.0	0.138	0.088	0.002

Chapter 4

New Nonparametric Tests when Treatment Level a is Small

4.1 Nonparametric Tests of No Covariate-Treatment Interaction and No Main Covariate Effects

The method discussed in Chapter 3 can be used to determine whether or not the response variable is independent of the covariate after adjusting for treatment effects. When it is found that Y is not independent of the covariate, the next step is to determine whether the dependence is through covariate-treatment interaction or through the main covariate variable. For example, in the ozone study in Section 3.3.1, the independence test was shown to be significant. The ozone (O_3) concentration is not independent of doy. The next question to be resolved is whether the dependence of O_3 and doy is through the doy and wind interaction or through the day of year (doy) alone. The tests developed in this chapter successfully address this issue.

There are many available methods to test for no main covariate and no covariate-treatment interaction effects. Some of them were discussed in Chapter 1. The methods such as likelihood ratio test, traditional ANCOVA and drop test are convenient, but are restricted to the presence of linear dependence of the response and covariate. Methods intended for discrete response variables, such as the Wald type test and the deviance test are

also restricted to linear dependence. In addition, methods just mentioned are also restricted to constant variance of the response within each treatment. In term of the nonparametric ANCOVA model approach, the test by Wang and Akritas (2006) discussed in Section 2.4.3 could be used to accommodate the heteroscedasticity treatment effect, except that it has been shown in the simulation studies that it is unreasonably computationally extensive. This chapter develops the theory of no main covariate and no main covariate-treatment interaction effects that are not restricted to constant variance, distributional assumption or linear relationship of covariate and response. The tests are developed using an approach similar to that in Chapter 3. Section 4.2 presents the nonparametric model and the hypotheses of interest that are used for the new tests and will also review and discuss two models that are relevant for the formulation of the nonparametric model.

4.2 Models and Hypotheses

Before discussing the formulation of the model to be used for the hypotheses of interest, the model and hypotheses for no covariate-treatment and no main covariate effect in the traditional one-way ANCOVA setting as well as in Wang and Akritas (2006) are presented for discussion.

First, recall that the one-way ANCOVA model is written as: $Y_{ij} = \mu_i + \varepsilon_{ij}$ where μ_i is the conditional mean, ε_{ij} is independent $N(0, \sigma^2)$ with $i = 1, ..., a; j = 1, ..., n_i$. Further, the conditional mean decomposes into $\mu_i = \mu_\cdot + \alpha_i + \beta x + \gamma_i x$ where $\mu_\cdot$ is a constant, α_i constitutes fixed treatment effects and $(\beta + \gamma_i)$ is the regression coefficient for the relation between X and Y. From the above model, the conditional mean of Y_{ij} given $X_{ij} = x$ is given by:

$$E(Y_{ij}|X_{ij} = x) = \mu_\cdot + \alpha_i + \beta x + \gamma_i x. \tag{4.2.1}$$

Therefore, the no covariate-treatment interaction and the no main covariate effects from the

model (4.2.1), correspond to the hypotheses:

$$H_{0\gamma} : \gamma_i = 0 \text{ for all } i \quad \text{and} \quad H_{0\beta} : \beta = 0 \text{ respectively.}$$

These hypotheses are similar to testing the equality of slopes and testing whether the slopes are 0 respectively. Under $H_{0\gamma}$, the conditional mean of Y given x in different treatment levels is modeled with parallel linear lines. Notice here that these hypotheses are restricted to the assumption of a linear association between X and Y. In general, when there is no covariate-treatment interaction effect the regression curves will be parallel in all the treatment levels.

On the other hand, Wang and Akritas (2006) used the fully nonparametric model (FNP) as in Akritas et al. (2000) which is based on the decomposition of the conditional CDF, $F_{ix}(y)$ of Y_{ij} given $X_{ij} = x$. The FNP model is written as

$$F_{ix}(y) = M(y) + A_i(y) + D_x(y) + C_{ix}(y), \tag{4.2.2}$$

as described in section 2.4.1. The decomposition in (4.2.2) was developed by mimicking the decomposition of the two-way ANOVA model, where the column factor is replaced by an artificial factor created from combining all the covariate levels in all of the row factor levels. Here, $F_{ix}(y)$ is the conditional distribution of Y_{ij} given $X_{ij} = x$. Define $\bar{F}_{i\cdot}(y) = \int_{-\infty}^{\infty} F_{ix}(y)dG(x)$ for any chosen cumulative density function $G(x)$ and $\bar{F}_{\cdot x}(y) = a^{-1}\sum_i F_{ix}(y)$, then the terms in (4.2.2) are $M(y) = a^{-1}\sum_{i=1}^{a} \bar{F}_{i\cdot}(y)$, $A_i(y) = \bar{F}_{i\cdot}(y) - M(y)$, $D_x(y) = \bar{F}_{\cdot}(y) - M(y)$ and $C_{ix}(y) = F_{ix}(y) - \bar{F}_{i\cdot}(y) - \bar{F}_{\cdot}x(y) + M(y)$. The hypotheses of no covariate-treatment interaction and no main covariate effects in Wang and Akritas (2006) are given as:

$$H_{0C} : C_{ix}(y) = 0 \text{ for all } i, x, \text{ and } y \quad \text{and} \quad H_{0D} : D_x(y) = 0 \text{ for all } x \text{ and } y.$$

respectively. When there is no covariate-treatment interaction effect, i.e H_{0C} in the FNP model is true, $F_{ix}(y)$ is a mixture distribution consisting of two components in which one

component depends only on i and the other component depends on x. The $M(y)$ term makes the right hand side of (4.2.2) a valid conditional CDF.

An in-depth look at the two sets of hypotheses above examines two examples using two different conditional distributions of Y given X, which are exponential and normal distributions to illustrate the hypothesis of no covariate-treatment interaction effect under the $H_{0\gamma}$ and under the H_{0C}. The examples presented below provide the motivation for the construction of the nonparametric model and hypotheses for the proposed tests of no covariate-treatment interaction and no main covariate effects.

Example 1a. Exponential Distribution with Quadratic Conditional Mean

Suppose the response variable Y follows an exponential distribution with the conditional mean in treatment i being

$$E(Y_{ij}|X_{ij} = x) = \frac{1}{\lambda_i} = m_i(x-5)^2 + b_i, i = 1, 2. \tag{4.2.3}$$

The cumulative conditional density function of Y_{ij} given X_{ij} then is written as:

$$F_{Y_i|X_{ij}=x}(y) = 1 - e^{-\frac{y}{m_i(x-5)^2+b_i}}. \tag{4.2.4}$$

When $m_i = m$, (4.2.4) becomes $F_{Y_i|X_{ij}=x}(y) = 1 - e^{-\frac{y}{(m(x-5)^2+b_i)}}$, and (4.2.3) becomes $E(Y_{ij}|X_{ij} = x) = \lambda_i^{-1} = m(x-5)^2 + b_i, i = 1, 2$. The hypothesis $H_{0\gamma}$ is clearly satisfied because the two conditional expectations differ only in the intercepts. The scatter plot in the top left panel of Figure 4.1 illustrates the case where the mean of the conditional exponential distribution is modeled with (4.2.3) for $m_i = 2$, $b_1 = 2$ and $b_2 = 10$. The scatter plot clearly shows that there is no covariate-treatment interaction when viewed from the behavior of the observations.

For the nonparametric hypotheses in Wang and Akritas (2006), when there is no covariate-treatment interaction effect i.e $H_{0C} : C_{ix}(y) = 0$, the cumulative conditional function becomes $F_{ix}(y) = M(y) + A_i(y) + D_x(y)$. The cumulative conditional distribution in (4.2.4) cannot be decomposed into the sum of a function that depends on i and y only and another

function that depends on x and y only. Therefore, the hypothesis H_{0C} is not satisfied for this specific example. Because it is difficult to infer from the graphs of the two conditional distributions when H_{0C} is true, the curve of the differences in the two conditional distributions $F_1(y|x) - F_2(y|x)$, was plotted which should not depend on x for all y under H_{0C}. The illustration in the top right hand panel of Figure 4.1, shows that the curve $F_1(y|x) - F_2(y|x)$ is not independent of x for each y under H_{0C}. This example shows that the H_{0C} hypothesis is not suitable to describe the no covariate-treatment interaction effect in terms of the behavior of observations. The next example shows a situation similar to the current example but where the response variable Y follows a normal distribution.

Example 1b: Normal Distribution with a Sinusoidal Conditional Mean

Suppose the response variable follows a normal distribution with the conditional mean in treatment i being

$$E(Y_{ij}|X_{ij} = x) = m_i \sin(10\pi x) + b_i, i = 1, 2. \tag{4.2.5}$$

The cumulative conditional density function of Y_{ij} given X_{ij} then is written as:

$$F_{Y_{ij}|X_{ij}=x}(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-[y-(m_i \sin(10\pi x)+b_i)]^2/(2\sigma^2)} dy. \tag{4.2.6}$$

When $m_i = m$, the scatter plot of the conditional means from the model (4.2.5) with $m = 4$ is displayed in the bottom left panel in Figure 4.1. The scatter plot based on the behavior of the observations exhibits a no covariate-treatment interaction effect. On the other hand, because the conditional CDF in (4.2.6) cannot be written in a closed form, it is difficult to determine whether it can be written in an additive form of a function that depends on i and y only and another function that depends on x and y only. Consequently, the curve of difference of the two conditional CDF's, $F_1(y|x) - F_2(y|x)$ according to (4.2.6) is plotted. The graph on the bottom right panel in Figure 4.1 indicates that $F_1(y|x) - F_2(y|x)$ still depends on X for each Y, contradicting the hypothesis H_{0C}.

From the two examples above, we see that the parametric effect of covariate-treatment

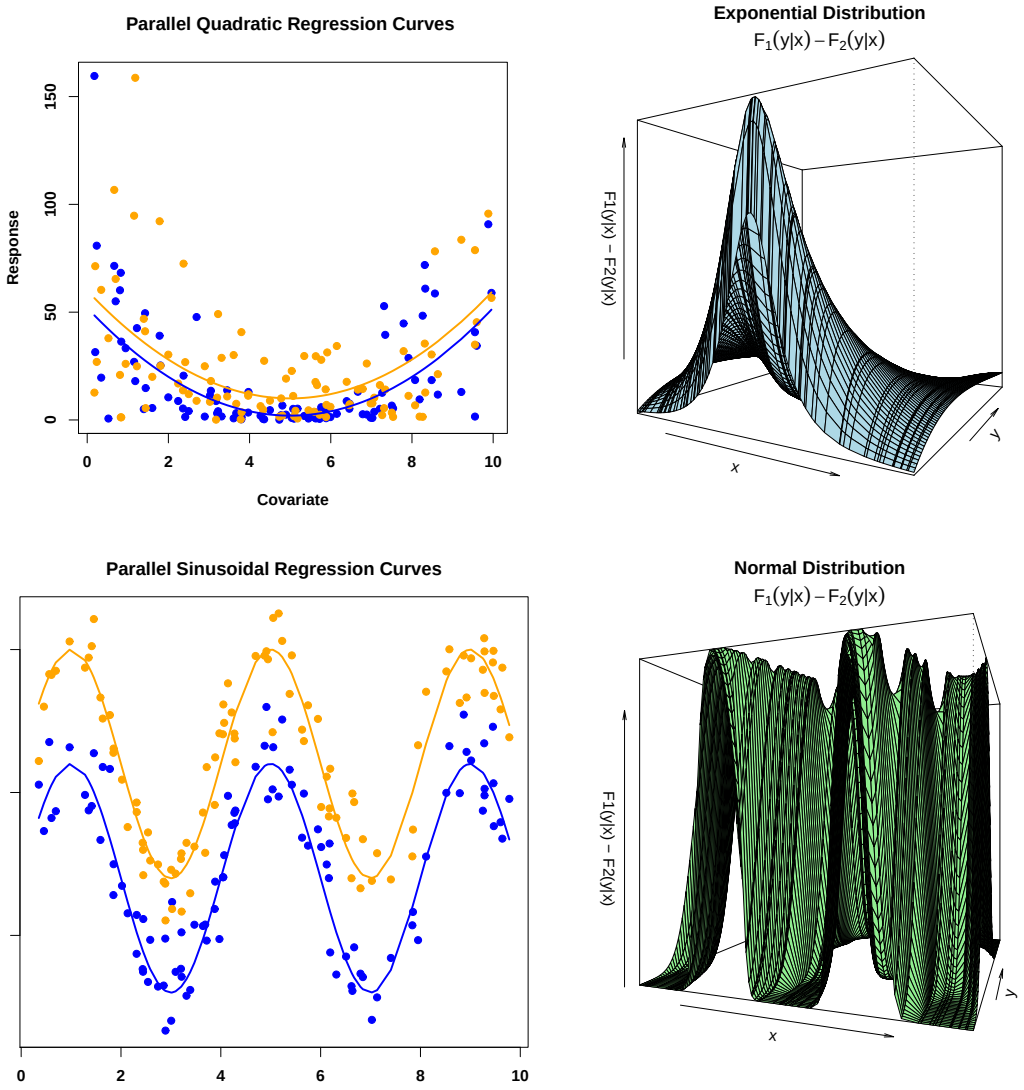


Figure 4.1: Scatter plots and graphs for the exponential with linear conditional mean and normal distribution with sinusoidal conditional mean

interaction using the traditional ANCOVA model based on decomposition of the conditional mean captures the behavior of observations better than using the hypothesis H_{0C} in the FNP model. As a suitable hypothesis regarding the behavior of the observations cannot be achieved by decomposing the conditional distribution function of Y_{ij} given $X_{ij} = x$ following Wang and Akritas (2006), an alternative decomposition is needed. To remedy the linear relationship restriction between the response and covariate, the decomposition of the conditional mean of Y_{ij} given $X_{ij} = x$ into nonparametric covariate-treatment interaction and no main covariate and no main treatment effects is considered. The decomposition will not restrict any linear association between Y and X. There are already some existing models that could be used, such as the general additive model (GAM) $Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$ from Hastie and Tibshirani (1990). The GAM does not require f_j to be linear, but their method not only uses scatter plot smoothers such as computer-intensive splines to estimate f_j, it is also restricted to response variables from the exponential family.

To construct a model that not only can accommodate nonlinear relationship between Y and X, and is also not restricted to exponential family and constant variance assumptions, assume the conditional distribution of Y_{ij} given $X_{ij} = x$ is $F_i(y|x)$, then decompose the conditional mean of Y_{ij} given $X_{ij} = x$, into

$$E(Y_{ij}|X_{ij} = x) = \mu_{ix} = \mu + \alpha_i + \eta(x) + \xi_i(x), \tag{4.2.7}$$

where for some continuous covariate X with probability distribution function $f_X(x)$ and its cumulative distribution function $F_X(x)$ and where $\mu = \frac{1}{a}\sum_{i=1}^{a}\int \mu_{ix}\ dF_X(x)$, $\mu_i = \int \mu_{ix}\ dF_X(x)$, $\mu_x = \frac{1}{a}\sum_i \mu_{ix}$, $\alpha_i = \mu_i - \mu$, $\eta(x) = \mu_x - \mu$, and $\xi_i(x) = \mu_{ix} - \mu_i - \mu_x + \mu$.

Following the one-way ANCOVA interpretation, μ is the overall mean, α_i is the treatment effects, $\eta(x)$ is the main covariate effect which is a function that depends on x, and $\xi_i(x)$ is the covariate-treatment interaction effect which is a function that depends on i and x. Here, there is no restriction to any specific distribution or constant variance of the response within each treatment. The error terms are independent with mean 0. The hypotheses to test the no covariate-treatment interaction and the no main covariate effects using the model (4.2.7)

are:

$$H_{0\xi} : \xi_i(x) = 0 \text{ for all } i, \text{ all } x \text{ and} \tag{4.2.8}$$

$$H_{0\eta} : \eta(x) = 0 \text{ for all } x \tag{4.2.9}$$

respectively.

4.3 Test Statistics

The same notation and conditional distribution assumptions in section 3.2.1 will be used in the construction of the test statistics to test the hypotheses of no main covariate effect $H_{0\eta}$ and no covariate-treatment interaction effects $H_{0\xi}$. Let (X_{ij}, Y_{ij}), $j = 1, \ldots, n_i$ be the original random sample from treatment i with conditional distribution of Y given X as $F_i(y|x)$. Following the approach in Chapter 3, combine all the covariate values and arrange them in ascending order and treat the covariate as a factor with many levels. This creates a two-way ANOVA setting without replication. The pseudo replications then are created using the same technique as in subsection 3.2.1 by augmenting each cell (i, c) with k nearest neighbors using observations from the i^{th} treatment level. Denote U_{ict} to be the observations in the augmented cell (i, c), then $\overline{U}_{ic\cdot} = k^{-1} \sum_t^k U_{ict}$ and $\overline{U}_{i\cdot\cdot} = N^{-1} \sum_{c=1}^N \overline{U}_{ic\cdot}$. Denote

$$\begin{aligned}
Q_N &= ak(N-1)^{-1} \sum_{c=1}^{N} \left(\overline{U}_{\cdot c\cdot} - \overline{U}_{\cdots}\right)^2, \\
G_N &= k(a-1)^{-1}(N-1)^{-1} \sum_{c=1}^{N} \sum_{i=1}^{a} \left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot} - \overline{U}_{\cdot c\cdot} + \overline{U}_{\cdots}\right)^2, \\
W_N &= \{Na(k-1)\}^{-1} \sum_{i=1}^{a} \sum_{c=1}^{N} \sum_{t=1}^{k} \left(U_{ict} - \overline{U}_{ic\cdot}\right)^2.
\end{aligned}$$

Note that W_N is the same as in subsection 3.2.1. Then, the test statistics $T_{cov} = \sqrt{N}(Q_N - W_N)$ and $T_{int} = \sqrt{N}(G_N - W_N)$ are used to test the hypotheses of no main covariate

effect $H_{0\eta}$ (4.2.9) and no covariate-treatment interaction effect $H_{0\xi}$ (4.2.8) respectively. The development of the asymptotic distribution for T_{cov} and T_{int} is similar to the development of the asymptotic distribution of the test statistic in Chapter 3. The lemmas and theorems in the next section provide the theoretical results.

4.4 Asymptotic Distribution of the Test Statistics

Following the procedure in Section 3.2.2, denote $Z_{ict} = U_{ict} - E(U_{ict}|\mathbf{X})$. Then Q_N and G_N defined in the previous section are projected onto the the space produced by the span of some functions of $\{\mathbf{Z}_c, c = 1, \ldots, N\}$, where $\mathbf{Z}_c = (Z_{1c1}, \ldots, Z_{ack})'$. In Section 4.4.1, Lemma 4.4.1 shows how the projection of Q_N is accomplished, followed by Theorem 4.4.2 showing the asymptotic distribution of T_{cov}. In Section 4.4.2 Lemma 4.4.3 shows how the projection of G_N is accomplished and Theorem 4.4.4 showing the asymptotic distribution of T_{int}.

4.4.1 No Main Covariate Effect

Lemma 4.4.1. *If the assumptions in subsection 3.2.1 are satisfied, let*

$$Q_N = P_Q(\mathbf{Z}) + R_Q(\mathbf{Z}) \tag{4.4.1}$$

$$\text{where } P_Q(\mathbf{Z}) = \frac{ak}{N}\sum_{c=1}^{N}\overline{Z}_{.c.}^2 \text{ and } R_Q(\mathbf{Z}) = \frac{ak}{N(N-1)}\sum_{c\neq c'}^{N}\overline{Z}_{.c.}\overline{Z}_{.c'.}. \tag{4.4.2}$$

Then, $\sqrt{N}R_Q(Z) \xrightarrow{p} 0$ *as* $N \to \infty$

Proof of Lemma 4.4.1

It is sufficient to show that $E(\sqrt{N}R_Q(Z)) \longrightarrow 0$ and $\text{Var}(\sqrt{N}R_Q(Z)) \longrightarrow 0$.

$$E(R_Q(Z)) = \frac{k}{aN(N-1)}E\left\{\sum_{i=1}^{a}\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right\} + \frac{k}{aN(N-1)}E\left\{\sum_{i\neq i'}^{a}\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}\right\}.$$

The second term of $E(R_Q(Z))$ vanishes because observations from different treatments are independent. The first term in $E(R_Q(Z))$ was shown in the proof of lemma 3.2.1 to go to 0 as $N \to \infty$

Next is to show that $Var(\sqrt{N}R_Q(Z))$ goes to 0 as $N \to \infty$. Because $E(\sqrt{N}R_Q(Z))$ goes to 0, it remains to show that $E(\sum_{c\neq c'}^{N} \overline{Z}_{\cdot c\cdot}\overline{Z}_{\cdot c'\cdot})^2$ also goes to 0. Write

$$E(\sum_{c\neq c'}^{N} \overline{Z}_{\cdot c\cdot}\overline{Z}_{\cdot c'\cdot})^2 = \sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} E(\overline{Z}_{\cdot c\cdot}\overline{Z}_{\cdot c'\cdot}\overline{Z}_{\cdot c_1\cdot}\overline{Z}_{\cdot c_1'\cdot}) = 3A + B$$

where

$$A = \sum_{i_1\neq i_2}^{a}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} E(\overline{Z}_{i_1c\cdot}\overline{Z}_{i_1c'\cdot}\overline{Z}_{i_2c_1\cdot}\overline{Z}_{i_2c_1'\cdot}),$$

$$B = \sum_{i_1=1}^{a}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} E(\overline{Z}_{i_1c\cdot}\overline{Z}_{i_1c'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_1c_1'\cdot}).$$

and when the cardinality of $\{i_1, i_2, i_3, i_4\}$ is greater than 2, $E(\overline{Z}_{i_1c\cdot}\overline{Z}_{i_2c'\cdot}\overline{Z}_{i_3c_1\cdot}\overline{Z}_{i_4c_1'\cdot}) = 0$. The expectation is not equal to 0 if the cardinality of $\{i_1, i_2, i_3, i_4\}$ is less than or equal to 2. A contains the three cases where the cardinality is equal to two; (i) when $i_1 = i_3$ and $i_2 = i_4$ (ii) when $i_1 = i_2$ and $i_3 = i_4$ and (iii) when $i_1 = i_4$ and $i_2 = i_3$. B corresponds to the case that the cardinality is one. It was shown in the proof of lemma 3.2.1 that B vanishes asymptotically. Hence we need only to consider A,

$$\begin{aligned} |A| &= \left|\sum_{i_1\neq i_2}^{a}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} E(\overline{Z}_{i_1c\cdot}\overline{Z}_{i_1c'\cdot}\overline{Z}_{i_2c_1\cdot}\overline{Z}_{i_2c_1'\cdot})\right| \qquad (4.4.3)\\ &\leq \sum_{i_1\neq i_2}^{a}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} |E(\overline{Z}_{i_1c\cdot}\overline{Z}_{i_1c'\cdot})E(\overline{Z}_{i_2c_1\cdot}\overline{Z}_{i_2c_1'\cdot})| \leq 4a(a-1)(N-1)^2k^2M_1^2 \leq O(N^2), \end{aligned}$$

where M_1 is finite and was defined in equation (3.2.8).

Therefore

$$\begin{aligned}\mathrm{Var}(\sqrt{N}R_Q(Z)) &= Var\left(\frac{\sqrt{N}ak}{N(N-1)}\sum_{c\neq c'}^{N}\overline{Z}_{\cdot c\cdot}\overline{Z}_{\cdot c'\cdot}\right)\\ &= \frac{a^2k^2N}{N^2(N-1)^2}Var\left(\sum_{c\neq c'}^{N}\overline{Z}_{\cdot c\cdot}\overline{Z}_{\cdot c'\cdot}\right)\\ &\leq \frac{a^2k^2}{N(N-1)^2}\{O[N^2]+O[N^2]\}\leq O(N^{-1}),\end{aligned}$$

which goes to zero when N goes to infinity.□

Theorem 4.4.2. *Assume that $H_{0\eta}$ is true and the assumptions in Theorem 3.2.2 are satisfied then, $\sqrt{N}(Q_N(\mathbf{U})-W_N(\mathbf{U}))\rightarrow N(0,\lim_{N\rightarrow\infty}\gamma_N^2+\lim_{N\rightarrow\infty}\varphi_N^2)$, where γ_N^2 is defined in Theorem 3.2.2 and*

$$\begin{aligned}\varphi_N^2 &= \frac{2}{Na^2k^2}\sum_{i\neq i'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}E\left\{\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'})\left[E^2(M_{iji'j'}|S_{ij,t},S_{i'j',t})+E(M_{iji'j'}|S_{ij,t},S_{i'j',t})\right]\right\}\\ &+ \Delta_{iji'j'}(t_1,t_2)\}+O(N^{-1}),\end{aligned}$$

$\sigma_i^2 = Var(Y_{ij}|X_{ij})$, $M_{iji'j'}=\sum_r^N I(j\in C_{ic})I(j'\in C_{i'c})$, $S_{ij,t}=(X_{ij},L_{ij}^{(t/2},U_{ij}^{(t/2)})$, $U_{ij}^{t/2}$ *and* $L_{ij}^{t/2}$ *be the upper and lower* $t/2$ *spacings from* X_{ij},
$\Delta_{iji'j'}(t_1,t_2)=I\left(\max\{X_{ij}-L_{ij}^{(t_1/2)},\ X_{i'j'}-L_{i'j'}^{(t_2/2)}\}\leq\min\{X_{ij}+U_{ij}^{(t_1/2)},\ X_{i'j'}+U_{i'j'}^{(t_2/2)}\}\right)$.

Proof of Theorem 4.4.2

After applying Lemma 4.4.1, it remains to show that $\sqrt{N}(P_Q(\mathbf{Z}))-W_N(\mathbf{Z}))$ is asymptotically normal. After algebraic simplification, $\sqrt{N}(P_Q(\mathbf{Z}))-W_N(\mathbf{Z}))$ is written as $\sqrt{N}(T_B+T_Q)$, where

$$T_B=\frac{1}{Na(k-1)}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}^{k}Z_{ict}Z_{ict'}\text{ same as equation (3.2.4)}$$

$$T_Q=\frac{k}{Na}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{i'c\cdot}. \tag{4.4.4}$$

The proof for the theorem is shown by verifying

$$\mathrm{Var}[\sqrt{N}(T_B + T_Q)|\mathbf{X}] \xrightarrow{p} \lim_{N\to\infty}(\gamma_N^2 + \varphi_N^2), \tag{4.4.5}$$

$$\frac{\sqrt{N}(T_B + T_Q)}{\sqrt{\sqrt{N}\mathrm{Var}[(T_B + T_Q)|\mathbf{X}]}} \to N(0,1). \tag{4.4.6}$$

First show that $\mathrm{Cov}(T_B, T_Q)|\mathbf{X})$=0.

$$\begin{aligned}
\mathrm{Cov}((T_B, T_Q)|\mathbf{X}) &= \frac{k}{N^2a^2(k-1)}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\sum_{i_1=1}^{a}\sum_{c_1=1}^{N}\sum_{t_1\neq t_{1'}}^{k}\mathrm{Cov}(\overline{Z}_{ic\cdot}\overline{Z}_{i'c\cdot}, Z_{i_1c_1t_1}Z_{i_1c_1t_{1'}}|\mathbf{X}) \\
&= \frac{k}{N^2a^2(k-1)}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\sum_{i_1=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}^{k}\frac{1}{k^2}\sum_{t}^{k}\sum_{t'}^{k}\mathrm{Cov}(Z_{ict}Z_{i'ct'}, Z_{i_1c_1t_1}Z_{i_1c_1t_1}|\mathbf{X}) \\
&= \frac{k}{N^2a^2(k-1)}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\sum_{i_1=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}^{k}\frac{1}{k^2}\sum_{t}^{k}\sum_{t'}^{k}\big(E(Z_{ict}Z_{i'ct'}Z_{i_1c_1t_1}Z_{i_1c_1t_{1'}}|\mathbf{X}) \\
&- E(Z_{ict}Z_{i'ct'}|\mathbf{X})E(Z_{i_1c_1t_1}Z_{i_1c_1t_{1'}}|\mathbf{X})\big) \\
&= 0.
\end{aligned} \tag{4.4.7}$$

The last equation is true because observations from different treatment are independent i.e. $E(Z_{ict}Z_{i'ct'}) = E(Z_{ict})E(Z_{i'ct'}) = 0$. Because $\mathrm{Cov}((T_B, T_Q)|\mathbf{X})$ is 0, the condition (4.4.5) is shown by demonstrating $Var(N^{1/2}T_B|\mathbf{X}) \to \lim_{N\to\infty}\gamma_N^2$ and $(E[\mathrm{Var}(N^{1/2}T_Q|\mathbf{X})] - \varphi_N^2) \to 0$. The convergence of $Var(N^{1/2}T_B|\mathbf{X})$ was shown in lemma 3.2.3.

Write $\mathrm{Var}(\sqrt{N}T_Q) = E(\mathrm{Var}(\sqrt{N}T_Q|\mathbf{X})) + \mathrm{Var}(\sqrt{N}E(T_Q|\mathbf{X}))$. Then $\mathrm{Var}(\sqrt{N}E(T_Q|\mathbf{X})) = 0$ and $E(\mathrm{Var}(\sqrt{N}T_Q|\mathbf{X})) - \lim_{N\to\infty}\varphi_N^2 \to 0$.

It is clear that $\mathrm{Var}(\sqrt{N}E(T_Q|\mathbf{X})) = 0$, since by the definition of T_Q in (4.4.4), the observations from different treatment are independent, thus $E(\sqrt{N}T_Q|\mathbf{X}) = 0$ $a.s.$ Next,

$$\begin{aligned}
&E(\mathrm{Var}(\sqrt{N}T_Q|\mathbf{X})) \\
&= E\left\{Var\left(\frac{k}{\sqrt{N}a}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}(\overline{U}_{ic\cdot} - E(\overline{U}_{ic\cdot}|\mathbf{X}))(\overline{U}_{i'c\cdot} - E(\overline{U}_{i'c\cdot}|\mathbf{X}))\,\middle|\,\mathbf{X}\right)\right\}.
\end{aligned} \tag{4.4.8}$$

Note that $U_{ict} = Y_{ij} I(j \in C_{ic})$ and $\overline{U}_{ic\cdot} = k^{-1} \sum_{j=1}^{n_i} Y_{ij}\ I(j \in C_{ic})$, and the indicator functions only depend on $\mathbf{X}_i$ and $\mathbf{X}_{i'}$, and not on $\mathbf{Y}$, so equation (4.4.8) is written as

$$E\left\{Var\left(\frac{k^{-1}}{\sqrt{N}a}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i'}(Y_{ij}-E(Y_{ij}|\mathbf{X}))I(j\in C_{ic})(Y_{i'j'}-E(Y_{i'j'}|\mathbf{X}))I(j'\in C_{i'c})\middle|\mathbf{X}\right)\right\}$$

$$=E\left\{Var\left(\sum_{i\neq i'}^{a}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i'}(Y_{ij}-E(Y_{ij}|\mathbf{X}))(Y_{i'j'}-E(Y_{i'j'}|\mathbf{X}))\middle|\mathbf{X}\right)\left[\sum_{c=1}^{N}\frac{I(j\in C_{ic})I(j'\in C_{i'c})}{\sqrt{N}ak}\right]^2\right\}.$$

Denote $t_{ii'jj'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{i'j'} - E(Y_{i'j'}|\mathbf{X}))$. Then the right hand side of the equation above becomes

$$E\left\{\sum_{i\neq i'}^{a}\sum_{i_1\neq i_1'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}\sum_{j_1'}^{n_{i_1}}\sum_{j_1'}^{n_{i_1'}}\mathrm{Cov}\left(t_{ii'jj'},t_{i_1i_1'j_1j_1'}\middle|\mathbf{X}\right)\left[\sum_{c=1}^{N}\frac{I(j\in C_{ic})I(j'\in C_{i'c})}{\sqrt{N}ak}\right]^2\right\}$$

$$=E\left\{\sum_{i\neq i'}^{a}\sum_{i_1\neq i_1'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}\sum_{j_1}^{n_{i_1}}\sum_{j_1'}^{n_{i_1'}}E\left(t_{ii'jj'}t_{i_1i_1'j_1j_1'}\middle|\mathbf{X}\right)\left[\sum_{c=1}^{N}\frac{I(j\in C_{ic})I(j'\in C_{i'c})}{\sqrt{N}ak}\right]^2\right\}, \quad (4.4.9)$$

where the equality (4.4.9) is true because $E(t_{ii'jj'}|\mathbf{X}) = E[Y_{ij} - E(Y_{ij}|\mathbf{X})|\mathbf{X}]E[Y_{i'j'} - E(Y_{i'j'}|\mathbf{X})|\mathbf{X}] = 0$ due to the fact that observations from different treatments are independent. Further, $E(t_{ii'jj'}\ t_{i_1i1'j_1j1'}|\mathbf{X}) = 0$ in the following four cases:

- Case 1. Either $\{i, i', i_1, i_1'\}$ or $\{j, j', j_1, j_1'\}$ has four different values.
- Case 2. Either $\{i, i', i_1, i_1'\}$ or $\{j, j', j_1, j_1'\}$ has three different values.
- Case 3. $i = i_1$ and $i' = i_1'$ but $j \neq j_1$ or $j' \neq j_1'$.
- Case 4. $i = i_1'$ and $i' = i_1$ but $j \neq j_1'$ or $j' \neq j_1$.

The remaining cases in the summations include $(i, j) = (i_1, j_1)$ and $(i', j') = (i_1', j_1')$ or $(i, j) = (i_1', j_1')$ and $(i', j') = (i_1, j_1)$. In such cases $E(t_{ii'jj'}t_{i_1i1'j_1j1'}|\mathbf{X}) = E[t_{ii'jj'}^2|\mathbf{X}]$. Note that $t_{ii'jj'} = t_{ii'j'j}$. Consequently, denoting $M_{iji'j'} = \sum_c^N I(j \in C_{ic})I(j' \in C_{i'c})$, equa-

tion(4.4.9) is written as

$$E\left\{\frac{2}{Na^2}\sum_{i\neq i'}^{a}\frac{1}{k^2}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}E[t^2_{ii'jj'}|\mathbf{X}]\left(\sum_{r}^{N}I(j\in C_{ic})I(j'\in C_{i'c})\right)^2\right\} \tag{4.4.10}$$

$$\begin{aligned}
&= E\left\{\frac{2}{Na^2}\sum_{i\neq i'}^{a}\frac{1}{k^2}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'})\left(M_{iji'j'}\right)^2\right\}\\
&= E\left\{\frac{2}{Na^2k^2}\sum_{i\neq i'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'})\left[E^2(M_{iji'j'}|S_{ij,t},S_{i'j',t})+\mathrm{Var}(M_{iji'j'}|S_{ij,t},S_{i'j',t})\right]\right\}\\
&= \frac{2}{Na^2k^2}\sum_{i\neq i'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}}E\left\{\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'})\left[E^2(M_{iji'j'}|S_{ij,t},S_{i'j',t})\right.\right.\\
&\quad \left.\left.+\mathrm{Var}(M_{iji'j'}|S_{ij,t},S_{i'j',t})\right]\Delta_{iji'j'}(t_1,t_2)\right\},
\end{aligned} \tag{4.4.11}$$

where $S_{ij,t}=(X_{ij},L_{ij}^{(t/2},U_{ij}^{(t/2})$, $U_{ij}^{t/2}$ and $L_{ij}^{t/2}$ are the upper and lower $t/2$ spacings from X_{ij} and

$$\Delta_{iji'j'}(t_1,t_2)=I\left(\max\{X_{ij}-L_{ij}^{(t_1/2)},\,X_{i'j'}-L_{i'j'}^{(t_2/2)}\}\leq\min\{X_{ij}+U_{ij}^{(t_1/2)},\quad X_{i'j'}+U_{i'j'}^{(t_2/2)}\}\right).$$

Denote X_c as $X_{i_1j_1}$ when X_c refers to the j_1th covariate value in group i_1. Because the augmentation for C_{ic} only uses observations from group i and that for $C_{i'c}$ only uses observations from group i', so event $\{j\in C_{ic}\}$ and $\{j'\in C_{i'c}\}$ are independent. Thus,

$$\begin{aligned}
E(M_{iji'j'}|S_{ij,t},S_{i'j',t}) &= \sum_{r}E[I(j\in C_{ic})I(j'\in C_{i'c})|S_{ij,t},S_{i'j',t}] &(4.4.12)\\
&= \sum_{i_1}\sum_{j_1}P(j\in C_{i,X_{i_1j_1}},j'\in C_{i',X_{i_1j_1}}|S_{ij,t},S_{i'j',t}) &(4.4.13)
\end{aligned}$$

Consider

$$\begin{aligned}
&P(j\in C_{i,X_{i_1j_1}},j'\in C_{i',X_{i_1j_1}}|S_{ij,t},S_{i'j',t})\\
&= \begin{cases}
\int_{\max\{X_{ij}-L_{ij}^{(t_1/2)},\quad X_{i'j'}-L_{i'j'}^{(t_2/2)}\}}^{\min\{X_{ij}+U_{ij}^{(t_1/2)},\quad X_{i'j'}+U_{i'j'}^{(t_2/2)}\}}g_{i_1}(x)dx\Delta_{iji'j'}(t_1,t_2), & i_1\neq i,i' \text{ or } j_1\neq j,j';\\
I(j'\in C_{i',X_{ij}}), & i_1=i,j_1=j;\\
I(j\in C_{i,X_{i'j'}}), & i_1=i',j_1=j',
\end{cases}
\end{aligned} \tag{4.4.14}$$

where $t_1 = t_2 = k$, if $i_1 \neq i, i'$; $t_1 = k-1, t_2 = k$, if $i_1 = i, j_1 \neq j$; $t_1 = k, t_2 = k-1$, if $i_1 = i', j_1 \neq j'$.

Note that

$$E[P(j \in C_{i,X_{i_1 j_1}}, j' \in C_{i',X_{i_1 j_1}} | S_{ij,t}, S_{i'j',t}) | X_{ij}, X_{i'j'}]$$
$$\leq \min\{P(j \in C_{i,X_{i_1 j_1}} | X_{ij}, X_{i'j'}),\ P(j' \in C_{i',X_{i_1 j_1}} | X_{ij}, X_{i'j'})\} = O_p(N^{-1}),$$

and

$$E[\mathrm{Var}(M_{iji'j'} | S_{ij,t}, S_{i'j',t}) | X_{ij}, X_{i'j'})]$$
$$= \sum_{c=1}^{N} [E(I(j \in C_{ic}) I(j' \in C_{i'c}) | S_{ij,t}, S_{i'j',t}) - E^2(I(j \in C_{ic}) I(j' \in C_{i'c}) | S_{ij,t}, S_{i'j',t})].$$

So

$$E[\mathrm{Var}(M_{iji'j'} | S_{ij,t}, S_{i'j',t}) \Delta_{iji'j'}(t_1, t_2) | X_{ij}, X_{i'j'}] = E[M_{iji'j'} \Delta_{iji'j'}(t_1, t_2) | X_{ij}, X_{i'j'}] + O_p(N^{-1}).$$

Note that the indicator function in (4.4.11) ensures that the summation over j and j' becomes a single summation of j from one to N and a summation of j' over finitely many values such that $X_{i'j'}$ is in the neighborhood of X_{ij}. So,

$$E(\mathrm{Var}(\sqrt{N} T_Q | \mathbf{X}))$$
$$= \frac{2}{Na^2k^2} \sum_{i \neq i'}^{a} \sum_{j}^{n_i} \sum_{j'}^{n_{i'}} E\left\{\sigma_i^2(X_{ij}) \sigma_{i'}^2(X_{i'j'}) \left[E^2(M_{iji'j'} | S_{ij,t}, S_{i'j',t}) + E(M_{iji'j'} | S_{ij,t}, S_{i'j',t})\right]\right.$$
$$\left. \Delta_{iji'j'}(t_1, t_2)\right\} + O(N^{-1}).$$

Therefore, $[E(\mathrm{Var}(\sqrt{N} T_Q | \mathbf{X})) - \varphi_N^2] \to 0$ where

$$\varphi_N^2 = \frac{2}{Na^2k^2} \sum_{i \neq i'}^{a} \sum_{j}^{n_i} \sum_{j'}^{n_{i'}} E\left\{\sigma_i^2(X_{ij}) \sigma_{i'}^2(X_{i'j'}) \left[E^2(M_{iji'j'} | S_{ij,t}, S_{i'j',t}) + E(M_{iji'j'} | S_{ij,t}, S_{i'j',t})\right]\right.$$
$$\left. \Delta_{iji'j'}(t_1, t_2)\right\}.$$

There is no closed form for the expression of $E(M_{iji'j'}|S_{ij,t}, S_{i'j',t})$. It is estimated by the total number of times the two covariate values $(X_{ij}, X_{i'j'})$ are both used in the augmentation of cells in the same column.

4.4.2 No Covariate-Treatment Interaction Effect

Lemma 4.4.3. *Under the assumptions in subsection 3.2.1, let* $G_N = P_G(\mathbf{Z}) + R_G(\mathbf{Z})$, *where*

$$P_G(\mathbf{Z}) = \frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2 - \frac{k}{a(a-1)N}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{i'c\cdot} \tag{4.4.15}$$

$$R_G(\mathbf{Z}) = -\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot} + \frac{k}{aN(N-1)(a-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}. \tag{4.4.16}$$

Then $\sqrt{N}R_G(\mathbf{Z}) \xrightarrow{p} 0$ *as* $N \to \infty$.

The expression of $R_G(\mathbf{Z})$ is similar to $R_Q(\mathbf{Z})$ in lemma 4.4.1 except for the multiplicative constant. Because the approaches are similar, the proof of the lemma is omitted.

Theorem 4.4.4. *Under* $H_{0\xi}$ *and the conditions in Theorem 3.2.2, then*

$$\sqrt{N}(G_N - W_N) \to N\left(0, \lim_{N\to\infty}(\gamma_N^2 + \frac{\varphi_N^2}{(a-1)^2})\right),$$

as $N \to \infty$, *where* γ_N^2 *and* φ_N^2 *are given in Theorem 3.2.2 and Theorem 4.4.2 respectively.*

By Lemma 4.4.3, $\sqrt{N}(G_N - W_N)$ has the same asymptotic distribution as $\sqrt{N}(P_G(\mathbf{Z}) - W_N) = T_B - T_Q/(a-1)$, where T_B and T_Q are defined in (3.2.4) and (4.4.4) respectively. The remaining proof follows arguments similar to those for Theorem 3.2.2, and thus is omitted.

4.5 Simulation studies

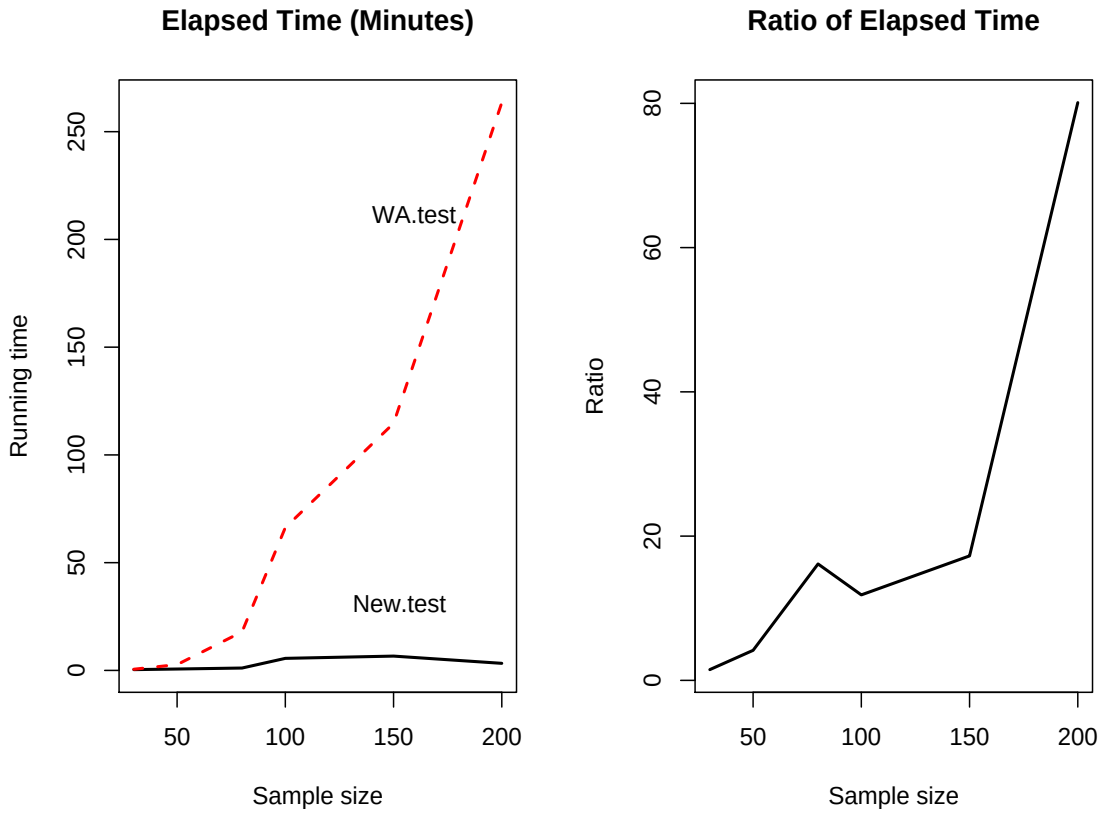
In this section, simulation studies were conducted to compare the performance of the proposed pNP tests to the WA, GAM Spline, GAM loess, drop and CF tests for testing no covariate-treatment interaction and no covariate effects. A situation where the response variable is discrete and from a Bernoulli random variable is also studied.

The simulations were conducted using different window sizes, k, different sample sizes n_i and different types of relationships between response and covariate. When the response variable was continuous, two case were considered; one where there existed a linear dependency and another where there existed a quadratic dependency between the response and the covariate. Unless it is specified, in all of the examples, the random variable X was generated from a uniform distribution $(0, 1)$, and ϵ_i is from $N(0, 1)$. 500 simulations were run for comparison with results reported in Wang and Akritas (2006). To determine the power of the tests, the simulations were run using different values of τ, one set where τ represents the slope of the linear function and another representing the vertical stretch or vertical shrink for the quadratic function. When $\tau = 0$ the proportion of rejections corresponds to the estimate of type I error rate for both cases, linear and quadratic.

4.5.1 Computational Time Comparison

The running time for the proposed pNP test is first compared with WA test. The simulation was conducted under the null hypothesis of no interaction effect for $a = 3$. The computer used for this comparison has an Intel (R) Pentium M processor 1.86GHz, 1GB of RAM. Figure 4.2 gives the running time (titled Elapsed Time) and the ratio of the running time (titled Ratio of Elapsed Time) for the pNP test over the WA test for sample sizes from $n_i = 10$ to $n_i = 200$. Comparing the running time, the pNP test is preferred over the WA test because the pNP test is computationally less extensive than the WA test. This is seen from the exponential increase in the ratio of elapsed time of WA and pNP tests. This comparison supports the proposed pNP test as a much more computationally efficient test compared to the WA test. An explanation for the observed phenomenon is that the asymptotic variance calculation for WA test has computational complexity of $O(N^2)$. On the other hand the new test has complexity O(N) for the asymptotic variance calculation.

Figure 4.2: Computational Time Comparison



4.5.2 Linear Alternative

Numerical results for covariate-treatment interaction effect

For this section, the study of the type I error estimate and power performance for the proposed covariate-treatment interaction test when the underlying relationship between response and covarite is linear is presented. The responses are generated from

$$Y_{1j} = 0.1\epsilon_{1j} \quad \text{and} \quad Y_{2j} = \tau X_{2j} + 0.1\epsilon_{2j}. \tag{4.5.1}$$

The type I error estimates and power performance at 0.05 level for the test of covariate-treatment interaction are summarized in Table 4.1. The WA test result is taken from Table 1 in Wang and Akritas (2006) because their results could not be reproduced. The estimated type I error corresponds to results when $\tau = 0$. The results indicate that the WA test of

no covariate-treatment interaction does not have a consistent control of the intended type I error for smaller k. For instance, for $n_i = 30$ and $k = 3$ the estimated type I error of 0.090 for the WA test is almost twice as much as α, while the estimated type I error for the pNP test is 0.030 which is an acceptable level. Further, when $n_i = 30$ and $k = 7$, the estimated type I error improved to 0.062. On the other hand, the type I error estimates for the pNP test are all acceptable for different values of k. Thus, in this situation, the pNP test of no covariate treatment interaction effect is preferable to the WA test.

Table 4.1 also shows the CF test for no covariate-treatment interaction has a good type I error estimate and power performance. The type I error estimates for case $n_i = 30$, is 0.058. This is mainly because the CF test is uniformly the most powerful invariant test (UMPI) in this data generation setting where the assumptions of normality, independence and homogeneity of variance are satisfied, aside from the linear association assumption. Like the CF test, the type I error estimate and the performance of the drop test is also good because the error distribution for this example satisfies the needed assumption for the drop test mentioned in section 2.1.4. In addition, the type I error estimate and power performance for GAM spline and GAM Loess are also included in the study. When $n_i = 30$, the two GAM tests have liberal type I error estimates even though the power performance is good.

For $n_i = 50$, the result from Table 4.1 shows that the pNP test for no treatment-covariate interaction effect has an acceptable level of the type I error estimate for the choice of $k = 5, 7$ and 9 and better estimates compared to the case where $n_i = 30$. On the other hand, the type I error estimate for the WA test for no covariate-treatment interaction effect is still slightly elevated when $k = 5$, but improved for $k = 7$ and 9. The power for the pNP test of no treatment-covariate interaction effect is comparable to that of the WA test. On the other hand, when $n_i = 50$, the type I error estimate and power performance for CF, drop, GAM spline and GAM Loess tests are better when compared to situation when $n_i = 30$.

In summary, the pNP test consistently produces acceptable type I errors for all sample sizes considered while maintaining comparable power to the WA test. In addition, for the pNP test, for the same sample sizes, a larger window size k yields a smaller type I error but

a higher power. On the contrary, the type I error for the WA test depends on both k and n_i and it is liberal for smaller n_i. This observation is consistent with the slow convergence rate of the test statistics used for the WA test. All tests have better power performance for larger sample sizes.

Table 4.1: Proportion of rejections at level 0.05 for testing no covariate-treatment interaction effect for WA, pNP, GAM Spline and GAM Loess, Drop and CF tests. The response variable is from the linear alternative model (4.5.1). The results are based on 500 simulations.

n_i	k	τ	WA.int	pNP.int	GAM Spline	GAM loessess	Drop.int	CF.int
30	3	0.000	0.090	0.030	0.070	0.076	0.052	0.058
		0.100	0.150	0.088	0.176	0.179	0.196	0.182
		0.200	0.395	0.260	0.553	0.556	0.550	0.588
		0.300	0.651	0.560	0.854	0.862	0.864	0.890
	5	0.000	0.076	0.026				
		0.100	0.145	0.090				
		0.200	0.413	0.296				
		0.300	0.694	0.678				
	7	0.000	0.062	0.020				
		0.100	0.141	0.104				
		0.200	0.427	0.304				
		0.300	0.707	0.696				
50	5	0.000	0.062	0.038	0.041	0.041	0.036	0.042
		0.100	0.172	0.120	0.260	0.268	0.266	0.274
		0.200	0.529	0.442	0.792	0.802	0.814	0.830
		0.300	0.872	0.872	0.992	0.992	0.992	0.992
	7	0.000	0.052	0.034				
		0.100	0.175	0.148				
		0.200	0.564	0.484				
		0.300	0.908	0.904				
	9	0.000	0.051	0.028				
		0.100	0.185	0.140				
		0.200	0.592	0.512				
		0.300	0.915	0.910				

Numerical results for main covariate effect in the absence of no covariate-treatment interaction

In this subsection the estimate of type I error and power performance of the pNP test for the test of no main covariate effect is presented. For the p values of the test of no main covariate

effect to be meaningful, the simulation needs to be conducted when there is no covariate-treatment interaction in the model. Therefore, the response variable was generated by the following model:

$$Y_{ij} = \tau X_{ij} + 0.1\epsilon_{ij}, \quad i = 1, 2. \tag{4.5.2}$$

The results of the simulation presented in Table 4.2 show that for $n_i = 30$, all the tests of interest except the GAM Spline have a good type I error estimate. The GAM Spline has a slightly inflated type I error estimate of 0.06. For $n_i = 50$, the type I error estimate for GAM Spline (0.084) and GAM Loess (0.076) are quite inflated at 0.05 level. On the other hand, the pNP, drop and CF tests have good type I error estimates and power performances. In addition, the CF test has the best performance using this model as the power approaches unity the fastest, which is to be expected for the same reasons mentioned in the above subsection, i.e., assumptions of normality, independence and homogeneity of variance and linear association are satisfied.

4.5.3 Quadratic Alternative

Covariate-Treatment Interaction

The study of the type I error and power performance for the proposed pNP test for covariate-treatment interaction when the underlying relationship between the covariate and response is quadratic is presented in this subsection. The responses are generated as follows:

$$Y_{1j} = 0.1\epsilon_{ij}, \qquad Y_{2j} = \tau(X_{2j}^2 - X_{2j} + 0.15) + 0.1\epsilon_{ij} \tag{4.5.3}$$

The results for the type I error estimates and power performances at 0.05 level for tests of no covariate-treatment interaction study are summarized in Table 4.3. The result of the WA test is taken from Table 2 of Wang and Akritas (2006).

Similar to the study of the linear alternative, because under the null hypothesis the estimated type I error is quite large (0.089), the power of the test of no treatment-covariate

Table 4.2: Proportion of rejections at level 0.05 for testing of no main covariate effect for WA Test, pNP test, GAM spline and GAM Loess, Drop test and CF test. The response variable is from linear alternative (4.5.2). The results are based on 500 simulations.

n	k	τ	WA.cov	pNP.cov	GAM Spline	GAM Loess	drop	CF
30	3	0.000	0.036	0.034	0.060	0.056	0.032	0.040
		0.100	0.224	0.218	0.416	0.414	0.532	0.572
		0.200	0.832	0.836	0.956	0.956	0.992	0.998
		0.300	0.994	0.994	1.000	1.000	1.000	1.000
	5	0.000	0.034	0.034				
		0.100	0.302	0.292				
		0.200	0.916	0.914				
		0.300	0.994	0.994				
	7	0.000	0.034	0.036				
		0.100	0.314	0.326				
		0.200	0.928	0.928				
		0.300	0.996	0.996				
50	5	0.000	0.042	0.046	0.084	0.076	0.044	0.042
		0.100	0.440	0.442	0.666	0.672	0.788	0.806
		0.200	0.980	0.982	0.998	0.998	1.000	0.998
		0.300	1.000	1.000	1.000	1.000	1.000	1.000
	7	0.000	0.044	0.046				
		0.100	0.484	0.474				
		0.200	0.990	0.992				
		0.300	1.000	1.000				
	9	0.000	0.038	0.038				
		0.100	0.516	0.520				
		0.200	0.994	0.992				
		0.300	1.000	1.000				

interaction effect for the WA test for $n_i = 30$ and $k = 3$ is not very reliable. On the other hand, for $n_i = 30$, the pNP test has an acceptable type I error and a good power performance for each value of k and n_i considered. The type I error estimate from the GAM Spline and GAM Loess tests are inflated when $n_i = 30$. In addition, the GAM spline and GAM Loess tests do not have power to detect the covariate-treatment interaction in this setting. Similarly, the drop and the CF tests also do not have power although the estimates of the type I error are acceptable.

When $n_i = 50$, the pNP test has conservative type I error estimates for all values of k used. The WA test has a high value type I error estimate when $k = 5$, but has an

acceptable estimate for $k = 7$ and $k = 9$. On the other hand, both the pNP and the WA tests for covariate-treatment interaction have comparable power for large k but the WA test is liberal for $k = 5$. Although all the other tests have acceptable type I error estimates, they do not have power at all for both $n_i = 30$ and $n_i = 50$.

Table 4.3: Proportion of rejections for testing no covariate-treatment interaction effect using WA, pNP, Drop and CF tests. The response variable is from the quadratic model (4.5.3). The results are based on 500 simulations.

n	k	τ	WA	pNP	GAM Spline	GAM Loess	drop	CF
30	3	0.000	0.089	0.032	0.070	0.076	0.058	0.058
		0.500	0.198	0.106	0.069	0.068	0.058	0.068
		1.000	0.498	0.372	0.064	0.062	0.070	0.076
		1.500	0.812	0.748	0.086	0.080	0.100	0.078
30	5	0.000	0.060	0.034				
		0.500	0.162	0.096				
		1.000	0.438	0.380				
		1.500	0.813	0.770				
	7	0.000	0.042	0.038				
		0.500	0.129	0.092				
		1.000	0.397	0.324				
		1.500	0.742	0.738				
50	5	0.000	0.070	0.028	0.041	0.041	0.036	0.042
		0.500	0.239	0.170	0.050	0.051	0.040	0.054
		1.000	0.670	0.656	0.056	0.063	0.050	0.060
		1.500	0.964	0.956	0.079	0.074	0.088	0.076
	7	0.000	0.053	0.028				
		0.500	0.236	0.178				
		1.000	0.687	0.656				
		1.500	0.964	0.942				
	9	0.000	0.042	0.034				
		0.500	0.213	0.174				
		1.000	0.666	0.648				
		1.500	0.961	0.940				

Main Covariate Effect

This subsection studies the type I error estimate and the power performance of the proposed pNP test of main covariate effect when there is no covariate-treatment interaction and the underlying relationship of response and covariate are quadratic. The data are generated as

Table 4.4: Proportion of rejections at level 0.05 for testing no main covariate effect for WA Test, pNP test, GAM spline and GAM Loess tests, Drop test and CF test. The response variable is from the quadratic model (4.5.4). The results are based on 500 simulations.

n	k	tau	WA.cov	pNP.cov	GAM Spline	GAM Loess	drop	CF
30	3	0.000	0.036	0.034	0.060	0.056	0.032	0.040
		0.500	0.350	0.348	0.636	0.646	0.028	0.042
		1.000	0.912	0.916	1.000	0.998	0.052	0.058
		1.500	0.998	0.998	1.000	1.000	0.076	0.072
	5	0.000	0.034	0.032				
		0.500	0.358	0.362				
		1.000	0.940	0.942				
		1.500	1.000	1.000				
	7	0.000	0.034	0.034				
		0.500	0.302	0.312				
		1.000	0.918	0.926				
		1.500	0.998	0.998				
50	5	0.000	0.042	0.042	0.084	0.076	0.044	0.042
		0.500	0.632	0.632	0.892	0.900	0.048	0.056
		1.000	1.000	1.000	1.000	1.000	0.062	0.068
		1.500	1.000	1.000	1.000	1.000	0.086	0.082
	7	0.000	0.044	0.046				
		0.500	0.660	0.668				
		1.000	1.000	1.000				
		1.500	1.000	1.000				
	9	0.000	0.040	0.038				
		0.500	0.634	0.654				
		1.000	1.000	1.000				
		1.500	1.000	1.000				

follows:

$$Y_{ij} = \tau(X_{ij}^2 - X_{ij} + 0.15) + 0.1\epsilon_{ij} \quad i = 1, 2 \tag{4.5.4}$$

The results for the simulation of main covariate effect are recorded in Table 4.4. As expected, the CF and drop tests did not have power at all to detect the quadratic relationship of response and covariate. On the other hand, the proposed pNP, WA, GAM Spline and GAM Loess tests all have good estimates of type I error and have high power to detect the underlying quadratic relationship of response and covariate.

4.5.4 Bernoulli Responses

Covariate-Treatment Interaction

This subsection of the simulation study discusses the type I error estimate and the power performance of the pNP test for the response variable emerging from Bernoulli trials. For testing the covariate-treatment interaction effect, data was generated as follows:

$$\begin{cases} y_{1j} = Bernoulli\left(\frac{\exp(\tau cos(2\pi X_{1j}))}{(1+\exp(\tau\cos(2\pi X_{1j})))}\right) & \tau = 0, 1, 2, 3, 4 \\ y_{ij} = Bernoulli\ (0.5) \quad \text{for } i = 2, 3. \end{cases} \tag{4.5.5}$$

The estimates of type I error and the performance of covariate-treatment interaction effect of the pNP test are compared with the WA, GLM Wald and GLM deviance tests. The pNP test is used when the response variable is discrete because there is no distributional restriction when calculating the asymptotic distribution of the test statistics. The GLM Wald and GLM deviance tests are those commonly used when the response variable is discrete, especially response variables from Bernoulli trials.

Table 4.5 summarizes the type I error estimates and power performance for testing the covariate-treatment interaction effect. The column labeled as "WA.int" is copied from Table 3 of Wang and Akritas (2006). The results show that the WA and pNP tests perform reasonably well for sample size $n_i = 30$. When $n_i = 50$, the performance of WA and pNP tests improve as window size increases. The GLM Wald and GLM deviance tests perform well under the null hypothesis but do not have power to detect the interaction when the alternative specified in equation (4.5.5) is true.

Main Covariate Effect

A simulation was conducted to study the type I error estimate and the power of the pNP test for main covariate effect when the response variable is discrete, and where interaction

does not exist between treatment and covariate under the following model:

$$y_{ij} = Bernoulli\left(\frac{\exp(\tau cos(2\pi X_{ij}))}{(1+\exp(\tau\cos(2\pi X_{ij})))}\right) \; i = 1, 2, 3 \tag{4.5.6}$$

In this model, the relationship of response and covariate is through the probability of success in the Bernoulli trials. The result of the simulation study is presented in Table 4.6. From the result, it is clear that the pNP test has a good type I error of 0.042 at α level of 0.05. It also has a very good power to detect the main covariate effect in the absence of covariate-treatment interaction effect. On the other hand both the GLM deviance and GLM Wald tests do not have any power to detect the main covariate effect in this simulation.

4.6 Data Analysis

This section continues the analysis of the Ozone concentration data from section 3.3.1 and the EFT data from section 3.3.2 testing for covariate-treatment interaction and main covariate effects. The tests being used for the data analysis comparison are the pNP, GAM Spline, GAM Loess, drop and CF tests.

4.6.1 Analysis of Ozone Concentration Data (continued from Chapter 3)

Recall that the variables of interest in the Ozone data are response variable ozone (O_3), covariate variable day of year (doy) and a factor wind speed which has 4 levels: low, medium, medium high and high. Table 4.7 gives the p-values for testing the doy-wind speed interaction effect. The results show that the only test that is significant is the pNP test.

On the other hand, the test of no main covariate effect is not meaningful in the presence of the covariate-treatment interaction effect. Because the interaction effect is not significant for GAM Spline, GAM Loess, drop and CF tests, the main doy effect is tested. The results are given in Table 4.8, showing significant main doy effect for GAM Spline and GAM Loess but not significant for the drop and the CF tests.

4.6.2 Application to EFT Study (continued from Chapter 3)

This section continues the EFT data analysis from section 3.3.2. The response variable of interest is the time in seconds for fifth grade students in a state primary school in a Sydney suburb to finish a test of the Wechsler Intelligence Scale for Children (WISC). The covariate variable is "field dependence" which is measured by providing an Embedded Figures Test (EFT) within the test. Subjects were assigned different places of instruction in either a corner group or a row group. In this dataset, there exist a pair of observations (139, 739) from the row group that is considered as an outlier. The EFT-group interaction test was performed using pNP, GAM Spline, GAM Loess, drop and CF tests from the original data and when the outlier was replaced by the median time to finish the WISC test. The p-values are recorded in Table 4.9. The results indicate nonsignificant interaction for covariate-treatment interaction for all the tests when the original data was used and when the outlier was replaced by the median time from the row group.

On the other hand, for testing main the EFT effect, when the original data was used, all the other tests except the pNP test are significant at $\alpha = 0.05$. When the outlier is replaced by the median to finish the WISC test, all the other tests became nonsignificant except the pNP test. This is also the case in reference to the results of tests of independence in Section 3.3.2. Therefore, the pNP test has a consistent result when the original observation is used and when the outlier is replaced by the median observation in the test of independence of response and the EFT adjusted for treatment effect (in Chapter 3) and in the test of no main EFT effect.

Table 4.5: Proportion of rejections at level 0.05 for testing no covariate-treatment interaction effect for WA, pNP, GLM Wald GLM and Deviance tests. The response variable is from the model in (4.5.5). The results are based on 500 simulations.

n_i	k	τ	WA.int	pNP.int	GLM.Dev	GLM.Wald
30	3	0	0.060	0.036	0.060	0.044
		1	0.090	0.100	0.072	0.034
		2	0.306	0.234	0.074	0.044
		3	0.486	0.402	0.070	0.058
		4	0.590	0.542	0.076	0.062
	5	0	0.026	0.016		
		1	0.074	0.084		
		2	0.266	0.248		
		3	0.458	0.486		
		4	0.596	0.628		
	7	0	0.016	0.022		
		1	0.062	0.062		
		2	0.206	0.208		
		3	0.380	0.428		
		4	0.518	0.586		
50	5	0	0.038	0.030	0.064	0.054
		1	0.094	0.094	0.050	0.038
		2	0.468	0.470	0.064	0.046
		3	0.746	0.750	0.052	0.048
		4	0.874	0.888	0.052	0.052
	7	0	0.030	0.024		
		1	0.096	0.096		
		2	0.490	0.504		
		3	0.800	0.816		
		4	0.890	0.926		
	9	0	0.028	0.026		
		1	0.078	0.110		
		2	0.482	0.506		
		3	0.788	0.828		
		4	0.910	0.930		

Table 4.6: Proportion of rejections at level 0.05 for testing no main covariate effect for WA, pNP, GLM Wald and GLM deviance tests. The response variable is from the model in (4.5.6). The results are based on 500 simulations.

n	k	tau	pNP.cov	GLM.Dev.cov	GLM.Wald.cov
30	3	0	0.042	0.052	0.052
		1	0.588	0.052	0.052
		2	0.992	0.056	0.056
		3	1.000	0.062	0.056
		4	1.000	0.066	0.058
	5	0	0.048		
		1	0.600		
		2	0.994		
		3	1.000		
		4	1.000		
	7	0	0.034		
		1	0.558		
		2	0.992		
		3	1.000		
		4	1.000		
50	5	0	0.042	0.056	0.050
		1	0.588	0.042	0.042
		2	0.992	0.060	0.056
		3	1.000	0.048	0.044
		4	1.000	0.046	0.040
	7	0	0.048		
		1	0.600		
		2	0.994		
		3	1.000		
		4	1.000		
	9	0	0.034		
		1	0.558		
		2	0.992		
		3	1.000		
		4	1.000		

Table 4.7: P values for test of no doy-wind speed interaction effect of the ozone data

pNP	GAM Spline	GAM Loess	drop	CF
0.000	0.707	0.761	0.445	0.501

Table 4.8: P values for test of no main doy effect of the ozone data

GAM Spline	GAM Loess	drop	CF
0.000	0.000	0.210	0.064

Table 4.9: P values for test of no EFT-group interaction effect of the EFT data

	pNP	GAM Spline	GAM Loess	drop	CF
Original data	0.862	0.675	0.404	0.192	0.128
Outlier replaced	0.773	0.496	0.390	0.724	0.654

Table 4.10: P values for test of no main EFT effect of the EFT data

	pNP	GAM Spline	GAM Loess	drop	CF
Original data	0.474	0.015	0.026	0.013	0.006
Outlier replaced	0.569	0.413	0.386	0.129	0.142

Chapter 5

New Nonparametric Tests when Treatment Level a is Large

5.1 Tests of No Covariate-Treatment Interaction, No Main Covariate and No Simple Covariate Effects when a is Large

Statistical studies are often conducted in a setting where the treatment levels are large. Because the method in the previous chapters is constructed under a fixed number of treatment levels, it is not suitable for data sets with a large number of treatment levels. For example, in the ozone data discussed in Section 3.3.1, the total number of wind levels is 11. Wind levels have to be combined in a manner that is appropriate in terms of their application that make them suitable to be analyzed using the proposed tests from the previous two chapters. This chapter discusses the asymptotic distributions theory, simulation studies and an application for testing of no covariate-treatment interaction, no main covariate and no covariate simple effects when the number of treatment level a and the covariate values n_i in each treatment level are large, thus making the total covariate values N also large.

The model (4.2.7) in Section 4.2 will be used for the construction of the hypotheses and test statistics for the case of large a and large N to test no main covariate, no covariate-treatment interaction and no simple covariate effect. Assume the conditional distribution

of Y_{ij} given $X_{ij} = x$ is $F_i(y|x)$, the conditional mean of Y_{ij} given $X_{ij} = x$, is decomposed into model (4.2.7) below;

$$E(Y_{ij}|X_{ij} = x) = \mu_{ix} = \mu + \alpha_i + \eta(x) + \xi_i(x),$$

where X is a continuous covariate with probability distribution function $f_X(x)$ and its cumulative distribution function $F_X(x)$, $\mu = \frac{1}{a}\sum_{i=1}^{a}\int \mu_{ix}\, dF_X(x)$, $\mu_i = \int \mu_{ix}\, dF_X(x)$, $\mu_x = \frac{1}{a}\sum_i \mu_{ix}$, $\alpha_i = \mu_i - \mu$, $\eta(x) = \mu_x - \mu$, and $\xi_i(x) = \mu_{ix} - \mu_i - \mu_x + \mu$. The hypotheses for no covariate-treatment interaction, no main covariate and no simple covariate effects using the above model are:

$$H_{0\xi} : \xi_i(x) = 0 \text{ for all } i \text{ and } x,$$

$$H_{0\eta} : \eta(x) = 0 \text{ for all } x,$$

$$H_{0\phi} : \eta(x) + \xi_i(x) = 0 \text{ for all } i, \text{ and } x,$$

respectively. The test statistics are constructed by treating the covariate as a factor with levels c ranging from the smallest to the largest covariate values as in sections 3.2.1 and 4.3. We still augment each cell (i, c) with k nearest neighbors using observations from the i^{th} treatment level and denote U_{ict} to be the observations in the augmented cell (i, c) defined in Section 3.2.1. Instead of having a standardized rate of $\sqrt{N}$, the test statistics for testing no covariate treatment interaction and no simple covariate effects have a standardizing rate of $\sqrt{aN}$, while the standardizing rate for testing the no main covariate effect is $\sqrt{N}$. The test statistics $T^1_{int} = \sqrt{aN}(G_N - W_N)$, $T^1_{cov} = \sqrt{N}(Q_N - W_N)$, and $T^1_{sim} = \sqrt{aN}(B_N - W_N)$ will be used to test the hypotheses of no covariate-treatment interaction $H_{0\xi}$, no main covariate $H_{0\eta}$ and the no simple covariate $H_{0\phi}$ effects respectively, where letting $\overline{U}_{ic\cdot} = k^{-1}\sum_{t=1}^{k} U_{ict}$,

$\overline{U}_{i\cdot\cdot} = N^{-1}\sum_{c=1}^{N}\overline{U}_{ic\cdot}$, then

$$G_N = k(a-1)^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot} - \overline{U}_{\cdot c\cdot} + \overline{U}_{\cdots}\right)^2,$$

$$Q_N = ak(N-1)^{-1}\sum_{c=1}^{N}\left(\overline{U}_{\cdot c\cdot} - \overline{U}_{\cdots}\right)^2,$$

$$B_N = ka^{-1}(N-1)^{-1}\sum_{i}\sum_{c=1}^{N}\left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot}\right)^2,$$

$$\text{and } W_N = \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t=1}^{k}\left(U_{ict} - \overline{U}_{ic\cdot}\right)^2.$$

The next section discusses the asymptotic distribution of the test statistics mentioned here.

5.2 Asymptotic Distribution of Test Statistics Under the Null Hypotheses

Before the asymptotic distribution of the test statistics T^1_{int}, T^1_{cov} and T^1_{sim} are obtained, the projection of the test statistics are needed. As in Section 3.2.2, denote $Z_{ict} = U_{ict} - E(U_{ict}|\mathbf{X})$, then G_N, Q_N and B_N defined in the previous section are projected onto the space span by function of $\{\mathbf{Z}_c, c = 1, \ldots, N\}$, where $\mathbf{Z}_c = (Z_{1c1}, \ldots, Z_{ack})'$. In the following subsections, lemmas for the projection of the test statistics will be presented preceding the theorems for the asymptotic distributions of the test statistics.

5.2.1 Test of No Covariate-treatment Interaction Effect.

This subsection consists of a lemma and a theorem toward the establishment of the asymptotic distribution of the test statistics T^1_{int} in the previous section. The following lemma derives the projection of G_N.

Lemma 5.2.1. *Write* $G_N = P_G(\mathbf{Z}) + R_G(\mathbf{Z})$, *where*

$$P_G(\mathbf{Z}) = \frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2 - \frac{k}{a(a-1)N}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{i'c\cdot}$$

and

$$R_G(\mathbf{Z}) = -\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot} + \frac{k}{aN(N-1)(a-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}$$

If the assumptions in Section 3.2.1 are satisfied, then as a *and* N *go to* ∞,

$$\sqrt{aN}G_N - \sqrt{aN}\frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2 = o_p(1)$$

The lemma states that only the first term in the four term decomposition of G_N is important. The other three terms are asymptotically negligible.

Proof of Lemma 5.2.1 First write $P_G(\mathbf{Z}) = P_G^{(1)} + P_G^{(2)}$, where $P_G^{(1)} = \frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2$ and $P_G^{(2)} = \frac{k}{a(a-1)N}\sum_{i\neq i'}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{i'c\cdot}$, and $R_G(\mathbf{Z}) = R_G^{(1)} + R_G^{(2)}$, where $R_G^{(1)} = -\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}$ and $R_G^{(2)} = \frac{k}{aN(N-1)(a-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}$. The proof of Lemma 5.2.1 involves showing the convergence in probability of $\sqrt{aN}P_G^{(2)}$, $\sqrt{aN}R_G^{(1)}$ and $\sqrt{aN}R_G^{(2)}$ to 0 as $a, N \to \infty$. First, $E(\sqrt{aN}P_G^{(2)}) \to 0$, because observations from different treatments are independent. Notice that $P_G^{(2)} = \frac{1}{a-1}T_Q$, where T_Q is defined in equation (4.4.4) in the proof of Theorem 4.4.2. Therefore,

$$\mathrm{Var}(\sqrt{aN}P_G^{(2)}) = \frac{a}{(a-1)^2}\mathrm{Var}(\sqrt{N}T_Q).$$

Following the argument in the proof of theorem 4.4.2, it can be seen that $\mathrm{Var}(\sqrt{aN}P_G^{(2)}) = O(a^{-1})$. Next, it will be shown that $\sqrt{aN}R_G^{(1)}$ converges to 0 in probability as a and N go

to infinity.

$$\begin{aligned} E(\sqrt{aN}R_G^{(1)}) &= E\left(-\frac{\sqrt{aN}k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right) \\ &= -E\left[\frac{\sqrt{N}k}{N(N-1)}\sum_{c\neq c'}^{N}\frac{\sqrt{a}}{a}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right] \\ &= O\left(\frac{\sqrt{a}}{\sqrt{N}}\right) = O\left(\frac{1}{\sqrt{\frac{1}{a}\sum_i^a n_i}}\right). \end{aligned}$$

Therefore, as a and n_i approaches infinity, $E(\sqrt{aN}R_G^{(1)})$ converges to 0. Moreover,

$$\begin{aligned} \mathrm{Var}(\sqrt{aN}R_G^{(1)}) &= \mathrm{Var}\left(-\frac{\sqrt{aN}k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right) \\ &= \frac{aNk^2}{a^2N^2(N-1)^2}\mathrm{Var}\left(\sum_{c\neq c'}^{N}\sum_{i=1}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right) \\ &= \frac{aNk^2}{a^2N^2(N-1)^2}\sum_{c\neq c'}^{N}\sum_{i}\sum_{c_1\neq c_{1'}}^{N}\sum_{i_1}E\left[\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_1c_{1'}\cdot}\right] \\ &= A + B \end{aligned}$$

where A is the case where $i \neq i_1$ and $A = O(\frac{Na}{(N-1)^2}) = O(\frac{a}{N})$, which converge to 0 as a and n_i go to infinity. On the other hand B is for the case when $i = i_1$ and $B = O(\frac{a^2N^2}{a^2N^2(N-1)^2}) = O(\frac{1}{(N-1)^2})$. So $\mathrm{Var}(\sqrt{aN}R_G^{(1)}) \to 0$ because $n_i \to \infty$.

Lastly, $E(\sqrt{aN}R_G^{(2)}) = 0$ because observations from different treatments are independent

and

$$\begin{aligned}
\mathrm{Var}(\sqrt{aN}R_G^{(2)}) &= \mathrm{Var}\left[\sqrt{aN}\frac{k}{aN(N-1)(a-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}^{a}\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}\right] \\
&= \frac{aNk^2}{a^2N^2(N-1)^2(a-1)^2}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_{1'}}^{N}\sum_{i\neq i'}^{a}\sum_{i_1\neq i_{1'}}^{a}E\left[\overline{Z}_{ic\cdot}\overline{Z}_{i'c'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_{1'}c_{1'}\cdot}\right] \\
&\leq \frac{aNk^2}{a^2N^2(N-1)^2(a-1)^2}O(a(a-1)N^2k^2) \\
&= O(a^{-1}N^{-1})
\end{aligned}$$

The inequality above comes from the arguments similar to the proof of lemma 4.4.1 in equation (4.4.3) and when $i = i_1$.

Theorem 5.2.2. *Assume that $H_{0\xi}$ is true and the assumptions in Section 3.2.1 are satisfied. Then as a and $N \to \infty$,*

$$\sqrt{aN}[G_N - W_N] \xrightarrow{\mathcal{L}} N(0, \lim_{a,N\to\infty} a\gamma_N^2)$$

where

$$\begin{aligned}
\gamma_N^2 &= \frac{4}{Na^2(k-1)^2}\sum_{i}^{a}\sum_{j<j'}^{n_i}E\left\{\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})\left[B_{ijj'}^2 + B_{ijj'}\right.\right. \\
&\quad \left.\left. -2I(j'_* - j_* \leq (k-1)/2)\right]\right\} I(j'_* - j_* \leq k-1) + O(N^{-1}),
\end{aligned}$$

with $B_{ijj'} = \sum_{i_1,i_1\neq i}^{a}\left(\frac{n_{i_1}}{n_i}d_{i_1i}(X_{ij}) + 1\right)[k-(j'_*-j_*)]I(j'_*-j_* \leq k-1)$, $d_{i_1i}(x) = f_{X,i_1}(x)/f_{X,i}(x)$ and $j_* < j'_*$, where j_*, j'_* are the ranks of X_{ij} and $X_{ij'}$ among covariate values in treatment i.

Proof of Theorem 5.2.2 By Lemma 5.2.1,

$$
\begin{aligned}
& \sqrt{aN}(G_N - W_N) \\
= \; & \sqrt{aN}\left[\left(\frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2\right) - \left(\frac{1}{aNk}\sum_{i=1}^{a}\sum_{c=1}^{N}s\sum_{t}Z_{ict}^2 - \frac{1}{Nak(k-1)}\sum_{i}^{N}\sum_{c}\sum_{t\neq t'}Z_{ict}Z_{ict'}\right)\right] \\
= \; & \sqrt{aN}\left[\left(\frac{1}{aNk}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t}Z_{ict}^2 + \frac{1}{aNk}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}Z_{ict}Z_{ict'}\right) - \right. \\
& \left.\left(\frac{1}{aNk}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t}Z_{ict}^2 - \frac{1}{Nak(k-1)}\sum_{i}^{N}\sum_{c}\sum_{t\neq t'}\overline{Z}_{ict}\overline{Z}_{ict'}\right)\right] \\
= \; & \sqrt{aN}(T_B)
\end{aligned}
$$

where $T_B = \frac{1}{aN(k-1)}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}Z_{ict}Z_{ict'}$ is defined in equation (3.2.4). First, the convergence of the asymptotic variance of $\sqrt{aN}(T_B)$ will be shown. Write $\mathrm{Var}(\sqrt{aN}T_B) = E(\mathrm{Var}(\sqrt{aN}T_B|\mathbf{X})) + \mathrm{Var}(\sqrt{aN}E(T_B|\mathbf{X}))$. Similar to equation (3.2.17), $E[\sqrt{aN}T_B|\mathbf{X}] \to 0$, and from equation (3.2.25), $[E[\mathrm{Var}(\sqrt{aN}T_B)] - a\gamma_N^2] \to 0$ where γ_N^2 is defined in Theorem 5.2.2. The proof for the asymptotic normality is presented in Lemma 5.2.3 below.

Lemma 5.2.3. *Under the null hypothesis of no covariate-treatment interaction $H_{0\xi}$ the test statistic $\sqrt{aN}(G_N - W_N)$ is asymptotically normal.*

Proof of Lemma 5.2.3 From the Lemma 5.2.1, $\sqrt{aN}(G_N - W_N)$ has the same asymptotic distribution as $\sqrt{aN}T_B$ as a and N go to infinity. It remains to prove that $\sqrt{aN}T_B$ is asymptotically normal. Similar to the proof of Theorem 3.2.2, let $t^{(2)}_{iji'j'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{i'j'} - E(Y_{i'j'}|\mathbf{X}))K_{ijj'}$, where $K_{ijj'}$ is defined in (3.2.6), and write

$$
\sqrt{aN}T_B = \frac{\sqrt{aN}}{Na(k-1)}\sum_{i,i',j,j'}t^{(2)}_{iji'j'}I(i=i')I(j\neq j') = \sum_{1\leq l_1\leq N}\sum_{1\leq l_2\leq N}V^G_{l_1l_2},
$$

where $l_1 = l(i,j)$ and $l_2 = l(i,j')$ are defined through a one to one index mapping function

$$
l(i,j) = \begin{cases} j & \text{for } i = 1 \\ \sum_{i_2=1}^{i-1}n_{i_2} + j & \text{for } i > 1, \end{cases} \tag{5.2.1}
$$

and

$$V^G_{l_1l_2} = \begin{cases} \frac{\sqrt{aN}}{Na(k-1)}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))\ K_{l_1l_2} & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise.} \end{cases} \tag{5.2.2}$$

Here $K_{l_1l_2}$ is same as $K_{ijj'}$ but using index l_1, l_2:

$$K_{l_1l_2} = \begin{cases} \sum_{i_1}^{a}\sum_{j_1}^{n_{i_1}} I(l_1 \in C_{iX_{i_1j_1}})I(l_2 \in C_{iX_{i_1j_1}}) & \text{for } i = 1 \\ \sum_{i_1}^{a}\sum_{j_1}^{n_{i_1}} I(\sum_{i_2=1}^{i-1} n_{i_2} + l_1 \in C_{iX_{i_1j_1}})I(\sum_{i_2=1}^{i-1} n_{i_2} + l_2 \in C_{iX_{i_1j_1}}) & \text{for } i > 1. \end{cases}$$

Notice $V^G_{l_1l_2} = V^G_{l_2l_1}$. Therefore,

$$\sqrt{aN}T_B = 2\sum_{1\le l_1<l_2\le N} V^G_{l_1l_2} \tag{5.2.3}$$

is a clean quadratic form as in de Jong (1987). In order to show that $\mathrm{Var}(\sqrt{aN}T_B)^{-1/2}\sqrt{aN}T_B \xrightarrow{\mathcal{L}} N(0,1)$, it will be shown that Proposition 3.2 in de Jong (1987) can be applied, i.e., shows that G^G_1, G^G_2 and G^G_3 (defined below) are of smaller order than that of $[\mathrm{Var}(\sqrt{aN}T_B)]^4 = O(1)$. Let $l_3 = l(i, j_3)$, and $l_4 = l(i, j_4)$. Define

$G^G_1 = \sum_{1\le l_1<l_2\le N} E(V^G_{l_1l_2})^4$,

$G^G_2 = \sum_{1\le l_1<l_2<l_3\le N}\{E(V^G_{l_1l_2}V^G_{l_1l_3})^2 + E(V^G_{l_2l_1}V^G_{l_2l_3})^2 + E(V^G_{l_3l_1}V^G_{l_3l_2})^2\}$, and

$G^G_3 = \sum_{1\le l_1<l_2<l_3<l_4\le N}\{E(V^G_{l_1l_2}V^G_{l_1l_3}V^G_{l_4l_2}V^G_{l_4l_3}) + E(V^G_{l_1l_2}V^G_{l_1l_4}V^G_{l_3l_2}V^G_{l_3l_4}) + E(V^G_{l_1l_3}V^G_{l_1l_4}V^G_{l_2l_3}V^G_{l_2l_4})\}$.

Following the detailed proof of Theorem 3.2.2, it is established that

$$G^G_1 = O(N^{-1}a^{-2}) = o(1),$$

$$G^G_2 = O(N^{-1}a^{-2})I(l^*_2 - l^*_1 \le k-1)I(l^*_3 - l^*_1 \le k-1) = o(1) \text{ and}$$

$$G^G_3 = O(N^{-1}a^{-2}) = o(1)$$

□

5.2.2 Test of No Main Covariate Effect

This subsection presents a lemma and a theorem for the development of the asymptotic distribution of the test statistic $T^1_{cov} = \sqrt{N}[Q_N - W_N]$ to test no main covariate effect when a and N are large. Similar to the traditional parametric effect, the test of no main covariate effect is only meaningful when the test for covariate-treatment interaction is not significant. The following lemma shows that $\sqrt{N}Q_N$ is partitioned into two sums of quadratic forms. It is shown in the lemma that one of the sums converges to 0 in probability as a and N go to ∞. The theorem for the asymptotic distribution of the test statistic T^1_{cov} follows the lemma.

Lemma 5.2.4. *Let* $Q_N = P_Q(\mathbf{Z}) + R_Q(\mathbf{Z})$ *where*

$$P_Q(\mathbf{Z}) = \frac{ak}{N}\sum_{c=1}^{N}\overline{Z}^2_{.c.} \text{ and } R_Q(\mathbf{Z}) = \frac{ak}{N(N-1)}\sum_{c\neq c'}^{N}\overline{Z}_{.c.}\overline{Z}_{.c'.}.$$

Then, $\sqrt{N}R_Q(Z) \xrightarrow{p} 0$ *as* a *and* $N \to \infty$.

Proof of Lemma 5.2.4

First write

$$R_Q(\mathbf{Z}) = \frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i}\overline{Z}_{ic.}\overline{Z}_{ic'.} + \frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}\overline{Z}_{ic.}\overline{Z}_{i'c'.} \tag{5.2.4}$$

It is sufficient to show that $E(\sqrt{N}R_Q(\mathbf{Z})) \to 0$ and $\text{Var}(\sqrt{N}R_Q(\mathbf{Z})) \to 0$ as a and N go to ∞. Then,

$$E(\sqrt{N}R_Q(\mathbf{Z})) = \sqrt{N}E\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i}\overline{Z}_{ic.}\overline{Z}_{ic'.} + \sqrt{N}E\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}\overline{Z}_{ic.}\overline{Z}_{i'c'.} = 0$$

The second term goes to 0 from the assumption that observations from different treatments are independent. From the first term,

$$E\left(\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i}\overline{Z}_{ic.}\overline{Z}_{ic'.}\right) = O(N^{-1}).$$

This case is similar to the proof of Lemma 3.2.1 stated in equation (3.2.7). Therefore

$\sqrt{N}E[R_Q(\mathbf{Z})] = O(N^{-1/2})$.

Next it will be shown that the variance of $\sqrt{N}R_Q(\mathbf{Z})$ goes to 0 as $a, N \to \infty$. Denote the two terms in the right hand side of (5.2.4) as $R_Q^1 + R_Q^2$ respectively. First it will be shown that $\mathrm{Cov}(R_Q^1, R_Q^2)$ is 0.

$$
\begin{aligned}
& \mathrm{Cov}(R_Q^1, R_Q^2) \\
= \; & \frac{k^2}{a^2N^2(N-1)^2} \sum_{c\neq c'}^{N} \sum_{i_1\neq i_{1'}} \sum_{c_1\neq c_{1'}}^{N} \sum_{i} \mathrm{Cov}(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}, \overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_{1'}c_{1'}\cdot}) \\
= \; & \frac{k^2}{a^2N^2(N-1)^2} \sum_{c\neq c'}^{N} \sum_{i_1\neq i_{1'}} \sum_{c_1\neq c_{1'}}^{N} \sum_{i} E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_{1'}c_{1'}\cdot}) - E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot})E(\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_{1'}c_{1'}\cdot}) \\
= \; & \frac{k^2}{a^2N^2(N-1)^2} \sum_{c\neq c'}^{N} \sum_{i_1\neq i_{1'}} \sum_{c_1\neq c_{1'}}^{N} \sum_{i} E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_{1'}c_{1'}\cdot}) = 0.
\end{aligned}
$$

Therefore $\mathrm{Var}(R_Q(\mathbf{Z})) = \mathrm{Var}(R_Q^1) + \mathrm{Var}(R_Q^2)$. This is similar to the argument in the proof of Lemma 4.3.1 with the exception that a is also tending to infinity together with N,

$$
\begin{aligned}
\mathrm{Var}(R_Q^1) &= \mathrm{Var}\left(\frac{k}{aN(N-1)} \sum_{c\neq c'}^{N} \sum_{i} \overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right) \\
&= \frac{k^2}{a^2N^2(N-1)^2} E\left[\sum_{c\neq c'}^{N} \sum_{i} \overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right]^2 \\
&= \frac{k^2}{a^2N^2(N-1)^2} \sum_{c\neq c'}^{N} \sum_{i} \sum_{c_1\neq c_{1'}}^{N} \sum_{i_1} E\left[\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{i_1c_1\cdot}\overline{Z}_{i_1c_{1'}\cdot}\right] \\
&< \frac{k^2}{a^2N^2(N-1)^2}[4a(a-1)(N-1)^2k^2M_1 + ak^2N^2M_2] \\
&= \frac{k^2}{a^2N^2(N-1)^2} O(a^2N^2),
\end{aligned}
$$

where M_1 and M_2 are finite and were defined as in equation (3.2.8). So $\mathrm{Var}(\sqrt{N}R_Q^1) = O(N^{-1})$.

Next,

$$\begin{aligned}
\text{Var}(R_Q^2) &= \text{Var}\left(\frac{k}{aN(N-1)}\sum_{c\neq c'}^{N}\sum_{i\neq i'}\overline{Z}_{ic.}\overline{Z}_{i'c'.}\right) \\
&= \frac{k^2}{a^2N^2(N-1)^2}E\left[\sum_{c\neq c'}^{N}\sum_{i\neq i'}\overline{Z}_{ic.}\overline{Z}_{i'c'.}\right]^2 \\
&= \frac{k^2}{a^2N^2(N-1)^2}\sum_{c\neq c'}^{N}\sum_{c_1\neq c_{1'}}^{N}\sum_{i\neq i'}\sum_{i_1\neq i_{1'}} E\left[\overline{Z}_{ic.}\overline{Z}_{i'c'.}\overline{Z}_{i_1c_1.}\overline{Z}_{i_{1'}c_{1'}.}\right] \\
&< \frac{k^2}{a^2N^2(N-1)^2}[4a(a-1)(N-1)^2k^2M_1] \\
&= \frac{k^2}{a^2N^2(N-1)^2}O(a^2N^2),
\end{aligned}$$

where M_1 is finite and was defined in equation (3.2.8). Therefore, $\text{Var}(\sqrt{N}R_Q^2) = O(N^{-1})$.

□

Theorem 5.2.5. *Assume that the hypothesis of no main covariate effect $H_{0\eta}$ is true and the assumptions in Section 3.2.1 are satisfied then as a and $N \to \infty$,*

$$\sqrt{N}[Q_N - W_N] \xrightarrow{\mathcal{L}} N(0, \lim_{a,N\to\infty}\varphi_N^2)$$

where

$$\begin{aligned}
\varphi_N^2 &= \frac{2}{Na^2k^2}\sum_{i\neq i'}^{a}\sum_{j}^{n_i}\sum_{j'}^{n_{i'}} E\{\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'})\left[E^2(M_{iji'j'}|S_{ij,t}, S_{i'j',t}) + E(M_{iji'j'}|S_{ij,t}, S_{i'j',t})\right] \\
&\quad \Delta_{iji'j'}(t_1,t_2)\} + O(N^{-1}),
\end{aligned}$$

and $\sigma_i^2 = Var(Y_{ij}|X_{ij})$, $M_{iji'j'} = \sum_r^N I(j \in C_{ic})I(j' \in C_{i'c})$, $S_{ij,t} = (X_{ij}, L_{ij}^{(t/2}, U_{ij}^{(t/2})$, $U_{ij}^{t/2}$ and $L_{ij}^{t/2}$ be the upper and lower $t/2$ spacings from X_{ij},
$\Delta_{iji'j'}(t_1,t_2) = I\left(\max\{X_{ij} - L_{ij}^{(t_1/2)}, X_{i'j'} - L_{i'j'}^{(t_2/2)}\} \leq \min\{X_{ij} + U_{ij}^{(t_1/2)}, \quad X_{i'j'} + U_{i'j'}^{(t_2/2)}\}\right)$.
The φ_N^2 above is also defined in Theorem 4.4.2.

Sketch of Proof of Theorem 5.2.5

Only the convergence of the asymptotic variance will be presented here. The asymptotic

normality will not be presented here because it will be similar to Lemma 5.2.3. After applying Lemma 5.2.4, $\sqrt{N}(Q_N - W_N)$ is equivalent to

$$\sqrt{N}(P_Q(Z) - W_N) = \sqrt{N}(T_B + T_D),$$

where

$$T_B = \frac{1}{Na(k-1)} \sum_{i=1}^{a} \sum_{c=1}^{N} \sum_{t \neq t'}^{k} Z_{ict} Z_{ict'},$$

$$T_D = \frac{k}{Na} \sum_{i \neq i'}^{a} \sum_{c=1}^{N} \overline{Z}_{ic\cdot} \overline{Z}_{i'c\cdot}$$

and Q_N and $P_Q(Z)$ is defined in Lemma 5.2.4. The asymptotic variance of $\sqrt{N}(T_B + T_D)$ will be shown here. First, $\text{Cov}(T_B, T_D)|\mathbf{X}) = 0$, by equation (4.4.7). Because $\text{Cov}((T_B, T_D)|\mathbf{X})$ is 0, the convergence of the asymptotic variance is shown by demonstrating $Var(N^{1/2}T_B|\mathbf{X}) \to 0$ and $Var(N^{1/2}T_D|\mathbf{X}) \to \varphi_N^2$. From the proof in Lemma 5.2.1 , $E[Var(N^{1/2}T_B|\mathbf{X})] \to \gamma_N^2$. Therefore as a and N go to ∞, $\gamma_N^2 \to 0$. As for $Var(N^{1/2}T_D|\mathbf{X})$, from the proof of Theorem 4.4.2,

$$\begin{aligned}
& E[Var(N^{1/2}T_D|\mathbf{X})] \\
= \; & \frac{2}{Na^2k^2} \sum_{i \neq i'}^{a} \sum_{j}^{n_i} \sum_{j'}^{n_{i'}} E\left\{\sigma_i^2(X_{ij})\sigma_{i'}^2(X_{i'j'}) \left[E^2(M_{iji'j'}|S_{ij,t}, S_{i'j',t}) + E(M_{iji'j'}|S_{ij,t}, S_{i'j',t})\right]\right. \\
& \left. \Delta_{iji'j'}(t_1, t_2)\right\} + O(N^{-1}) \\
= \; & \varphi_N^2.
\end{aligned}$$

Therefore, as a and $N \to \infty$, $\text{Var}[\sqrt{N}(P_Q(Z) - W_N)] \to \varphi_N^2$

5.2.3 Test for No Simple Covariate Effect

This subsection presents a lemma and a theorem for the development of the asymptotic distribution of test statistics $\sqrt{aN}(B_N - W_N)$ to test for the simple covariate effect.

Lemma 5.2.6. *Write*

$$B_N \;=\; ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\left(\overline{Z}_{ic\cdot}-\overline{Z}_{i\cdot\cdot}\right)^2 = P_B(\mathbf{Z})+S_B(\mathbf{Z})$$

where

$$P_B(\mathbf{Z})=\frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2 \quad \text{and } S_B(\mathbf{Z})=-\frac{k}{aN(N-1)}\sum_{i=1}^{a}\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}.$$

When the assumptions in Section 3.2.1 are satisfied, then as a and $N\to\infty$,

$$\sqrt{aN}B_N \to \sqrt{aN}\frac{k}{aN}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2$$

Proof of Lemma 5.2.6

The proof of the lemma involves showing that $\sqrt{aN}S_B \to 0$. Notice that $S_B = -P_G^{(2)}$ from the proof of Lemma 5.2.1. Thus $\sqrt{aN}S_B \to 0$.

Theorem 5.2.7. *Assume that $H_0(B)$ is true and the assumptions in Section 3.2.1 are satisfied, then as a and $N\to\infty$,*

$$\sqrt{aN}(B_N - W_N) \xrightarrow{\mathcal{L}} N(0, \lim_{a,N\to\infty} a\gamma_N^2)$$

where γ_N^2 is defined in Theorem 5.2.2.

Proof of Theorem 5.2.7

The proof of the theorem comes from Applying Lemma 5.2.6, and following the proof of Theorem 5.2.2, thus it is omitted.

5.3 Numerical studies

5.3.1 Simulation Studies Setting

This section reports a simulation study to investigate the type I error and the power performance of the pNP test of no covariate-treatment interaction, no main covariate and no simple covariate effects. The pNP tests are compared to the traditional F test (CF), the

drop test and the GAM tests using spline and loess smoothing methods. The simulation study was conducted for a case where the number of treatment level a is set to be 20 and the number of covariate values in each treatment level (n_i) is also 20. In this setting, the covariate values X_{ij} are iid and were generated from a mixture distribution with 3 components f_1, f_2 and f_3 as follows:

$$\begin{cases} f_1(x) = Uniform(-0.5, 0) & \text{with probability } (1 - prop)/2 \\ f_2(x) = Uniform(0, b) & \text{with probability } prop, \\ f_3(x) = Uniform(b, 1) & \text{with probability } (1 - prop)/2. \end{cases} \tag{5.3.1}$$

The parameter *prop* ranges from 0.1 to 0.9 are given in Table 5.1. The values of b are 0.5 and 0.7, which will be described in the later part of this section. The responses in each treatment level i are generated according to equation (5.3.2) below.

$$Y_{ij} = X_{ij} \ \tan(\theta_i) \ I(0 < X_{ij} \leq b_i) + b_i \ \tan(\theta_i) \ I(X_{ij} > b_i) - (10 \ \theta_i / \ \tau)\sqrt{|X_{ij}|} \ \varepsilon_{ij} \tag{5.3.2}$$

where $\varepsilon_{ij} \sim Weibull \ (shape = 2, scale = 5 \ |X_{ij} - 0.5|)$.

Because the pNP test was constructed without any specific distributional and constant variance assumptions, this model is particularly suitable to evaluate the performance of the proposed test under a nonconstant variance setting with a nonnormal distribution. The nonconstant conditional variance term is described by the term $10 \ \theta/\tau\sqrt{|x_{ij}|} \ \varepsilon_{ij}$. This is a covariate-treatment dependent error terms. Given $X_{ij} = x$, the conditional variance of the error term is written as:

$$\begin{aligned} \text{Var}\left(10 \ \theta_i/\tau\sqrt{|x_{ij}|} \ \varepsilon_{ij} \ \ |X_{ij} = x\right) \quad &= 100 \ \theta_i^2/\tau^2|x| \ \text{Var}(\varepsilon_{ij}) \\ &= 625 \ \theta_i^2/ \ \tau^2|x|(x - 0.5)^2(4 - \pi). \end{aligned} \tag{5.3.3}$$

Notice that the variance of the observations is directly proportional to a third degree polynomial in x and with θ_i^2 but inversely proportional with τ^2. Therefore, the conditional variance of responses not only depend on X and τ, but also on θ_i, which are distinct at the different treatment levels.

The conditional mean of Y_{ij} given X_{ij} for model (5.3.2) is written as

$$\begin{aligned} E(Y_{ij}|X_{ij}=x) &= x\ \tan(\theta_i)\ I(0<x\leq b_i) + b_i\ \tan(\theta_i)\ I(x>b_i) \\ &\quad -(10\ \theta_i/\ \tau)\sqrt{|x|}\ E(\varepsilon_{ij}) \\ &= I + II - III \end{aligned} \tag{5.3.4}$$

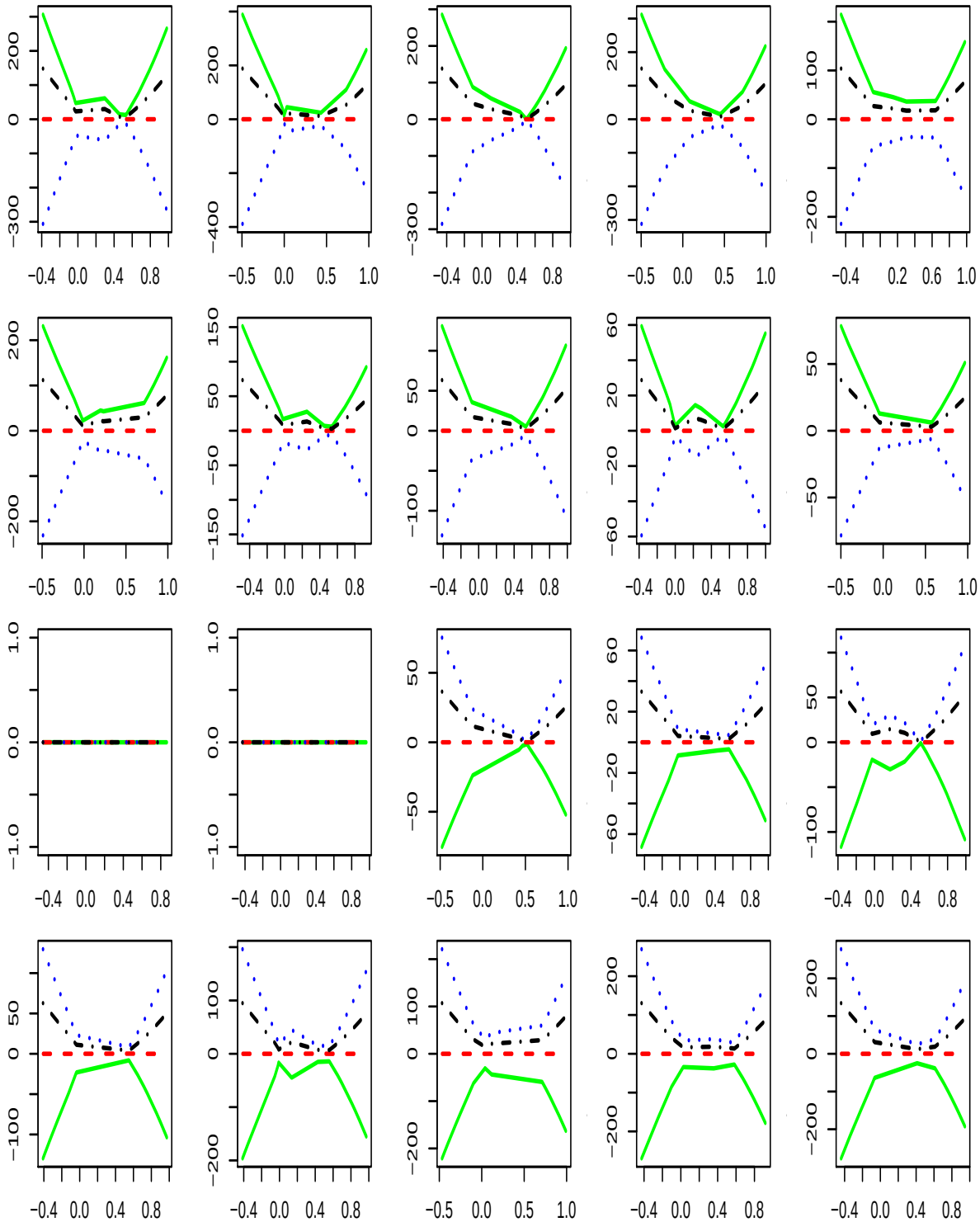
where $I = x\ \tan(\theta_i)\ I(0<x\leq b_i)$, $II = b_i\ \tan(\theta_i)\ I(x>b_i)$ and $III = 25\,\theta_i/\tau\sqrt{\pi}\sqrt{|x|}\ |x-0.5|$. The interaction effect exists if values of θ_i and b_i are non identical for different treatments. When θ_i and b_i are the same for all treatment levels, these correspond to a null hypothesis of no covariate-treatment interaction. Figures 5.1 and 5.2 illustrate the graph of the true mean component $I + II - III$, component III, component I and the variance component from the equation (5.3.3) when $\tau = 0.0625$ and $prop = 0.1$ and $prop = 0.9$ respectively. In addition, Figure 5.2 which details $\tau = 0.0625$ and $prop = 0.9$ also includes the plot of the observations. From the graph it is seen that the linear component (I) does not have a very strong influence in the true conditional mean for this model. The most critical component is part III of the model.

5.3.2 Covariate-treatment Interaction

Type I error estimates for the test of no covariate-treatment interaction

The type I error estimates for the tests considered above are obtained for the no covariate-treatment interaction effect when X_{ij} are generated by the mixture distribution (5.3.1) with $b = 0.5$. The responses are generated following the equation (5.3.2) with $\theta = \pi/4$. The type I error estimates were calculated for different values of $prop$ and τ, where $prop$ values range from 0.1 to 0.9 and τ values range from 0.0625 to 1. The scatter plots in Figure 5.3 and Figure 5.4 illustrate the data generation for the cases when $\tau = 0.25$ and $prop = 0.5$ and $\tau = 0.0625$ and $prop = 0.9$ respectively. The scatter plot of data in Figure 5.3 exhibits nonlinear associations, while the data in Figure 5.4 appears to exhibit some linear associations. In both plots, there is no significant covariate-treatment interaction

Figure 5.1: Plot of the conditional mean components (5.3.4) and variance (5.3.3) for the simulation $prop = 0.1$, and $\tau = 0.0625$. The green line corresponds to component $I+II-III$, the blue dotted line corresponds to component III, the red dashed line depicts component I of (5.3.4) and black dotted line give the variance component (5.3.3).



effect present but there is a main covariate effect present.

The results for the estimates of type I error at 0.01 level for this setting are presented in Table 5.1. The table shows that the type I error estimates from GAM tests, the drop test and the CF test are inflated with the exception of the pNP test. This is most likely due to the fact that in this setting, the errors which correspond to component III of equation (5.3.2) depend on the covariate values and have a nonconstant variance in different treatment levels. Thus this result is to be expected because all the other tests except the pNP test assume that the error has constant variance. Moreover, the type I error estimates for these tests seem to increase as the value of *prop* increases for all values of τ. On the other hand, the estimates for type I error from the pNP test for no covariate-treatment interaction are very conservative for all values of τ and *prop*. Therefore, the pNP test has acceptable type I error estimates under the complicated heteroscedastic conditional data setting.

Power performance for covariate-treatment interaction

To study the power performance of the proposed test, the covariate values were generated using the mixture distribution in the model (5.3.1). The responses are generated following the equation (5.3.2) with θ ranges from -0.25π to 0.2π and $b = 0.5, 0.7$.

The graph of the equation (5.3.2) was constructed to illustrate the relation between the response Y and X at different values of b and θ. The covariate values X were generated from (5.3.1) where $prop = \{0.1, 0.26, 0.33, 0.5, 0.9\}$. Figure 5.5 illustrates the setting for $\tau = 0.25$ and $prop = 0.33$ while Figure 5.6 illustrates for $\tau = 0.25$ and $prop = 0.9$. The scatter plots show the existence of a covariate-treatment interaction effect. In Figure 5.5, there appears to be a second order association between the covariate X and the response Y. In Figure 5.6, the association between X and Y is not evident. The proportion of rejections of no covariate-treatment interaction at the 0.01 level are presented in Table 5.3 for p from 0.1 to 0.9 and τ from 0.0625 to 0.5. In general, the results indicate that the performance of the pNP test is better than the performance of the other four comparable tests when the range of *prop* goes from 0.1 to 0.5 and for all the τ values considered. When *prop* varies from 0.76 to 0.90 the pNP seems unable to detect the covariate-treatment interaction compared with the other four tests. Table 5.4 shows the gradual change in performance of pNP test

Table 5.1: Proportion of rejections at level 0.01 under the null hypothesis of no covariate-treatment interaction effect when $n_i = 20$, $a = 20$, $\theta = \pi/4$ and $b = 0.5$ from (5.3.2). The results are based on 1000 simulations.

τ	*prop*	pNP.5	GAM Spline	GAM Loess	drop	CF
0.0625	0.10	0	0.226	0.224	0.055	0.096
	0.26	0	0.404	0.397	0.129	0.177
	0.33	0	0.437	0.435	0.216	0.240
	0.50	0	0.687	0.700	0.390	0.435
	0.76	0	0.934	0.948	0.521	0.881
	0.90	0	0.951	0.970	0.434	0.980
0.125	0.10	0	0.226	0.224	0.055	0.096
	0.26	0	0.404	0.397	0.129	0.177
	0.33	0	0.437	0.435	0.216	0.240
	0.50	0	0.687	0.715	0.390	0.435
	0.76	0	0.934	0.948	0.521	0.881
	0.90	0	0.951	0.970	0.434	0.980
0.25	0.10	0	0.226	0.224	0.055	0.097
	0.26	0	0.404	0.397	0.129	0.178
	0.33	0	0.437	0.435	0.216	0.240
	0.50	0	0.687	0.700	0.390	0.435
	0.76	0	0.935	0.948	0.521	0.881
	0.90	0	0.951	0.970	0.434	0.980
0.5	0.10	0	0.226	0.224	0.055	0.097
	0.26	0	0.404	0.397	0.129	0.178
	0.33	0	0.437	0.435	0.216	0.241
	0.50	0	0.687	0.697	0.390	0.434
	0.76	0	0.935	0.948	0.521	0.879
	0.90	0	0.951	0.970	0.434	0.980

for additional values of *prop* between 0.5 and 0.76. Although GAM Spline, GAM Loess and CF seem to perform quite well when *prop* = 0.9, with p-values of 0.964, 0.697, 0.977 and 0.963 respectively. However, this could not be the correct power for these tests, because the type I errors corresponding to this power simulation study are very inflated. Those errors are 0.951, 0.970, 0.430 and 0.980 respectively (from Table 5.1).

Because of the inflated values of the estimated type I error for the four tests and the very conservative values for the pNP test, one might want to see the bootstrap power level for these tests. The results for the bootstrap power performance for the covariate-treatment interaction tests are reported in Table 5.5. These results are based on the cutting point

Table 5.2: The values b's and θ's that generate the treatment levels combination for the simulation study for power performance of test of no covariate-treatment interaction effect.

trt	1	2	3	4	5	6	7	8	9	10
b	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7
θ	$\frac{-5}{20}\pi$	$\frac{-5}{20}\pi$	$\frac{-4}{20}\pi$	$\frac{-4}{20}\pi$	$\frac{-3}{20}\pi$	$\frac{-3}{20}\pi$	$\frac{-2}{20}\pi$	$\frac{-2}{20}\pi$	$\frac{-1}{20}\pi$	$\frac{-1}{20}\pi$
trt	11	12	13	14	15	16	17	18	19	20
b	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7
θ	0	0	$\frac{1}{20}\pi$	$\frac{1}{20}\pi$	$\frac{2}{20}\pi$	$\frac{2}{20}\pi$	$\frac{3}{20}\pi$	$\frac{3}{20}\pi$	$\frac{4}{20}\pi$	$\frac{4}{20}\pi$

(threshold) being the lower 1 percentile of the p values under the null hypothesis. Instead of rejecting the null hypothesis when the p value is less than 0.01, the null hypothesis is rejected when the p value is less than the corresponding threshold for all τ and $prop$ values. From the results in Table 5.5, the bootstrap power performance for the pNP test is better than the result from the nonbootstrap performance. The bootstrap power for the pNP test are all very high for all value of τ's and $prop$'s. From Table 5.3, the pNP test has very low power for $prop = 0.76$ and $prop = 0.9$, but not in Table 5.5. The power for the pNP test when $prop = 0.76$ and $prop = 0.9$ improved significantly when the lower 1% threshold of p-values was used. On the other hand, the bootstrap power performance for the GAM Spline, GAM Loess and the CF tests were very weak, especially for large values of $prop$, i.e. $prop = 0.76$ and $prop = 0.9$. The bootstrap power performance for the drop test was moderate for $prop = 0.9$ and $prop = 0.1$.

5.3.3 When There is No Covariate-treatment Interaction : Test of No Simple Covariate effect and Test of No Main Covariate Effect

This subsection presents type I error analysis and power performance for the tests of no simple covariate and no main covariate effects when there is no covariate-treatment interaction effect in the model when compared with the other four tests stated in the previous subsection. The setting for the simulation to estimate the type I error for the no covariate-

Table 5.3: (Power performance) Proportion of rejections at level 0.01 for testing of no covariate-treatment interaction $n_i = 20$ and $a = 20$ using the model (5.3.2). The results are based on 1000 simulations.

τ	prop	pNP.5	GAM Spline	GAM Loess	Drop	CF
0.0625	0.10	0.999	0.701	0.698	0.856	0.697
	0.26	0.999	0.655	0.656	0.728	0.652
	0.33	0.999	0.797	0.802	0.863	0.799
	0.50	0.990	0.754	0.759	0.795	0.750
	0.76	0.099	0.761	0.759	0.709	0.739
	0.90	0.059	0.964	0.967	0.977	0.963
0.125	0.10	0.999	0.708	0.709	0.845	0.701
	0.26	0.999	0.660	0.657	0.731	0.659
	0.33	0.999	0.801	0.809	0.853	0.804
	0.50	0.991	0.763	0.765	0.804	0.763
	0.76	0.098	0.763	0.760	0.686	0.741
	0.90	0.059	0.966	0.967	0.976	0.963
0.25	0.10	0.999	0.716	0.719	0.842	0.714
	0.26	1.000	0.673	0.672	0.689	0.669
	0.33	1.000	0.815	0.820	0.850	0.813
	0.50	0.990	0.772	0.773	0.780	0.776
	0.76	0.112	0.766	0.766	0.656	0.761
	0.90	0.049	0.967	0.969	0.969	0.964
0.50	0.10	1.000	0.744	0.742	0.804	0.745
	0.26	1.000	0.696	0.700	0.670	0.695
	0.33	1.000	0.839	0.837	0.823	0.831
	0.50	0.989	0.790	0.797	0.762	0.790
	0.76	0.104	0.775	0.772	0.633	0.752
	0.90	0.067	0.968	0.972	0.970	0.967

treatment interaction effect in the subsection above will also be used in the simulation to study the power performance for simple covariate and main covariate effects. The power performance for testing no simple covariate and no main covariate effects was conducted under the setting for the null hypothesis of no covariate-treatment interaction in subsection 5.3.2 described above. Before discussing the power performance for these tests, the type I error analysis will be discussed that corresponds to the power performance setting.

The simulation setting for the null hypotheses of no simple covariate and no main covariate effects is such that $a = 20$, $n_i = 20$ and X_{ij} were generated following the model in

Table 5.4: Power for no covariate-treatment interaction, no covariate simple and no main covariate effects for pNP test with additional values of *prop*.

τ	prop	pNP.int	pNP.simp	pNP.cov
0.0625	0.6	0.763	0.716	0.067
	0.63	0.655	0.614	0.062
	0.65	0.649	0.608	0.057
	0.68	0.215	0.181	0.026
	0.7	0.216	0.183	0.025
	0.73	0.166	0.145	0.030
0.125	0.6	0.756	0.719	0.069
	0.63	0.661	0.615	0.059
	0.65	0.662	0.615	0.061
	0.68	0.228	0.193	0.029
	0.7	0.223	0.189	0.025
	0.73	0.175	0.152	0.030
0.25	0.6	0.760	0.720	0.065
	0.63	0.667	0.623	0.061
	0.65	0.668	0.625	0.058
	0.68	0.239	0.197	0.029
	0.7	0.237	0.197	0.029
	0.73	0.181	0.159	0.032
0.5	0.6	0.779	0.742	0.067
	0.63	0.681	0.655	0.068
	0.65	0.685	0.660	0.071
	0.68	0.239	0.210	0.029
	0.7	0.241	0.212	0.030
	0.73	0.197	0.182	0.036

(5.3.1) with $b = 0.5$. The responses were generated following

$$Y_{ij} = 0.25\tan(\theta_i) - 25/\tau\ \theta_i\sqrt{\pi}(0.3 - 0.2\ prop) + 10/\tau\ \theta_i[\varepsilon_{ij} - E(\varepsilon_{ij})] \tag{5.3.5}$$

where $\varepsilon_{ij} = Weibull\ (shape = 2, scale = 5(0.5 - 0.25\ prop)) + \cos(\theta_i)$ and the $E(\varepsilon_{ij}) = [5(0.5 - 0.25\ prop) + \cos(\theta_i)]\ \sqrt{\pi}/2$. Notice that the equation (5.3.5) does not depend on X for all i, thus there is no simple covariate and no main covariate effect. The twenty values of θ that contribute to the treatment effect in the model in each treatment group for this simulation are $\{-3.000\ \pi/8, -2.684\ \pi/8, -2.368\ \pi/8, -2.053\ \pi/8, -1.737\ \pi/8, -1.421\ \pi/8, -1.105\ \pi/8, -0.789\ \pi/8, -0.474\ \pi/8, -0.158\ \pi/8, 0.158\ \pi/8, 0.474\ \pi/8, 0.790\ \pi/8,$

Table 5.5: (Bootstrap Power performance) Proportion of rejections at level 0.01 for testing of no covariate-treatment interaction using the lower 1 percentile of the corresponding p-values of estimate of type I error as the cut off point for rejecting the null when $n_i = 20$ and $a = 20$ using the model (5.3.2). The results are based on 1000 simulations.

τ	prop	pNP	GAM Spline	GAM Loess	drop	CF
0.0625	0.1	1.000	0.298	0.287	0.727	0.474
	0.26	1.000	0.101	0.109	0.390	0.282
	0.33	1.000	0.160	0.184	0.518	0.405
	0.5	1.000	0.048	0.060	0.274	0.200
	0.76	0.977	0.004	0.003	0.131	0.022
	0.9	0.900	0.062	0.038	0.736	0.053
0.125	0.1	1.000	0.305	0.297	0.721	0.481
	0.26	1.000	0.101	0.112	0.404	0.288
	0.33	1.000	0.166	0.192	0.551	0.413
	0.5	1.000	0.049	0.058	0.274	0.203
	0.76	0.974	0.004	0.003	0.115	0.023
	0.9	0.915	0.062	0.040	0.688	0.053
0.25	0.1	1.000	0.313	0.308	0.666	0.495
	0.26	1.000	0.105	0.117	0.373	0.299
	0.33	1.000	0.181	0.204	0.510	0.425
	0.5	1.000	0.054	0.063	0.240	0.212
	0.76	0.936	0.004	0.003	0.080	0.024
	0.9	0.953	0.064	0.045	0.669	0.054
0.5	0.1	1.000	0.347	0.335	0.659	0.525
	0.26	1.000	0.119	0.135	0.325	0.334
	0.33	1.000	0.209	0.233	0.458	0.452
	0.5	1.000	0.064	0.067	0.222	0.236
	0.76	0.945	0.005	0.003	0.077	0.026
	0.9	0.963	0.073	0.051	0.660	0.059

$1.105\ \pi/8, 1.421\ \pi/8, 1.737\ \pi/8, 2.053\ \pi/8, 2.368\ \pi/8, 2.684\ \pi/8, 3.000\ \pi/8\}$. The variance of ε_{ij} in this simulation setting range between $(25(0.5 - 0.25prop)^2)[1 - \pi/4]$ and $(5(0.5 - 0.25prop) + 1)[1 - \pi/4]$. Notice that the first two terms of the right side of equation (5.3.5) resulted from taking the expectation of the conditional mean of Y_{ij} given X_{ij} in the equation (5.3.4). The results for the estimates of type I error and power performance of the tests of simple covariate and main covariate effects for pNP, GAM, drop and CF tests are presented in the following subsections.

Test of No Simple Covariate effect

The estimates of type I error at 0.01 level for simple covariate effect from the model (5.3.5) are presented in Table 5.6. The values of τ and *prop* used are those used in the previous section. It is seen that the type I error estimate for the pNP test is very conservative for all values of τ and *prop*. For the GAM tests and the CF test, the type I error estimates range from 0.025 to 0.048 for all τ and *prop*. These results are quite inflated.

Table 5.6: Proportion of rejections at level 0.01 under the null hypothesis of no simple covariate effect in 5.3.5 for $n_i = 20$ and $a = 20$. The results are based on 1000 simulations.

τ	prop	pNP	GAM Spline	GAM Loess	Drop	CF
0.0625	0.1	0.001	0.033	0.030	0.172	0.033
	0.26	0.000	0.031	0.029	0.180	0.033
	0.33	0.000	0.048	0.048	0.170	0.046
	0.5	0.000	0.030	0.029	0.184	0.025
	0.76	0.002	0.035	0.034	0.158	0.031
	0.9	0.000	0.042	0.039	0.157	0.040
0.125	0.1	0.000	0.034	0.034	0.195	0.036
	0.26	0.000	0.031	0.029	0.182	0.033
	0.33	0.000	0.048	0.048	0.179	0.046
	0.5	0.000	0.030	0.029	0.169	0.025
	0.76	0.003	0.035	0.034	0.157	0.031
	0.9	0.000	0.042	0.039	0.166	0.040
0.25	0.1	0.000	0.033	0.030	0.187	0.033
	0.26	0.000	0.031	0.029	0.185	0.033
	0.33	0.000	0.048	0.048	0.171	0.046
	0.5	0.000	0.030	0.029	0.179	0.025
	0.76	0.001	0.035	0.034	0.158	0.031
	0.9	0.000	0.042	0.039	0.166	0.040
0.5	0.1	0.001	0.034	0.034	0.188	0.036
	0.26	0.000	0.031	0.029	0.174	0.033
	0.33	0.000	0.048	0.048	0.179	0.046
	0.5	0.000	0.030	0.029	0.184	0.025
	0.76	0.002	0.035	0.034	0.156	0.031
	0.9	0.000	0.042	0.039	0.167	0.040

To study the power performance for a simple covariate effect, the data were generated under the model (5.3.2) for $\theta = \pi/4$ and $b = 0.5$ which is also used for simulation study to estimate the type I error for the test of no covariate-treatment interaction effect in the

previous section. The estimated power for the simple covariate effect is presented in Table 5.7. The table demonstrates that in general, all the tests of main covariate effect discussed here performed well. However, since the type I error estimates from Table 5.6 for the GAM tests, drop test and CF test are inflated, the bootstrap power analysis is calculated. The bootstrap analysis means the cutoff threshold used to determine the rejection of the null hypotheses is the lower 1 percentile of the corresponding p-value under the null hypothesis instead of $\alpha = 0.01$. The results for the bootstrap power analysis for simple covariate effect are presented in Table 5.8. It is seen that the bootstrap power of GAM Spline, GAM Loess and CF test does not differ from the estimated power without the bootstrap. On the other hand, the bootstrap power performance of the drop test becomes weaker compared to the one without bootstrap in Table 5.7. For the pNP test, the bootstrap performance slightly improved when the 1 percentile threshold was used from the corresponding empirical p values under the null hypothesis.

Test of No Main Covariate Effect

Similar to the simulation study for testing the no simple covariate effect, the simulation study to test the main covariate effect was conducted under the model (5.3.5) to estimate the type I error rate and under the model (5.3.2) for the power performance of the pNP test. Table 5.9 shows the estimates of type I error to test for the no main covariate effect. All of the tests being considered here have a good estimate of type I error for all combinations of τ and *prop* values considered. The results for the power performance for the test of the main covariate effect in Table 5.10 indicate that all the tests considered have good power under this simulation setting.

Table 5.7: Proportion of rejections at level 0.01 to test for simple covariate effect under the model (5.3.2) for $\theta = \pi/4$ and $b = 0.5$, $n_i = 20$ and $a = 20$. The results are based on 1000 simulations.

τ	*prop*	pNP	GAM Spline	GAM Loess	drop	CF
0.0625	0.1	1.000	1.000	1.000	0.913	0.951
	0.26	1.000	1.000	1.000	0.828	0.906
	0.33	1.000	1.000	1.000	0.979	0.980
	0.5	0.999	1.000	1.000	0.958	0.966
	0.76	0.003	1.000	1.000	0.839	0.953
	0.9	0.013	1.000	1.000	0.994	0.997
0.125	0.1	1.000	1.000	1.000	0.913	0.954
	0.26	1.000	1.000	1.000	0.828	0.908
	0.33	1.000	1.000	1.000	0.979	0.981
	0.5	0.999	1.000	1.000	0.958	0.968
	0.76	0.004	1.000	1.000	0.839	0.955
	0.9	0.014	1.000	1.000	0.994	0.998
0.25	0.1	1.000	1.000	1.000	0.913	0.957
	0.26	1.000	1.000	1.000	0.828	0.914
	0.33	1.000	1.000	1.000	0.979	0.983
	0.5	1.000	1.000	1.000	0.958	0.972
	0.76	0.003	1.000	1.000	0.839	0.957
	0.9	0.023	1.000	1.000	0.994	0.998
0.5	0.1	1.000	1.000	1.000	0.913	0.967
	0.26	1.000	1.000	1.000	0.828	0.932
	0.33	1.000	1.000	1.000	0.979	0.984
	0.5	1.000	1.000	1.000	0.958	0.974
	0.76	0.006	1.000	1.000	0.839	0.959
	0.9	0.032	1.000	1.000	0.994	0.999

Figure 5.2: Plot of the conditional mean components (5.3.4) and variance (5.3.3) for the simulation $b = 0.7$, $prop = 0.9$, and $\tau = 0.0625$. with simulated data. The green line corresponds to component $I + II - III$, the blue dotted line corresponds to component III, thed red dashed line depicts component I of (5.3.4) and black dotted line give the variance component (5.3.3).

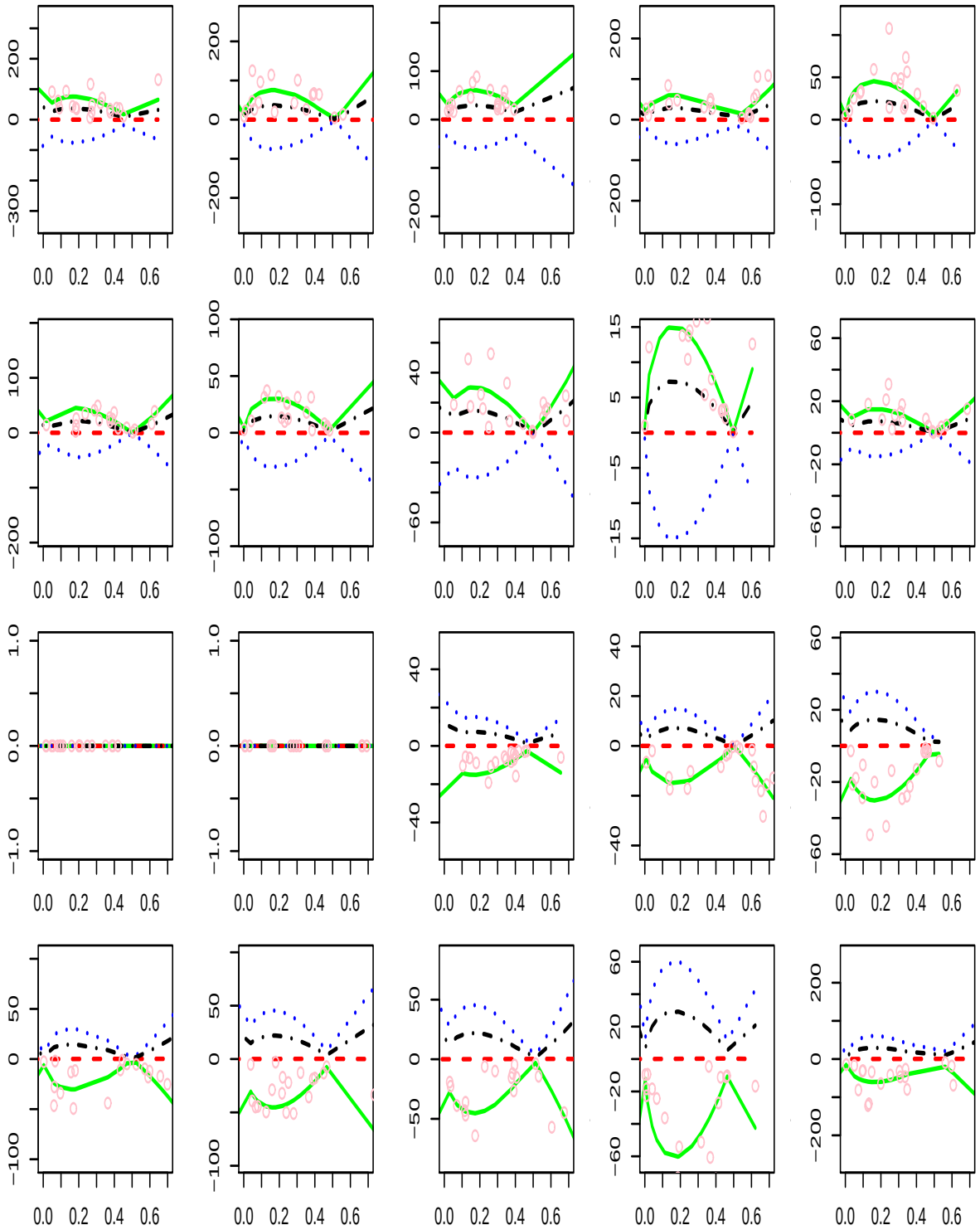


Figure 5.3: Scatter plot for data generated under the null hypothesis of no covariate-treatment interaction with $\theta = \pi/4$, $b = 0.5$, $\tau = 0.25$ and $prop = 0.5$ following the model (5.3.2).

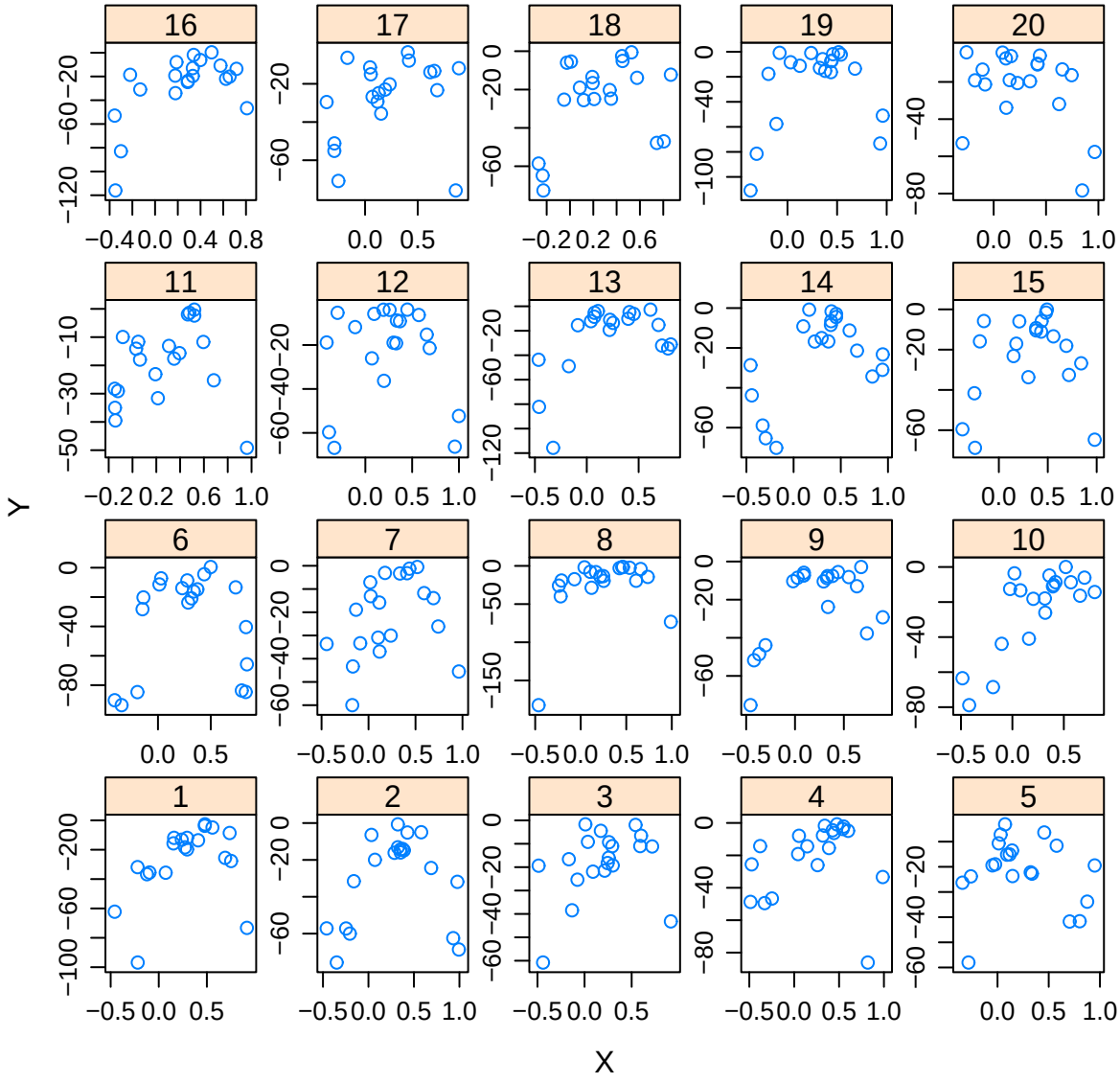


Figure 5.4: Scatter plot for data generated under the null hypothesis of no covariate-treatment interaction with $\theta = \pi/4$, $b = 0.5$, $\tau = 0.0625$ and $prop = 0.9$ following the model (5.3.2).

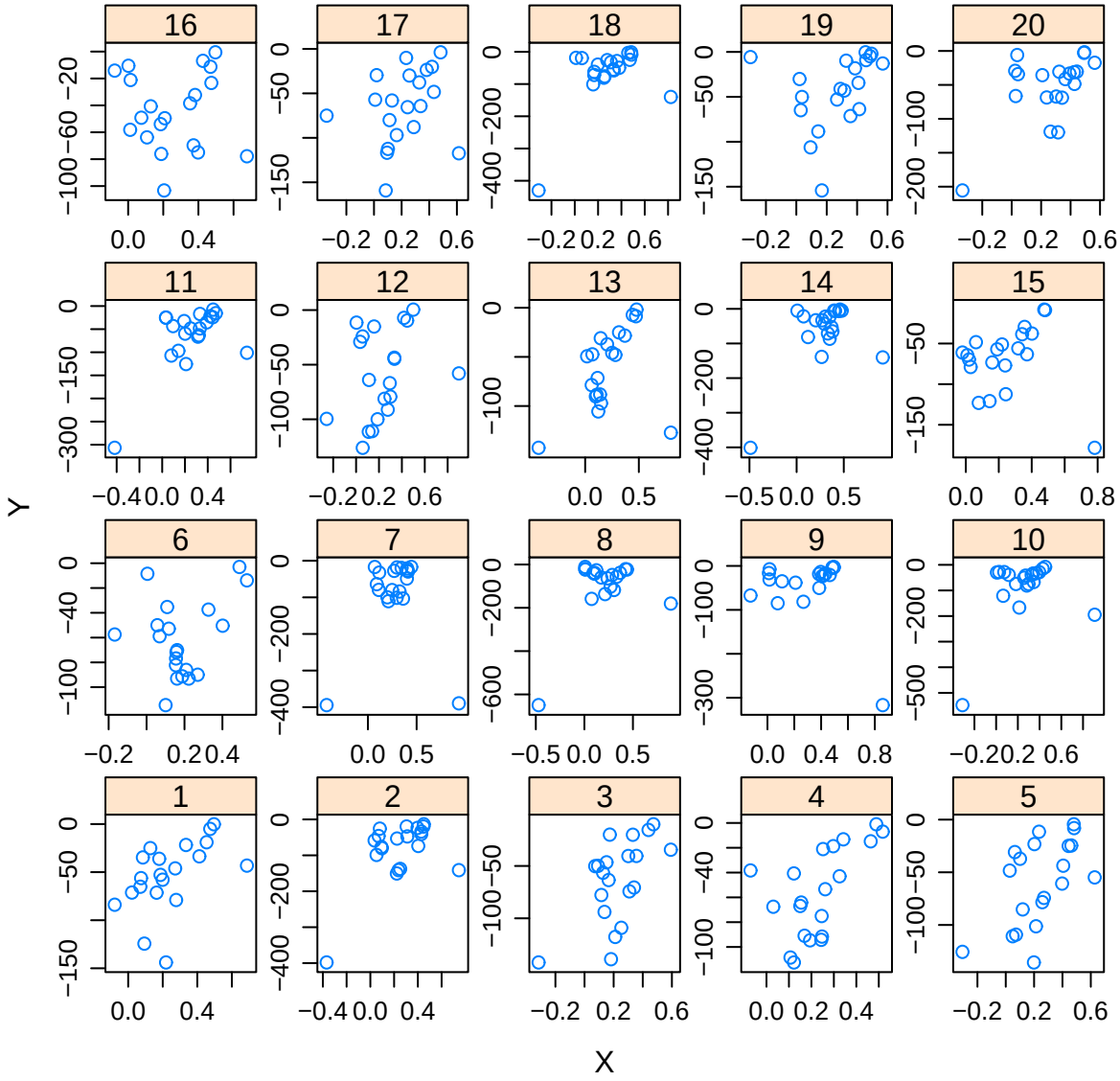


Figure 5.5: Scatter plot of data generated from equation (5.3.2) when $\tau = 0.25$ $prop = 0.33$ for power estimation of no covariate-treatment interaction test in which the treatment level were generated by θ_i and b_i values described in Table 5.2. This figure illustrates the existence of covariate-treatment interaction in the data.

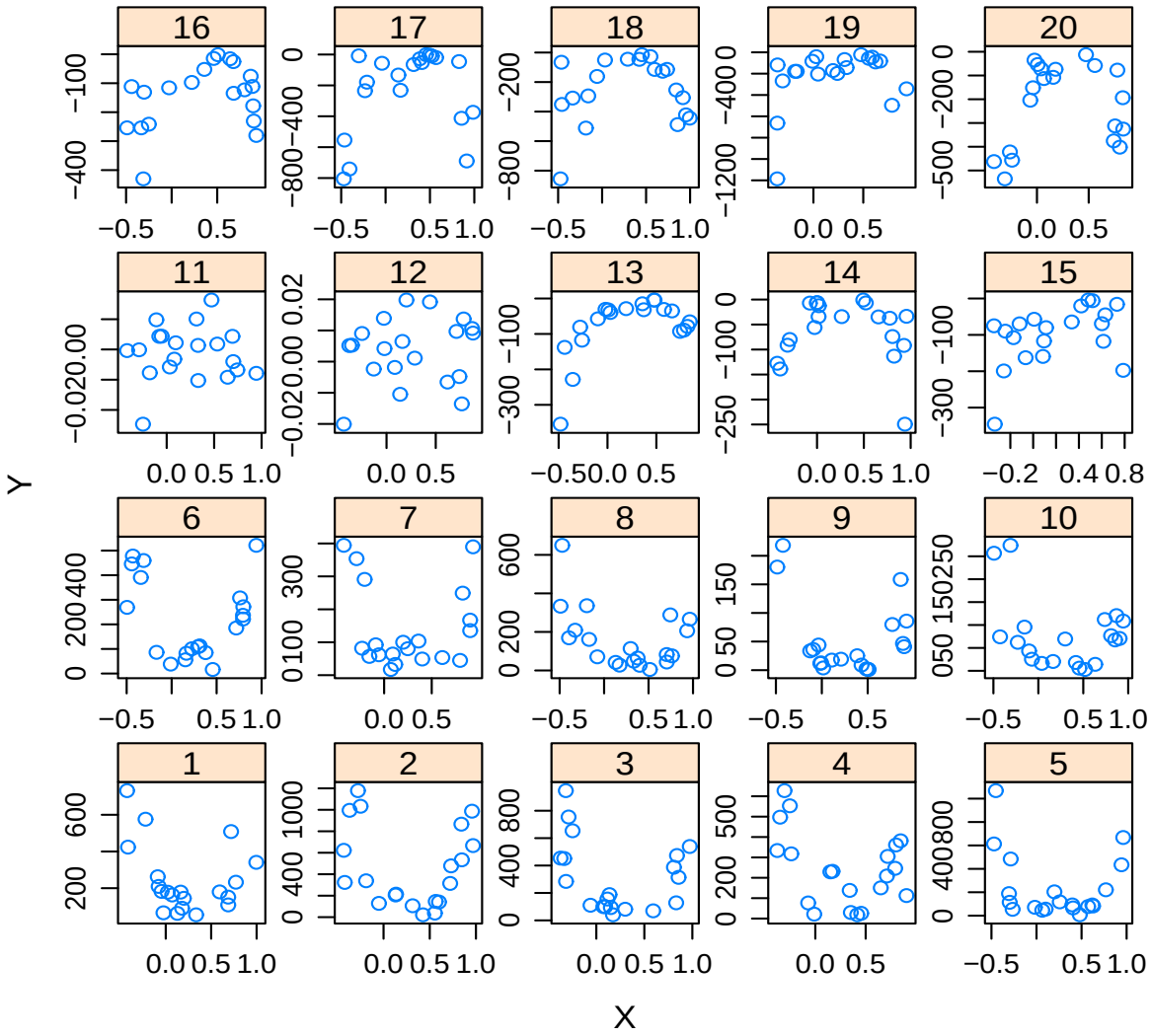


Figure 5.6: Scatter plot of data generated from equation (5.3.2) when $\tau = 0.25$ $prop = 0.9$ for power estimation of no covariate-treatment interaction test in which the treatment level were generated by θ_i and b_i values described in Table 5.2. This figure illustrates the existence of covariate-treatment interaction in the data.

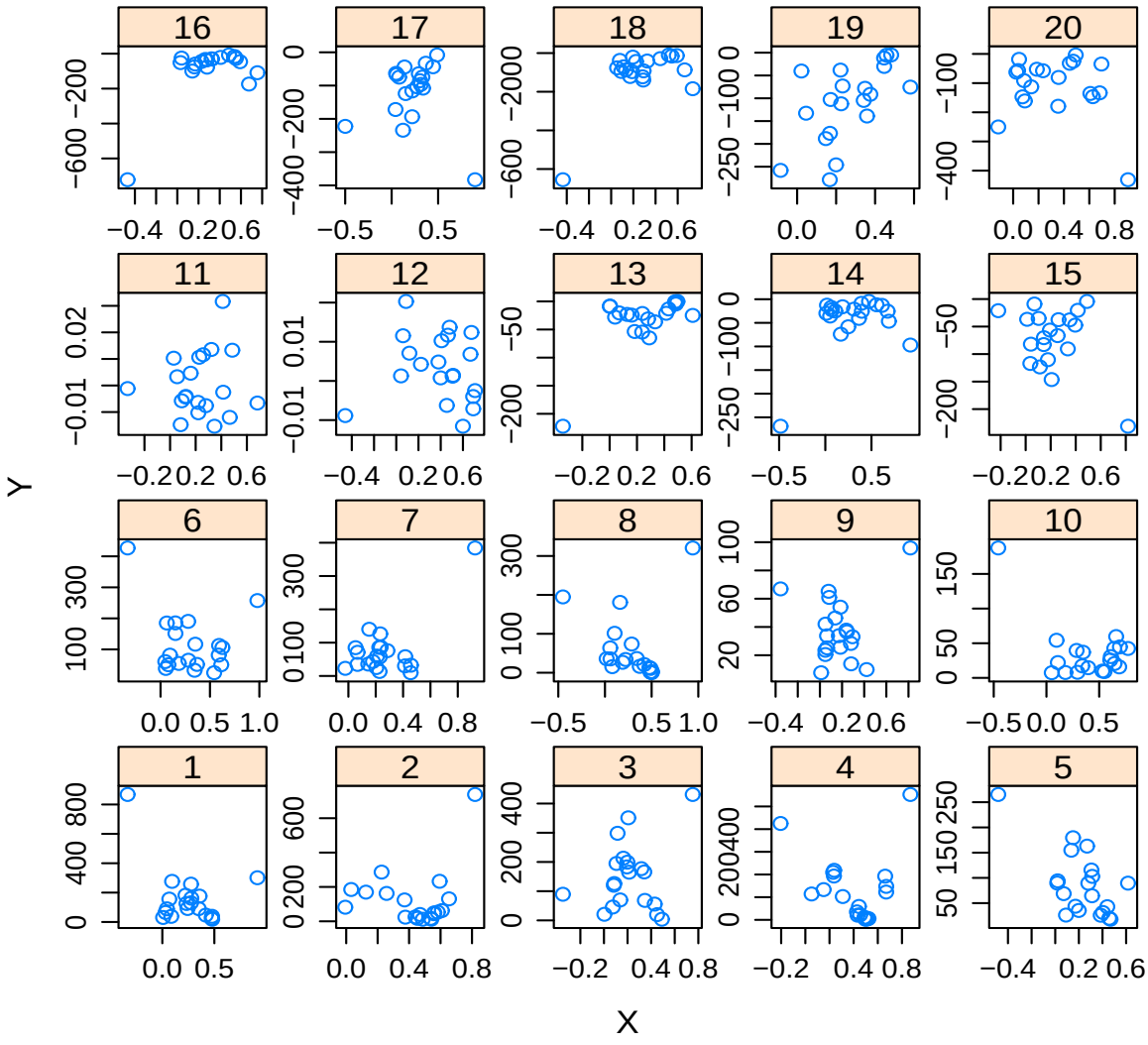


Table 5.8: (Bootstrap Power performance) Proportion of rejections at level 0.01 for testing of no simple covariate effect using the lower 1 percentile of the corresponding p-values under the null hypothesis as the cut off point for rejecting the null when $n_i = 20$ and $a = 20$ using the model (5.3.2). The results are based on 1000 simulations

τ	prop	pNP	GAM Spline	GAM Loess	drop	CF
0.0625	0.1	1.000	1.000	1.000	0.434	0.923
	0.26	1.000	1.000	1.000	0.261	0.855
	0.33	1.000	1.000	1.000	0.450	0.920
	0.5	1.000	1.000	1.000	0.357	0.926
	0.76	0.180	1.000	1.000	0.296	0.912
	0.9	0.259	1.000	1.000	0.870	0.993
0.125	0.1	1.000	1.000	1.000	0.468	0.907
	0.26	1.000	1.000	1.000	0.258	0.858
	0.33	1.000	1.000	1.000	0.491	0.922
	0.5	1.000	1.000	1.000	0.358	0.929
	0.76	0.138	1.000	1.000	0.267	0.914
	0.9	0.258	1.000	1.000	0.847	0.993
0.25	0.1	1.000	1.000	1.000	0.423	0.932
	0.26	1.000	1.000	1.000	0.262	0.864
	0.33	1.000	1.000	1.000	0.469	0.923
	0.5	1.000	1.000	1.000	0.317	0.935
	0.76	0.209	1.000	1.000	0.216	0.915
	0.9	0.303	1.000	1.000	0.830	0.994
0.5	0.1	1.000	1.000	1.000	0.352	0.926
	0.26	1.000	1.000	1.000	0.220	0.883
	0.33	1.000	1.000	1.000	0.437	0.942
	0.5	1.000	1.000	1.000	0.296	0.940
	0.76	0.187	1.000	1.000	0.196	0.918
	0.9	0.365	1.000	1.000	0.824	0.995

Table 5.9: Proportion of rejections at level 0.01 under the null hypothesis of no main covariate effect in the model (5.3.5) when $n_i = 20$ and $a = 20$. The results are based on 1000 simulations.

τ	prop	pNP	GAM Spline	GAM Loess	Drop	CF
0.0625	0.1	0.019	0.012	0.013	0.012	0.007
	0.26	0.015	0.013	0.015	0.014	0.010
	0.33	0.009	0.013	0.014	0.015	0.013
	0.5	0.013	0.011	0.008	0.015	0.007
	0.76	0.009	0.012	0.015	0.006	0.006
	0.9	0.009	0.012	0.009	0.007	0.010
0.125	0.1	0.019	0.015	0.019	0.009	0.011
	0.26	0.012	0.013	0.015	0.011	0.010
	0.33	0.013	0.013	0.014	0.013	0.013
	0.5	0.015	0.011	0.008	0.012	0.007
	0.76	0.011	0.012	0.015	0.006	0.006
	0.9	0.013	0.012	0.009	0.009	0.010
0.25	0.1	0.020	0.012	0.013	0.010	0.007
	0.26	0.013	0.013	0.015	0.011	0.010
	0.33	0.011	0.013	0.014	0.012	0.013
	0.5	0.014	0.011	0.008	0.012	0.007
	0.76	0.013	0.012	0.015	0.008	0.006
	0.9	0.009	0.012	0.009	0.008	0.010
0.5	0.1	0.016	0.015	0.019	0.011	0.011
	0.26	0.012	0.013	0.015	0.011	0.010
	0.33	0.011	0.013	0.014	0.013	0.013
	0.5	0.013	0.011	0.008	0.011	0.007
	0.76	0.011	0.012	0.015	0.011	0.006
	0.9	0.011	0.012	0.009	0.008	0.010

Table 5.10: (Power performance) Proportion of rejections at level 0.01 for testing of main covariate effect when $n_i = 20$ and $a = 20$ using the model 5.3.2 with $\theta = \pi/4$ and $b = 0.5$.

tau	prop	pNP	GAM Spline	GAM Loess	drop	CF
0.0625	0.1	1.000	1.000	1.000	0.995	0.999
	0.26	1.000	1.000	1.000	0.981	0.989
	0.33	1.000	1.000	1.000	1.000	0.997
	0.5	1.000	1.000	1.000	0.995	0.988
	0.76	1.000	1.000	1.000	0.833	0.652
	0.9	0.999	1.000	1.000	1.000	0.935
0.125	0.1	1.000	1.000	1.000	0.999	0.999
	0.26	1.000	1.000	1.000	0.985	0.991
	0.33	1.000	1.000	1.000	1.000	0.997
	0.5	1.000	1.000	1.000	0.999	0.989
	0.76	1.000	1.000	1.000	0.836	0.663
	0.9	1.000	1.000	1.000	1.000	0.937
0.25	0.1	1.000	1.000	1.000	0.997	1.000
	0.26	1.000	1.000	1.000	0.987	0.995
	0.33	1.000	1.000	1.000	1.000	0.998
	0.5	1.000	1.000	1.000	0.998	0.990
	0.76	1.000	1.000	1.000	0.871	0.675
	0.9	1.000	1.000	1.000	1.000	0.948
0.5	0.1	1.000	1.000	1.000	1.000	1.000
	0.26	1.000	1.000	1.000	0.989	0.997
	0.33	1.000	1.000	1.000	1.000	0.999
	0.5	1.000	1.000	1.000	0.997	0.992
	0.76	1.000	1.000	1.000	0.895	0.711
	0.9	0.999	1.000	1.000	1.000	0.954

5.4 Data Analysis

5.4.1 Ozone Data Revisited

In this section, the ozone data which was discussed in Section 3.3.1 and Section 4.6.1 is revisited. Here the change of ozone concentration (O_3) is considered with the day of year (doy), while different temperature levels are observed. The temperature is divided into 20 intervals and is considered as a categorical factor (trt) with 20 levels. The factor levels and their corresponding temperature values and the number of covariate values (doy) in each level is summarized in Table 5.11.

Table 5.11: Table of levels of temperature (trt) and its corresponding temperature values and the number of covariate values in each level

trt	1	2	3	4	5	6	7	8	9	10
temp	[24.5, 38.5)	[38.5, 41.5)	[41.5, 46.5)	[46.5, 49.5)	[49.5, 51.5)	[51.5, 53.5)	[53.5, 55.5)	[55.5, 58.5)	(58.5, 60.5)	[60.5, 62.5)
n_i	16	17	17	16	19	17	21	15	15	19
trt	11	12	13	14	15	16	17	18	19	20
temp	[62.5, 63.5)	[63.5, 65.5)	[65.5, 67.5)	[67.5,70.5)	[70.5, 72.5)	[72.5, 75.5)	[75.5, 78.5)	[78.5, 81.5)	[81.5, 86.5)	[86.5, 93.5)
n_i	10	16	16	21	14	18	15	19	14	15

Tests of no covariate-treatment interaction, no main covariate and no simple covariate effects in the data where the covariate variable is doy and the treatment (group) is temperature level are performed. The quadratic like relationship between O_3 and doy was depicted in Figure 3.1 in Section 3.3.1. The scatter plot (a) in Figure 5.7 shows the relationship between O_3 and temperature. The plot shows heteroscedasticity as the variance of O_3 increases with the temperature level. Graph (b) in Figure 5.7 strengthens the evidence of heteroscedasticity in graph (a). Figure 5.8 shows the relationship of O_3 and doy within each treatment level. The graph shows doy-temperature interaction effect in the data. These were conveyed by the difference pattern of relationship between doy and O_3 in different temperature level. For example, in trt level 15, 17 and 18 the relationship of doy and O_3 appears to be negatively exponential, while other at temperature levels, the relationship between doy and O_3 is not apparent.

All the tests being considered in the previous section which are the proposed pNP test,

GAM Spline, GAM Loess, drop and CF test are used to analyze the data. The results for testing covariate-treatment interaction, simple covariate and main covariate effect are presented in Table 5.12. The table shows that the only test that has a significant covariate-treatment interaction at $\alpha = 0.05$ is the pNP test. This result is parallel to the results of the simulation study for the performance of the pNP test in Table 5.3 and in Table 5.5 where the pNP test has the highest power to detect the covariate-treatment interaction especially in the presence of heteroscedasticity of variance in different treatments level as they appear in Figure 5.11. All tests for the simple covariate effect of doy on O_3 after adjusting for temperature effect, show significant result at $\alpha = 0.05$. However at $\alpha = 0.01$, the drop (p-value = 0.012) and the CF (p-value=041) tests do not have a significant result.

Figure 5.7: Scatter plot of ozone vs temp and box plot of ozone vs temp

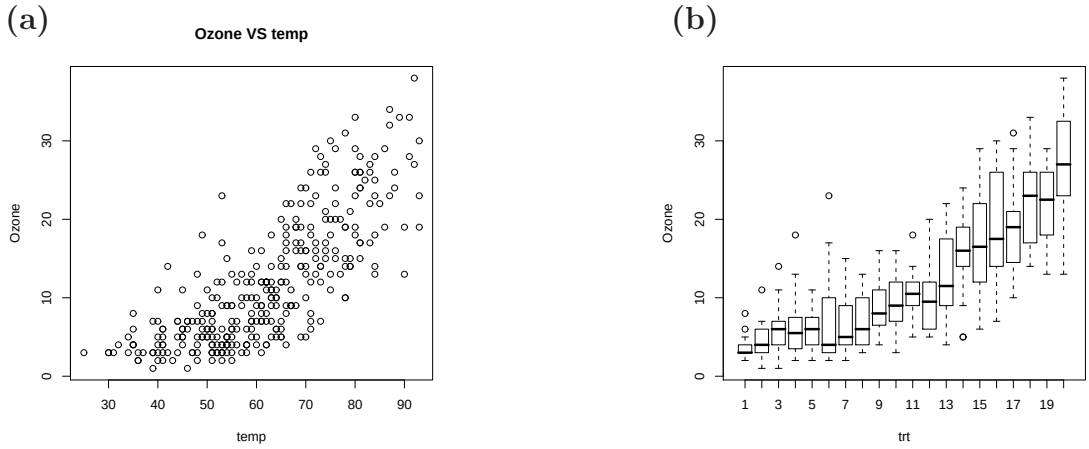
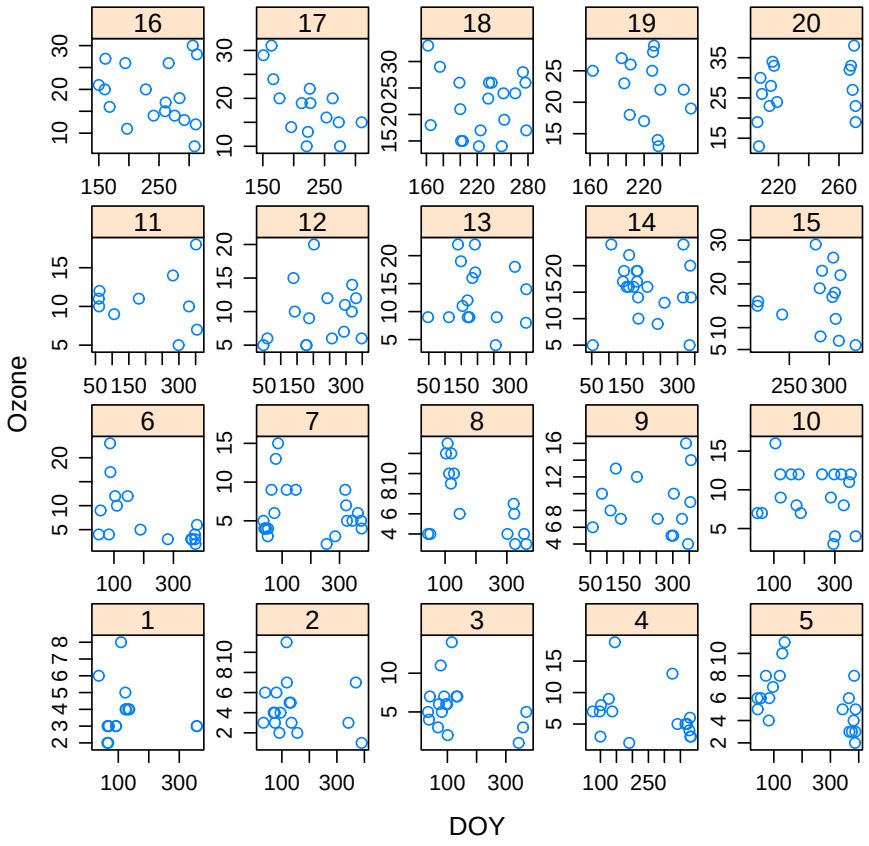


Table 5.12: P-values for test of no doy-temperature interaction, no simple doy and no main doy effects.

Tests	doy-temp int	simple doy	main doy
pNP	0.027	0.003	0.000
GAM Spline	0.414	0.000	0.000
GAM Loess	0.393	0.000	0.000
drop	0.099	0.012	0.002
CF	0.243	0.041	0.002

Figure 5.8: Ozone Vs doy within temperature levels



Chapter 6

Conclusion and Post-dissertation Research

6.1 Conclusion

This dissertation developed nonparametric tests to study the relationship between a response variable and a continuous covariate in the presence of categorical factors. Traditional approaches like general linear models (GLM) and generalized linear model (GLMz) cater to specific types of response variables, e.g., GLM is suitable for a continuous response variable and GLMz is suitable for either a continuous or a discrete response variable that comes from an exponential family. The tests developed in this dissertation were not restricted to any particular type of response variable. The models, hypotheses and test statistics were formulated in a general form to incorporate both continuous and discrete response variables. Further, the asymptotic results were obtained without any restrictions on distributional assumptions, any particular link function, any constant variance or any explicit relationship of the conditional mean of response with the given covariate.

A model employing a conditional distribution function was used to formulate the hypothesis of independence in the first part of the contribution. The second and the third parts use a model that was based on the decomposition of a conditional mean of regression function that is potentially nonlinear. The foundation of the method developed involves augmenting

each pair of the data for all treatments with a fixed number of nearest neighbors as pseudo-replicates. The test statistics were constructed by taking the difference of two quadratic forms multiplied by an appropriate standardizing rate. The asymptotic distributions of the test statistics were obtained under a setting in which the number of nearest neighbors is small and the number of covariate values is large. Simulation studies were presented to evaluate the performance of the pNP test and compared to several benchmark methods. Real applications of two data sets were also discussed.

The first part of the contribution (Chapter 3) was devoted to the development of theory for the test of independence between a continuous covariate and a continuous or discrete response variable after adjusting for the heteroscedastic treatment effect. In this case, the test statistic is equivalent to the average lagged correlations between the response and nearest neighbor local estimates of the conditional mean of response given the covariate for each treatment group. The parametric standardizing rate was obtained for the proposed test statistics. Numerical studies showed that the new test procedure not only maintains the intended type I error rate, but also has robust power to detect nonlinear dependency in the presence of outliers that might result from highly skewed distributions.

Chapter 4, the second part of the contribution, presented the theory and numerical studies for tests of no covariate-treatment interaction and no main covariate effects specified through a decomposition of a conditional mean of regression function that can possibly be nonlinear. In depth discussion on the effects defined through traditional decomposition of the mean regression function and a nonparametric decomposition of conditional distribution function from Wang and Akritas (2006) favored the former in a possibly nonlinear form to allow generality and appropriate interpretation based on the data. The test for no covariate-treatment interaction effect developed in this chapter has demonstrated superior performance in computing time, estimates of type I error and power performance compared to the test from Wang and Akritas (2006).

Due to the need for tests that can accommodate data containing a large number of factors or factor levels, the third part of the contribution (Chapter 5) extended the theory in the previous two parts to the case where the number of treatment levels go to infinity.

Results were obtained in this asymptotic setting. Simulation studies and an application were presented.

6.2 Post-dissertation Research

The new tests currently are applicable when there is only one continuous covariate. Using an approach similar to that used in constructing the statistics of the new tests, these new tests could be extended to cater to the existence of more than one covariate. We will define nearest neighbors through multivariate spacings used in Li and Liu (2008).

The new method tests only the dependency of two variables using the original observations. These results rely on finite fourth moment and the asymptotic variances are functions of the conditional variances of the responses. Estimation of variances for skewed or heavy tailed data often has very poor performance. A competing set of rank results may be developed and are expected to perform better.

In addition, the methods used here could also be extended to high dimensional data by combining asymptotic theory for the construction of the new tests with a shrinkage method.

Bibliography

Aitkin, M., D. Anderson, B. Francis, and J. Hinde (1989). *Statistical Modelling in GLIM*. Oxford University Press: New York.

Akritas, M., E. Antoniou, and L. Wang (2003). Fully nonparametric ancova with fixed window sizes. *manuscript*.

Akritas, M., S. Arnold, and Y. Du (2000). Nonparametric models and methods for non-linear analysis of covariance. *Biometrica 87*, 507–526.

Akritas, M. and N. Papadatos (2004). Heteroscedastic one-way ancova and lack-of-fit tests. *Journal of the American Statistical Association 99*, 368–382.

Azzalini, A. and A. Bowman (1993). On the use of nonparametric regression for checking linear relationship. *Journal of royal Statistical Society B 55*, 549–557.

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc. 80*, 580–598.

Butte, A. and I. S. Kohane (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 418–429.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory 75*, 261–277.

Deheauvels, P. (1981). An asymptotic decomposition for multivariate distribution free test of independence. *J Multivariate Anal. 11*, 102 –113.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *The Annals Of Statistics 27*, 1012–1040.

Dette, H. and A. Munk (1998). Validation of linear regression models. *The Annals of Statistics 26*, 778–800.

D'haeseleer, P., X. Wen, S. Fuhrman, and R. Somogyi (1998). Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. *Proceedings of the second international workshop on Information processing in cell and tissues*, 203–212.

Dobson, A. (2002). *An Introduction To Generelized Linear Models*. Boca Raton, London, New York and Washington D.C.: Chapman and Hall/CRC.

Eubank, R. and J. Hart (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics 20*, 1412–1425.

Faraway, J. (2006). *Extending the Linear Model with R*. Chapman and Hall/CRC: Boca Raton, FL.

Genest, C. and B. Rémilland (2004). Tests of independence and randomness based on the empirical copula process. *Sociedad de Estadistica e Investigacion Operativa 13*, 335–369.

Hardle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*, 1926–1947.

Harrel, F. E. (2002). *Regression Modelling Strategies With Aplication to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning, Data Mining. Inference and Prediction*. New York: Springer.

Hettmansperger, T. and J. McKean (1998). Robust nonparametric statistical methods. pp. 170–175.

Koenker, R. and G. Basset (1978). Regression quantiles. *Econometrica 46(1)*, 33–50.

Kraskov, A., H. Stogbauer, and P. Grassberger (2004). Estimating mutual information. *Physical Review 69*.

Kutner, M. H., C. Nacthtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models*. New York: McGraw - Hill Irwin.

Li, J. and R. Y. Liu (2008). Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *Ann. Statist. Volume 36*, 1299–1323.

McCullagh, P. and J. Nelder (1993). *Generealized Linear Model*. Chapman and Hall/CRC.

McKean, J. and R. M. Schrader (1980). The geometry of robust procedures in linear models. *Journal of Royal Statistical Sociey 42*, 366–371.

Müller, H.-G. (1992). Goodness-of-fit diagnostics for regression models. *Scand. J. Statist. 19*, 157–172.

Muller, H. G. (1992). Goodness of fit diagnostics for regression models. *Scandinavian Journal of Statistics 19*, 157–172.

Parson, L., E. Haque, and H. Liu (2004). Subspace clustering for high dimensional data:a review. *Sigkdd Explorations 6*, 90–105.

Pyke, R. (1965). Spacing (with discusssion). *Journal of Royal Statistical Society 27*, 395–449.

Steuer, R., J.Kurths, C.O.Daub, J.Weise, and J.Selbig (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics 18*, S231–S240.

Terpstra, J. T. and J. Mckean (2005). Rank-based analyses of linear models using r. *Journal of Statistical Software 14(7)*.

Tsangari, H. and M. Akritas (2004). Nonparametric ancova with two and three covariates. *Journal of Multivariate Analysis 88*(2), 298–319.

Venables, W. and B. Ripley (1997). *Modern Applied Statistics with S-Plus Second Edition*. New York: Springer.

Wang, L. and M. Akritas (2006). Testing for covariate effects in the fully nonparametric analysis of covariance model. *Journal of the American Statistical Association 101*,

507–526.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *JRSSB 62*, 413–428.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *JRSSB 70*, 495–518.

Appendix A

R code

A.1 pNP Tests

```
#The following R code is the code to do the proposed test pNP when a
#is small. dat is a dataframe with three columns; one column has
#name Y; one column has name X; the other column has name trt;

#map the index over r=1,...,N to i=1, ...a, j=1, .. n_i , r is an
#integer; n is a vector of the sample sizes

mapindex=function(r, n){
    aaa=length(n)
    sumn=numeric()
    for ( i in 1:aaa)     sumn=c(sumn, sum(n[seq(i)]) )
    imap=sum(sumn<r)+1
    jmap=r-sum(n[ seq(aaa)[(sumn<r)]] )
    c(imap, jmap)
} # mapindex function that works for vector; first column gives the
#i's and 2nd gives the j's.

mapindexV=function(r, n){
aaa=length(n)
```

```
sumn=numeric()

for ( i in 1:aaa)
sumn=c(sumn, sum(n[seq(i)]) )
imap=unlist(tapply(r, seq(length(r)),function(x) sum(sumn<x)+1  )  )
jmap=unlist(tapply(r, seq(length(r)),function(x)
     x-sum(n[seq(aaa)[(sumn<x)]])))
     cbind(imap, jmap)
      }
##################
#i1 is the i1 th group; n is the vector of
#sample sizes; position.i1 function gives the starting and end
#position of covariate values in the i1th group among the vector
#listing all covariate values. e.g., covariate values in group 1
#start from 1st value to the n1 th value; those in group 2 start
#from n1+1 and end at n1+n2 th value.

position.i1=function(i1, n){ if (i1==1) lower=1 else
lower=sum(n[1:(i1-1)])+1 upper=sum(n[1:i1]) c(lower, upper) }

##################
# k is the number of nearest neighbors used

NPtest.new= function( dat, k) {
  X=dat$X;  trt=dat$trt; Y1=dat$Y;
  #Y1=unlist(tapply(dat$Y, trt,standard) )
  alltrt=rbind(Y1, X, unlist(tapply(X, trt, rank) ) )
  n=unlist(tapply(rep(1, nrow(dat)), trt, sum))
  N=sum(n); a=length(n)
  for (i1 in 1:a){
  locationi1=position.i1(i1,n);
  orderwant=order(alltrt[2,locationi1[1]:locationi1[2]])+locationi1[1]-1;
```

```
  alltrt[,locationi1[1]:locationi1[2] ]=  alltrt[,orderwant]
     }
     psudo<-array(0, c(a, sum(n), k))
     index<-array(0, c(a,sum(n), k))
     ### Augment observations for each cell
     ########################################
for (i in 1:a){
for (j in 1:N){
if (i==1){
if ( j<= n[1] ) { newtrt<-alltrt[,1:n[1]]
total<-ncol(newtrt)
jj<-j
       }
if  (j>=n[1]+1) { newtrt<-cbind(alltrt[,1:n[1]], alltrt[, j])
total<-jj<- ncol(newtrt)
      }
      }
if (i>1) {
if ((j<=sum(n[1:i]))& (j>=sum(n[1:(i-1)])+1) ) {
newtrt<-alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])]
total<- ncol(newtrt)
jj<- j-sum(n[1:(i-1)])

    } else {
newtrt<-cbind(alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])],alltrt[,j] )
total<-jj<-ncol(newtrt)
                  }
               }
newtrt[3, ]<-rank(newtrt[2, ])
 flag<-((jj==total)& (jj>n[i])&
 c(rep(T, total-1), F)  )  | (jj<=n[i])
```

```
if ((jj==total) & (jj>n[i]) ) {newtrt[3, -jj]<- rank(newtrt[2, -jj])
total<-total-1 }
target<-newtrt[3, jj ]
newtrt<-newtrt[, flag]

if (trunc(target) <= ((k-1)/2) ) {
psudo[i,j,]<-newtrt[1,order(newtrt[3, ])[1:k]]
index[i,j,  ]<-seq(1,
total)[ order(newtrt[3, ])[1:k]]}

 if (trunc(target) > (total- ((k-1)/2)))
 {psudo[i,j,  ]<- newtrt[1, order(total-newtrt[3, ])[1:k] ]
 index[i,j,  ]<- seq(1, total)[ order(total-newtrt[3, ])[1:k]]
              }
 if ((trunc(target)<=(total-(k-1)/2 ))&(trunc(target) >((k-1)/2) ))
               {
 psudo[i,j,]<-newtrt[1,order((abs(newtrt[3,]-trunc(target))))[1:k] ]
 index[i,j,]<-seq(1,total)[order((abs(newtrt[3,]-trunc(target))))[1:k]]
              }
               }  #end of j
               }   #end of i

##*************************************
 cellmean<-apply(psudo,c(1,2), mean)
 colmean=apply(psudo, 2, mean)
 sig<- cov(t(cellmean))
 # diagonal part gives the \hat\sigma_{1,i}^2
 #and off-diagonal part gives \hat\sigma_{1,i_1, i_2}

 sigXij<-apply(psudo, c(1, 2), var)
 # get a axN matrix with \hat\sigma_i^2(X_{ij}) =sigXij[i, j]
```

```
MSTd=k*a*sum( (colmean-mean(psudo)  )^2       )  /(sum(n)-1)
meanrk=apply(psudo,1, mean)
MSTphi=k*sum((cellmean-matrix(rep(meanrk,N),ncol=N))^2)
       /((sum(n)-1)*a)
MSTc=k*sum((cellmean-matrix(rep(meanrk,N),ncol=N)-
     matrix(rep(colmean,a),
     ncol=N, byrow=T)+mean(cellmean) )^2  )/((a-1)*(N-1))
MSE=sum((psudo-array(rep(cellmean, k), c(a, sum(n), k))   )^2  )
     / (sum(n)*a*(k-1)  )

Tsinter=(sqrt(sum(n)) * (MSTc-MSE))
Tsc=(sqrt(sum(n)) * (MSTd-MSE))
Tss=(sqrt(sum(n)) * (MSTphi-MSE))

## calculate Td1 and Td2 for diagnostics
offdiagsum=function(x){
sum(matrix(x)%*%matrix(x,ncol=length(x)))-sum(x^2)
}
 Td1=k*mean( apply(cellmean, 2, offdiagsum) )/a
 Td2=mean(apply(psudo, c(1,2), offdiagsum ) )/(k-1)

 ##************************************
 #Calculate estimate of variance for test statistics
 # count is a matrix; first three columns give the value of
 # i1, j2,i; the last column
 # gives the number of times X_{ij_2} is used in construction
 #of windows for all covariate values in group i_1

 count<-matrix(-1, a^2*N, 4)
 whereini=0
 for( i1 in 1:a){
 for (j2 in 1:N){
```

```
 for (i in 1:a){
 whereini=whereini+1
 if (i1==1) lower=1 else lower=sum(n[1:(i1-1)])+1
 upper=sum(n[1:i1])
 counti1j2i=sum(index[i,lower:upper, ]==((mapindex(j2, n)[1]==i)
            *mapindex(j2,n)[2]))

 count[whereini, ]=c(i1, j2, i, counti1j2i)
                }}}
 # to calculate the first term of asym variance
   tau1.2=Exi1calc(alltrt, index, a, n, k, sigXij)
# to calculate the third term of asym variance: tau3^2
subcount= count[count[,1]!=count[,3], ]
# entries in count that i1 \ne i
prodcount1.1=tapply(subcount[,4],list(subcount[,2],subcount[,3]),sum)
# prodcount1.1 is \sum_{i_1, i_1 \ne i}^a \frac{n_{i_1}}{n_i}
#d_{i_1i}(X_{ij})
  tau3=0
  for (i in 1:a){
  starti=position.i1(i,n)[1]-1
  for (jp in starti+(2:n[i])){
  for (j in (max(1, (jp-k+1)): (jp-1)) ){
  Bijjp= (prodcount1.1[j,i]/k+1 ) * (k-jp+j)*(jp-j<=k-1)
     tau3=tau3+ (  Bijjp^2+Bijjp-2*(jp-j<=((k-1)/2) )  )
     *(jp-j<=k-1) * sigXij[i,j] *sigXij[i,jp] *(j!=jp)
   } } }

tau3=tau3*4/(sum(n)*a^2*(k-1)^2)
tauAsyc=tau1.2+tau3
tauAsys=tau3
tauAsyinter=(tau1.2)/((a-1)^2) + tau3
   pvalue.inter=1-pnorm(Tsinter/sqrt(tauAsyinter))
```

```
   pvalue.cov=1-pnorm(Tsc/sqrt(tauAsyc))
   pvalue.sim=1-pnorm(Tss/sqrt(tauAsys))

list(Tsc=Tsc,Tss=Tss,Tsinter, tauAsyc=tauAsyc, tauAsys=tauAsys,
tauAsyinter=tauAsyinter, Td1=Td1, Td2=Td2, pvalue.cov=pvalue.cov,
pvalue.sim=pvalue.sim,pvalue.inter=pvalue.inter,
pvalues=c(pvalue.cov,pvalue.sim, pvalue.inter),
tau1.2=tau1.2,tau3=tau3)
        }

#makepseudo creates pseudo observations

makepseudo=function(N,n, k, a, alltrt){
          psudo<-array(0, c(a, sum(n), k))
          index<-array(0, c(a,sum(n), k))
# Augment observations for each cell
#****************************************
 for (i in 1:a){
 for (j in 1:N){
    if (i==1){
    if ( j<= n[1] ) {
    newtrt<-alltrt[,1:n[1]]
    total<-ncol(newtrt)
    jj<-j
             }
    if  (j>=n[1]+1) {
       newtrt<-cbind(alltrt[,1:n[1]], alltrt[, j])
       total<-jj<- ncol(newtrt)
                    }
             }
    if (i>1) {
    if ((j<=sum(n[1:i]))& (j>=sum(n[1:(i-1)])+1) ) {
       newtrt<-alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])]
```

```
     total<- ncol(newtrt)
     jj<- j-sum(n[1:(i-1)])
     } else {
newtrt<-cbind(alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])],alltrt[,j])
total<-jj<-ncol(newtrt)
                  }
           }
 newtrt[3, ]<-rank(newtrt[2, ])
 flag<-((jj==total)& (jj>n[i])& c(rep(T, total-1), F) )|(jj<=n[i])
 if ((jj==total) & (jj>n[i]) ) {
 newtrt[3, -jj]<- rank(newtrt[2, -jj])
 total<-total-1
                         }
 target<-newtrt[3, jj ]
 newtrt<-newtrt[, flag]
 if (trunc(target) <= ((k-1)/2) )
 {psudo[i,j,  ]<- newtrt[1, order(newtrt[3, ])[1:k]]
 index[i,j,  ]<- seq(1, total)[ order(newtrt[3, ])[1:k]]
               }
 if (trunc(target) > (total- ((k-1)/2)))
 {psudo[i,j,  ]<- newtrt[1, order(total-newtrt[3, ])[1:k] ]
 index[i,j,  ]<- seq(1, total)[ order(total-newtrt[3, ])[1:k]]
           }
 if ((trunc(target) <=(total-(k-1)/2 ) )&(trunc(target) >((k-1)/2)))
              {
 psudo[i,j,]<-newtrt[1, order((abs(newtrt[3,]-trunc(target))))[1:k]]
 index[i,j, ]<- seq(1, total)[
              order((abs(newtrt[3,]-trunc(target) ))) [1:k]]
              }
   }  #end of j
   }   #end of i
 psudo }
```

```
Exi1calc=function(alltrt, index, a, n, k, sigXij) {
 countijipjp=numeric()
 for( i in 1:a){
 whichparti=position.i1(i,n)[1]: position.i1(i,n)[2]
 for (ip in ((1:a)[-i]) ){
 whichpartip=position.i1(ip,n)[1]: position.i1(ip,n)[2]

 for (j in whichparti){

 tmp.jp=alltrt[2, whichpartip]-alltrt[2,j]
 morejp=T; nextone=1
 while ((morejp==T)||(nextone<=min(c(5*k, n[ip]))))  {
 jp=jprange.all[nextone]
 ij=mapindex(j, n)
 ipjp=mapindex(jp, n)
 Pjir.jpipr.vec=numeric()

 for (i1 in 1:a){
 bound=low.up(i, ip,j,jp,ij,ipjp,alltrt,k,whichparti,whichpartip,index)
 lowbound=bound[1]
 upbound=bound[2]
 morejp=(lowbound <= upbound)
 if (morejp==T){
 #Pjir.jpipr gives the proportion of covariates values in i_1 group fall
 #in interval (max, min).
 #That is \int_{max}^{min} g_{i_1}(x) dx
 probvector=(alltrt[2,position.i1(i1,n)[1]:position.i1(i1,n)[2]]>lowbound)
          *(alltrt[2,position.i1(i1,n)[1]:position.i1(i1,n)[2]]<=upbound )
 if (i1==i) probvector[ij[2]]= sum(index[ip,jp ,]==( (ij[1]==i) *ij[2]) )
 # indicator function of whether X_{ij} is used in W_{i',X_{i'j'}}
 if(i1==ip)probvector[ipjp[2]]= sum(index[i,j,]==((ipjp[1]==ip) *ipjp[2]))
 # indicator function of whether X_{ij} is used in W_{i',X_{i'j'}}
```

```
 Pjir.jpipr=mean(probvector)
 } else Pjir.jpipr=0       #end of (morejp==T)
 Pjir.jpipr.vec=c(Pjir.jpipr.vec, Pjir.jpipr*n[i1] )
 } #end of i1
 countijipjp=rbind(countijipjp, c(i,j,ip,jp, sum(Pjir.jpipr.vec),
                                    sigXij[i,j]*sigXij[ip,jp]) )
 nextone=nextone+1
 morejp=morejp&(nextone<=n[ip])
                    }  # end while
 }}} # end of j, ip, and i
 # the 5th column of countijipjp gives E(M_{iji'j'|C_{ij,t}, C_{i'j',t})
 countijipjp=countijipjp[countijipjp[,5]>0,]
 Exi1=sum((countijipjp[,5]^2 +countijipjp[,5])
              *countijipjp[,6])*2/(sum(n)*a^2*k^2)
 Exi1
                  }
# low.up gives the lower bound max{,} and upper bound min{, }
low.up=function(i, ip, j, jp, ij, ipjp, alltrt, k, whichparti,
whichpartip, index){
        ni=length(whichparti)
        nip=length(whichpartip)
        where1=ifelse((j-k > whichparti[1]), j-k, whichparti[1])
        where2=ifelse((j+k<=whichparti[ni]), j+k, whichparti[ni])
Xijused.i=ifelse(apply(index[i, where1:where2, ]==(   (ij[1]==i)
          *ij[2]) , 1, sum), T, F)
# give T or F to tell whether X_{ij} is used in window W_{ir} for
#all r in group i
where1p=ifelse((jp-k > whichpartip[1]), jp-k, whichpartip[1])
where2p=ifelse((jp+k<=whichpartip[nip]), jp+k, whichpartip[nip])
Xijpused.ip=ifelse(apply(index[ip, where1p:where2p, ]==(
            (ipjp[1]==ip) *ipjp[2]) , 1, sum) , T, F)
lp1=(where1:where2)[Xijused.i][1]; Lij=(alltrt[2, lp1] + alltrt[2,
```

```
    ifelse((lp1>whichparti[1]), lp1-1, lp1) ]  )/2

up1=(where1:where2)[Xijused.i][length((where1:where2)[Xijused.i]) ];
Uij=(alltrt[2, up1] + alltrt[2, ifelse((up1<whichparti[ni]), up1+1,
up1) ]  )/2

lp2=(where1p:where2p)[Xijpused.ip][1]; Lipjp=(alltrt[2, lp2] +
alltrt[2, ifelse((lp2>whichpartip[1]), lp2-1, lp2) ]  )/2
up2=(where1p:where2p)[Xijpused.ip][length((where1p:where2p)
                                          [Xijpused.ip])];
Uipjp=(alltrt[2, up2] + alltrt[2, ifelse((up2<whichpartip[nip]),
      up2+1, up2) ]  )/2
lowbound=max(Lij,  Lipjp) ; upbound=min(Uij, Uipjp)
c(lowbound, upbound) }
```

A.2 pNP Tests When a and N Are Large

```
#The following R code is the code to do the proposed test pNP when a
#and N are large. The program follows that of the pNP test for small
#except for different test statistics and asymptotic variance
#Test statistics
Tsinter=(sqrt(a* sum(n)) * (MSTc-MSE))
Tsc=(sqrt(sum(n))*(MSTd-MSE))
Tss=(sqrt(a* sum(n)) * (MSTphi-MSE))
#Asymptotic Variance
   tauAsyc=tau1.2     #covariate
   tauAsys=a*tau3     #simple
   tauAsyinter=a*tau3 #interaction
```

A.3 WA Tests

```
# Lan Nonparametric test # i1 is the i1 th group; n is the vector of
# sample sizes;
# position.i1 function gives the starting and end position of
#covariate values in the i1th group among the vector listing all
#covariate values. e.g., covariate values in group 1 start from 1st
#value to the n1 th value; those in group 2 start from n1+1 and end
#at n1+n2 th value.
position.i1=function(i1, n){ if (i1==1) lower=1 else
lower=sum(n[1:(i1-1)])+1 upper=sum(n[1:i1]) c(lower, upper) }
#dat is a dataframe with three columns;one column has name Y
#onecolumn has name X; the other column has name trt # k is the
#number of nearest neighbors used

lan.NP2= function( dat, k) {
  X=dat$X;  trt=dat$trt; Y=dat$Y
  alltrt=rbind(Y, X, unlist(tapply(X, trt, rank) ) )
  n=unlist(tapply(rep(1, nrow(dat)), trt, sum))
  N=sum(n); a=length(n)
  for (i1 in 1:a){
  locationi1=position.i1(i1,n);
  orderwant=order(alltrt[2,locationi1[1]:locationi1[2]])+locationi1[1]-1;
  alltrt[,locationi1[1]:locationi1[2] ]=  alltrt[,orderwant]
                    }
  psudo<-array(0, c(a, sum(n), k))
  index<-array(0, c(a,sum(n), k))
  ### Augment observations for each cell
  #****************************************
  for (i in 1:a){
  for (j in 1:N){
  if (i==1){
```

```
if ( j<= n[1] ) {
newtrt<-alltrt[,1:n[1]]
total<-ncol(newtrt)
jj<-j
                    }
if  (j>=n[1]+1) {
    newtrt<-cbind(alltrt[,1:n[1]], alltrt[, j])
    total<-jj<- ncol(newtrt)
            }
      }
if (i>1) {
if ((j<=sum(n[1:i]))& (j>=sum(n[1:(i-1)])+1) ) {
newtrt<-alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])]
total<- ncol(newtrt)
jj<- j-sum(n[1:(i-1)])
     } else {
newtrt<-cbind(alltrt[,(sum(n[1:(i-1)])+1): sum(n[1:i])], alltrt[,j] )
total<-jj<-ncol(newtrt)
         }
      }
newtrt[3, ]<-rank(newtrt[2, ])
flag<-((jj==total)& (jj>n[i])& c(rep(T, total-1), F)  )  | (jj<=n[i])
if ((jj==total) & (jj>n[i]) ) {
newtrt[3, -jj]<- rank(newtrt[2, -jj])
total<-total-1
                   }
target<-newtrt[3, jj ]
newtrt<-newtrt[, flag]
if (trunc(target) <= ((k-1)/2) )
   {psudo[i,j,  ]<- newtrt[1, order(newtrt[3, ])[1:k]]
    index[i,j,  ]<- seq(1, total)[ order(newtrt[3, ])[1:k]]
    }
```

```
if (trunc(target) > (total- ((k-1)/2)))
   {psudo[i,j,  ]<- newtrt[1, order(total-newtrt[3, ])[1:k] ]
       index[i,j,  ]<- seq(1, total)[ order(total-newtrt[3, ])[1:k]]
    }
if ((trunc(target)<=(total-(k-1)/2 ) ) & (trunc(target) >((k-1)/2) ))
   {psudo[i,j,]<- newtrt[1,
              order((abs(newtrt[3,]-trunc(target) ) ))[1:k] ]
   index[i,j, ]<- seq(1, total)
                  [ order((abs(newtrt[3,]-trunc(target) ) )) [1:k]]
    }
   }  #end of j
   }   #end of i
##**************************************
cellmean<-apply(psudo, c(1,2), mean)
colmean=apply(psudo, 2, mean)
 # diagonal part gives the \hat\sigma_{1,i}^2
 # and off-diagonal part gives \hat\sigma_{1,i_1, i_2}
 sigXij<-apply(psudo, c(1, 2), var)
 # get a axN matrix with \hat\sigma_i^2(X_{ij}) =sigXij[i, j]
 MSTd=k*a*sum( (colmean-mean(psudo)  )^2       )  /(sum(n)-1)
 meanrk=apply(psudo,1, mean)
 MSTphi=k*sum( (cellmean-matrix(rep(meanrk,N),ncol=N)   )^2   )
         /((sum(n)-1)*a)
 MSTc= k*sum((cellmean-matrix(rep(meanrk,N),ncol=N) -
      matrix(rep(colmean,a), ncol=N,
      byrow=T)+mean(cellmean) )^2  )/((a-1)*(N-1))
 MSE= sum((psudo-array(rep(cellmean, k), c(a, sum(n), k)) )^2 )/
      (sum(n)*a*(k-1)  )
 Tsinter=(sqrt(sum(n)) * (MSTc-MSE))/sqrt(k)
 Tsc=(sqrt(sum(n)) * (MSTd-MSE))/sqrt(k)
 Tss=(sqrt(sum(n)) * (MSTphi-MSE))
##**************************************
```

```
 #Calculate estimate of variance for test statistics
 # countlan gives a matrix; first four columns are indices i1,i2,l1,l2
 #respectively fifth column gives the count of how many times (i1, l1)
 #value and (i2, l2) value are  used together in the same window among
 #all windows constructed. The sixth column gives the product of
 #\hat{\sigma} (X_{i1l1} \hat{\sigma}(X_{i2l2}
 countlan<-numeric()
 for( i1 in 1:a){
 for (i2 in 1:a){
 for (l1 in position.i1(i1, n)[1]: position.i1(i1, n)[2] ){
 for (l2 in position.i1(i2, n)[1]: position.i1(i2, n)[2] ){
 whereused.l1=apply((index[i1,1:N, ]==  mapindex(l1, n)[2]), 1, sum)
 # gives a vector of size N with values 1 or 0;
 # The jjjth element is 1 if l1 th covariate value is used in window
 #construction for the jjjth window
 whereused.l2=apply((index[i2,1:N, ]==  mapindex(l2, n)[2]), 1, sum)
 counti1i2l1l2=sum(whereused.l1*whereused.l2)
 # how many times (i1, l1),(i2,l2) are both used in the same windows
 # for all covariate values
 countlan=rbind(countlan, c(i1, i2, l1, l2, counti1i2l1l2,
         sigXij[i1, l1]*sigXij[i2, l2] ) )
       }}}}

sigbycount2lan=countlan[,5]^2 * countlan[,6]
xi.index=(countlan[,1]==countlan[,2])*(countlan[,3]!=countlan[,4])
eta.index=(countlan[,1]!=countlan[,2])
xi4=3*sum(sigbycount2lan*xi.index  )/(2*sum(n)*k*(k-1)^2)
eta4=3*sum(sigbycount2lan*eta.index)/(2*sum(n)*k^3)
 #####
 tauAsyc=(xi4+eta4)*4/(3*a^2)
 tauAsyinter=(xi4+eta4/((a-1)^2))*4/(3*a^2)
 pvalue.inter=1-pnorm(Tsinter/sqrt(tauAsyinter))
```

```
  pvalue.cov=1-pnorm(Tsc/sqrt(tauAsyc))
    list(Tsc=Tsc,Tsinter=Tsinter, pvalues=c(pvalue.cov, pvalue.inter),
         xi4=xi4,eta4=eta4)
        }
```

A.4 Classical F Test (CF Test)

The next code run the CF test.

```
classical.int.simp.cov=function(dat){
    try.ancova= lm(Y ~ X*factor(trt), data = dat)
    drop1.int=drop1(try.ancova,scope=~.,test="F")
    try.ancova1=lm(Y~factor(trt), data = dat)
    try.ancova2= lm(Y~X+factor(trt),data = dat)
    int.pvalue=anova(try.ancova,try.ancova2,test="F")[2,6]
    sim.pvalue=anova(try.ancova,try.ancova1,test="F")[2,6]
    cov.pvalue=anova(try.ancova2,try.ancova1,test="F")[2,6]
    c(int.pvalue,sim.pvalue,cov.pvalue)
 }
```

A.5 Drop Test

The following code do drop test for no covariate-treatment interaction, no simple covariate and no main covariate effect

```
source("mckean.r")
 all.droptest=function(dat){
              fit1=lm(Y~X*as.factor(trt), data=dat)
              my.mat= model.matrix(fit1)[,-1]
 # my.amat is the matrix for hypothesis.   H0: my.amat %x% beta=0,
 #where beta is the regression
 #parameter vector without intercept
 a=length(levels(as.factor(dat$trt)) )
```

```
 my.amat=matrix(0, a, ncol(my.mat))
 my.amat[,-(2:a)]=rbind(cbind(diag(a-1),-rep(1,a-1)),c(rep(0,a-1), 1));
 p.drop.simple=droptest(xmat=my.mat, y=dat$Y, amat=my.amat)$pval
 #no simple effect of covariate
 inter.amat=my.amat[-1,]
 p.drop.inter=droptest(xmat=my.mat, y=dat$Y, amat=inter.amat)$pval
 cov.amat=matrix(c(1,rep(0,a-1)), nrow=1)
 ## test of covariate effect when no interaction exists
 p.drop.cov=try(droptest(xmat=my.mat[,1:a],y=dat$Y,amat=cov.amat)$pval,T)
 result=c(p.drop.simple, p.drop.inter, p.drop.cov)
 names(result)=c("drop.simple", "drop.inter", "drop.cov")
 result
}
```

A.6 GAM Models (Spline and Loess)

```
#Method is spline
gam.spline=function(dat){
 library(gam)
 gam10=gam(Y ~ s(X) * factor(trt), data = dat)
 gam9=gam(Y ~ s(X) + factor(trt), data = dat)
 gam8=gam(Y ~factor(trt), data = dat)
 int.pvalue=anova(gam10,gam9,test="F")[2,6]
 cov.pvalue=anova(gam8,gam9,test="F")[2,6]
 sim.pvalue=anova(gam10,gam8,test="F")[2,6]
 c(int.pvalue,sim.pvalue,cov.pvalue)
}
#Method is loess
gam.loess=function(dat){
               library(gam)
               gaml10=gam(Y ~ lo(X) * factor(trt), data = dat)
               gaml9=gam(Y ~ lo(X) + factor(trt), data = dat)
```

```
            gaml8=gam(Y ~factor(trt), data = dat)
        int.pvalue=anova(gaml10,gaml9,test="F")[2,6]
        cov.pvalue=anova(gaml9,gaml8,test="F")[2,6]
        sim.pvalue=anova(gaml10,gaml8,test="F")[2,6]
        c(int.pvalue,sim.pvalue,cov.pvalue)
}
```

A.7 GAM Pspline

```
library(mgcv) gam.mgcv=function(data){
fit1=gam(Y~s(X)+ factor(trt), data=data,family=quasi)
fit2=gam(Y~factor(trt), data=data, family=quasi)
GAM.mgcv=anova(fit1, fit2, test= "F")[2,6]
GAM.mgcv }
```

A.8 Correlation Based Tests

```
cor.test(x, y, alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95)
```

A.9 ACE

```
library(acepack)
ACEtest=function(data){ acefit=ace(x=data$X,data$Y)
       fit1=lm(acefit$ty~ acefit$tx *trt, data=data  )
       fit2=lm(acefit$ty~trt, data=data  )
       ace.p=anova(fit1, fit2, test="F")[2,6]
       ace.p }
```

A.10 Wald and Deviance Tests

```
#Wald and Deviance Tests
my.glm=function(dat){
      glmfit=glm(Y~factor(trt)*X, data=dat,family=binomial)
      glmfit1=glm(Y~factor(trt)+ X, data=dat,family=binomial)
      glmfit2=glm(Y~factor(trt), data=dat,family=binomial)
      dev.test.int=anova(glmfit,glmfit1,test="Chisq")[2,5]
      dev.test.sim=anova(glmfit,glmfit2,test="Chisq")[2,5]
      dev.test.cov=anova(glmfit1,glmfit2,test="Chisq")[2,5]
      library(lmtest)
      galm.wald.int=waldtest(glmfit,glmfit1, test="Chisq")[2,4]
      galm.wald.sim=waldtest(glmfit,glmfit2, test="Chisq")[2,4]
      galm.wald.cov=waldtest(glmfit1,glmfit2, test="Chisq")[2,4]
     }
```

A.11 Comparing Computational Time

```
# compare running time for the whole test used
 result=numeric()
 ni.list=c( 50, 80, 100, 150, 200, 500)
 for (ni in ni.list){
 set.seed(1)
  theta=0
  n=rep(ni, 3)
  n1= n[1];  n2= n[2]; n3=n[3]
  x1=runif(n1);  x2=runif(n2) ;  x3=runif(n3)
  y1= rnorm(n1)
  y2= theta*(x2^2-x2+0.15)+ rnorm(n2)
  y3=  rnorm(n3)
  dat=data.frame(X=c(x1, x2, x3), Y=c(y1, y2, y3), trt=c(rep(1,
      n1), rep(2, n2), rep(3, n3) ))
```

```
  k=5
  thisresult=system.time(NPtest.new(dat, k) )
  lan2=  system.time(lan.NP2(dat, k) )
  reportT=c(thisresult, lan2)
  cat("ni=", ni, reportT, "\n", file="run.time.n.txt", append=T)
  result=rbind(result, reportT)
}
result
```

A.12 Code for Simulation in Chapter 3

```
source("mckean.r"); source("NPtest.new.r");
source("functions.new.r");
source("compare.test.indept.functions.r")
dat.EFT=read.table("EFT.data.txt", header=T)
dat2=dat.EFT
colnames(dat2)=c("Y", "X","trt")
ranges.EFT=tapply(dat2[-10,2],dat2[-10,3], range)
ranges.time.EFT=tapply(dat2[-10,1], dat2[-10,3], range)
n=12
trt=gl(2, n, labels=c("Row","Corner") )
library(gam)
for (tau in c(0.01, 0.03,0.04, 0.09)){
 perc=0.1
 H0result=numeric()
 for (i in 1:2000){
 corner.EFT=runif(n, ranges.EFT$Corner.group[1],
 ranges.EFT$Corner.group[2])
corner.time.EFT= tau*(corner.EFT-67.5)^2+runif(n, -5,15)
row.EFT=ifelse(runif(n)<perc, rbeta(n,1.2,
3)*(ranges.EFT$Row.group[2]-ranges.EFT$Row.group[1]), rlnorm(n, 1.2,
2)*(145-135) )
```

```
row.time.EFT=ifelse(runif(n)<0.6,rbeta(n,1.2,
3)*(ranges.time.EFT$Row.group[2]-ranges.time.EFT$Row.group[1]),
rlnorm(n, 1.2, 2)*(745-735) )
thistest=alltests( data.frame(X=c(row.EFT, corner.EFT),
Y=c(row.time.EFT, corner.time.EFT ), trt) )
H0result=rbind(H0result,thistest) cat(thistest, "\n",
file=paste("no.Sin.tau", tau, ".power.txt",sep=""), append=T) }
levels005=apply(H0result, 2, function(x) mean(x<=0.05,  na.rm=T) )
levels001=apply(H0result, 2, function(x) mean(x<=0.01,  na.rm=T) )
cat(tau, levels005, "\n", file="no.Sin.EFT.power005.desktop.txt",
append=T)
cat(tau, levels001, "\n",
file="no.Sin.EFT.power001.desktop.txt", append=T) }
# rerun for gam becuase previous runs did not include interaction
#effect for gam
for (tau in c( 0.07, 0.08, 0.09)){
 perc=0.1
 H0result=numeric() for (i in 1:2000){
 corner.EFT=runif(n,ranges.EFT$Corner.group[1],
           ranges.EFT$Corner.group[2])
corner.time.EFT= tau*(corner.EFT-67.5)^2+runif(n, -5,15)
row.EFT=ifelse(runif(n)<perc, rbeta(n,1.2,
3)*(ranges.EFT$Row.group[2]-ranges.EFT$Row.group[1]), rlnorm(n, 1.2,
2)*(145-135) ) row.time.EFT=ifelse(runif(n)<0.6,rbeta(n,1.2,
3)*(ranges.time.EFT$Row.group[2]-ranges.time.EFT$Row.group[1]),
rlnorm(n, 1.2, 2)*(745-735) ) thistest=GAM Loess.sp(
data.frame(X=c(row.EFT, corner.EFT), Y=c(row.time.EFT,
corner.time.EFT ), trt) )
 H0result=rbind(H0result,thistest) cat(thistest, "\n",
 file=paste("gam.no.Sin.tau", tau,".power.txt",sep=""), append=T) }
levels005=apply(H0result, 2, function(x) mean(x<=0.05,  na.rm=T) )
levels001=apply(H0result, 2, function(x) mean(x<=0.01,  na.rm=T) )
```

```
cat(tau, levels005, "\n",
file="gam.no.Sin.EFT.power005.desktop.txt", append=T)
cat(tau, levels001, "\n",
file="gam.no.Sin.EFT.power001.desktop.txt", append=T) }
# actual run is on desktop without gam and drop test
tau=2.5
n=20
trt=gl(2, n, labels=c("Row","Corner") )
perc=0.1
H0result=numeric()
 for (i in 1:2000){
 corner.EFT=runif(n,ranges.EFT$Corner.group[1],
            ranges.EFT$Corner.group[2])
corner.time.EFT= tau*(corner.EFT-67.5)^2+runif(n, -5,15)
row.EFT=ifelse(runif(n)<perc, rbeta(n,1.2,
3)*(ranges.EFT$Row.group[2]-ranges.EFT$Row.group[1]), rlnorm(n, 1.2,
2)*(145-135) )
row.time.EFT=ifelse(runif(n)<0.6,rbeta(n,1.2,
3)*(ranges.time.EFT$Row.group[2]-ranges.time.EFT$Row.group[1]),
rlnorm(n, 1.2, 2)*(745-735) )
thistest=alltests( data.frame(X=c(row.EFT, corner.EFT),
Y=c(row.time.EFT, corner.time.EFT ), trt) )
H0result=rbind(H0result,thistest) cat(thistest, "\n",
file=paste("no.Sin.tau", tau, ".n.power.txt",sep=""), append=T) }
levels005=apply(H0result, 2, function(x) mean(x<=0.05,  na.rm=T) )
levels001=apply(H0result, 2, function(x) mean(x<=0.01,  na.rm=T) )
cat(tau, levels005, "\n", file="no.Sin.EFT.power005.n.txt", append=T)
cat(tau, levels001, "\n",
file="no.Sin.EFT.power001.n.txt", append=T)
```

A.13 Data Generation in Chapter 4

```
#To estimate type I error when the association is linear.
####################
f11=function(x1,tau) {tau*x1};
f22=function(x2,tau) {tau*x2}
gen.dat=function(tau, n ){
x1=runif(n); x2=runif(n)
f1=f11(x1,tau); f2=f22(x2,tau)
y1= f1 + 0.1* rnorm(n)
y2= f2 + 0.1* rnorm(n)
dat=data.frame(X=c(x1, x2), Y=c(y1, y2 ), trt=c(rep(1, n), rep(2, n))  )
 #put test here
}
################
 for (n.values in c(30,50)){
 for (tau.range in c(0,0.1,0.2,0.3) ) {
 res1=numeric()
 set.seed(400)
 repli=500
 for (b1 in 1:repli) {
 res1=rbind(res1, gen.dat(tau=tau.range,n=n.values ))
} }
}

#Data generation to estimate power for the simulation when
#association is linear f22=function(x2,tau) {tau*x2}
gen.dat=function( n ,k,tau){
  x1=runif(n); x2=runif(n)
  f1=f11(x1,tau); f2=f22(x2,tau)
  y1= 0.1* rnorm(n)
  y2= f2 + 0.1* rnorm(n)
  dat=data.frame(X=c(x1, x2),Y=c(y1,y2),trt=c(rep(1, n),rep(2,n)) )
```

```
  #put test here
}

#Data generation to estimate type I for the simulation when
#association is quadratic
f11=function(x1,tau) {tau*(x1^2 - x1 + 0.15}
f22=function(x2,tau) {tau*(x2^2 - x2 + 0.15}
gen.dat=function( n ,k,tau){
  x1=runif(n); x2=runif(n)
  f1=f11(x1,tau); f2=f22(x2,tau)
  y1= f1 + 0.1* rnorm(n)
  y2= f2 + 0.1* rnorm(n)
  dat=data.frame(X=c(x1, x2),Y=c(y1,y2),trt=c(rep(1, n),rep(2, n)))
  #put test here
}

#Data generation to estimate power for the simulation when
#association is quadratic
f22=function(x2,tau) {tau*(x2^2 - x2 + 0.15}
gen.dat=function( n ,k,tau){
  x1=runif(n); x2=runif(n)
  f1=f11(x1,tau); f2=f22(x2,tau)
  y1= 0.1* rnorm(n)
  y2= f2 + 0.1* rnorm(n)
  dat=data.frame(X=c(x1, x2),Y=c(y1,y2),trt=c(rep(1,n),rep(2,n)))
  #put test here
}

# Data generation to estimate Type I error for binomial data
f11=function(x1,tau) {tau*cos(2*pi*x1)}
f22=function(x2,tau){tau*cos(2*pi*x2)}
f33=function(x3,tau){tau*cos(2*pi*x3)}
glm.alpha=function(n,k,tau){
 x1=runif(n);    x2=runif(n); x3=runif(n)
```

```
 f1=f11(x1,tau); f2=f22(x2,tau); f3=f33(x3,tau)
 y1=rbinom(n,1,  exp(f1 )/(1+exp(f1))  )
 y2=rbinom(n,1, exp(f2 )/(1+exp(f2))  )
 y3=rbinom(n,1, exp(f3 )/(1+exp(f3))  )
 X=c(x1, x2, x3)
 Y=c(y1, y2,y3)
 trt=c(rep(1, n), rep(2, n),rep(3, n)
 dat=data.frame(X, Y, trt )
 # Put test here
  }

# Data generation to estimate power for binomial data
f11=function(x1,tau) {tau*cos(2*pi*x1)}
glm.power=function(n,k,tau){
 x1=runif(n);    x2=runif(n); x3=runif(n)
 f1=f11(x1,tau); f2=f22(x2,tau); f3=f33(x3,tau)
 y1=rbinom(n,1,  exp(f1 )/(1+exp(f1))  )
 y2=rbinom(n,1, 0.5 )
 y3=rbinom(n,1, 0.5 )
 X=c(x1, x2, x3)
 Y=c(y1, y2, y3)
 trt=c(rep(1, n), rep(2, n),rep(3, n)
 dat=data.frame(X, Y, trt )
 # Put test here
  }
```

A.14 Data Generation for Simulation Study in Chapter 5

```
y7f=function(x,theta,b,tau){
 multiplyterm=abs(x)
 myscale=5*abs((x-0.5))
```

```
 res=tan(theta)*x *(ifelse((x> 0) & (x< b),1,0))  +
 b* tan(theta)* ifelse((x> b) ,1,0) -
 theta*10/tau*sqrt(multiplyterm)*rweibull(length(x),shape=2,scale=myscale)
 res
}
gen.dat=function(n,prop=0.5,tau){ dat=numeric(); k=1
#for (theta in (-5:4)/20*pi){ #for power
 for (theta in rep(pi/4,10)){
 for (b in rep(0.5,2)){
#for (b in c(0.5, 0.7)){ for power
 n1=round((1-prop)/2*n )
 n2=round(prop*n)
 n3=n-n1-n2
 x=c(runif(n1, min=-0.5, max=0), runif(n2, min=0, max=b),
 runif(n3, min=b, max=1))
 y=y7f(x, theta, b,tau)
 dat=data.frame(rbind(dat, cbind(X=x, Y=y, trt=rep(k, length(x)))) )
 k=k+1
 }
 }
#run tests here }
#for (tau.range in c(0.0625,0.125,0.25,0.5)){ for
#(prop.range in #c(0.10,0.26,0.33,0.50,0.76, 0.90) ){
#res1=numeric() #set.seed(400) repli= 1000 for (b1 in 1:repli){
#res1=rbind(res1,gen.dat(n=20,prop=prop.range,tau=tau.range)) } } }
```

A.15 Code for examples in Chapter 4

```
## for exponential distribution
## produce scatterplot with regression curves
b1=2; b2=10; a=2
x=runif(100, min=0, max=10)
```

```
yf=function(x,b){
rexp(100,rate=1/((a*(x-5)^2+b)) )}
y1=yf(x,b1)
y2=yf(x,b2)
matplot(x=x,y=cbind(y1,y2), pch=c(19, 19), col=c("blue", "orange"),
font=2, font.axis=2, font.lab=2, main="Parallel Quadratic Regression
Curves ", ylab="Response", xlab="Covariate")
regression.curve=function(x, b)  (a*(x-5)^2+b) lines(sort(x),
regression.curve(sort(x), b1), col="blue", lwd=2) lines(sort(x),
regression.curve(sort(x), b2), col="orange", lwd=2)}

# plot the difference of two conditional cdf funtions and see if
#it depends on x.
cdf.exp=function(y, b, x){
 pexp(y,rate=1/((a*(x-5)^2+b)) )}
 F1.minusF2.YgivenX= function(x, y) {
 cdf.exp(y, b1, x)  - cdf.exp(y, b2, x) }
 temp=range(y1, y2)
 y=seq(100)/100* (temp[2]-temp[1])
 x0=sort(x)
 par(mar=c(2,1,3,1))
 persp(x0, y, z=outer(x0, y, F1.minusF2.YgivenX ), theta=30, phi=0,
 xlab="x", zla="F1(y|x) - F2(y|x)",col="light blue" ,main=NA)
 ep1=expression(F[1]("y|x") - F[2]("y|x")) ep2= "Exponential
 Distribution" mtext(text=ep2, side=3, line=1, font=2, cex=1.2 )
 mtext(text=ep1, side=3, line=-0.5 , font=2, cex=1.2)

## For Normal distribution
b3=2;b4=6;a1=0.5
ynorm=function(x,b){
rnorm(100,mean=4*sin(a1*pi*x)+ b,sd=1)}
y3=ynorm(x,b3)
```

```
y4=ynorm(x,b4)
matplot(x=x,y=cbind(y3,y4), pch=c(19, 19),
col=c("blue", "orange"), font=2, font.axis=2, font.lab=2,
main="Parallel Sinusoidal Regression Curves", ylab="Response",
xlab="Covariate")} regression.curve.norm=function(x, b)  {
 4*sin(a1*pi*x)+ b
 lines(sort(x), regression.curve.norm(sort(x), b3), col="blue",
 lwd=2) lines(sort(x), regression.curve.norm(sort(x), b4),
 col="orange", lwd=2)


## plot the difference of two conditional cdf funtions and see if it
#depends on x.
cdf.norm=function(y, b, x) {
pnorm(y,mean=4*sin(a1*pi*x)+ b )}
norm.F1.minusF2.YgivenX= function(x, y){
cdf.norm(y, b3, x) - cdf.norm(y, b4, x) }
temp34=range(y3, y4)
y=seq(100)/100* (temp34[2]-temp34[1])
x0=sort(x) par(mar=c(2,1,3,1))
persp(x0, y, z=outer(x0, y,
norm.F1.minusF2.YgivenX ), theta=20, phi=0, xlab="x", zlab="F1(y|x)
- F2(y|x)",col="light green",main=NA)

ep1=expression(F[1]("y|x") - F[2]("y|x"))
 ep2= "Normal Distribution" mtext(text=ep2, side=3, line=1,
  font=2, cex=1.2 ) mtext(text=ep1,
side=3, line=-0.5 , font=2, cex=1.2)
```