# Some Nonparametric Methods for Clinical Trials and High Dimensional Data

**Xiaoru Wu**

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

**COLUMBIA UNIVERSITY**

2011

# ABSTRACT

# Some Nonparametric Methods for Clinical Trials and High Dimensional Data


## Xiaoru Wu



This dissertation addresses two problems from novel perspectives. In chapter 2, I propose an empirical likelihood based method to nonparametrically adjust for baseline covariates in randomized clinical trials and in chapter 3, I develop a survival analysis framework for multivariate $K$-sample problems.

(I): Covariate adjustment is an important tool in the analysis of randomized clinical trials and observational studies. It can be used to increase efficiency and thus power, and to reduce possible bias. While most statistical tests in randomized clinical trials are nonparametric in nature, approaches for covariate adjustment typically rely on specific regression models, such as the linear model for a continuous outcome, the logistic regression model for a dichotomous outcome, and the Cox model for survival time. Several recent efforts have focused on model-free covariate adjustment. This thesis makes use of the empirical likelihood method and proposes a nonparametric approach to covariate adjustment. A major advantage of the new approach is that it automatically utilizes covariate information in an optimal way without fitting a nonparametric regression. The usual asymptotic properties, including the Wilks-type result of convergence to a $\chi^2$ distribution for the empirical likelihood ratio based test, and asymptotic normality for the cor-

responding maximum empirical likelihood estimator, are established. It is also shown that the resulting test is asymptotically most powerful and that the estimator for the treatment effect achieves the semiparametric efficiency bound. The new method is applied to the Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO)-I trial. Extensive simulations are conducted, validating the theoretical findings. This work is not only useful for nonparametric covariate adjustment but also has theoretical value. It broadens the scope of the traditional empirical likelihood inference by allowing the number of constraints to grow with the sample size.

(II): Motivated by applications in high-dimensional settings, I propose a novel approach to testing equality of two or more populations by constructing a class of intensity centered score processes. The resulting tests are analogous in spirit to the well-known class of weighted log-rank statistics that is widely used in survival analysis. The test statistics are nonparametric, computationally simple and applicable to high-dimensional data. We establish the usual large sample properties by showing that the underlying log-rank score process converges weakly to a Gaussian random field with zero mean under the null hypothesis, and with a drift under the contiguous alternatives. For the Kolmogorov-Smirnov-type and the Cramér-von Mises-type statistics, we also establish the consistency result for any fixed alternative. As a practical means to obtain approximate cutoff points for the test statistics, a simulation based resampling method is proposed, with theoretical justification given by establishing weak convergence for the randomly weighted log-rank score process. The new approach is applied to a study of brain activation measured by functional magnetic resonance imaging when performing two linguistic tasks and also to a prostate cancer DNA microarray data set.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

Foremost, I owe my deepest gratitude to my advisor, Professor Zhiliang Ying, whose incredible breadth and depth of knowledge have been an inspiration to me and of great importance in shaping the research contained in this thesis. My graduate studies would not have been so rewarding without the interactions I have had with Zhiliang. Zhiliang provides insight into my research through many insightful questions (some of which I still haven't fully understood) and he is readily available to answer all of my questions as well. I have learned a lot from attending the reading groups he organized every week, most of which happened during weekends. Through the enlightening discussions with him, I have not only accumulated knowledge of statistics but also deepened my understanding of doing research. I thank him for having always been patient, supportive and approachable.

I have also had the honor of working with Professor Ian McKeague. I first got to know Ian in 2007 when I took his course "Asymptotic Statistics", from which I had my first lecture in empirical likelihood. At that time, my classmate Xiaodong Li and I wanted to develop an empirical likelihood method to test for conditional independence. Ian actively provided us with references for ideas and encouraged us when we met all kinds of difficulties. Since then, Ian has always been there for me when I have questions. About a year ago, Ian and I began to collaborate on projects. During my many visits to the Biostatistics Department and numerous emails, Ian provides me with invaluable guidance.

In addition to my interactions with Zhiliang and Ian, the other collaborators also play a pivotal role in my graduate studies. Professor Tian Zheng, who was my mentor during the first year at Columbia, in addition to offering numerous thoughtful suggestions and timely help, has always been a good friend in life. At the National Cancer Institute, I had many thought-provoking discussions with Dr. Jing Qin, Dr. Kai Yu and Dr. Nilanjan Chatterjee, with whom I look forward to continuing the collaboration. I would also like to thank Dr. Lu Cui at Eisai Medical Research and Martin Carlsson and Cheryl Feiner at Pfizer, who enabled me to get some hands-on experience in clinical trials and understand the motivation of biostatistics research better. I must highlight Wen Yu, who was a visiting student in our department and now is an assistant professor at Fudan University, for carefully listening to my presentations each time I made little progress on this work.

I am very grateful to Professor Michael Sobel, Professor Ian Mckeague, Dr. Lu Cui, Professor Tian Zheng and Professor Zhiliang Ying for serving on my defense committee. I am particularly indebted to Professor Michael Sobel for his extensive editing of papers that comprise the main part of this dissertation.

I am very thankful to all the faculty and students in the Statistics Department at Columbia University. The coursework at Columbia has been key to the development and accumulation of my knowledge of statistics in the past five years. And all the good times with my fellow students, numerous parties, retreats, barbecues, post-qual trip to Puerto Rico, graduation trip to Miami and so on, greatly enriched my life at Columbia and will be my warmest memory forever.

On a more personal note, I would especially like to thank my parents for their never-ending support for my pursuits. Since I left home nine years ago to start my undergraduate studies in Beijing , the time I spend with them has been less

and less. However, they are always standing by my side although the geographical distance between us increases from 700 miles to 7000 miles. I also must thank all of my friends, especially Xiaohui Wang and Xingshi Wang, for their encouragement, patience, and distractions through the years. Finally, this endeavor would have been much more challenging without the love and support of Jerry Fu.

# Chapter 1

# Introduction

Formally suggested by R.A. Fisher in the 1920s (Box 1980), randomization was first used in medical studies by Bradford Hill and Richard Doll in Great Britain in the 1940s. In the United States randomization was advocated by early trialists such as Tom Chalmers and Paul Meier (Piantadosi 2005). Nowadays, randomization is widely adopted in the design of clinical trials and other experiments for comparing treatments (Hill 1960; Byar et al. 1976). Without randomization, the possibility that the allocation of patients to treatments is consciously or unconsciously based on patients' prognostic factors can not be precluded (Efron 1971). Because selection bias can influence outcomes as strongly as many treatment effects, treatments yielding differences of a clinically important size could then be due only to the bias in selection. Thus, randomization prevents confounding of the effects of the therapy with the prognostic factors. A more far-reaching benefit of randomization is that it provides a valid basis for testing the null hypothesis of no treatment effect without any assumption of a population model and without the necessity of measurements on all the possibly important covariates, which gives randomized studies their high degree of inferential directness and reliability

(Fisher 1966; Kempthorne 1977; Lehmann 1975).

The primary objective of many randomized clinical trials is to compare the difference in mean outcome among two or more treatments. In addition to the primary outcome and treatment assignment, substantial amounts of baseline data on each subject prior to randomization are routinely collected. These data may include patient demographic characteristics, previous medical history, lifestyle measurements, current medical condition, baseline measure of the outcome and other assessments. Some of these baseline covariates may be related to the primary outcome and may exhibit chance imbalances between the two treatment groups.

There exists a vast literature on whether or not and how to adjust the analysis of treatment difference for the effects of covariates. Two schools of thought exist on the role of covariate adjustment. One, looking at the treatment comparison conditional on the allocation of covariates, regards covariate adjustment as required to characterize the potential benefit an individual might accrue from treatment, e.g. Hauck et al. (1998) "recommend that the primary analysis adjust for important prognostic covariates in order to come as close as possible to the clinically most relevant subject-specific measure of treatment effect". The other, a long-standing idea originated with R.A.Fisher (1932), is concerned with the overall treatment difference and sees the role of covariate adjustment as a means of increasing precision and thereby statistical power. There is no doubt that both conditional and unconditional (on covariates) treatment effects are of considerable and complementary importance in developing a comprehensive understanding of how treatments compare. Inference on the former can reveal interactions between treatment and patient characteristics, which may have critical implications for use of the treatment in certain subpopulations. The latter provides a measure of overall effect useful for broad policy recommendations, which explains its role as primary focus

of regulatory authorities.

With continuous outcome, the above debate rarely receives explicit mention because (a) in randomized clinical trials the independence between the treatment assignment and baseline covariates leads to the coincidence of the unconditional and conditional treatment differences regardless of whether or not the regression model between the outcome and covariates is correctly represented (b) the adjusted analysis is generally more precise than the competing methods which do not take the covariates into account. Nevertheless, the conventional wisdom that such unbiasedness and efficiency gains will also be achieved with respect to regression models other than classical linear regression does not apply to all the situations. On the one hand, Gail (1984) shows that the regression of the response variable on treatment and covariates being linear or exponential is the sufficient and necessary condition for the coincidence of the conditional and unconditional treatment effects. On the other hand, unlike the classical linear regression case, covariate adjustment in nonlinear regression models does not necessarily guarantee the efficiency improvement, e.g. Robinson and Jewell (1991) prove that adjustment for covariates always leads to a loss (or at best no gain) of precision in logistic regression models (Mantel and Haenszel 1959; Mantel 1989). A similar point is also made by Breslow and Day (1987).

The discrepancy between the adjusted and unadjusted analyses has inspired considerable controversy among numerous researchers (Pocock et al. 2002; Assmann et al. 2000; Raab et al. 2000; Senn 2000) and regulatory authorities (Lewis 1999; Grouin et al. 2004). Since covariate adjustment may involve a *post hoc* selection of baseline covariates and different choices of covariates can lead to different treatment effects, analysts are often tempted to find "the covariate model that best accentuates the estimate and/or statistical significance of the treatment

difference" (Pocock et al. 2002). Thus, trialists and regulatory agencies are reluctant to endorse adjusted analyses, and current guidelines assert strongly that, if adjustment is undertaken, only a few such covariates should be used, chosen on the basis of prior knowledge or their prognostic value; these should be prespecified in the protocol or analysis plan, as should be the form of the model relating covariates to outcome to be used for adjustment(Lewis 1999; Grouin et al. 2004). However, associations between covariates and outcome may not be appreciated at the design stage (Pocock et al. 2002), particularly if such information was not collected systematically in previous studies, but may be evident only at the analysis stage, subsequent to unblinding. An unfortunate consequence of these recommendations is that a critical opportunity to enhance the efficiency and reveal important real effects may be lost.

To this end, it is desirable to find approaches that make best use of the baseline data while supporting objective incorporation of covariate effects. In other words, approaches that can obviate subjective modeling of the covariate-outcome relationships while simultaneously exploiting these relationships to improve the precision of treatment effect inference are needed. In the spirit above, an early development that has drawn a great deal of interest was proposed by Koch et al. (1998). Koch's sampling-based method corrects for random imbalances for various types of responses and designs and does not require regression modeling of covariates effect. The resulting algorithm always reduces the variance and is computationally straightforward through the application of weighted least squares. Nonetheless, a general strategy for achieving the goal was not available until Tsiatis et al. (2008). Based on some previous work in the literature of missing data problems by Leon et al. (2003) and Davidian et al. (2005), Tsiatis et al. show that all unbiased estimators for the overall difference in mean outcome between two groups can

be obtained through augmenting the difference in sample means by an auxiliary term. This characterization summarizes many familiar estimators, including ANCOVA and Koch's estimator, in a general framework, which facilitates comparison among existing methods and provides insights into the precision improvement by incorporating covariates. The general representation also allows us to identify the most efficient estimator(which outperforms Koch's estimator) and thus derive the semiparametric efficiency bound.

In Chapter 2, we overview existing model-free methods and a more general approach by Zhang et al. (2008), based upon which we have developed an empirical likelihood based method for nonparametric covariate adjustment. Emerging elegantly from empirical likelihood theory developed by Owen (1988, 1990) and Qin and Lawless (1994) is the proposed adjustment method that not only supports the objective incorporation of baseline information but also reduces the computational complexity caused by fitting a nonparametric regression, which is unavoidable in the implementation of Tsiatis et al.(2008) and Zhang et al. (2008). More ideally, the asymptotic variance of the maximum empirical likelihood estimator will be monotone decreasing with more moment constraints and the semiparametric efficiency bound can be achieved if the number of constraints grows to infinity. This result is not only useful for nonparametric covariate adjustment but also broadens the scope of the traditional empirical likelihood method, where the inferential efficiency of the parameter of interest is well studied only under finitely many constraints.

Addressing a different problem, Chapter 3 is also devoted to developing nonparametric methods. In statistics, it is of fundamental interest to test whether two data sets come from the same underlying distribution. In one dimension, one would have little hesitation in using a rank based method, such as Kolmogorov-Smirnov

and Cramér-von Mises tests. A main ingredient in those tests has been the ranks of the observations in the pooled sample. Since a monotone transformation can make univariate data follow a uniform distribution on the unit interval while the ranks remain invariant, the corresponding test statistics are distribution free under the null hypothesis. However, there is no natural order on the multidimensional space and therefore it is difficult to define a parallelism for "rank". There exists a rich literature extending the notion of ranking to multivariate cases, based on the concepts of "data depth". Early contributions include those of Tukey (1975), Liu (1988,1990), Donoho and Gasko (1992), Dümbgen (1992) and Liu and Singh (1992,1993). Nevertheless, to our knowledge, the multivariate two-sample problem has never been connected with survival analysis. In Chapter 3, by converting the original observations in the multivariate space into survival times, we construct a class of intensity censored score processes under varied censoring schemes, which give us power to detect the localized differences in nonlinear problems. In fact, the proposed approach can easily be extended to functional spaces as long as a reasonable distance measure can be introduced. When the sample size involved is small, a permutation based test can be used to obtain the cutoff values. While the sample size is large, we can employ a random weighting scheme which greatly reduces the computational intensity. We close this dissertation by summarizing Chapter 2 and 3 and discussing some future directions in Chapter 4.

# Chapter 2

# Nonparametric covariate adjustment

## 2.1 Introduction

Testing for the statistical significance of treatment differences is a key element in the analysis of randomized clinical trials. In its simplest form, patients are randomly allocated to either a treatment or control group and their responses are recorded. Many statistical methods are available for testing whether there is convincing evidence that a treatment difference exists between the two groups; cf. Pocock (1983) and Friedman, Furberg and DeMets (1998). In addition to treatment allocation and outcome values, baseline covariate information is often collected in such clinical studies. Classical analysis of covariance (ANCOVA) and other regression model-based tests may be used to handle covariate adjustment; cf. Scheffe (1959), Simon (1984), McCullagh and Nelder (1989) and Rutter and Elashoff (1994). When properly used, covariate adjustment can increase efficiency and, in the case of an observational study, reduce bias (Armitage 1981).

Due to randomization, most two-sample (multi-sample if more than two treatment groups are involved) tests are valid without any parametric assumption. Therefore, these tests are nonparametric in nature, a feature of great importance in a clinical trial. Standard methods for covariate adjustment, however, require that a specific regression model be assumed; see, for example, Piantadosi (2005, Chapter 17).

Adjusting for covariates without assuming a regression model has been studied by Koch (1998), Tsiatis, Davidian, Zhang and Lu (2008) among others. In particular, Koch (1998) proposed a weighted least squares method to include covariate information for estimating the treatment difference. This method always leads to a variance reduction, thus an increase in power. By appealing to semiparametric efficiency theory, Tsiatis et al. (2008) developed a general approach to covariate adjustment that circumvents modeling the covariate-outcome relationship. Their approach allows for nonlinear terms in relating the auxiliary covariates to the outcome variable, thereby further reducing the variability. They showed that the method is semiparametrically efficient by deriving the semiparametric information bound and by showing the bound is attained with their approach.

An essential ingredient in the approach by Tsiatis et al. (2008) is the use of the independence of treatment assignment and baseline covariates to construct estimating equations. These equations can be viewed as constraints that, when properly utilized, may lead to further reduction in variability of the outcome variable. How to optimally use these constraints is therefore crucial for efficiency improvement.

Empirical likelihood (Owen 1988) is a general method for efficiently utilizing constraints or estimating equations. Specifically, it maximizes the nonparametric likelihood (Kiefer and Wolfowitz 1956) subject to certain constraints that are

specific to the problem of interest. It can be used to obtain empirical likelihood ratio tests as well as confidence intervals. Examples include testing and interval estimation for population means and for regression coefficients. Qin and Lawless (1994) showed that the constraints can be used more liberally in the sense that the number of constraints may exceed the number of parameters of interest. They also showed that the empirical likelihood utilizes the information in the constraints in an optimal way.

Because baseline covariate information for a randomized clinical trial generates constraints, it is natural to consider the empirical likelihood as a means to improve efficiency for the primary problem of testing and estimating treatment difference. To that end, this chapter proposes a general approach to covariate adjustment by making use of the empirical likelihood and suitably choosing constraints. The new approach does not require any model assumption on the relationship between the outcome variable and baseline covariates. It is shown that such an empirical likelihood based method automatically results in efficiency improvement. For testing, it is asymptotically most powerful; for estimation, it achieves the semiparametric information bound.

The rest of the chapter is organized as follows. In Section 2.2 we introduce some notation. Section 2.3 and 2.4 briefly discuss existing model-based and model-free methods, respectively. We apply the empirical likelihood method for covariate adjustment and extend it to inference with an increasing number of constraints in Section 2.5. The design and results of simulation studies are described in Section 2.6. In Section 2.7, the method is applied to a study of acute myocardial infarction.

## 2.2 Notation and model specification

In a $(K+1)$-arm $(K \geq 1)$ randomized clinical trial, for subject $i$, let $Y_i$, $Z_i$ and $\boldsymbol{X}_i$ denote the outcome, treatment allocation and available auxiliary baseline covariates, respectively. Assume that $(Y_i, Z_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) and that the random allocation probabilities $\pi_k = P(Z = k)$, $k = 0, \ldots, K$, where $\sum_{k=0}^{K} \pi_k = 1$, are known.

Throughout, $G^k$ denotes the conditional distribution of the outcome variable $Y$ given treatment allocation $Z = k$, $k = 0, \ldots, K$. Then the usual null hypothesis of no treatment difference is given by

$$H_0 : G^0 = G^1 = \ldots = G^K.$$

Note that there is no assumption on the form of $\{G^k, k = 0, \ldots, K\}$.

To study treatment effects, one may choose certain contrasts among the treatment groups in terms of their population characteristics, for example, the difference in mean outcomes between two treatment groups. Following Zhang et al. (2008), the treatment effect can be identified by considering

$$\beta_1 = E(Y|Z = 0), \quad \beta_2 = E(Y|Z = 1) - E(Y|Z = 0), \tag{2.1}$$

or equivalently, by formulating

$$E(Y|Z) = \beta_1 + \beta_2 Z. \tag{2.2a}$$

Clearly, such an approach does not require model assumption on the underlying distribution functions $G^k$, $k = 0, \ldots, K$. If there are more than two treatment

groups, equation (2.2a) becomes

$$E(Y|Z) = \beta_1 + \beta_2 1_{(Z=1)} + \ldots + \beta_{K+1} 1_{(Z=K)}, \tag{2.2b}$$

where $1_{(.)}$ is the indicator function and $\beta_{k+1}$ represents the difference in mean outcome between group $k$ and group 0. For a binary outcome, an alternative formulation is via the log-odds ratios:

$$\text{logit}\{P(Y=1|Z)\} = \log\left\{\frac{P(Y=1|Z)}{P(Y=0|Z)}\right\} = \beta_1 + \beta_2 1_{(Z=1)} + \ldots + \beta_{K+1} 1_{(Z=K)}. \tag{2.3}$$

Under this formulation, testing the null hypothesis of no treatment difference is tantamount to testing $H_0 : \beta_2 = \ldots = \beta_{K+1} = 0$, and estimating the treatment effect is tantamount to estimating values of the $\beta_k, k = 2, \ldots, K+1$. For notational convenience, we use $\boldsymbol{\beta}$ to denote the parameter vector $(\beta_1, \ldots, \beta_{K+1})^T$.

Besides the outcome variable and treatment assignment, relevant baseline covariates, which may comprise patients' demographic information, medical history, lifestyle measurements, etc., may be recorded as well. Their association with and impact on the outcome variable can then be explored for efficiency gains in testing and estimation of treatment effects.

## 2.3 Model-based covariate adjustment

A common approach to making use of the covariate information is to postulate a certain regression model, which gives treatment comparisons conditional on values of the covariates. Without loss of generality, we first assume the randomized clinical trial is two-armed, i.e., $Z = 1$ or 0 with probabilities $\pi_1$ or $1 - \pi_1$ and there is a

single covariate $X$, which is related to the outcome $Y$. Randomization guarantees statistical independence of Z and X.

### 2.3.1 Covariate inclusion in the linear model

If the outcome $Y$ is continuous, consider the following two models:

$$\text{Model 1} \quad Y = \beta_1 + \beta_2 Z + \varepsilon_1, \quad \mathrm{E}(\varepsilon_1) = 0, \quad \mathrm{Var}(\varepsilon_1) = \sigma_1^2 \tag{2.4}$$

$$\text{Model 2} \quad Y = \beta_1 + \beta_2^* Z + \gamma X + \varepsilon_2, \quad \mathrm{E}(\varepsilon_2) = 0, \quad \mathrm{Var}(\varepsilon_2) = \sigma_2^2. \tag{2.5}$$

Model 2, usually called the "analysis of covariance" (ANCOVA) model, incorporates the covariate information by adding a linear term of $X$ in the linear regression model 1. Because of the underlying linearity of the problem and the independence of $Z$ and $X$, the magnitude of the treatment effect in model 1, $\beta_2$, is exactly the same as that in model 2, $\beta_2^*$, which enables us to infer the unconditional treatment effect using the conditional model. Besides, we have the following formulae of variances:

$$\begin{aligned} \text{Model 1} \quad \mathrm{Var}(\hat{\beta}_2) &= \sigma_1^2 / \sum_{i=1}^{n} (z_i - \bar{z})^2 \\ \text{Model 2} \quad \mathrm{Var}(\hat{\beta}_2^*) &= \{\sigma_2^2 / \sum_{i=1}^{n} (z_i - \bar{z})^2\} / [1 - \{\mathrm{corr}(z, x)\}^2] \end{aligned}$$

where $\bar{z} = n^{-1} \sum_{i=1}^{n} z_i$ and $\mathrm{corr}(z, x)$ denotes the sample correlation coefficient between the treatment assignment $Z$ and the covariate $X$. The sample correlation $\mathrm{corr}(z, x)$ should be close to zero because of randomization. Since the covariate $X$ is correlated with the outcome $Y$, there will be a reduction of residual variance, i.e. $\sigma_2^2$ will be less than $\sigma_1^2$. Consequently, the ANCOVA model improves precision

in the linear regression case.

### 2.3.2 Covariate inclusion in the nonlinear model

The success of the ANCOVA model in preserving the unconditional treatment differences as well as improving the precision gave rise to the popularity of regression based covariate adjustments. However, conditional and unconditional treatment effects are not always equal. In fact, there is a vast literature that examines in what situations the magnitude of the treatment effect changes with the inclusion of covariates. Even in randomized clinical trials with perfectly balanced covariates, the conditional and unconditional effects may be unequal. This can be seen from a simple hypothetical example. Suppose in a randomized clinical trial, half of the patients are randomized to treatment A ($Z = 1$) and the other half to treatment B ($Z = 0$). The outcome of interest $Y$ is the survival status ($Y = 0$ for survival and $Y = 1$ for death) 30 days after treatment. The data are summarized in Tables 2.1 and 2.2. The baseline covariate in consideration, gender ($X$), is fully balanced between treatments A and B, but it is a strong predictor of the mortality rate. In particular, 16% of the men died and 84% of the women died. Simple algebra shows that the mortality rate odds ratio is 0.5 in both strata, but 0.7 in the whole population. Since gender is dichotomous, the following two models are both valid.

$$\text{logit}\{P(Y = 1|Z)\} = \beta_1 + \beta_2 Z \tag{2.6}$$

$$\text{logit}\{P(Y = 1|Z, X)\} = \beta_1^* + \beta_2^* Z + \gamma X \tag{2.7}$$

In the hypothetical example, the unconditional treatment effect estimate $\hat{\beta}_2$ from (2.6) is $\log(0.7)$ while the conditional estimate $\hat{\beta}_2^*$ from (2.7) is $\log(0.5)$. In fact, Gail (1984) demonstrates that "the asymptotic bias from omitting needed covariates is

Table 2.1: Hypothetical example: by strata

| | Male: OR= 0.5 | | Female: OR= 0.5 | |
|---|---|---|---|---|
| Treatment | Dead | Survival | Dead | Survival |
| A | 10 | 80 | 72 | 18 |
| B | 18 | 72 | 80 | 10 |
| Total | 28(16%) | 152(84%) | 152(84%) | 28(16%) |

Table 2.2: Hypothetical example: pooled

| | Total: OR= 0.7 | | |
|---|---|---|---|
| Treatment | Dead | Survival | Total |
| A | 82 | 98 | 180 |
| B | 98 | 82 | 180 |
| Total | 180(50%) | 180(50%) | |

zero if the regression of the outcome on treatment assignment and covariates is linear or exponential, and, in regular cases, this is a necessary condition for zero bias".

According to Gail (1984), randomization leads to consistent treatment estimates in a large number of nonlinear models, however, it excludes many important ones such as logistic regression with binary data and Cox models with censored data. For logistic regression, the conditional and unconditional treatment effect estimates coincide when $Z$ and $X$ are conditionally independent given $Y$. In that case, people tend to believe that adjustment for $X$ will also improve the precision in logistic models. Surprisingly, Robinson and Jewell (1991) demonstrate that adjustment for predictive covariates will always result in a loss of precision. Whereas when testing for a treatment effect in randomized studies greater efficiency is still achieved by adjusting for predictive covariates, just as for classic linear regression. Lagakos and Schoenfeld (1984) and Morgan (1986) consider a related issue. They find a loss of efficiency for the test of no treatment effect when covariates are omitted from Cox regressions, with the magnitude of the loss increasing with the

magnitude of the effect of the covariates. However, the regression parameters are biased toward zero in the unconditional model, as compared with the Cox model including covariates, and the bias depends on the amount of censoring.

In summary, regression based covariate adjustments lead to inconsistent estimates when the regression of outcome on the treatment assignment and covariates is not linear or exponential. Even though the estimates are consistent in some circumstances, the conditional model may not necessarily improve precision.

## 2.4 Existing model-free methods

To make covariate adjustment completely objective, model-free adjustment approaches are needed. The aim is to efficiently utilize the baseline information to estimate the unconditional treatment difference, without modeling of the relationship between the outcome and covariates. In this section, we will outline two existing methods meeting this goal.

### 2.4.1 Koch's method

In pioneering work, Koch et al. (1998) constructed a consistent estimator of the unconditional treatment effect that always reduces the asymptotic variance (relative to using the difference between means in the treatment and control groups). The strategy therein can be used to compare continuous, ordinal and binary responses in a randomized clinical trial and does not require regression modeling of covariates effect.

Suppose $n_1 = \sum_{i=1}^{n} Z_i$ subjects are randomized to treatment ($Z = 1$) and $n_0 = \sum_{i=1}^{n}(1-Z_i)$ subjects are randomized to control ($Z = 0$), $n_1+n_0 = n$. Let $\beta$ denote the difference in mean outcome between treatment and control. A natural estimate

for $\beta$ is the difference in observed average responses under treatment and control, i.e., $\hat{\beta}_{unadj} = \bar{y}_1 - \bar{y}_0$, where $\bar{y}_1 = n_1^{-1} \sum_{i=1}^{n} Z_i Y_i$ and $\bar{y}_0 = n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i$. The rationale for Koch's method is based on a weighted least-squares procedure to estimate the treatment difference. More specifically, the linear model

$$\boldsymbol{d} = \begin{pmatrix} \hat{\beta}_{unadj} \\ \boldsymbol{d}_x \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \bar{y}_0 \\ \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0 \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{0} \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_y \\ \boldsymbol{\varepsilon}_{\bar{x}} \end{pmatrix}$$

is assumed, where $\bar{\boldsymbol{x}}_1 = n_1^{-1} \sum_{i=1}^{n} Z_i \boldsymbol{X}_i$, $\bar{\boldsymbol{x}}_0 = n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) \boldsymbol{X}_i$ are the sample averages of the covariate vectors for the treatment and control groups, respectively. The treatment difference $\beta$ is estimated with weights based on a consistent estimate $\boldsymbol{V}$ for the covariance matrix of $\boldsymbol{d}$, denoted as

$$\boldsymbol{V} = \begin{pmatrix} v_{yy} & \boldsymbol{v}_{xy}^T \\ \boldsymbol{v}_{yx} & \boldsymbol{v}_{xx} \end{pmatrix}.$$

The resulting adjusted estimator of the treatment difference

$$\hat{\beta}_{adj} = \bar{y}_1 - \bar{y}_0 - \boldsymbol{v}_{yx}^T \boldsymbol{v}_{xx}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0). \tag{2.8}$$

The asymptotic variance of $\hat{\beta}_{adj}$, $v_{adj}$, is the limit of $v_{yy} - \boldsymbol{v}_{yx}^T \boldsymbol{v}_{xx}^{-1} \boldsymbol{v}_{yx}$. Note that $v_{adj}$ is never larger than the asymptotic variance of the unadjusted estimator $v_{unadj}$, which is the limit of $v_{yy}$. Koch et al. also show that both the unadjusted test statistic $Q_{unadj} = \hat{\beta}_{unadj}^2 / v_{yy}$ and the adjusted test statistic $Q_{adj} = \hat{\beta}_{adj}^2 / (v_{yy} - \boldsymbol{v}_{yx}^T \boldsymbol{v}_{xx}^{-1} \boldsymbol{v}_{yx})$ have a $\chi_1^2$ distribution under the null hypothesis that $\beta = 0$. Under the alternative hypothesis that the two means are different, $Q_{unadj}$ and $Q_{adj}$ have asymptotic non-central $\chi_1^2$ distributions with non-centrality parameters $\beta^2 / v_{unadj}$ and $\beta^2 / v_{adj}$,

respectively. Thus this nonparametric approach increases the statistical power through variance reduction; see Koch et al. (1998) for the details.

### 2.4.2 Semiparametric inference

A more general strategy achieving the goal of separating modeling of the covariate-outcome relationship from evaluation of the treatment effect is elucidated by Tsiatis et al. (2008). An original motivation for building a semiparametric framework is to conceptualize inference on the treatment effect $\boldsymbol{\beta}$ as a "missing data problem" (Leon et al. 2003). Ideally, if we could observe the outcome on each subject under both treatment and control conditions, we would have complete sample information on the treatment effect, which should lead to the most efficient inference. However, this is not possible: for subjects randomized to treatment, the outcome they would have if assigned to the control group is "missing", and vice versa. Nevertheless, randomization still enables a valid comparison. Hence, covariate adjustment may be viewed as an attempt to use covariates that are correlated with the outcome to recover the "missing" information to some extent and thus improve efficiency. By making this analogy to missing-data problems and using the semiparametric missing-data theory of Robins et al. (1994), Leon et al. (2003) and Davidian et al. (2005) derive the class of all consistent estimators for $\boldsymbol{\beta}$ when one of the elements of $\boldsymbol{X}$ is a baseline observation on $Y$.

#### 2.4.2.1 A semiparametric framework

Following the notation in Section 2.4.1, let $\bar{Z} = n^{-1}\sum_{i=1}^{n} Z_i$ denote the sample proportion randomized to treatment. If we make no assumptions about the joint distribution of $(Y, Z, \boldsymbol{X})$ except that $Z$ is independent of $\boldsymbol{X}$, as implied by ran-

domization, then it follows from Leon et al. (2003) and Davidian et al. (2005) that all reasonable consistent and asymptotically normal estimators of $\boldsymbol{\beta}$ can be expressed either exactly or are asymptotically equivalent to an expression of the form

$$\bar{Y}_1 - \bar{Y}_0 - \sum_{i=1}^{n}(Z_i - \bar{Z})\{n_1^{-1}h_1(\boldsymbol{X}_i) + n_0^{-1}h_0(\boldsymbol{X}_i)\} \tag{2.9}$$

where $h_k$, $k = 0, 1$ are arbitrary scalar functions of $\boldsymbol{X}$.

When $h_1(\cdot) = h_0(\cdot) = 0$, (2.9) reduces to the sample mean difference $\bar{Y}_1 - \bar{Y}_0$, the standard unadjusted estimator. According to (2.9), all consistent and asymptotically normal estimators for $\beta$ can be obtained by augmenting this standard estimator by the second term, which incorporates the covariates. Since $Z$ and $\boldsymbol{X}$ are independent, the "augmentation" term converges in probability to zero for any $h_k$, so that the resulting estimators will always be consistent for $\beta$. Besides, varied choices for $h_k$ provide insight into the nature of the improvement in precision.

#### 2.4.2.2 Distinction and relative precision of common estimators

Tsiatis et al. (2008) demonstrate that many familiar estimators can be written in the form of (2.9) asymptotically and thus are consistent for $\beta$. The least square estimator for $\beta_2^*$ in ANCOVA model (2.5), which we denote as $\hat{\beta}_{ANCOVA1}$, is asymptotically equivalent to (2.9) with

$$h_1(\boldsymbol{X}_i) = h_0(\boldsymbol{X}_i) = \boldsymbol{\Sigma}_{XY}^T \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{X}_i \tag{2.10}$$

$$\boldsymbol{\Sigma}_{XY} = \mathrm{E}[\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}\{Y - \mathrm{E}(Y)\}] \tag{2.11}$$

$$\boldsymbol{\Sigma}_{XX} = \mathrm{E}[\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}^T]. \tag{2.12}$$

Note that model (2.5) from which $\hat{\beta}_{ANCOVA1}$ is derived need not be correctly specified for $\mathrm{E}(Y|\boldsymbol{X}, Z)$ in order for the above results to hold. In practice, we can substitute the corresponding sample covariance matrices for (2.11) and (2.12) and the resulting estimator will have the same asymptotic distribution as if (2.11) and (2.12) were known. In general, replacing $h_k$ with their consistent estimators will not affect the large sample properties.

From (2.10), $h_k$, associated with $\hat{\beta}_{ANCOVA1}$, are the same for $k = 0, 1$ and are linear in $\boldsymbol{X}_i$. Define

$$\boldsymbol{\Sigma}_{XY}^{(k)} = \mathrm{E}[\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}\{Y - \mathrm{E}(Y)\}|Z = k], k = 0, 1. \tag{2.13}$$

Since $\boldsymbol{\Sigma}_{XY} = \pi_0\boldsymbol{\Sigma}_{XY}^{(0)} + \pi_1\boldsymbol{\Sigma}_{XY}^{(1)}$, (2.10) can be expressed equivalently as

$$h_1(\boldsymbol{X}_i) = h_0(\boldsymbol{X}_i) = \{\pi_0\boldsymbol{\Sigma}_{XY}^{(0)} + \pi_1\boldsymbol{\Sigma}_{XY}^{(1)}\}^T\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{X}_i. \tag{2.14}$$

Other popular estimators, linear in $\boldsymbol{X}$, may also be written in the form of (2.9) asymptotically, with $h_1 = h_0$. Consider another ANCOVA model including an interaction term between $Z$ and $\boldsymbol{X}$, which can be expressed in a centered version as

$$\mathrm{E}\{Y - E(Y)|Z, \boldsymbol{X}\} = \gamma_X^T\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\} + \gamma_{XZ}^T\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}\{Z - E(Z)\} + \beta_Z\{Z - E(Z)\}. \tag{2.15}$$

It can be fitted by least-squares regression of $Y_i - \bar{Y}$ on $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ and $Z_i - \bar{Z}$, where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$. The least-squares estimator for $\beta_Z$, denoted as $\hat{\boldsymbol{\beta}}_{ANCOVA2}$, can be expressed in the form of (2.9) asymptotically

with

$$h_1(\boldsymbol{X}_i) = h_0(\boldsymbol{X}_i) = \{\pi_1 \boldsymbol{\Sigma}_{XY}^{(0)} + \pi_0 \boldsymbol{\Sigma}_{XY}^{(1)}\}^T \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{X}_i. \tag{2.16}$$

Hence, $\hat{\boldsymbol{\beta}}_{ANCOVA2}$ is a consistent and asymptotically normal estimator for $\beta_2$ in (2.4) regardless of whether (2.15) is a correct representation for $\mathrm{E}(Y|\boldsymbol{X}, Z)$.

Expressions (2.14) and (2.16) are identical if either $\pi_0 = \pi_1 = 0.5$ or $\boldsymbol{\Sigma}_{XY}^{(0)} = \boldsymbol{\Sigma}_{XY}^{(1)}$. Otherwise, $\hat{\boldsymbol{\beta}}_{ANCOVA2}$ is more precise than $\hat{\boldsymbol{\beta}}_{ANCOVA1}$. In fact, $\hat{\boldsymbol{\beta}}_{ANCOVA2}$ has the smallest asymptotic variance among all the estimators for which $h_k(\boldsymbol{X}_i)$, $k = 0, 1$, are linear in $\boldsymbol{X}_i$. In other words, estimators with $h_k$ (possibly different for $k = 0$ and $k = 1$) linear in $\boldsymbol{X}_i$ can not be more precise than $\hat{\boldsymbol{\beta}}_{ANCOVA2}$.

The estimate (2.8) proposed by Koch et al. (1998), here denoted as $\hat{\beta}_{KOCH}$, can easily be seen to have the form (2.9) by replacing $v_{xy}$ and $v_{xx}$ with their limits in probability. Since for $k = 0, 1$

$$\begin{aligned}
v_{xy} &= n_0^{-1} \hat{\boldsymbol{\Sigma}}_{XY}^{(0)} + n_1^{-1} \hat{\boldsymbol{\Sigma}}_{XY}^{(1)} \\
v_{xx} &= n_0^{-1} \hat{\boldsymbol{\Sigma}}_{XX}^{(0)} + n_1^{-1} \hat{\boldsymbol{\Sigma}}_{XX}^{(1)} \\
\hat{\boldsymbol{\Sigma}}_{XY}^{(k)} &= (n_k - 1)^{-1} \sum_{i=1}^{n} 1_{(Z_i = k)} (Y_i - \bar{Y}^{(k)})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^{(k)}) \\
\hat{\boldsymbol{\Sigma}}_{XX}^{(k)} &= (n_k - 1)^{-1} \sum_{i=1}^{n} 1_{(Z_i = k)} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^{(k)})^T \\
\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_0 &= (n_0^{-1} + n_1^{-1}) \sum_{i=1}^{n} (Z_i - \bar{Z}) \boldsymbol{X}_i,
\end{aligned}$$

it can be shown that $\hat{\beta}_{KOCH}$ is asymptotically equivalent to $\hat{\beta}_{ANCOVA2}$. Consequently, Koch's estimator is asymptotically the most precise one among all the estimators having $h_k$ linear in $\boldsymbol{X}_i$.

It is natural to ask if there are estimators that outperform these linear candidates. By appealing to semiparametric theory or a direct argument (Tsiatis et al.

2008), among all the estimators exactly equal to or asymptotically equivalent to an expression of form (2.9), that with the smallest variance asymptotically has

$$h_k(\boldsymbol{X}_i) = E(Y_i|Z_i = k, \boldsymbol{X}_i), \text{k=0,1.}$$

That is, the "optimal" $h_k$, $k = 0, 1$, are the true regression relationships of $Y$ on $\boldsymbol{X}$ for the corresponding treatment group, which may not be linear in $\boldsymbol{X}$. By virtue of being in the form (2.9), failure to identify these true regression relationships will not affect the consistency and asymptotic normality of the corresponding estimators and their asymptotic variances will usually be smaller than that of the unadjusted estimator.

#### 2.4.2.3 Generalized framework

The above framework for comparing two mean outcomes can be extended to general measures of treatment effects, such as an odds ratio associated with a binary outcome, a hazards ratio associated with a censored time-to-event outcome and so on. Let $\boldsymbol{\beta}$ denote the $q$-vector of parameters involved in making treatment comparisons under the marginal model between $Y$ and $Z$ (e.g. log-odds ratio) and let $\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)$ be the corresponding $q$-vector of estimating equations which usually has been well studied and leads to a consistent and asymptotically normal estimator, unadjusted for covariates (see Section 2.5.2 for more details). In the same spirit, the estimator for $\boldsymbol{\beta}$ with the smallest asymptotic variance can be obtained by solving the following estimating equations

$$\widetilde{\boldsymbol{m}} = \boldsymbol{m}(\boldsymbol{\beta}; Y, Z) - \sum_{g=0}^{K} (1_{\{Z=g\}} - \pi_g) \mathrm{E}(\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)|Z = g, \boldsymbol{X}). \tag{2.17}$$

See Zhang et al. (2008) for detailed derivations.

The optimal estimator from solving a sample version of (2.17) depends on the form of the conditional expectations

$$\mathrm{E}(\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)|Z = g, \boldsymbol{X}), g = 0, \ldots, K, \tag{2.18}$$

which are unknown. Although we have argued that falsely specifying the true regression relationships won't affect the consistency and asymptotic normality, we do hope to postulate a regression model as close as possible to (2.18) to achieve optimality. To that end, the following adaptive strategy is proposed in Zhang et al. (2008).

1. Solve the original estimating equations $\sum_{i=1}^{n} \boldsymbol{m}(\boldsymbol{\beta}; Y_i, Z_i) = 0$ to obtain the unadjusted estimator $\hat{\boldsymbol{\beta}}_{un}$. For each subject i, obtain the values $\boldsymbol{m}(\hat{\boldsymbol{\beta}}_{un}; Y_i, g)$ for each $g = 0, \ldots, K$.

2. For each $g$ and for each component of the $q$-vector $\boldsymbol{m}(\hat{\boldsymbol{\beta}}_{un}; Y_i, g)$, develop a parametric regression model

$$\mathrm{E}\{m_u(\hat{\boldsymbol{\beta}}_{un}; Y, Z)|\boldsymbol{X}, Z = g\} = r_{gu}(\boldsymbol{X}, \boldsymbol{\zeta}_{gu}), u = 1, \ldots, q,$$

   where $\boldsymbol{m} = (m_1, \ldots, m_q)^T$. The regression models $r_{gu}(\boldsymbol{X}, \boldsymbol{\zeta}_{gu})$ are represented as $\boldsymbol{c}_{gu}(\boldsymbol{X})^T\boldsymbol{\zeta}_{gu}$, $u = 1, \ldots, q$, where $\boldsymbol{c}_{gu}(\boldsymbol{X})$ are vectors of basis functions in $\boldsymbol{X}$ that may include polynomial terms, splines, interactions, etc. and $\boldsymbol{\zeta}_{gu}$ are the coefficients. Let $\boldsymbol{r}_g(\boldsymbol{X}, \hat{\boldsymbol{\zeta}}_g)$ denote the vector of prediction values of the conditional expectations $(\boldsymbol{r}_{g1}(\boldsymbol{X}, \hat{\boldsymbol{\zeta}}_{g1}), \ldots, \boldsymbol{r}_{gq}(\boldsymbol{X}, \hat{\boldsymbol{\zeta}}_{gq}))^T$, where $\hat{\boldsymbol{\zeta}}_g = (\hat{\boldsymbol{\zeta}}_{g1}, \ldots, \hat{\boldsymbol{\zeta}}_{gq})$ are the least-squares estimates for $\boldsymbol{\zeta}_g = (\boldsymbol{\zeta}_{g1}, \ldots, \boldsymbol{\zeta}_{gq})$.

3. Using the predicted values from step 2, form the augmented estimating equation

$$\sum_{i=1}^{n}\Big[\boldsymbol{m}(\boldsymbol{\beta};Y_i,Z_i)-\sum_{g=0}^{K}\{1_{(Z=g)}-\pi_g\}\boldsymbol{r}_g(\boldsymbol{X}_i,\hat{\boldsymbol{\zeta}}_g)\Big]=0 \tag{2.19}$$

and solve for $\boldsymbol{\beta}$ to obtain the adjusted estimator.

Undoubtedly, approaches that can avoid fitting nonparametric regressions while achieving the semiparametric efficiency are highly desirable. Thus, the empirical likelihood method, which is not only nonparametric and constraint-based but can automatically summarize data in the most efficient way as well, becomes a natural and promising candidate to meet all the needs.

## 2.5 Empirical likelihood based methods for nonparametric covariate adjustment

This section is devoted to the development of an empirical likelihood based method for nonparametric covariate adjustment arising from a typical randomized clinical trial. To begin with, Section 2.5.1 reviews some basic concepts and results from empirical likelihood, upon which the contributions herein are made. Section 2.5.2 develops an empirical likelihood ratio based test and establishes its asymptotic properties. The subsequent subsection deals with the dual problem of estimating treatment effects via maximizing the empirical likelihood when the number of constraints exceeds the number of parameters. Section 2.5.4 extends the results of §2.5.2 and §2.5.3 to the situation in which the number of constraints increases with the sample size. Asymptotic normality and Wilks type $\chi^2$ approximations as well as asymptotic efficiency are established for all the cases under suitable regularity conditions.

### 2.5.1 Background

#### 2.5.1.1 Definition

Being first implicitly used in Thomas and Grunkemeier (1975), empirical likelihood was developed into a general methodology by Owen (1988, 1990). Given $\boldsymbol{W}_i$, $i = 1, \ldots, n$, assumed to be independent with a common cumulative distribution function (CDF) $F_0$, the empirical likelihood function is a nonparametric likelihood function of the CDF $F$

$$L(F) = \prod_{i=1}^{n} dF(\boldsymbol{w}_i) = \prod_{i=1}^{n} p_i, \tag{2.20}$$

where $\boldsymbol{w}_i$ is the observed value of $\boldsymbol{W}_i$, $p_i = dF(\boldsymbol{w}_i) = P(\boldsymbol{W}_i = \boldsymbol{w}_i)$, $i = 1, \ldots, n$. To be a valid probability measure, we have

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1.$$

Without additional constraints, it is well known that the empirical distribution function is the nonparametric maximum likelihood estimate of $F_0$, denoted as $F_n$, or $p_i$ are estimated to be all equal to $n^{-1}$. Then the empirical likelihood ratio is defined as

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i.$$

Note that the formulae here and elsewhere in this dissertation do not require that the observations $\boldsymbol{W}_i$, $i = 1, \ldots, n$, are distinct.

#### 2.5.1.2 Empirical likelihood with moment constraints

Most empirical likelihood problems are concerned with estimating or testing for a parameter $\boldsymbol{\mu}$ associated with $F$, with $\boldsymbol{\mu}$ defined through moment constraints, or, estimating equations. Below we walk through a simple but illuminating example, in which the mean $\boldsymbol{\mu}$ of $F$ is considered. Although this example seems oversimplified, it is straightforward to extend the methodology illustrated there to more general cases.

To obtain confidence regions for $\boldsymbol{\mu}$, we incorporate the first order moment constraint $\mathrm{E}(\boldsymbol{W}) = \boldsymbol{\mu}$ in defining the profile empirical likelihood function

$$L_E(\boldsymbol{\mu}) = \sup_{p_1,\ldots,p_n} \Big\{ \prod_{i=1}^{n} p_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(\boldsymbol{w}_i - \boldsymbol{\mu}) = 0 \Big\}. \tag{2.21}$$

As noted by Owen(1988,1990), a unique value for the right-hand side of (2.21) exists, provided that $\boldsymbol{\mu}$ is inside the convex hull of the observations $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$. To test the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the empirical likelihood ratio statistic is defined as

$$R_E(\boldsymbol{\mu}_0) = \frac{L_E(\boldsymbol{\mu}_0)}{\sup_{\boldsymbol{\mu}} L_E(\boldsymbol{\mu})}.$$

Since $L(F)$ is maximized at $F_n$, it follows that $L_E(\boldsymbol{\mu})$ is maximized at $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{w}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{w}_i$ and $p_i = n^{-1}$. Using a Lagrange multiplier argument (see the next part for details), the maximum for the numerator is attained when

$$p_i = p_i(\boldsymbol{\mu}_0) = n^{-1}\{1 + \boldsymbol{\lambda}^T(\boldsymbol{w}_i - \boldsymbol{\mu}_0)\}^{-1},$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu}_0)$ is the solution to

$$\sum_{i=1}^{n} \{1 + \boldsymbol{\lambda}^T(\boldsymbol{w}_i - \boldsymbol{\mu}_0)\}^{-1}(\boldsymbol{w}_i - \boldsymbol{\mu}_0) = 0.$$

As a result,

$$R_E(\boldsymbol{\mu}_0) = \sum_{i=1}^{n} \{1 + \boldsymbol{\lambda}^T(\boldsymbol{w}_i - \boldsymbol{\mu}_0)\}^{-1}$$

and the empirical likelihood ratio test statistic, defined as $T_E(\boldsymbol{\mu}_0) = -2 \log R_E(\boldsymbol{\mu}_0)$, equals

$$T_E(\boldsymbol{\mu}_0) = 2 \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^T(\boldsymbol{w}_i - \boldsymbol{\mu}_0)\}.$$

Owen (1988,1990) has proved under mild conditions that under the null hypothesis $H_0$, the empirical likelihood ratio statistic $T_E(\boldsymbol{\mu}_0)$ converges in distribution to a $\chi^2$ random variable with $q$ (=dim($\boldsymbol{\mu}_0$)) degrees of freedom as $n$ goes to infinity. Hence, approximate $\alpha$-level confidence regions for $\boldsymbol{\mu}$ may be obtained as the set of points $\boldsymbol{\mu}$ such that $T_E(\boldsymbol{\mu}) \leq c_\alpha$, where $c_\alpha$ is defined such that $P(\chi^2_q \leq c_\alpha) = \alpha$. Similarly, the profile empirical likelihood ratio statistic can also be used to construct confidence regions for subsets of $\boldsymbol{\mu}$. With an arbitrary parameter $\boldsymbol{\mu}$ defined through estimating functions $\mathrm{E}\boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}) = 0$, we simply replace $\boldsymbol{w}_i - \boldsymbol{\mu}_0$ in the above steps with $\boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)$ and all the results still hold.

#### 2.5.1.3 A two-stage Newton algorithm

Maximizing the empirical likelihood function is a constrained optimization problem, where a Lagrange multiplier approach is commonly used. Let

$$l = \sum_{i=1}^{n} \log p_i + n\boldsymbol{\lambda}^T \sum_{i=1}^{n} p_i \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i) + \gamma(\sum_{i=1}^{n} p_i - 1), \tag{2.22}$$

where $\boldsymbol{\lambda}$ and $\gamma$ are Lagrange multipliers. In this primal problem, the parameters are $p_i, i = 1 \ldots, n$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and $\gamma$. Since empirical likelihood is a nonparametric likelihood, the number of parameters is essentially infinite. Setting the derivatives of $l$ with respect to all the nuisance parameters to zero, we have

$$\frac{\partial l}{\partial p_i} = \frac{1}{p_i} + n\boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i) + \gamma = 0 \quad i = 1, \ldots, n \tag{2.23}$$

$$\frac{\partial l}{\partial \boldsymbol{\lambda}} = n \sum_{i=1}^{n} p_i \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i) = 0 \tag{2.24}$$

$$\frac{\partial l}{\partial p_i} = \sum_{i=1}^{n} p_i - 1 = 0. \tag{2.25}$$

From (2.23), (2.24) and (2.25),

$$0 = \sum_{i=1}^{n} p_i \frac{\partial l}{\partial p_i} = n + \gamma,$$

and therefore $\gamma = -n$. Substituting $\gamma = -n$ into (2.23), we have

$$p_i = \frac{1}{n\{1 - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)\}}, \tag{2.26}$$

where $\boldsymbol{\lambda}$ satisfies

$$\sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)}{n\{1 - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)\}} = 0. \tag{2.27}$$

By (2.26), the log-empirical likelihood function $l_E(\boldsymbol{\mu})$ can be written as

$$l_E(\boldsymbol{\mu}) = -\log[n\{1 - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)\}], \tag{2.28}$$

which is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ only. To obtain the maximum empirical likelihood estimator for $\boldsymbol{\mu}$, a two-stage Newton algorithm is commonly used. In the first

stage, we fix $\boldsymbol{\mu}$ at some initial value and solve for $\boldsymbol{\lambda}$ from (2.27). Note that this is equivalent to a dual approach, in which we minimize $l_E$ with respect to $\boldsymbol{\lambda}$. The dual approach is a convex problem in two ways. First, the Hessian matrix of $l_E$ with respect to $\boldsymbol{\mu}$,

$$\sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)\boldsymbol{g}^T(\boldsymbol{\mu}; \boldsymbol{w}_i)}{[1 - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i)]^2}$$

is positive semidefinite. Second, since we require all $p_i$ to be nonnegative, the optimization domain

$$\bigcap_{i=1}^{n} \{1 - \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{\mu}; \boldsymbol{w}_i) \geq 0\}$$

is the intersection of $n$ hyperplanes, which is nonempty because $\boldsymbol{\lambda} = \mathbf{0}$ is always in it. Due to the convex duality, it is fast and robust to solve for $\boldsymbol{\lambda}$ when $\boldsymbol{\mu}$ is fixed. The dual approach not only reduces the dimensionality of the primal problem by writing the likelihood function in terms of a finite number of parameters, it also sheds light on the underlying structure of the Lagrange multiplier approach. In fact, the Lagrange multiplier $\boldsymbol{\lambda}$ is the nuisance score in the semiparametric framework. In the second stage, we fix $\boldsymbol{\lambda}$ at its obtained value from the first stage and maximize (2.28) to solve for $\boldsymbol{\mu}$, then iterating the two-stage process until convergence. Since the Hessian matrix of $l_E$ with respect to $\boldsymbol{\mu}$ is not necessarily negative definite, the resulting algorithm is not robust. To that end, modern optimization techniques developed to tackle the computational issues in empirical likelihood optimization are further discussed in Section 2.6.

### 2.5.2 Testing treatment differences

Empirical likelihood methodology for inference is based on maximizing the nonparametric likelihood (2.20) subject to appropriately formulated and problem-

specific constraints. For the two-arm randomized clinical trial specified by (2.2a), the constraints are generated by

$$\boldsymbol{m}(\boldsymbol{\beta}; Y, Z) = (1, Z)^T (Y - \beta_1 - \beta_2 Z). \tag{2.29a}$$

For general $K$ specified by (2.2b), it becomes

$$\boldsymbol{m}(\boldsymbol{\beta}; Y, Z) = (1, 1_{(Z=1)}, \ldots, 1_{(Z=K)})^T (Y - \beta_1 - \beta_2 1_{(Z=1)} - \ldots - \beta_{(K+1)} 1_{(Z=K)}). \tag{2.29b}$$

The zero-mean property of $\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)$ uniquely determines the value of $\boldsymbol{\beta}$ and can be used to obtain estimators through the sample-generated estimating equations. The resulting inference involves only the $Y_i$ and $Z_i$.

The availability of baseline covariates $\boldsymbol{X}_i$ should enable us to obtain additional estimating equations, thereby additional constraints. Indeed, Davidian et al. (2005) and Leon et al. (2003) found that the following form gives a general family of estimating equations:

$$\sum_{k=0}^{K} (1_{(Z=k)} - \pi_k) h_k(\boldsymbol{X}), \tag{2.30}$$

where $h_k$, $k = 0, 1, \ldots, K$ are arbitrary functions. The independence of $Z$ and $\boldsymbol{X}$ guarantees the zero-mean property of the resulting estimating equations.

It is clear now that the number of zero-mean estimating equations as provided by (2.29) and (2.30) exceeds the number of parameters which specify the treatment effect. In fact, the number of possible equations that can be generated from (2.30) can be unlimited when the baseline covariates $\boldsymbol{X}$ are continuous. Suppose we fix the choice of $h_k$ and consider how to make use of $\boldsymbol{X}$ for efficiency improvement. For

notational simplicity, we use $\boldsymbol{g}_r(\boldsymbol{\beta}; Y, Z, \boldsymbol{X})$ to denote an $r$-vector of the resultant estimating equations that include both (2.29) and (2.30). Here $r \geq 2$ in the two-sample case and $r \geq K+1$ for the general $(K+1)$-sample case.

It is well known that the empirical likelihood approach links inference for certain parameters and the available estimating equations to form a constrained optimization problem. With constraints given by $\boldsymbol{g}_r$, $L(F)$ is maximized in (2.20) subject to the following constraints:

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{g}_r(\boldsymbol{\beta}; Y_i, Z_i, \boldsymbol{X}_i) = 0. \tag{2.31}$$

This optimization problem has a unique maximizer provided that 0 is inside the convex hull of $\{\boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ for a given $\boldsymbol{\beta}$ (Owen 2001). By applying the Lagrange multiplier argument (Lang 1987), we can easily get $p_i = \{n[1 + \widehat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta})\boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i)]\}^{-1}$, where $\widehat{\boldsymbol{\lambda}}$, which is a function of $\boldsymbol{\beta}$, is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i)}{1 + \widehat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta})\boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i)} = 0. \tag{2.32}$$

Therefore, the resulting profile empirical log-likelihood, as a function of $\boldsymbol{\beta}$, takes the form

$$l_E(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left[1 + \widehat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta})\boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i)\right]. \tag{2.33}$$

**Theorem 2.5.1.** *Let $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $q_1$- and $q_2$-vectors. Define*

$$T_E = 2l_E(\widehat{\boldsymbol{\beta}}_{10}, 0) - 2l_E(\widehat{\boldsymbol{\beta}}), \tag{2.34}$$

*the logarithmic empirical profile likelihood ratio for testing $\widetilde{H_0} : \boldsymbol{\beta}_2 = 0$, where $\widehat{\boldsymbol{\beta}}_{10}$ minimizes $l_E(\boldsymbol{\beta}_1, 0)$ with respect to $\boldsymbol{\beta}_1$ and $\widehat{\boldsymbol{\beta}}$ minimizes $l_E(\boldsymbol{\beta})$. Then, under some*

*mild regularity conditions, $T_E$ converges to a $\chi^2_{(q_2)}$ random variable in distribution under $\widetilde{H_0}$.*

Theorem 2.5.1 is a direct adaptation of Corollary 5 in Qin and Lawless (1994). It enables us to get the $p$-value in testing the null hypothesis of no treatment difference and to invert the test to obtain the confidence limits. A numerical way to find $\widehat{\boldsymbol{\beta}}$, and similarly for $\widehat{\boldsymbol{\beta}}_{10}$, is to use a two-stage Newton algorithm. We first specify an initial value $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$ and solve (2.32) to obtain $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(0)})$. Next, we fix $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ in (2.33) at $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(0)})$ and minimize (2.33) over $\boldsymbol{\beta}$ to obtain a new value $\boldsymbol{\beta}^{(1)}$. We iterate the process until convergence.

From Qin and Lawless (1994), it follows that the empirical likelihood ratio test incorporating covariate information through constraints $\boldsymbol{g}_r(\boldsymbol{\beta}; Y, Z, \boldsymbol{X})$ is always more powerful than the one with $\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)$ only. Moreover, the more constraints we put into $\boldsymbol{g}_r$, the more powerful the test becomes. Because the net effect of the empirical likelihood method with more constraints than parameters is an optimal linear combination of the constraints, additional constraints should be chosen so as to avoid redundancy. However, it is not necessary to model the relationship between the covariates and the outcome, as is evident from equation (2.30); this is a very desirable feature with important practical implications.

For a binary outcome variable, if we are interested in using the log-odds ratio, then we can replace (2.29b) with

$$\boldsymbol{m}(\boldsymbol{\beta}; Y, Z) = (1, 1_{(Z=1)}, \ldots, 1_{(Z=K)})^T [Y - \phi(\beta_1 + \beta_2 1_{(Z=1)} + \ldots + \beta_{(K+1)} 1_{(Z=K)})],$$

where $\phi(\cdot) = \exp(\cdot)/[1 + \exp(\cdot)]$ is the logistic function. We can then follow the same steps to construct the empirical likelihood ratio test. As before, the large sample properties given by Theorem 3.1 continue to hold.

### 2.5.3 Maximum empirical likelihood estimate of treatment effect

Without adjusting for baseline covariates, the number of estimating equations, derived from the score functions, equals the number of parameters. Solving equations $\sum_{i=1}^{n} m(\boldsymbol{\beta}; Y_i, Z_i) = 0$ gives us the M-estimator for $\boldsymbol{\beta}$, which is known to be consistent and asymptotically normal (Huber 1981). With covariate adjustment, we have additional estimating equations containing auxiliary information through (2.30). Since the number of all available estimating equations $r$ exceeds the number of parameters $q = q_1 + q_2$, we cannot obtain the estimators simply by finding zeros of those estimating equations. One way to handle the additional constraints is to form $q$-dimensional linear combinations of all available estimating equations so that the resulting set of equations has a unique solution. One can further evaluate the limiting covariance matrix of the estimator to identify the optimal choice of such linear combinations; cf. Goldambe and Heyde (1987). Because the empirical likelihood method with overly constrained estimating equations can result in the optimal combination (Qin and Lawless 1994), it provides a natural alternative. The following result follows directly from Qin and Lawless (1994).

**Theorem 2.5.2.** *Let* $\boldsymbol{D}_r = E[\partial \boldsymbol{g}_r(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T]$ *and* $\boldsymbol{\Sigma}_r = E(\boldsymbol{g}_r \boldsymbol{g}_r^T)$. *Then, under certain regularity conditions, we have*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N\Big(0, (\boldsymbol{D}_r^T \boldsymbol{\Sigma}_r^{-1} \boldsymbol{D}_r)^{-1}\Big), \tag{2.35}$$

*where* $\widehat{\boldsymbol{\beta}}$ *is the maximum empirical likelihood estimate (MELE).*

The theorem above allows us to construct Wald-type confidence intervals using the robust variance estimate. From Corollary 2 of Qin and Lawless (1994), it fol-

lows that $\widehat{\boldsymbol{\beta}}$ has the smallest asymptotic variance among all the $q$-dimensional linear combinations of $\boldsymbol{g}_r(\boldsymbol{\beta}; Y, Z, \boldsymbol{X})$. In particular, when $r = q$, the maximum empirical likelihood estimator $\widehat{\boldsymbol{\beta}}$ will be asymptotically equivalent to the M-estimator. Furthermore, Corollary 1 of Qin and Lawless (1994) ensures that the more constraints being put into the optimization problem, the more precision one can achieve.

As an example, consider again a two-arm clinical trial with a binary outcome variable and a continuous covariate $X$, and suppose the log-odds ratio is of interest. We can incorporate both linear and quadratic terms of $X$ by using constraints

$$\boldsymbol{g}_r(\boldsymbol{\beta}; Y, Z, X) = \Big((1, Z)[Y - \phi(\beta_1 + \beta_2 Z)], (Z - \pi_1), (Z - \pi_1)X, (Z - \pi_1)X^2\Big)^T.$$

The resulting estimator will be more efficient than the M-estimator from $(1, Z)^T[Y - \phi(\beta_1 + \beta_2 Z)]$. Note that, for regression model based covariate adjustment, Robinson and Jewell (1991) demonstrated that including predictive covariates in the logit will always result in a loss of precision. In contrast, for our empirical likelihood approach, including predictive covariates in the constraints will never lead to an increase in the asymptotic variance. The fact that incoporating additional estimating equations always improves efficiency makes the empirical likelihood approach advantageous and convenient.

### 2.5.4 Empirical likelihood with an increasing number of constraints

Since we can achieve more precision by increasing the number of constraints, it is intuitive that semiparametric efficiency may be attained when the number of constrains grows with the sample size. In this connection, we consider in this

subsection empirical likelihood based covariate adjustment when the number of constraints grows to infinity as $n \to \infty$. Note here that the dimension of $\boldsymbol{\beta}$, which is of primary concern, remains fixed.

Suppose besides the $q$-dimensional score $\boldsymbol{m}(\boldsymbol{\beta}; Y, Z)$, the auxiliary information is contained in an $r_n$-vector of estimating equations $\boldsymbol{g}^*_{r_n}(\boldsymbol{\beta}) = (\boldsymbol{m}^T(\boldsymbol{\beta}; Y, Z), \boldsymbol{V}_n^T)^T$. Instead of a fixed number $r$, $r_n$ here will grow to infinity with $n$ at a certain rate. The $j^{th}$ component of $\boldsymbol{V}_n$ has the form $(1_{(Z=k)} - \pi_k)h_j(\boldsymbol{X})$ for $j = 1, \ldots, r_n - q$, where $h_j$ is a real-valued function. The following conditions will be used.

(C1) There exists a non-random $(r_n - q) \times (r_n - q)$ matrix $\boldsymbol{W}_n$ such that (i)-(iii) below are satisfied for $\boldsymbol{g}_{r_n}(\boldsymbol{\beta}) = (\boldsymbol{m}^T(\boldsymbol{\beta}; Y, Z), (\boldsymbol{W}_n V_n)^T)^T$.

(i) Components of $\boldsymbol{g}_{n,i}$, $i = 1, \ldots, n$, are uniformly bounded by a finite constant $M > 0$, where $\boldsymbol{g}_{n,i}(\boldsymbol{\beta}) = \boldsymbol{g}_{r_n}(\boldsymbol{\beta}; Y_i, Z_i, \boldsymbol{X}_i)$.

(ii) Eigenvalues of $\boldsymbol{\Sigma}_{n,g} = E(\boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0)\boldsymbol{g}^T_{r_n}(\boldsymbol{\beta}_0))$ are bounded away from zero and infinity.

(iii) There exists a $q \times (r_n - q)$ non-random matrix $\boldsymbol{A}_n$ such that

$$\boldsymbol{A}_n \boldsymbol{W}_n \boldsymbol{V}_n \to \sum_{k=0}^{K} (1_{(Z=k)} - \pi_k) E(\boldsymbol{m}(\boldsymbol{\beta}; Y, Z) | Z = k, \boldsymbol{X}) \quad \text{in } \mathbb{L}^2.$$

(C2) The growth rate of $r_n$ is limited to $r_n^3 = o(n)$.

(C3) Matrix $\widetilde{\boldsymbol{\Sigma}} = E(\widetilde{\boldsymbol{m}}\widetilde{\boldsymbol{m}}^T)$ is positive definite, where

$$\widetilde{\boldsymbol{m}} = \boldsymbol{m}(\boldsymbol{\beta}; Y, Z) - \sum_{k=0}^{K} (1_{(Z=k)} - \pi_k) E(\boldsymbol{m}(\boldsymbol{\beta}; Y, Z) | Z = k, \boldsymbol{X}).$$

Throughout, $\|\cdot\|$ is used to denote the Euclidean norm. For notational conve-

nience, let

$$\begin{aligned} G_n(\boldsymbol{\beta}) &= \max_{1\le i\le n} \left\|\boldsymbol{g}_{n,i}(\boldsymbol{\beta})\right\|, \quad \boldsymbol{D}_{r_n} = E(\partial \boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T), \\ \boldsymbol{D}_{m_n} &= E(\partial \boldsymbol{m}_{r_n}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T), \quad \boldsymbol{D}_{opt} = E(\partial \boldsymbol{m}^{opt}_{r_n}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T), \\ \boldsymbol{\Sigma}_{n,m} &= E(\boldsymbol{m}_{r_n}(\boldsymbol{\beta}_0)\boldsymbol{m}^T_{r_n}(\boldsymbol{\beta}_0)), \quad \boldsymbol{\Sigma}_{n,opt} = E(\boldsymbol{m}^{opt}_{r_n}(\boldsymbol{\beta}_0)(\boldsymbol{m}^{opt}_{r_n}(\boldsymbol{\beta}_0))^T). \end{aligned}$$

**Lemma 1.** *The probability that zero is outside the convex hull spanned by* $\{\boldsymbol{g}_{n,i}, i = 1,\ldots,n\}$ *goes to zero as* $n \to \infty$.

*Proof.* This follows from Lemma 4.2 in Hjort et al. (2009) and discussions thereof. □

**Lemma 2.** *Under (i),(ii) and C2, the eigenvalues of* $\boldsymbol{S}_n(\boldsymbol{\beta}_0)$ *are bounded away from 0 and* $\infty$.

*Proof.* It can be shown by making use of proofs of condition (D4) and Lemma 4.5 in Hjort et al. (2009). □

**Lemma 3.** *Under (i),(ii) and C2,*

$$\left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta}_0)\right\| = O_p(n^{-1/2} r_n^{1/2}) \tag{2.36}$$

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le n^{-1/3}} \left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\| = O_p(n^{-1/3}) \tag{2.37}$$

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le n^{-1/3}} \left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta}) - \boldsymbol{S}_n(\boldsymbol{\beta})^{-1}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta})\right\| = O_p(n^{-2/3} r_n^{1/2}). \tag{2.38}$$

*Proof.* Under (i),(ii) and C2, we can apply results in Portnoy (1988) to get

$$\left\|n^{1/2}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)\right\| = O_p(r_n^{1/2}). \tag{2.39}$$

Under (i),

$$G_n(\boldsymbol{\beta}) \leq M r_n^{1/2} = O_p(r_n^{1/2}). \tag{2.40}$$

Write $\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta}) = \left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\| \boldsymbol{u}_n(\boldsymbol{\beta})$, where $\|\boldsymbol{u}_n(\boldsymbol{\beta})\| = 1$. Then similarly to (2.32), we can show that

$$0 = \boldsymbol{u}_n^T(\boldsymbol{\beta})\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{g}_{n,i}(\boldsymbol{\beta})}{1+\widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta})\boldsymbol{g}_{n,i}(\boldsymbol{\beta})} \leq \boldsymbol{u}_n^T(\boldsymbol{\beta})\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) - \frac{\left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\|}{1+\left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\| G_n(\boldsymbol{\beta})} mineig(\boldsymbol{S}_n(\boldsymbol{\beta})),$$

where $mineig(\boldsymbol{M})$ stands for the minimum eigenvalue of the matrix $\boldsymbol{M}$. Therefore, we have

$$\left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\| (mineig(\boldsymbol{S}_n(\boldsymbol{\beta})) - \boldsymbol{u}_n^T(\boldsymbol{\beta})\bar{\boldsymbol{g}}_n(\boldsymbol{\beta})G_n(\boldsymbol{\beta})) \leq \boldsymbol{u}_n^T(\boldsymbol{\beta})\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}), \tag{2.41}$$

from which we know that (2.36) holds due to (2.39), (2.40) and Lemma 2.

When $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq n^{-1/3}$, define

$$L_n = \max_{j,k} \left|\boldsymbol{S}_{n,j,k}(\boldsymbol{\beta}) - \boldsymbol{S}_{n,j,k}(\boldsymbol{\beta}_0)\right|. \tag{2.42}$$

Using the same technique as in Lemma 2, $r_n L_n = o_p(1)$ ensures that the minimum eigenvalue of $S_n(\boldsymbol{\beta})$ is bounded away from zero. Since there are only finitely many terms in $\boldsymbol{g}_{r_n}$ containing $\boldsymbol{\beta}$, due to the $\delta$-method, this can be further reduced to $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = o(r_n^{-1})$, which is true under C2. By expanding $\bar{\boldsymbol{g}}_n(\boldsymbol{\beta})$ in the $n^{-1/3}$ neighborhood of $\boldsymbol{\beta}_0$, we obtain $\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) = O_p(n^{-1/3})$ uniformly in $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq n^{-1/3}$. Then (2.37) follows from equation (2.41).

We know that $\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})$ satisfies the constraint

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_{n,i}(\boldsymbol{\beta})/\{1+\widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta})\boldsymbol{g}_{n,i}(\boldsymbol{\beta})\} = 0,$$

which implies

$$\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta}) = \boldsymbol{S}_n(\boldsymbol{\beta})^{-1}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) + \boldsymbol{S}_n(\boldsymbol{\beta})^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_{n,i}(\boldsymbol{\beta})\frac{\boldsymbol{u}_n^T(\boldsymbol{\beta})\boldsymbol{g}_{n,i}(\boldsymbol{\beta})\boldsymbol{g}_{n,i}^T(\boldsymbol{\beta})\boldsymbol{u}_n(\boldsymbol{\beta})}{1+\widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta})\boldsymbol{g}_{n,i}(\boldsymbol{\beta})}\left\|\widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta})\right\|^2. \tag{2.43}$$

By the triangle inequality and some simple algebra, the final term in (2.43) is bounded by $O_p(n^{-2/3}r_n^{1/2})$. Since $\|\boldsymbol{S}_n(\boldsymbol{\beta})^{-1}\| = O_p(1)$, (2.38) follows from (2.43). □

**Lemma 4.** *Under Conditions C1-C3,*

$$\left\|\boldsymbol{D}_{r_n}^T\boldsymbol{\Sigma}_{n,g}^{-1}\boldsymbol{D}_{r_n} - \boldsymbol{D}_m^T\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{D}_m\right\| = o(1).$$

*Proof.* Let $\boldsymbol{m}_{r_n} = \boldsymbol{m}(\boldsymbol{\beta}; Y, Z) + \boldsymbol{A}_n\boldsymbol{W}_n\boldsymbol{V}_n$. Since $\boldsymbol{A}_n\boldsymbol{W}_n\boldsymbol{V}_n$ does not involve $\boldsymbol{\beta}$, we have

$$\boldsymbol{D}_{m_n}^T\boldsymbol{\Sigma}_{n,m}^{-1}\boldsymbol{D}_{m_n} = \boldsymbol{D}_m^T\boldsymbol{\Sigma}_{n,m}^{-1}\boldsymbol{D}_m, \tag{2.44}$$

which by (iii), converges to $\boldsymbol{D}_m^T\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{D}_m$.

Second, following Qin and Lawless (1994), for any $n$, we have

$$\boldsymbol{D}_{r_n}^T\boldsymbol{\Sigma}_{n,g}^{-1}\boldsymbol{D}_{r_n} = \boldsymbol{D}_{opt}^T\boldsymbol{\Sigma}_{n,opt}^{-1}\boldsymbol{D}_{opt},$$

where $\boldsymbol{m}_{r_n}^{opt} = \boldsymbol{A}_{opt}(\boldsymbol{\beta})\boldsymbol{g}_{r_n}$ is a $q$-vector and $\boldsymbol{A}_{opt}(\boldsymbol{\beta})$ is the optimal linear combination of $\boldsymbol{g}_{r_n}$. So it suffices to show the following difference is zero:

$$\boldsymbol{D}_{opt}^T\boldsymbol{\Sigma}_{n,opt}^{-1}\boldsymbol{D}_{opt} - \boldsymbol{D}_m^T\boldsymbol{\Sigma}_{n,m}^{-1}\boldsymbol{D}_m. \tag{2.45}$$

Given (2.44), (2.45) is positive definite due to optimality. Furthermore,

$$\begin{aligned} & \boldsymbol{D}_{opt}^T \boldsymbol{\Sigma}_{n,opt}^{-1} \boldsymbol{D}_{opt} - \boldsymbol{D}_m^T \boldsymbol{\Sigma}_{n,m}^{-1} \boldsymbol{D}_m \\ = \; & \boldsymbol{D}_{opt}^T \boldsymbol{\Sigma}_{n,opt}^{-1} \boldsymbol{D}_{opt} - \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m + \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m - \boldsymbol{D}_m^T \boldsymbol{\Sigma}_{n,m}^{-1} \boldsymbol{D}_m. \end{aligned}$$

By Zhang et al. (2008), we know that $\boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m$ is the semiparametric efficiency bound, which implies the first difference is non-positive definite. Since the second difference is $o_p(1)$, we know (2.45) is nonpositive definite. □

**Lemma 5.** *Under (i), (ii) and C2,* $\left\| \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right\| < n^{-1/3}$.

*Proof.* We first consider $\boldsymbol{\beta}$ on the $n^{-1/3}$ sphere of $\boldsymbol{\beta}_0$, i.e. $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = \boldsymbol{u} n^{-1/3}$, where $\boldsymbol{u}$ is a unit vector. On the one hand, by the Taylor series expansion and Lemma 3,

$$2 \sum_{i=1}^{n} \log \left( 1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}) \boldsymbol{g}_{n,i}(\boldsymbol{\beta}) \right) = 2n \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}) \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) - n \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}) \boldsymbol{S}_n(\boldsymbol{\beta}) \widehat{\boldsymbol{\lambda}}_n(\boldsymbol{\beta}) + O_p(r_n^{1/2}).$$

By (2.38), this is equivalent to $n \bar{\boldsymbol{g}}_n^T(\boldsymbol{\beta}) \boldsymbol{S}_n^{-1}(\boldsymbol{\beta}) \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) + O_p(r_n^{1/2})$. By taking the Taylor series expansion at $\boldsymbol{\beta}_0$, this equals

$$\boldsymbol{u}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{D}_{r_n} \boldsymbol{u} n^{1/3} + o_p(n^{1/3}),$$

which is bounded below by $O_p(n^{1/3})$ by Lemma 4. On the other hand, $2 \sum_{i=1}^{n} \log \Big( 1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}_0) \boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0) \Big) = O_p(r_n)$, which is strictly less than $O_p(n^{1/3})$ by condition C2. Therefore, $\left\| \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right\| < n^{-1/3}$. □

**Lemma 6.** *Under conditions C1-C3, we have the asymptotic normality of the "influence function"*

$$\boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} n^{1/2} \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) \to N(0, \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m).$$

*Proof.* We can reduce the problem to the unidimensional case by noting that it suffices to show that for any $q \times 1$ vector $\boldsymbol{t}$,

$$\boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} n^{1/2} \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) \to N(0, \boldsymbol{t}^T \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m \boldsymbol{t}). \tag{2.46}$$

First, the variance of the left hand side of (2.46) is $\boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{D}_{r_n} \boldsymbol{t}$, which converges to $\boldsymbol{t}^T \boldsymbol{D}_m \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m \boldsymbol{t}$ by Lemma 4.

Second, we verify the Lindeberg condition (Billingsley 1986)

$$\begin{aligned} & \sum_{i=1}^{n} E\Big\{ \Big[ n^{-1/2} \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0) \Big]^2 \mathbf{1}_{\left[ \left| n^{-1/2} \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0) \right| > \varepsilon \right]} \Big\} \\ = \; & E\Big\{ \Big[ \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0) \Big]^2 \mathbf{1}_{\left[ \left| \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0) \right| > n^{1/2} \varepsilon \right]} \Big\} \to 0, \end{aligned}$$

where the last step comes from

$$P\Big( \left| \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0) \right| > n^{1/2} \varepsilon \Big) \leq E\Big( \boldsymbol{t}^T \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} \boldsymbol{g}_{r_n}(\boldsymbol{\beta}_0) \Big)^2 \Big/ n\varepsilon^2,$$

which goes to 0 since the numerator is asymptotically bounded. Hence Lemma 6 holds by the Lindeberg-Feller Central Limit Theorem. □

**Theorem 2.5.3.** *Let $\widehat{\boldsymbol{\beta}}_n$ be the maximum empirical likelihood estimate based on constraints $\boldsymbol{g}_{r_n}^*(\boldsymbol{\beta})$ and $\boldsymbol{D}_m = E(\partial \boldsymbol{m}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^T)$. Then, under Conditions C1-C3,*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N\Big(0, (\boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m)^{-1}\Big). \tag{2.47}$$

*Proof.* Let $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}_{n,i}(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^T \boldsymbol{g}_{n,i}(\boldsymbol{\beta})}$ and $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{\lambda} \partial \boldsymbol{g}_{n,i}^T(\boldsymbol{\beta})/\partial \boldsymbol{\beta}}{1 + \boldsymbol{\lambda}^T \boldsymbol{g}_{n,i}(\boldsymbol{\beta})}$. We know that $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\lambda}}_n)$ satisfies $\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\lambda}}_n) = 0$ and $\boldsymbol{V}_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\lambda}}_n) = 0$. By taking the Taylor series

expansion, we have

$$\begin{aligned} 0 &= \boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\lambda}}_n) \\ &= \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + \hat{\boldsymbol{D}}^T(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) - \boldsymbol{S}_n(\boldsymbol{\beta}_0)\widehat{\boldsymbol{\lambda}}_n + O_p(n^{-2/3} r_n^{1/2}), \quad \text{and} (2.48) \\ 0 &= \boldsymbol{V}_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\lambda}}_n) \\ &= \hat{\boldsymbol{D}}^T(\boldsymbol{\beta}_0)\widehat{\boldsymbol{\lambda}}_n + O_p(n^{-2/3}). \qquad (2.49) \end{aligned}$$

Solving (2.48) and (2.49) for $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$, we get,

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -n^{1/2}(\hat{\boldsymbol{D}}(\boldsymbol{\beta}_0)^T \boldsymbol{S}_n^{-1}(\boldsymbol{\beta}_0)\hat{\boldsymbol{D}}(\boldsymbol{\beta}_0))^{-1}\hat{\boldsymbol{D}}(\boldsymbol{\beta}_0)\boldsymbol{S}_n^{-1}(\boldsymbol{\beta}_0)\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_p(1). \tag{2.50}$$

Using the triangle inequality and Lemma 4, we can show that

$$\left\| (\hat{\boldsymbol{D}}^T(\boldsymbol{\beta}_0)\boldsymbol{S}_n^{-1}(\boldsymbol{\beta}_0)\hat{\boldsymbol{D}}(\boldsymbol{\beta}_0))^{-1} - (\boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m)^{-1} \right\| = o_p(1). \tag{2.51}$$

By Lemma 6,

$$\begin{aligned} \hat{\boldsymbol{D}}^T(\boldsymbol{\beta}_0)\boldsymbol{S}_n^{-1}(\boldsymbol{\beta}_0) n^{1/2} \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) &= \boldsymbol{D}_{r_n}^T \boldsymbol{\Sigma}_{n,g}^{-1} n^{1/2} \bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_p(n^{-1/2+\varepsilon} r_n^{1/2}) \\ &\to N(0, \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m). \end{aligned}$$

Then Theorem 2.5.3 follows from (2.50), (2.51) and Slutsky's Theorem. □

Minimizing the asymptotic variance of the M-estimator from the class of arbitrary $q$-dimensional unbiased estimating equations, Zhang et al. (2008) derived the semiparametric efficiency bound for the estimators of treatment effect. From Zhang et al. (2008) and Theorem 2.5.3, we have the following result.

**Corollary 2.5.4.** *The limiting variance-covariance matrix,* $(\boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m)^{-1}$,

*achieves the semiparametric efficiency bound, i.e., $\widehat{\boldsymbol{\beta}}_n$ in Theorem 2.5.3 is asymptotically efficient.*

In practice, in order to construct the Wald type confidence interval for $\boldsymbol{\beta}_0$, we need to estimate the asymptotic variance expressed in (2.47). Let $\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{n,i}(\boldsymbol{\beta})$, $\boldsymbol{S}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{n,i}(\boldsymbol{\beta})\boldsymbol{g}_{n,i}^T(\boldsymbol{\beta})$ and $\hat{\boldsymbol{D}}(\boldsymbol{\beta}) = \partial\bar{\boldsymbol{g}}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T$. Theorem 2.5.5 below shows that a consistent estimate of the limiting variance-covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ is $[\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{S}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_n)]^{-1}$.

**Theorem 2.5.5.** *Under Conditions C1-C3,* $\left\|\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{S}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_n) - \boldsymbol{D}_m^T\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{D}_m\right\| = o_p(1)$.

*Proof.* Since there are only finitely many terms in $\bar{\boldsymbol{g}}_n$ and $\boldsymbol{S}_n$ that contain $\boldsymbol{\beta}$, by the $\delta$-method, we have

$$\left\|(\hat{\boldsymbol{D}}^T(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{S}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{D}}(\widehat{\boldsymbol{\beta}}_n))^{-1} - (\hat{\boldsymbol{D}}^T(\boldsymbol{\beta}_0)\boldsymbol{S}_n^{-1}(\boldsymbol{\beta}_0)\hat{\boldsymbol{D}}(\boldsymbol{\beta}_0))^{-1}\right\| = o_p(1).$$

Then the result follows from (2.51). □

Theorem 2.5.3 states that the listed conditions are sufficient to ensure standard asymptotic properties of the MELE. Moreover, Corollary 2.5.4 states that when the number of constraints grows to infinity at a certain rate, the MELE achieves the semiparametric efficiency bound as derived in Zhang (2008). In Theorem 2.5.3, $\boldsymbol{g}_{r_n}$ is essentially a linear transformation of $\boldsymbol{g}_{r_n}^*$. Since a linear transformation does not change the constraints, the estimator using $\boldsymbol{g}_{r_n}$ will be the same as that using $\boldsymbol{g}_{r_n}^*$. The fact that the MELE will not be affected by a linear transformation of the constraints greatly facilitates the applicability of the empirical likelihood approach because we can just throw in all the constraints we have without forming the appropriate combination of them. For example, $E[\boldsymbol{g}_{r_n}^*(\boldsymbol{g}_{r_n}^*)^T]$ might be ill

conditioned but we can still use it as long as there exists a $\boldsymbol{W}_n$ such that the corresponding $\boldsymbol{\Sigma}_{n,g}$ is better conditioned. For this reason, we will not distinguish among linear transformations of constraints in the following discussion.

Theorem 2.5.3 holds for a general $q$-dimensional score $\boldsymbol{m}$ as long as some regularity conditions in the case of fixed number of constraints (Qin and Lawless 1994) are satisfied, including $E\Big(\partial \boldsymbol{m}(\boldsymbol{\beta}; Y, Z)/\partial \boldsymbol{\beta}^T\Big)$ is of full rank $p$, $\big\|\partial \boldsymbol{m}(\boldsymbol{\beta}; Y, Z)/\partial \boldsymbol{\beta}^T\big\|$ and $\big\|\partial^2 \boldsymbol{m}(\boldsymbol{\beta}; Y, Z)/\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T\big\|$ can be bounded by some integrable function in a neighborhood of $\boldsymbol{\beta}_0$ and $\partial \boldsymbol{m}(\boldsymbol{\beta}; Y, Z)/\partial \boldsymbol{\beta}$ and $\partial^2 \boldsymbol{m}(\boldsymbol{\beta}; Y, Z)/\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T$ are continuous in this neighborhood.

Condition C2 imposes an upper bound on the growth rate of the number of constraints at which a well-behaved MELE can be obtained. In practice, the number of constraints need not be large. In fact, we find that additional gain by including an extra constraint diminishes quickly, due to the optimal use of constraints by the empirical likelihood method. It is important to note that the asymptotic normality and efficiency are not affected by the choice of $r_n$, as long as C2 is satisfied. It is certainly of theoretical interest to find the sharp upper bound for $r_n$ to grow such that the resulting estimate is still asymptotically normal and efficient. But we will not get into this complication here since finding the optimal rate is not our main concern. If we knew the conditional expectations in Condition C3, the optimal estimating equations $\widetilde{\boldsymbol{m}}$ would be the constraints that lead to the optimal estimator. Although they are unknown in practice, it is clear that Condition C3 is fairly mild.

For Condition C1, we need to make use of the orthogonality and boundedness of certain basis functions to properly design $h(\boldsymbol{X})$ in the constraints. Suppose $Z = 0, 1, 2$ and the empirical CDF of the one dimensional auxiliary covariate $X$ is $F_n(x) = n^{-1}\sum_{i=1}^{n} 1_{\{X_i \le x\}}$. By making use of multivariate Fourier

expansion, the arguments can be generalized to the high dimensional auxiliary covariate case. Let $\boldsymbol{g}^*_{r_n}(\boldsymbol{\beta}) = (\boldsymbol{m}^T(\boldsymbol{\beta}; Y, Z), (1_1 - \pi_1), \widehat{s}_{11}, \widehat{c}_{11}, \ldots, \widehat{s}_{1d_n}, \widehat{c}_{1d_n}, (1_2 - \pi_2), \widehat{s}_{21}, \widehat{c}_{21}, \ldots, \widehat{s}_{2d_n}, \widehat{c}_{2d_n})^T$, where $1_k = 1_{(Z=k)}$, $r_n = 4d_n + q + 2$, $\widehat{s}_{ij} = (1_i - \pi_i)\sin(2\pi j F_n(X))$, $\widehat{c}_{ij} = (1_i - \pi_i)\cos(2\pi j F_n(X))$, $i = 1, 2$, $j = 1, \ldots, d_n$. It can be shown that, when $d_n = o(n^{1/4})$, (i)-(iii) are satisfied. For example, we can apply the fact that those basis functions are orthogonal when their arguments are $U[0, 1]$ and they are bounded to show (i) and (ii) hold. Because the procedure is invariant under linear transformations, the eigenvalues can grow with $n$ if all of them grow at the same rate. However, we do not believe in general they can grow at different rates since the covariance matrix is sandwiched in the variance-covariance expression, which needs to be well-conditioned. Furthermore, (iii) can be verified by taking the expansion of the conditional expectations. Likewise, we may apply other orthogonal basis functions that are bounded. For example, we can use the Legendre polynomials of $(2F_n(X) - 1)$ which are bounded by 1 on [-1,1]. Legendre polynomials, i.e. $1, x, (3x^2 - 1)/2, \ldots$, are linear transformations of polynomial terms $1, x, x^2, \ldots$. Therefore we can also use polynomial terms of $(2F_n(X) - 1)$ in the auxiliary constraints due to linear transformation invariance of the empirical likelihood. Note that the standard independence assumption for empirical likelihood is violated due to the plug-in estimator $F_n$. Intuitively, the validity of using $F_n$ instead of $F$ relies on the fact that those constraints are still zero-mean conditioning on all the covariates. A rigorous proof is provided below.

*Proof.* We verify that $\boldsymbol{g}^*_{r_n}$ in the examples following Corollary 2.5.4 satisfies Condition C1. The other conditions are satisfied trivially. Since the Fourier basis functions are naturally bounded by 1, the uniform boundedness reduces to the boundedness of $\boldsymbol{m}$ which is of finite dimension and usually holds. So (i) is satis-

fied. Let

$$\boldsymbol{V}_n = ((1_1 - \pi_1)/\pi_1, s_{11}, c_{11}, \ldots, s_{1d_n}, c_{1d_n}, (1_2 - \pi_2)/\pi_2, s_{21}, c_{21}, \ldots, s_{2d_n}, c_{2d_n})^T$$

and $\boldsymbol{g}_{r_n}(\boldsymbol{\beta}) = (\boldsymbol{m}^T(\boldsymbol{\beta}; Y, Z), \boldsymbol{V}_n^T)^T$, where $1_k = 1_{(Z=k)}$, $s_{ij} = \sqrt{2}(1_i - \pi_i)\sin(2\pi j F(X))/\pi_i$, $c_{ij} = \sqrt{2}(1_i - \pi_i)\cos(2\pi j F(X))/\pi_i$, $i = 1, 2$, $j = 1, \ldots, d_n$. For notational simplicity, we omit $\boldsymbol{W}_n$ in $\boldsymbol{W}_n\boldsymbol{V}_n$ when there is no ambiguity. Then letting $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix, we have the following matrix partition

$$\boldsymbol{\Sigma}_{n,g} = \left[\begin{array}{c|cc} E(\boldsymbol{m}\boldsymbol{m}^T) & E(\boldsymbol{m}\boldsymbol{V}_n^T) & \\ \hline & \frac{1-\pi_1}{\pi_1}\boldsymbol{I}_{2d_n+1} & -\boldsymbol{I}_{2d_n+1} \\ E(\boldsymbol{V}_n\boldsymbol{m}^T) & -\boldsymbol{I}_{2d_n+1} & \frac{1-\pi_2}{\pi_2}\boldsymbol{I}_{2d_n+1} \end{array}\right].$$

Thus, by some simple algebra and C3, we can show that the eigenvalues of $\boldsymbol{\Sigma}_{n,g}$ are bounded away from 0 and $\infty$. However, since F is unknown in practice, we typically use $F_n(x) = n^{-1}\sum_{i=1}^n 1_{\{X_i \le x\}}$ instead. Let

$$\widehat{\boldsymbol{V}}_n^T(z, x) = ((1_1 - \pi_1)/\pi_1, \widehat{s}_{11}, \widehat{c}_{11}, \ldots, \widehat{s}_{1d_n}, \widehat{c}_{1d_n}, (1_2 - \pi_2)/\pi_2, \widehat{s}_{21}, \widehat{c}_{21}, \ldots, \widehat{s}_{2d_n}, \widehat{c}_{2d_n})$$

and $\widehat{\boldsymbol{g}}_{r_n}(\boldsymbol{\beta}) = (\boldsymbol{m}^T(\boldsymbol{\beta}; Y, Z), \widehat{\boldsymbol{V}}_n^T(Z, X))^T$, where $\widehat{s}_{ij} = \sqrt{2}(1_i - \pi_i)\sin(2\pi j F_n(x))/\pi_i$, $\widehat{c}_{ij} = \sqrt{2}(1_i - \pi_i)\cos(2\pi j F_n(x))/\pi_i$, $i = 1, 2$, $j = 1, \ldots, d_n$. Define $\boldsymbol{\varepsilon}_n = \widehat{\boldsymbol{g}}_{r_n}\widehat{\boldsymbol{g}}_{r_n}^T - \boldsymbol{g}_{r_n}\boldsymbol{g}_{r_n}^T$. Then

$$\begin{aligned} r_n \max_{j,k} |\boldsymbol{\varepsilon}_{n,j,k}| &\le 2M^2 r_n \left|\sin \pi d_n (F_n(X) - F(X))\right| \\ &= O_p(r_n^2 n^{-1/2}). \end{aligned}$$

Following the argument in Lemma 2, when we let $r_n = o(n^{\frac{1}{4}})$, we know the eigen-

values of $E(\widehat{\boldsymbol{g}}_{r_n}(\boldsymbol{\beta}_0)\widehat{\boldsymbol{g}}_{r_n}^T(\boldsymbol{\beta}_0))$ are also bounded away from zero and infinity. So (ii) holds.

Moreover, let $f(z,x) = \sum_{k=0}^{K}(1_k - \pi_k)E(\boldsymbol{m}(\boldsymbol{\beta};Y,Z)|Z=k,x)$ and $\boldsymbol{A}_n$ be the Fourier coefficients in the Fourier expansion of $f(z,x)$ with the Fourier basis specified in $\widehat{\boldsymbol{V}}_n(z,x)$. We know from Fourier approximation theory that $\boldsymbol{A}_n\widehat{\boldsymbol{V}}_n(z,x) \to f(z,x)$ uniformly. Thus, by Condition C3 and the Dominated Convergence Theorem, (iii) is satisfied.

Proof of the validity of the plug-in estimator $F_n$. Checking the derivation of all the theorems, we find that the following two conditions will guarantee the validity of the theorems when $F$ is replaced by $F_n$

$$\left\| n^{-1/2}\sum_{i=1}^{n}(\hat{\boldsymbol{g}}_{n,i} - \boldsymbol{g}_{n,i}) \right\| = o_p(1) \tag{2.52}$$

$$\left\| n^{-1}\sum_{i=1}^{n}\left\{\hat{\boldsymbol{g}}_{n,i}\hat{\boldsymbol{g}}_{n,i}^T - \boldsymbol{g}_{n,i}\boldsymbol{g}_{n,i}^T\right\} \right\| = o_p(1), \tag{2.53}$$

where $\boldsymbol{g}_{n,i}$ and $\hat{\boldsymbol{g}}_{n,i}$ are $\boldsymbol{g}_{r_n}$ and $\hat{\boldsymbol{g}}_{r_n}$ evaluated at the $i^{th}$ sample. The norm of a matrix $\boldsymbol{M}$ is defined to be $\sup_{\boldsymbol{u}} \|\boldsymbol{M}\boldsymbol{u}\|$, where $\boldsymbol{u}$ is a unit vector. The sufficiency of the above two conditions when the number of constraints is fixed can be seen from the existing literature (see, for example, Hjort et al. (2009)).

Denote the $j^{th}$ component of a vector $\boldsymbol{g}$ by $\boldsymbol{g}^j$. Then, for any $j$, we have

$$E\left[\left\{n^{-1/2}\sum_{i=1}^{n}(\hat{\boldsymbol{g}}_{n,i}^j - \boldsymbol{g}_{n,i}^j)\right\}^2 \middle| \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\right] \le C_1 r_n^2 \left\|F_n - F\right\|_\infty^2,$$

where $C_1$ is a universal constant. Therefore,

$$E\left\{\left\|n^{-1/2}\sum_{i=1}^{n}(\hat{\boldsymbol{g}}_{n,i} - \boldsymbol{g}_{n,i})\right\|^2 \middle| \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\right\} \le C_1 r_n^3 \left\|F_n - F\right\|_\infty^2 = O_p(r_n^3/n),$$

which converges to 0 in probability due to C2. By Chebyshev's inequality, we know that for any $\varepsilon > 0$,

$$P\Big\{ \Big\| n^{-1/2} \sum_{i=1}^{n} (\hat{\boldsymbol{g}}_{n,i} - \boldsymbol{g}_{n,i}) \Big\| \geq \varepsilon \Big| \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \Big\} = o_p(1),$$

which implies (2.52) due to the dominated convergence theorem.

Denote $\boldsymbol{\varepsilon}_u = n^{-1} \sum_{i=1}^{n} \Big\{ \hat{\boldsymbol{g}}_{n,i} \hat{\boldsymbol{g}}_{n,i}^T - \boldsymbol{g}_{n,i} \boldsymbol{g}_{n,i}^T \Big\} \boldsymbol{u}$. Then we have $E\{(\boldsymbol{\varepsilon}_u^j)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \leq O_p(r_n^3 n^{-2})$ uniformly for $\boldsymbol{u}$ and $j$. Therefore, $E\{\|\boldsymbol{\varepsilon}_u\|^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \leq O_p(r_n^4 n^{-2}) \leq o_p(1)$, which implies (2.53). □

Analogous to the case with a fixed number of constraints, let $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \Big( 1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}) \boldsymbol{g}_{n,i}(\boldsymbol{\beta}) \Big)$. The empirical likelihood ratio statistic for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ is

$$T_{1n} = 2l(\boldsymbol{\beta}_0) - 2l(\widehat{\boldsymbol{\beta}}_n). \tag{2.54}$$

Then under Conditions C1-C3, the Wilks type theorem of convergence to the $\chi^2$ distribution is still valid for testing the null hypothesis of no treatment effect.

**Theorem 2.5.6.** *Suppose that Conditions C1-C3 are satisfied. Then, under the null hypothesis $H_0$, $T_{1n}$ converges in distribution to a $\chi^2_{(q)}$ random variable as $n \to \infty$.*

*Proof.* Taking the Taylor series expansion, we get

$$\begin{aligned} T_{1n} &= n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \Big[ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \frac{1}{n} \sum_{i=1}^{n} \log \Big( 1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}_0) \boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0) \Big) \Big] n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(1) \\ &= n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \boldsymbol{A} n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(1). \end{aligned}$$

Then Theorem 2.5.3 implies $T_{1n} \to \chi^2_q$ as $n \to \infty$, when $H_0$ is true. □

More generally, we can test hypotheses concerning a subset of treatment effects

$\boldsymbol{\beta}$. For instance, we may be interested in testing whether $\boldsymbol{\beta}_2 = 0$ in the simple example (2.2a). Specifically, let $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $q_1$- and $q_2$-vectors, respectively. For $\widetilde{H_0} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$, the profile empirical likelihood ratio test statistic is simply

$$T_{2n} = 2l(\boldsymbol{\beta}_{10}, \widehat{\boldsymbol{\beta}}_{20}) - 2l(\widehat{\boldsymbol{\beta}}_n), \tag{2.55}$$

where $\widehat{\boldsymbol{\beta}}_{20}$ minimizes $l(\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_2)$ with respect to $\boldsymbol{\beta}_2$. The following result shows that a Wilks type $\chi^2$ approximation still holds.

**Corollary 2.5.7.** *Suppose that Conditions C1-C3 are satisfied. Then, under the null hypothesis, $T_{2n}$ converges in distribution to a $\chi^2_{(q_1)}$ random variable as $n \to \infty$.*

*Proof.* When only $\boldsymbol{\beta}_1$ is specified in the null hypothesis, we write the likelihood ratio statistic as the sum of two differences, each of which can be expanded in a manner similar to that in Theorem 2.5.6, and we have

$$\begin{aligned} T_{2n} &= \Big[2\sum_{i=1}^{n} \log\Big(1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0)\Big) - 2\sum_{i=1}^{n} \log\Big(1 + \widehat{\boldsymbol{\lambda}}_n^T(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{g}_{n,i}(\widehat{\boldsymbol{\beta}}_n)\Big)\Big] \\ &\quad - \Big[2\sum_{i=1}^{n} \log\Big(1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{g}_{n,i}(\boldsymbol{\beta}_0)\Big) - 2\sum_{i=1}^{n} \log\Big(1 + \widehat{\boldsymbol{\lambda}}_n^T(\boldsymbol{\beta}_{10}, \widehat{\boldsymbol{\beta}}_{20})\boldsymbol{g}_{n,i}(\boldsymbol{\beta}_{10}, \widehat{\boldsymbol{\beta}}_{20})\Big)\Big] \\ &= n^{1/2}(\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{1n})^T(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})n^{1/2}(\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{1n}) + o_p(1). \end{aligned}$$

The last equation comes from $\widehat{\boldsymbol{\beta}}_{20} - \boldsymbol{\beta}_{20} = \widehat{\boldsymbol{\beta}}_{2n} - \boldsymbol{\beta}_{20} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\widehat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) + o_p(1)$. Thus Corollary 2.5.7 holds because $n^{1/2}(\widehat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})$ converges in distribution to a $N(0, (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1})$ distribution under $\widetilde{H_0}$. □

Auxiliary information can be used to not only increase the precision of estimated treatment effects, but to also increase power in hypothesis testing. To evaluate power, we need to derive the asymptotic distribution of the test statistic under the alternative hypothesis. We shall consider the contiguous alternative

which deviates from the null by the order of $O(n^{-1/2})$; cf. Hajek, Sidak and Sen (1999) and Serfling (1980). For notational convenience, let $\boldsymbol{A} = \boldsymbol{D}_m^T \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}_m$ and write

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix},$$

where $\boldsymbol{A}_{ij} = E(\partial \boldsymbol{m}^T(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}_i) \widetilde{\boldsymbol{\Sigma}}^{-1} E(\partial \boldsymbol{m}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}_j^T)$, $i = 1, 2$ and $j = 1, 2$.

**Theorem 2.5.8.** *Suppose that Conditions C1-C3 are satisfied. Then under the sequence of contiguous alternatives* $H_a : \boldsymbol{\beta} = \boldsymbol{\beta}_a = \boldsymbol{\beta}_0 + \boldsymbol{h}/\sqrt{n}$, *the empirical likelihood ratio test statistic* $T_{1n}$ *converges in distribution to a noncentral* $\chi^2$ *with degrees of freedom* $q$ *and noncentrality parameter* $\boldsymbol{h}^T A \boldsymbol{h}$.

*Proof.* Following the same steps as in the proof of Theorem 2.5.3, we can show that

$$n^{1/2} \boldsymbol{A}^{-1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(\boldsymbol{A}^{-1/2} \boldsymbol{h}, \boldsymbol{I}).$$

Taking the Taylor series expansion of the empirical likelihood ratio test statistic at $\boldsymbol{\beta}_0$, we have

$$T_{1n} = n^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_a + \boldsymbol{h}/\sqrt{n})^T \boldsymbol{A}(\boldsymbol{\beta}_0) n^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_a + \boldsymbol{h}/\sqrt{n}) + o_p(1),$$

where the second equality comes from $\boldsymbol{\beta}_a = \boldsymbol{\beta}_0 + \boldsymbol{h}/\sqrt{n}$ being a sequence of contiguous alternatives. Therefore, $T_{1n} \to \chi_q^2$ with noncentrality parameter $\boldsymbol{h}^T \boldsymbol{A} \boldsymbol{h}$ as $n \to \infty$ under the alternative $H_a : \boldsymbol{\beta} = \boldsymbol{\beta}_a = \boldsymbol{\beta}_0 + \boldsymbol{h}/\sqrt{n}$. □

Similarly, the noncentrality parameter of the limiting $\chi^2$ distribution becomes the projected Fisher information when there are nuisance parameters.

**Corollary 2.5.9.** *Under the same assumptions as those in Theorem 2.5.8 and with $H_a$ replaced by $\widetilde{H_a} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1a} = \boldsymbol{\beta}_{10} + \boldsymbol{h}_1/\sqrt{n}$, the empirical likelihood ratio test statistic $T_{2n}$ in (2.55) converges in distribution to a noncentral $\chi^2$ with degrees of freedom $q_1$ and noncentrality parameter $\boldsymbol{h}_1^T(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})\boldsymbol{h}_1$.*

*Proof.* Similarly to the preceding proof, we have under the contiguous alternative

$$T_{2n} = n^{1/2}(\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{1n})^T(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})n^{1/2}(\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{1n}) + o_p(1).$$

Similarly to Theorem 2.5.3, we can show that when $\widetilde{H_a} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1a} = \boldsymbol{\beta}_{10} + \boldsymbol{h}_1/\sqrt{n}$ is true, $(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1/2}n^{1/2}(\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{1n})$ converges in distribution to $N((\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1/2}\boldsymbol{h}_1, \boldsymbol{I})$, which implies Corollary 2.5.9. □

It can be seen that the empirical likelihood approach reproduces the standard asymptotic results in parametric likelihood theory (Cox and Hinkley 1974). Similar to the estimation problem, adding more constraints will result in more powerful tests. When the number of constraints goes to infinity, the corresponding tests become asymptotically most powerful.

## 2.6 Numerical studies

In this section, we discuss computational issues arising from implementing the constrained optimization problems and report simulation results associated with the empirical likelihood based covariate adjustment method.

The primary step in computing the empirical likelihood is to maximize (2.20) subject to constraints (2.31). The Lagrangian is

$$\mathbb{P}_\star(p, \boldsymbol{\beta}, \boldsymbol{\lambda}, \gamma) = \sum_{i=1}^{n} \log_\star(p_i) + n\boldsymbol{\lambda}^T \sum_{i=1}^{n} p_i \boldsymbol{g}_r(\boldsymbol{\beta}; y_i, z_i, \boldsymbol{x}_i) + n\gamma(\sum_{i=1}^{n} p_i - 1),$$

where $\boldsymbol{\lambda}$ and $\gamma$ are the Lagrange multipliers. In practice we use a modified natural logarithm $\log_\star$ as defined in Owen (2001) instead of log. Thus, we obtain estimators for $p$ and $\boldsymbol{\beta}$ by differentiating $\mathbb{P}_\star$ with respect to $p$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\gamma$ and setting them to 0.

Working directly with $n+q+r+1$ free variables involves gradient and Hessian matrices of daunting dimensions. Alternatively we may use the two-stage Newton algorithm as discussed in Section 2.5.1 that can eliminate some parameters. Nonetheless, unlike the usual testing case where $\boldsymbol{\beta}$ is fixed at $\boldsymbol{\beta}_0$, the outer stage in the two-stage Newton algorithm, i.e. minimization over $\boldsymbol{\beta}$ while keeping $\boldsymbol{\lambda}$ fixed, is difficult in practice because of the possibility of a non-positive definite Hessian matrix. Zedlewski (2008) points out that "Concentrating out some parameters leads to a smaller optimization problem, but it can make it more difficult. Thus the two-stage Newton algorithm is fast but unreliable and can lead to frustrating convergence problems. In most cases $n$ is much greater than $q+r$, so the largest block of the Hessian is an $n \times n$ diagonal matrix.". Further, the largest block is a negative definite matrix as well. In our implementation, we use a Matlab package "matElike", which solves the primal problem by including modern optimization codes exploiting matrix sparsity. We find the package to be both robust and fast. The link to the Matlab package and the code to implement our method can be found at `http://www.stat.columbia.edu/~xwu/software.html`.

### 2.6.1 Estimation

The simulation results reported below are all based on 5000 Monte Carlo replications. The sample size is chosen to be 200 throughout. We consider the case of two treatment groups with the treatment indicator $Z$ generated with $P(Z=0) =$

$P(Z = 1) = 0.5$. The response variable $Y$ is binary with $\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 Z$. The parameter of interest is either $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ or $\beta_2$.

In the first scenario, the auxiliary covariate $X$ is generated as a one dimensional Normal random variable with mean 0 and different variances. The magnitude of the variance correlates with the influence of X on the response. Given $Z$ and $X$, $Y$ is then generated as Bernoulli according to $\text{logit}\{P(Y = 1|Z = g, X)\} = \alpha_{0g} + \alpha_g X$, where $\alpha_{00} = 0.3, \alpha_{01} = 1, \alpha_0 = 1, \alpha_1 = 1.5$ and $g = 0$ or 1.

From Table 3.1 we see that when the standard deviation of X is 2, the Monte Carlo standard errors gradually decrease and approach the optimal ones. From "marginal" to "5 Fourier", the standard errors drop significantly. However, additional constraints beyond "5 Fourier" do not appear to have much impact on further variance reduction. Note that a large number of additional constraints require substantially more computing time. Thus, we will only compare the results of "marginal" with "5 Fourier" in the other cases. A single (i.e., nonparallel) process that calculates the maximum empirical likelihood estimate and the p-value for testing the null hypothesis of no treatment difference takes, on average, less that 2 seconds to run for a data set of 200 samples using 5 constraints. The computation time is estimated using a 2.33GHz processor on a server with 8GB RAM.

Table 3.1 also shows that the means of Monte Carlo estimates differ from the true value of $\boldsymbol{\beta}$ at the third decimal place and the coverage probabilities are around 0.95. The Monte Carlo standard errors of estimates from five estimating equations are generally smaller than those from marginal models. The improvement becomes more pronounced when the variance of X becomes larger. Also, the average length of 95% Wald confidence intervals are smaller than those of marginal models.

In the second scenario, the link function is quadratic in X, i.e., $\text{logit}\{P(Y = 1|Z = g, X)\} = \alpha_{0g} + \alpha_g X^2$, with the same $\alpha_0 g$ and $\alpha_g$ values, $g = 0, 1$. From

Table 2.3: Bias and Standard Error Comparisons When Logit is Linear in X.

| Method | True $\boldsymbol{\beta}$ | MC Bias | OptStd | MC Std | CovProb | avlen |
|---|---|---|---|---|---|---|
| | | | $X \sim N(0, 0.5^2)$ | | | |
| marginal | 0.2832 | 0.0033 | 0.1992 | 0.2025 | 0.9520 | 0.7960 |
| | 0.6096 | 0.0063 | 0.2872 | 0.3007 | 0.9486 | 1.1801 |
| 5 Fourier | 0.2832 | 0.0036 | 0.1992 | 0.2027 | 0.9500 | 0.7870 |
| | 0.6096 | 0.0056 | 0.2872 | 0.2968 | 0.9468 | 1.1536 |
| | | | $X \sim N(0, 1^2)$ | | | |
| marginal | 0.2479 | 0.0010 | 0.1929 | 0.2025 | 0.9520 | 0.7940 |
| | 0.4634 | 0.0063 | 0.2585 | 0.2988 | 0.9472 | 1.1562 |
| 5 Fourier | 0.2479 | 0.0011 | 0.1929 | 0.1992 | 0.9496 | 0.7718 |
| | 0.4634 | 0.0049 | 0.2585 | 0.2812 | 0.9424 | 1.0785 |
| | | | $X \sim N(0, 2^2)$ | | | |
| marginal | 0.1814 | 0.0040 | 0.1800 | 0.1995 | 0.9526 | 0.7912 |
| | 0.2792 | 0.0003 | 0.2110 | 0.2951 | 0.9452 | 1.1324 |
| 5 Fourier | 0.1814 | 0.0043 | 0.1800 | 0.1873 | 0.9518 | 0.7337 |
| | 0.2792 | -0.0018 | 0.2110 | 0.2439 | 0.9418 | 0.9292 |
| 7 Fourier | 0.1814 | 0.0030 | 0.1800 | 0.1860 | 0.9494 | 0.7186 |
| | 0.2792 | 0.0008 | 0.2110 | 0.2341 | 0.9442 | 0.8846 |
| 9 Fourier | 0.1814 | 0.0032 | 0.1800 | 0.1857 | 0.9464 | 0.7101 |
| | 0.2792 | 0.0008 | 0.2110 | 0.2311 | 0.9384 | 0.8631 |
| 11 Fourier | 0.1814 | 0.0032 | 0.1800 | 0.1852 | 0.9448 | 0.7037 |
| | 0.2792 | 0.0007 | 0.2110 | 0.2293 | 0.9340 | 0.8490 |

NOTE: In all the tables, 'marginal' means using empirical likelihood method with 2 marginal estimating equations $Y - \phi(\beta_1 + \beta_2 Z)$ and $Z\Big(Y - \phi(\beta_1 + \beta_2 Z)\Big)$, while "5 Fourier" has three additional estimating equations $2Z - 1$, $\sqrt{2}(2Z-1)\sin[2\pi F_n(X)]$ and $\sqrt{2}(2Z-1)\cos[2\pi F_n(X)]$, where $F_n(X)$ is the empirical cumulative distribution function of X. MC Bias is Monte Carlo bias, OptStd is the asymptotic standard error obtained according to the sandwich formula, MC Std is the Monte Carlo standard error, CovProb is the coverage probability of 95% Wald confidence intervals and avlen is the average length of those confidence intervals.

Table 2.4: Bias and Standard Error Comparisons When Logit is Quadratic in X.

| Method | True $\boldsymbol{\beta}$ | MC Bias | OptStd | MC Std | CovProb | avlen |
|---|---|---|---|---|---|---|
| | | | $X \sim N(0, 0.5^2)$ | | | |
| marginal | 0.5298 | 0.0057 | 0.2059 | 0.2093 | 0.9516 | 0.8160 |
| | 0.7758 | 0.0094 | 0.3169 | 0.3266 | 0.9536 | 1.2683 |
| 5 Fourier | 0.5298 | 0.0061 | 0.2059 | 0.2088 | 0.9480 | 0.8090 |
| | 0.7758 | 0.0088 | 0.3169 | 0.3257 | 0.9524 | 1.2523 |
| | | | $X \sim N(0, 1^2)$ | | | |
| marginal | 0.9664 | 0.0106 | 0.2182 | 0.2307 | 0.9476 | 0.8845 |
| | 0.8105 | 0.0182 | 0.3466 | 0.3795 | 0.9494 | 1.4450 |
| 5 Fourier | 0.9664 | 0.0111 | 0.2182 | 0.2254 | 0.9448 | 0.8604 |
| | 0.8105 | 0.0156 | 0.3466 | 0.3648 | 0.9502 | 1.3800 |

Table 3.2, we see that the coverage probabilities are satisfactory and close to their nominal levels as in the first scenario. The biases are slightly larger, however, they are still small relative to the standard errors. As expected, the Monte Carlo standard errors and the average lengths of 95% Wald confidence intervals from five estimating equations are smaller than those from the two marginal ones.

In the third scenario, there are two auxiliary covariates $X_1$ and $X_2$ and the response Y is generated as $\text{logit}\{P(Y = 1|Z = g, \boldsymbol{X})\} = \alpha_{0g} + \alpha_{1g}X_1 + \alpha_{2g}X_2$, g=0,1, with $\alpha_{00} = 0.3, \alpha_{01} = 1, \alpha_{10} = 1, \alpha_{11} = 1.5, \alpha_{20} = 2, \alpha_{21} = 1.5$. The estimating equations for the marginal method remain the same since there is no covariate adjustment involved. Let $\kappa(Z) = \sqrt{2}(2Z - 1)$ and $W_k = 2\pi F_n(X_k)$, $k = 1, 2$. The empirical likelihood method with constraints, $\kappa(Z)$, $\kappa(Z)\sin(W_1)$, $\kappa(Z)\cos(W_1)$, $\kappa(Z)\sin(W_2)$, $\kappa(Z)\cos(W_2)$, except the marginal estimating equations is denoted by "7 Fourier". From Table 3.3, the performance of the estimates is similar to that in the previous two scenarios.

Table 2.5: Bias and Standard Error Comparisons When Logit Contains Two Covariates.

| Method | True $\boldsymbol{\beta}$ | MC Bias | OptStd | MC Std | CovProb | avlen |
|---|---|---|---|---|---|---|
| | | | $X_1 \sim N(0,1^2), X_2 \sim N(0,2^2)$ | | | |
| marginal | 0.1061 | -0.0003 | 0.1649 | 0.2005 | 0.9558 | 0.7883 |
| | 0.3157 | 0.0043 | 0.1828 | 0.2933 | 0.9494 | 1.1282 |
| 7 Fourier | 0.1061 | -0.0009 | 0.1649 | 0.1761 | 0.9526 | 0.6813 |
| | 0.3157 | 0.0051 | 0.1828 | 0.2311 | 0.9438 | 0.8716 |
| | | | $X_1 \sim N^2(0,1^2), X_2 \sim N(0,2^2)$ | | | |
| marginal | 0.4389 | 0.0063 | 0.1688 | 0.2032 | 0.9550 | 0.8069 |
| | 0.5493 | 0.0056 | 0.1985 | 0.3072 | 0.9486 | 1.2023 |
| 7 Fourier | 0.4389 | 0.0052 | 0.1688 | 0.1825 | 0.9494 | 0.7012 |
| | 0.5493 | 0.0041 | 0.1985 | 0.2490 | 0.9458 | 0.9396 |
| | | | $X_1 \sim N^2(0,0.5^2), X_2 \sim N^2(0,1^2)$ | | | |
| marginal | 1.4746 | 0.0149 | 0.2482 | 0.2594 | 0.9562 | 1.0201 |
| | 0.5813 | 0.0233 | 0.3857 | 0.4310 | 0.9498 | 1.6363 |
| 7 Fourier | 1.4746 | 0.0144 | 0.2482 | 0.2512 | 0.9518 | 0.9771 |
| | 0.5813 | 0.0224 | 0.3857 | 0.4126 | 0.9486 | 1.5485 |

NOTE: The logit is either quadratic ($X \sim N^2(\cdot,\cdot)$) or linear ($X \sim N(\cdot,\cdot)$) in each covariate.

Table 2.6: Power Comparison When Logit is Linear in X.

| | | | marginal | | 5 Fourier | |
|---|---|---|---|---|---|---|
| X | $\beta_{10}$ | $\beta_{20}$ | CovProb | Power | CovProb | Power |
| $N(0, 0.5^2)$ | 0.2125 | 0.8304 | 0.9498 | 0.7928 | 0.9492 | 0.8216 |
| $N(0, 1^2)$ | 0.1379 | 0.8207 | 0.9494 | 0.7826 | 0.9458 | 0.8682 |
| $N(0, 2^2)$ | 0.0386 | 0.8182 | 0.9486 | 0.7938 | 0.9436 | 0.9568 |

Table 2.7: Power Comparison When Logit is Quadratic in X.

| | | | marginal | | 5 Fourier | |
|---|---|---|---|---|---|---|
| X | $\beta_{10}$ | $\beta_{20}$ | CovProb | Power | CovProb | Power |
| $N(0, 0.5^2)$ | 0.8511 | 1.0599 | 0.9442 | 0.8428 | 0.9448 | 0.8498 |
| $N(0, 1^2)$ | 0.9662 | 0.9359 | 0.9464 | 0.7356 | 0.9482 | 0.7724 |

### 2.6.2 Testing

With the same data generating process as in the preceding subsection, the corresponding hypothesis testing results are presented in Tables 3.4, 2.7 and 2.8. In each scenario, the profile empirical likelihood ratio test is used to test the null hypothesis $\widetilde{H_0} : \beta_2 = 0$. CovProb denotes coverage probabilities for testing $\beta_2 = \beta_{20}$. We have the following observations. First, in all three tables, both coverage probabilities of the profile empirical likelihood ratio tests are close to the nominal 95% level. Second, the attained power from 5 estimating equations is larger than that from marginal estimating equations. Third, when X is one dimensional, the gain in power is more significant as the standard deviation of X increases. Note that in each scenario, $\beta_{10}$ and $\beta_{20}$ are the true values. The profile empirical likelihood ratio test is used to test the null hypothesis $\widetilde{H_0} : \beta_2 = 0$. CovProb are the coverage probabilities of tests $\beta_2 = \beta_{20}$.

Table 2.8: Power Comparison When Logit Contains Two Covariates.

| | | | marginal | | 7 Fourier | |
|---|---|---|---|---|---|---|
| $X_1$, $X_2$ | $\beta_{10}$ | $\beta_{20}$ | CovProb | Power | CovProb | Power |
| $N(0,1^2)$, $N(0,2^2)$ | 0.0694 | 0.8461 | 0.9488 | 0.8166 | 0.9430 | 0.9308 |
| $N^2(0,1^2)$, $N(0,2^2)$ | 0.2468 | 0.7012 | 0.9418 | 0.6584 | 0.9438 | 0.8636 |
| $N^2(0,0.5^2)$, $N^2(0,1^2)$ | 1.1701 | 0.8342 | 0.9496 | 0.5873 | 0.9478 | 0.6140 |

## 2.7 Application

We apply the proposed empirical likelihood based approach to the Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO)-I trial data, which were kindly provided to us by Karen Pieper from the Duke Clinical research Institute. The primary endpoint was 30-day death, which occurred in 6.29% of 10366 patients randomly assigned to tissue plasminogen activator (TPA) (g=1), 7.32% of 10354 patients randomly assigned to skreptokinase (SK) with IV heparin (g=2), 6.99% of 10303 patients randomly assigned to a combination of SK and TPA (g=3) and 7.24% of 9773 patients randomly assigned to SK with SQ heparin (g=4). Besides treatment assignment and outcome, some baseline auxiliary covariates concerning demographics (age, sex, weight, height), risk factors (hypertension, diabetes, smoking, hypercholesterolemia), other history (family history of MI, previous MI, previous angina, previous revascularization) and presenting characteristics (blood pressure, tachycardia, anterior infarct location, killip class, ST elevation on electrocardiography) were recorded on each subject. In Steyerberg et al. (2000), the relative prognostic strength of 17 baseline covariates was evaluated by their univariate $\chi^2$ model, which was calculated as the difference in -2 log-likelihood between a univariate logistic regression model with and without the characteristic. The strongest prognostic factor was age and this was further confirmed by the $R^2$

measure on the log-likelihood scale, which approximately indicated the percentage of variance explained (see Figure 2.1). Except calculating correlations, adjustment for important predictors such as age is always recommended in the case of short-term death after acute myocardial infarction. Thus, we will compare unadjusted and age-adjusted results for the four treatment groups.

The marginal model between the 30-day death (Y) and treatment assignment (Z) is given by $\text{logit}\{E(Y|Z)\} = \beta_1 + \beta_2 1_{(Z=2)} + \beta_3 1_{(Z=3)} + \beta_4 1_{(Z=4)}$. For the age(X) adjustment, we use 9 auxiliary constraints $(1_{(Z=g)} - 0.25)$, $(1_{(Z=g)} - 0.25)F_n(x)$ and $(1_{(Z=g)} - 0.25)F_n^2(x)$, $g = 2, 3, 4$, where $F_n(x)$ is the empirical c.d.f. of age.

The unadjusted estimates $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)$ are (-2.7014, 0.1630, 0.1129, 0.1517) with standard errors (0.04109, 0.05619, 0.05670, 0.05557). Estimates adjusted for age are (-2.7014, 0.1628, 0.1126, 0.1521) with standard errors (0.04109, 0.05619, 0.05670, 0.05556). The p-values for the unadjusted and adjusted hypothesis testing of $\beta_2 = \beta_3 = \beta_4 = 0$ are 0.0136 and 0.0135, respectively.

The unadjusted test is already significant, so the additional improvement in $p$-value after covariate adjustment only reconfirms the scientific conclusion. However, if the sample size were smaller, the change in $p$-value might be more consequential. For illustrative purposes, we randomly draw a subsample of size 20000 from the complete data and pretend that is what we had in reality. In one of these cases, the $p$-values for the unadjusted test of $\beta_2 = \beta_3 = \beta_4 = 0$ is 0.0391 while it becomes 0.0362 after adjusting for age. In another case, it changes from 0.0508 to 0.0458.

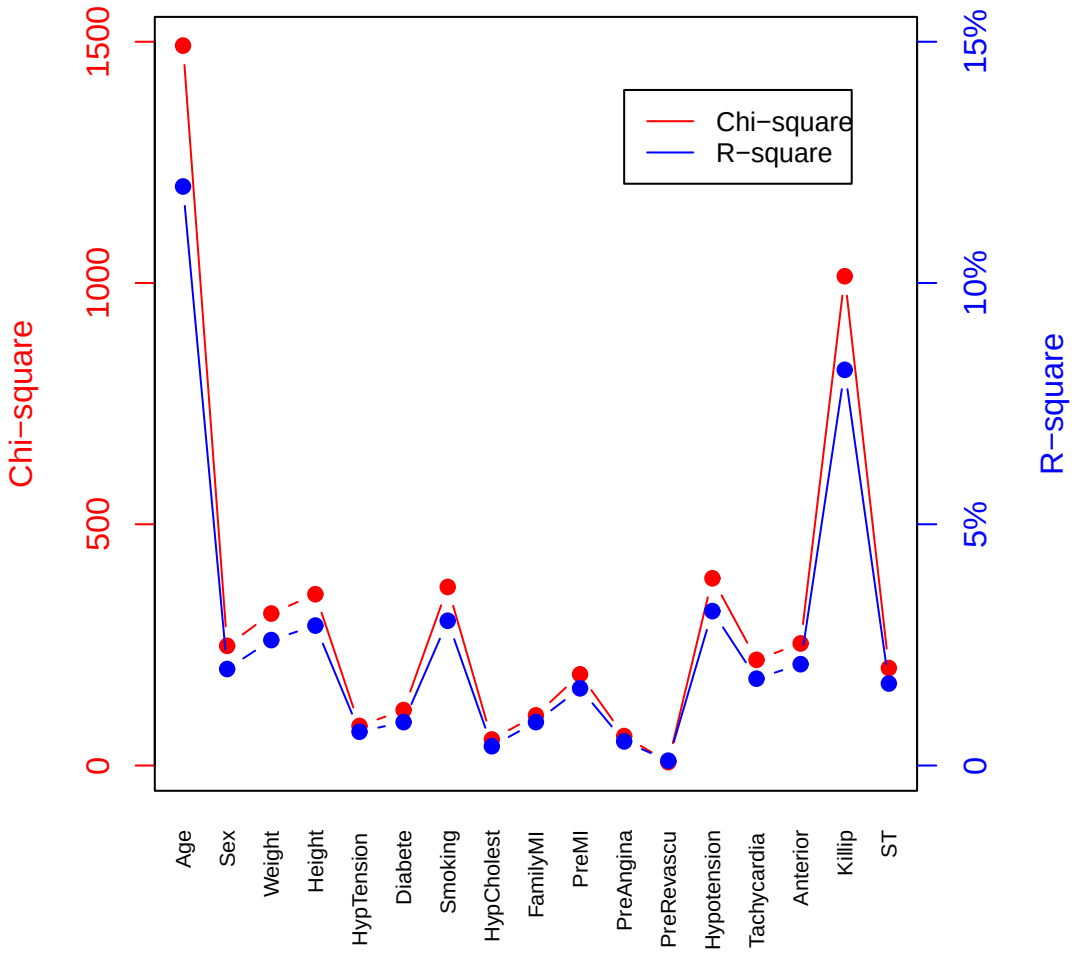



Figure 2.1: Relative prognostic strength of 17 baseline covariates.

# Chapter 3

# Multivariate $K$-sample problem

## 3.1 Introduction

Testing for differences between two populations is one of the classical problems in nonparametric inference. For univariate data, there exists an extensive literature (Hájek & Šidák, 1967; Hollander & Wolfe, 1973; Lehmann, 1975). A common feature in most approaches is that the test statistics are based on the ranks of the observations, which are invariant under monotone transformations, in the "pooled" sample that combines the sample data from each of the two populations. As a result, the test statistics are distribution free under the null hypothesis. An intrinsic difficulty in extending these tests to the multivariate case is the fact that there is no natural order on $p$-dimensional ($p \geq 2$) space, and monotone transformation of each coordinate does not necessarily lead to the uniform distribution on the unit $p$-cube. Thus, the resulting test statistics are no longer distribution free under the null hypothesis in the multivariate case. Since the multivariate case has been studied far less thoroughly, there has always been interest in finding nonparametric approaches. As a generalization of the univariate ordering of the pooled sample,

Friedman & Rafsky (1979) proposed a multivariate run test based on the minimal spanning tree of the observations. Based on the number of nearest neighbor type coincidences, Henze (1988) developed a procedure which can be implemented using an approximate permutation test. Liu & Singh (1993) introduced the concepts of multivariate median and data depth for nonparametric analysis of multivariate data.

More recently, by examining the number of points from different groups that fall into local neighborhoods, Hall & Tajvidi (2002) developed a permutation based approach to testing for group differences. Rosenbaum (2005), on the other hand, introduced the concept of optimal non-bipartite matching and proposed what he called the cross-match test. Both approaches make use of certain inter-point distances so that the high-dimensional structure is reduced to the univariate situation.

Survival analysis is one of the most active areas where nonparametric tests are the gold standard, and the properties of these tests are well understood. Most notably, the log-rank test and its weighted versions are commonly used for testing for treatment effects when survival time is the outcome variable. We refer to Mantel & Haenszel (1959), Mantel (1966), Gehan (1965), Peto & Peto (1972) and Prentice (1978) for some of the initial developments and Gill (1980) for the counting process and associated martingale representations of the weighted log-rank test statistics.

The nonparametric nature, the flexibility of the weight functions, the well-understood theoretical properties and the widely available software tools suggest the utility of using the class of weighted log-rank statistics to connect the problem of testing for differences in high-dimensional data to survival analysis. To that end, we propose a survival analysis approach to the multivariate $K$-sample problem, by converting multivariate data into survival data. Using this conversion, we are able to make use of powerful weighted log-rank tests to develop a class of nonparamet-

ric tests. This approach is simple to implement and to adapt to different space configurations.

We develop large sample properties for the proposed statistics by studying the underlying score processes. By centering the scores at their marginal intensities, we establish their weak convergence to Gaussian random fields under the null and contiguous alternative hypotheses. For the Kolmogorov-Smirnov-type and Cramér-von Mises-type tests, we also establish consistency against any fixed alternative. As a practical means to obtain approximate cutoff points, we propose a simulation based resampling method that is easy to implement and has rigorous theoretical justification.

The rest of the chapter is organized as follows. The proposed tests and their theoretical properties are given in the Section 3.2. Section 3.3 is devoted to simulation studies. In Section 3.4, the method is applied to two real data sets.

## 3.2 A class of weighted log-rank based test statistics

### 3.2.1 Weighted log-rank score process and test statistics

We consider first the case of two populations. Let $X_1, \ldots, X_{n_1}$ be a $p$-dimensional random sample from a population with distribution function $F_1$ and $X_{n_1+1}, \ldots, X_{n_1+n_2}$ be a second random sample from distribution $F_2$. Suppose that the null hypothesis of interest is

$$H_0 : F_1 = F_2, \tag{3.1}$$

and that the alternative hypothesis is its complement. For any fixed point $x$ in $\mathbb{R}^p$, let $T_i(x) = d(X_i, x)$, where $d$ is a certain distance metric on $\mathbb{R}^p$. Typically $d$ is taken to be the Euclidean distance, $d(X_i, x) = \|X_i - x\|$. Furthermore, define $T_i(x)$-induced counting process

$$N_i(x;t) = I(T_i(x) \leq t), \quad t \geq 0, \;\; i = 1, \ldots, n,$$

where $I(\cdot)$ denotes the indicator function and $n = n_1 + n_2$.

By converting the original observations $X_1, \ldots, X_n$ into $T_1(x), \ldots, T_n(x)$, we may regard $T_i$ as survival times and connect the two sample problem with the well-studied weighted log-rank tests for survival data. Define the weighted log-rank score process

$$U_w(x;t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t W_x(s) \Big\{ Z_i - \frac{\sum_{j=1}^n Z_j I(T_j(x) \geq s)}{\sum_{j=1}^n I(T_j(x) \geq s)} \Big\} dN_i(x;s), \qquad (3.2)$$

where $Z_i = I(i > n_1)$ and $W_x(s)$ is a weight function. A widely used class of weight functions in survival analysis corresponds to the $G$-$\rho$ class (Harrington and Fleming 1982) of weighted log-rank test statistics, where $W_x(t) = w(\hat{F}(x; t-))$, $w(u) = (1-u)^\rho$, $1 - \hat{F}(x; t-)$ is the Kaplan-Meier estimate of the survival function from $T_1(x), \ldots, T_n(x)$ and $\rho \geq 0$ is a tuning parameter. Let $\tilde{t}$ be an upper bound on the distances $T_i, i = 1, \ldots, n$. For a fixed $x$, $U_w(x; \tilde{t})$ is the usual log-rank test statistic for $\rho = 0$ and it becomes the Peto-Prentice extension of the Wilcoxon statistic when $\rho = 1$. The choice of $\rho$ depends on the projected alternatives: if the difference in the two corresponding hazards is more pronounced for smaller $t$ values, then a larger $\rho$ is preferred. Thus, if the group difference lies in "local features", the Peto-Prentice statistic achieves greater efficiency. For any fixed $x$,

since (3.1) implies that the $T_i$ have a common distribution, it follows from the usual counting process-martingale approach to survival analysis (Gill, 1980; Flemming & Harrington, 2005; Anderson et al., 1993) that $U_w(x;t)$ is a zero-mean martingale in $t$ for a suitable $\sigma$-filtration under $H_0$. Viewed as a two-parameter process (of $x$ and $t$), Theorem 3.2.1 below shows that $U_w$ converges weakly to a zero-mean Gaussian random field on $\mathbb{R}^{p+1}$. Moreover, we will use the notation $\widehat{\Gamma}_k(x;t) = n^{-1}\sum_{i=1}^n Z_i^k I(T_i(x) \geq t)$ and $\Gamma_k(x;t) = \lim_{n\longrightarrow\infty} \widehat{\Gamma}_k(x;t)$, $k = 0, 1$ throughout.

**Theorem 3.2.1.** *Under $H_0 : F_1 = F_2$, the weighted log-rank score process $U_w(\cdot;\cdot)$ converges weakly to a zero-mean Gaussian random field $G_0(\cdot;\cdot)$ with covariance function $C(x_1,t_1;x_2,t_2) =$*

$$\lim_{n\longrightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}\int_0^{t_1} W_{x_1}(s)\Big\{Z_i - \frac{\Gamma_1(x_1;s)}{\Gamma_0(x_1;s)}\Big\}dM_i(x_1;s)\times\int_0^{t_2} W_{x_2}(s)\Big\{Z_i - \frac{\Gamma_1(x_2;s)}{\Gamma_0(x_2;s)}\Big\}dM_i(x_2;s),$$

*where $M_i(x;t) = N_i(x;t) - \int_0^t I(T_i(x) \geq s)d\Lambda(x;s)$ and $\Lambda(x;s)$ is the common cumulative hazard function of $T_i(x)$.*

*Proof.* Under $H_0$, for each fixed $x$, $T_i(x), i = 1,\ldots,n$, are iid random variables. By the standard counting processes theory, $M_i(x;t), i = 1,\ldots,n$, as processes in $t$, are zero-mean martingales. Hence, we can write

$$U_w(x;t) = n^{-1/2}\sum_{i=1}^{n}\int_0^{t} W_x(s)\Big\{Z_i - \frac{\hat{\Gamma}_1(x;s)}{\hat{\Gamma}_0(x;s)}\Big\}dM_i(x;s),$$

which, again for each fixed $x$, is a martingale with respect to $t$. Similarly, another process in $t$,

$$n^{-1/2}\sum_{i=1}^{n}\int_0^{t} W_x(s)\Big\{\frac{\hat{\Gamma}_1(x;s)}{\hat{\Gamma}_0(x;s)} - \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)}\Big\}dM_i(x;s) \tag{3.3}$$

is also a martingale. By the law of large numbers, the predictable variation of (3.3)

converges to 0. Therefore (3.3) is of order $o_p(1)$ and consequently,

$$U_w(x;t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t W_x(s) \Big\{ Z_i - \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)} \Big\} dM_i(x;s) + o_p(1). \tag{3.4}$$

The first term on the right-hand side of (3.4) is a sum of independent zero-mean random variables. By the classical multivariate central limit theorem, $U_w(\cdot;\cdot)$ converges in finite dimensional distributions to a Gaussian random field, whose covariance function is given by $C(x_1, t_1; x_2, t_2)$ in Theorem 3.2.1.

It remains to prove the "tightness" of $U_w(\cdot;\cdot)$. This can be done by applying modern empirical process theory, as given in Pollard (1990) and van der Vaart and Wellner (2000). □

Using (3.2) as the basic vehicle, we can construct a variety of test statistics. The main idea is to combine the weighted log-rank statistics at different $x$ locations and to introduce censoring to the observed survival times so that the resulting test statistic has more robust power. Two common ways to summarize over $x$ correspond to the Kolmogorov-Smirnov (K-S) sup-type and the Cramér-von Mises (C-vM) integral-type approaches. To those ends, we propose the following statistics: $U_1 = \sup_{x \in D} |U_w(x;\tilde{t})|$ and $U_2 = \int_{x \in D} U_w^2(x;\tilde{t}) \pi(x,\tilde{t}) dx$, where $D$ is a suitably chosen subset of $\mathbb{R}^p$ and $\pi(x,t)$ is a weight function. Here $\tilde{t}$ is typically an upper bound, which is commonly used in survival analysis to control possible tail instability. Let $T_{(1)}(x), \ldots, T_{(n)}(x)$ denote the order statistics from $\{T_1(x), \ldots, T_n(x)\}$. We may set $\tilde{t}$ to be the $k$th order statistic $T_{(k)}(x)$, mimicking the type II censoring in survival analysis. The use of $T_{(k)}(x)$ localizes the discrepancy between the two populations. For example, if $x$ is surrounded locally by the first group of $X$'s, then, with a small $k$, $U_w(x; T_{(k)}(x))$ tends to take a negative value. Taking a small $k$ is in the spirit of $k$-nearest neighbor method in nonparametric regression. This

idea can be generalized to other censoring schemes to accommodate different data patterns.

The weak convergence of $U_w$ can be used to derive limiting distributions of $U_1$ and $U_2$, which are functionals of $U_w$. The following corollary can be shown by applying continuous mapping theorem to Theorem 3.2.1.

**Corollary 3.2.2.** *Let $G_0$ be the Gaussian random field introduced in Theorem 1. Then, under $H_0 : F_1 = F_2$, $U_1$ and $U_2$ converge in distribution to $\sup_{x\in D}|G_0(x;\tilde{t})|$ and $\int_{x\in D} G_0^2(x;\tilde{t})\pi(x,\tilde{t})dx$, respectively.*

REMARK. It is intuitive that selection of $D$ should not be beyond the support of the underlying multivariate distributions. Choosing $D$ to be the support is asymptotically equivalent to taking $D$ to be the observed sample points, due to the weak convergence of $U_w$. The latter enables us to avoid maximizing a statistic over an unbounded high dimensional space, which could be difficult and time consuming. This is especially so when the dimension of $X$ is much larger than the sample size $n$. Note that the zero-mean property of $U_w$ under the null hypothesis remains the same, irrespective of the dimension.

Because of their intractable forms, the limiting distributions in Corollary 3.2.2 do not immediately lead to the cutoff points for the corresponding tests. An alternative way is to simulate replicates of processes $U_w^*$, which have asymptotically the same limiting distribution as that of $U_w$, thereby constructing respective functionals of $U_w^*$ that numerically approximate the distribution of $U_1$ and $U_2$. Such an approximation is theoretically justified by the following result.

**Theorem 3.2.3.** *Let $U_w^*(x;t) = n^{-1/2}\sum_{i=1}^n V_i \int_0^t W_x(s)\Big\{Z_i - \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)}\Big\}d\widehat{M}_i(x;s)$, where $\widehat{M}_i(x;t)$ is the same as $M_i(x;t)$ except with $\Lambda(x;t)$ being replaced by the Nelson-Aalen estimator (Anderson et al., 1993) and $V_i$ are independent standard*

*normal random variables that are independent of* $\{(X_i, Z_i)\}_{i=1}^n$*, the observed data. Then, the conditional distribution of* $U_w^*$ *given data* $\{(X_i, Z_i)\}_{i=1}^n$ *converges to the same limiting Gaussian random field* $G_0$ *as that of* $U_w$*.*

*Proof.* Conditioning on the data $\{(X_i, Z_i)\}_{i=1}^n$, $\{V_i\}_{i=1}^n$ are the only random components in $U_w^*$. Since $V_i$ are generated to be iid standard normals and independent of the data, it follows from the central limit theorem and a straightforward covariance calculation that $U_w^*$ converges in finite dimensional distributions to a zero-mean Gaussian process with the same covariance function $C(x_1, t_1; x_2, t_2)$ defined in Theorem 3.2.1. Thus, it suffices to prove the tightness of $U_w^*$, which can be done by making use of the functional central limit theorem (Pollard 1990); see Lin et al. (2000) for arguments in a more general case. □

Note that the conditional distribution of $U_w^*$ given $\{(X_i, Z_i)\}_{i=1}^n$ can be simulated by repeatedly generating random sequences $\{V_i\}_{i=1}^n$. Therefore, we can approximate the distribution of any functional of $U_w$ by that of the corresponding functional of $U_w^*$, which can be obtained via simulations. This random weighting method is computationally efficient because $\int_0^t W_x(s)\Big\{Z_i - \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)}\Big\}d\widehat{M}_i(x;s), i = 1, \ldots, n$, are fixed at their observed values for each sample of $U_w^*$. We refer to Lin et al. (2002) for a comprehensive discussion of such an approach and related methods. Specifically, to obtain the cutoff value for $U_1$ or $U_2$, we can either use a permutation test by randomly dividing $n$ observations into two groups of sizes $n_1$ and $n_2$ or utilize the simulation based resampling by repeatedly generating $U_1^* = \sup_{x\in D}|U_w^*\{x;\tilde{t}\}|$ or $U_2^* = \int_{x\in D}\{U_w^*(x;\tilde{t})\}^2\pi(x,\tilde{t})dx$ when the sample size is relatively large.

### 3.2.2 Extension to the $K$-sample case

Suppose for each $k$, $\{X_{ki}\}_{i=1}^{n_k}$ is a random sample from a population with distribution function $F_k$ on $\mathbb{R}^p$, $k = 1, \ldots, K$. We wish to test the null hypothesis that $F_1 = \ldots = F_K$ against the complementary alternative. Similarly to the two-sample case ($K = 2$), for each fixed point $x \in \mathbb{R}^p$, let $T_{ki}(x) = d(X_{ki}, x)$ and $N_{ki}(x;t) = I(T_{ki}(x) \leq t)$. In the $K$-sample case, the weighted log-rank score process is $U_w = (U_{w1}, \ldots, U_{w(K-1)})^T$, where

$$U_{wk}(x;t) = n^{-1/2} \int_0^t W_x(s)\{dN_k(x;s) - \frac{\Gamma_k(x;s)}{\Gamma_\cdot(x;s)} dN_\cdot(x;s)\},$$

where $N_k(x;t) = \sum_{i=1}^{n_k} N_{ki}(x;t)$, $N_\cdot(x;t) = \sum_{k=1}^{K} N_k(x;t)$, $\Gamma_k(x;t) = \sum_{i=1}^{n_k} I(T_{ki}(x) \geq t)$ and $\Gamma_\cdot(x;t) = \sum_{k=1}^{K} \Gamma_k(x;t)$. Under some mild regularity conditions, we have the following result.

**Theorem 3.2.4.** *Under $H_0 : F_1 = F_2$, the weighted log-rank score process $U_w$ converges weakly to a zero-mean multivariate Gaussian random field with covariance matrix $C(x_1, t_1; x_2, t_2)$, whose $(l, j)$th element is*

$$\lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^{t_1} W_{x_1}(s)\{\delta_{kl} - \frac{\Gamma_l(x_1;s)}{\Gamma_\cdot(x_1;s)}\} dM_{ki}(x_1;s) \times \int_0^{t_2} W_{x_2}(s)\{\delta_{kj} - \frac{\Gamma_j(x_2;s)}{\Gamma_\cdot(x_2;s)}\} dM_{ki}(x_2;s),$$

*where $n = \sum_{k=1}^{K} n_k$, $\delta_{kl} = I(k = l)$, $M_{ki}(x;t) = N_{ki}(x;t) - \int_0^t I(T_{ki}(x) > s) d\Lambda(x;s)$ and $\Lambda(x;t)$ is the common cumulative hazard function of $T_{ki}(x)$.*

For statistics defined as functionals of $U_w$, it is again difficult to obtain analytic forms of their limiting distributions. We therefore propose to use a simulation based random weighting method to approximate the limiting distributions and to obtain cutoff points. For theoretical justification, we need the following generalization of Theorem 3.2.3.

**Theorem 3.2.5.** *Let* $U^*_{wl}(x;t) = n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k} V_{ki}\int_0^t W_x(s)\{\delta_{kl} - \frac{\Gamma_l(x;s)}{\Gamma_\cdot(x;s)}\}d\hat{M}_{ki}(x;s)$, *where* $\hat{M}_{ki}(x;t)$ *is the same as* $M_{ki}(x;t)$ *with* $\Lambda(x;t)$ *replaced by the Nelson-Aalen estimator,* $V_{ki}$ *are independent standard normals and are independent of the data. Then, the conditional distribution of* $U^*_w = (U^*_{w1}, \ldots, U^*_{w(K-1)})^T$ *given the data converges to the same limiting multivariate Gaussian random field as in Theorem 3.2.4.*

### 3.2.3 Asymptotic properties under alternative hypotheses

In this section, we establish asymptotic properties for the proposed test statistics under alternative hypotheses. We first consider the case of a fixed alternative and show that the Kolmogorov-Smirnov-type and von Mises-type tests are consistent against any such alternative.

**Theorem 3.2.6.** *Under any fixed alternative* $F_2 \neq F_1$, *for any sequence of random variables* $\{c_n : n \geq 1\}$ *converging to a constant* $c > 0$, *we have*

$$\lim_{n\longrightarrow\infty} P\Big(\sup_{x\in\mathbb{R}^p, t\in[0,\infty)} |U_w(x;t)| > c_n\Big) = 1$$

*and if,* $\pi(x,t) > 0$ *for all* $x$ *and* $t$,

$$\lim_{n\longrightarrow\infty} P\Big(\int_{x\in\mathbb{R}^p}\int_{t\geq 0} U^2_w(x;t)\pi(x,t)dtdx > c_n\Big) = 1.$$

*Proof.* Based on the fact that there exists a $t \in [0,\infty)$ making the weighted log-rank test consistent, it suffices to show that if $X$ and $Y$ differ in distribution, then there exists an $x$ such that $d(X,x)$ and $d(Y,x)$ follow different distributions. We prove this by contradiction. Suppose for any $x$, $d(X,x)$ and $d(Y,x)$ have the same

distribution, then for any $\Delta > 0$,

$$\frac{P(d(X,x) \le \Delta)}{\Delta^p} = \frac{P(d(Y,x) \le \Delta)}{\Delta^p}.$$

Letting $\Delta$ go to zero on both sides, we know that $f_X(x) = f_Y(x)$ at any $x$, where $f_X$ and $f_Y$ are respective densities. This contradicts the assumption. □

REMARK Because $U_w$ is standardized and converges weakly, the critical values for these tests, for example those obtained via the random weighting method discussed in §3.2.1, should converge in probability to constants under the null and under the alternative. Therefore, the theorem above shows that the two types of tests are consistent when summarized over all $x$ and $t$. However, for a single pair of $x$ and $t$, it is not hard to see that $U_w(x,t)$ itself may not have any power. For instance, if the first population follows $N_p(\mu, \Sigma)$ and the second follows $N_p(-\mu, \Sigma)$ and $x$ is fixed at the origin, then $U_w(x,t)$ will have no power due to symmetry.

It is customary to study the asymptotic power of test statistics by considering contiguous alternatives. We next derive the limiting distributions of the proposed tests under such contiguous alternatives. Following Hájek and Šidák (1967), let $P_n$ be the sequence probability measures representing the null and $Q_n$ its contiguous alternatives. Denote by $\frac{dQ_n}{dP_n}$ the Radon-Nikodym derivative between the two measures.

**Theorem 3.2.7.** *Suppose that, for any $x$ and $t$, the random vector $\left(U_w(x;t), \log \frac{dQ_n}{dP_n}\right)$ converges in distribution to $N\left(\begin{pmatrix} 0 \\ -0.5\sigma^2 \end{pmatrix}, \begin{pmatrix} C(x;t) & \tau(x;t) \\ \tau(x;t) & \sigma^2 \end{pmatrix}\right)$ under $P_n$. Then, under $Q_n$, $U_w(x;t)$ converges weakly to a Gaussian random field $G_Q(x;t)$ with mean $\tau(x;t)$ and covariance function $C(x;t)$.*

*Proof.* By Le Cam's third lemma (Hájek and Šidák 1967), $U_w(x;t)$ converges in

finite dimensional distribution to $G_Q(x;t)$ under $Q_n$. Moreover, we can show that $\left(U_w(x;t), \log \frac{dQ_n}{dP_n}\right)$ is tight by applying empirical process theory, as given in Pollard (1990) and van der Vaart and Wellner (2000). □

**Corollary 3.2.8.** *Under the assumption in Theorem 3.2.7, we have for any* $\tilde{t}$

(i) $\sup_{x\in D}|U_w(x;\tilde{t})|$ converges in distribution to $\sup_{x\in D}|G_Q(x;\tilde{t})|$.
(ii) $\int_{x\in D} U_w^2(x;\tilde{t})\pi(x,\tilde{t})dx$ converges in distribution to $\int_{x\in D} G_Q^2(x;\tilde{t})\pi(x,\tilde{t})dx$.

Corollary 3.2.8 can be shown by applying continuous mapping theorem to Theorem 3.2.7. Comparing the limiting distribution in Theorem 3.2.7 with that in Theorem 3.2.1, we know that the power of the respective tests is governed by the mean drift $\tau(x;t)$ of the Gaussian random field. To gain some insight into the power properties, we calibrate $\tau(x;t)$ in two concrete examples. Let $\delta_i = I(\|X_i - x\| \le t)$, $p_z = P(Z = 1)$ and let $H_x$ be the distribution function of $\|X - x\| \wedge t$.

*Example 1.* Consider two populations differing in location:

$$H_1 : X \sim f(\cdot - n^{-1/2}cZ),$$

where $f$ is the density function of $X$ under the null and $c$ is a $p$-vector of constants. The alternative is clearly contiguous to $H_0$ (Hájek and Šidák 1967). It can be shown that the drift function has expression

$$\tau(x;t) = p_z(1-p_z)E[\{\delta\phi \circ H_x(\|X-x\|) + (1-\delta)\Phi \circ H_x(t)\}c^T f'(X)/f(X)],$$

where $\phi(u) = w(u) - \int_0^u w(v)/(1-v)dv$ and $\Phi(u) = \phi(u) - w(u)$. In particular, if $f$ is the density for $N_p(\mu, I)$, where $I$ is the identity matrix, and $t = +\infty$, then $\tau(x;t) = -p_z(1-p_z)E\{\phi \circ H(\|X-x\|)c^T(X-\mu)\}$. When $\mu = 0$ and $D = \{0\}$, i.e., if under the null hypothesis, $X$ follows a $p$-variate normal distribution with

mean zero and if we construct our test statistics using $D = \{0\}$, it is easily seen that $\tau(x;t) = 0$. Thus, the corresponding tests have no power. This is somewhat surprising since the two populations do differ by a contiguous location shift. This phenomenon is also validated by simulation studies in section 3.3.2.

*Example 2.* We next consider two populations differing in scale:

$$H_2 : X \sim \exp(-n^{-1/2} c^T 1_p Z) f(\exp(-n^{-1/2} Zc) * \cdot),$$

where $1_p$ is a $p$-vector consisting of all ones and $\vec{a} * \vec{b} = (a_1 b_1, \ldots, a_p b_p)$ for $p$-vectors $\vec{a} = (a_1, \ldots, a_p)^T$ and $\vec{b} = (b_1, \ldots, b_p)^T$. We can get

$$\tau(x;t) = 2p_z(1-p_z)E[\{\delta\phi \circ H_x(\|X - x\|) + (1-\delta)\Phi \circ H_x(t)\}(c * X)^T f'(X)/f(X)].$$

If we let $t = +\infty$, $f$ be the density of $N_p(0, \sigma_0^2 I)$ and $c = c_0 1_p$, then it simplifies to

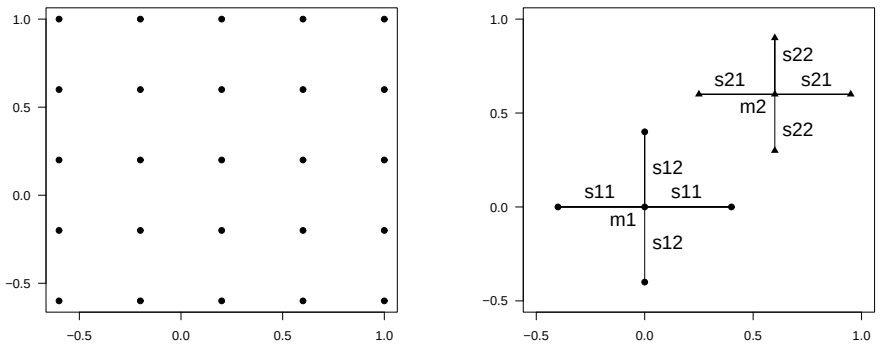
$$\tau(x;t) = -2p_z(1-p_z)c_0\sigma_0^{-2}E\{\phi \circ H_x(\|X - x\|)\,\|X\|^2\}.$$

In this case, simulation results in section 3.3.2 show that if we let $D$ contain only one point $x$, then $x = 0$ gives the highest power and the farther $x$ moves away from 0, the lower the power is. In Figure 3.3(d), we can see that when $x = 0$, the proposed tests have approximately the same power as the likelihood ratio test, which is the asymptotically most powerful one.

## 3.3 Simulations

We conducted extensive simulations to assess the performance of the proposed method. The results reported here are based on 1000 Monte Carlo data sets. For

simplicity, we only consider the case of two groups of data points on $\mathbb{R}^2$. Two approaches, fixed and data dependent, are used for choosing $D$, defined in the previous section. For the former, $D$ is set to be $\{(-1+0.4i, -1+0.4j), i, j = 1, \ldots, 5\}$ (see Figure. 3.1(a)). For the latter, $D$ is data dependent and we use ten points $\{m_g + (0, js_{g2}), m_g + (is_{g1}, 0), g = 1, 2, i, j = -1, 0, 1\}$, where $m_g$ is the sample mean of the $g$th group, $s_{gr}$ is the standard deviation of the $g$th group in the $r$th direction (see Figure. 3.1(b)). The definition of $\rho$ in Tables 1-4 is the same as that in the $G$-$\rho$ class of weighted log-rank tests discussed in §3.2.1. In the simulation study, we explore both small and large sample cases. When the total sample size is small, we apply permutation tests and when it is large, we use the random weighting method.




(a) (b)

Figure 3.1: Graphs showing (a) the fixed $D$, and (b) the data dependent $D$ (•: group 1; ▲: group 2).

### 3.3.1 Permutation-based tests

We simulate 15 samples for each group and use 5000 permutations to obtain the cutoff value. When $k$ is small, there are many ties among the 5000 test statistics.

Thus, randomization is applied to deal with ties on the boundary of the rejection region. For example, suppose $C$ is the cutoff value obtained from 5000 permutations and there might be many test statistics equal to $C$. Then we reject the null hypothesis if the observed test statistic is greater than $C$ and we accept the null if it is smaller than $C$. If the observed test statistic is equal to $C$, then we reject the null with probability $q$, where $q$ is calculated as $(0.05 \times 5000 - C_{>})/C_{=}$, with $C_{>}$ being the number of test statistics among the 5000 permutations that are greater than $C$ and $C_{=}$ being the number equaling to $C$. The $g$th group of observations is simulated with distribution function $F_g$, which is specified as follows

$$F_1 = N\Big(\tbinom{0}{0}, \tbinom{1}{0.5}\ \tbinom{0.5}{1}\Big) \quad H_1 : F_2 = N\Big(\tbinom{1}{1}, \tbinom{1}{0.5}\ \tbinom{0.5}{1}\Big) \quad H_2 : F_2 = N\Big(\tbinom{0}{0}, \tbinom{4}{0.5}\ \tbinom{0.5}{4}\Big)$$

Here, $H_0$ is the null hypothesis, $H_1$ and $H_2$ are the alternative hypotheses with location shift and scale difference, respectively.

From Table 3.1, when $D$ is fixed, the type-I error rate is close to its nominal level 0.05. For data dependent $D$, the type-I error is inflated to be around 0.06, which is understandable because the sample size for each group is not large enough to secure a stable $D$ for each Monte Carlo data set. Under the alternatives, the performance of Cramér-von Mises-type integral tests is generally slightly better than the K-S-type sup test. For each fixed $k$, the power of the test increases with the increase of $\rho$ (see Table 3.2 and 3.3). This is because as $\rho$ increases, observations close to points in $D$ are given more weight in the test statistics; thus, the procedure becomes more powerful at detecting the local discrepancy between the two groups. When $\rho$ is fixed, the relationship between power and the value of $k$ is not necessarily monotone. Since the determination of $k$ depends on the pattern of the observations, there is no unique approach for determining the optimal censoring

Table 3.1: Type I error

| | Fixed $D$ | | | | | | Data dependent $D$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Integral | | | Sup | | | Integral | | | Sup | | |
| | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ |
| $n = 30$ | | | | | | | | | | | | |
| $k = 10$ | 0.052 | 0.053 | 0.051 | 0.048 | 0.054 | 0.050 | 0.480 | 0.625 | 0.635 | 0.421 | 0.609 | 0.674 |
| $k = 20$ | 0.051 | 0.057 | 0.049 | 0.049 | 0.059 | 0.057 | 0.486 | 0.602 | 0.637 | 0.416 | 0.578 | 0.623 |
| $k = 30$ | 0.053 | 0.055 | 0.056 | 0.050 | 0.054 | 0.056 | 0.488 | 0.585 | 0.593 | 0.401 | 0.542 | 0.571 |
| $n = 200$ | | | | | | | | | | | | |
| $k = 50$ | 0.048 | 0.047 | 0.053 | 0.045 | 0.044 | 0.050 | 0.058 | 0.052 | 0.050 | 0.054 | 0.039 | 0.044 |
| $k = 100$ | 0.048 | 0.050 | 0.052 | 0.047 | 0.043 | 0.041 | 0.053 | 0.050 | 0.049 | 0.058 | 0.038 | 0.043 |
| $k = 200$ | 0.048 | 0.050 | 0.048 | 0.050 | 0.044 | 0.045 | 0.056 | 0.042 | 0.048 | 0.059 | 0.044 | 0.042 |

Table 3.2: Power under $H_1$

| | Fixed $D$ | | | | | | Data dependent $D$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Integral | | | Sup | | | Integral | | | Sup | | |
| $n = 30$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ |
| $k = 10$ | 0.278 | 0.370 | 0.351 | 0.247 | 0.479 | 0.551 | 0.480 | 0.625 | 0.635 | 0.421 | 0.609 | 0.674 |
| $k = 20$ | 0.271 | 0.375 | 0.390 | 0.251 | 0.473 | 0.518 | 0.486 | 0.602 | 0.637 | 0.416 | 0.578 | 0.623 |
| $k = 30$ | 0.257 | 0.344 | 0.359 | 0.247 | 0.420 | 0.439 | 0.488 | 0.585 | 0.593 | 0.401 | 0.542 | 0.571 |
| $n = 200$ | | | | | | | | | | | | |
| $k = 50$ | 0.484 | 0.740 | 0.912 | 0.466 | 0.814 | 0.979 | 0.714 | 0.882 | 0.973 | 0.639 | 0.850 | 0.974 |
| $k = 100$ | 0.475 | 0.708 | 0.866 | 0.475 | 0.806 | 0.940 | 0.715 | 0.868 | 0.943 | 0.646 | 0.849 | 0.943 |
| $k = 200$ | 0.465 | 0.675 | 0.763 | 0.458 | 0.773 | 0.871 | 0.717 | 0.844 | 0.904 | 0.647 | 0.823 | 0.902 |

Table 3.3: Power under $H_2$

| | Fixed $D$ | | | | | Data dependent $D$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Integral | | | Sup | | Integral | | | Sup | |
| $n = 30$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ |
| $k = 10$ | 0.722 | 0.874 | 0.916 | 0.577 | 0.837 | 0.880 | 0.678 | 0.872 | 0.913 | 0.528 | 0.783 | 0.844 |
| $k = 20$ | 0.708 | 0.854 | 0.884 | 0.577 | 0.798 | 0.839 | 0.670 | 0.850 | 0.887 | 0.530 | 0.740 | 0.779 |
| $k = 30$ | 0.695 | 0.820 | 0.832 | 0.562 | 0.739 | 0.760 | 0.657 | 0.800 | 0.826 | 0.504 | 0.702 | 0.716 |
| $n = 200$ | | | | | | | | | | | | |
| $k = 50$ | 0.542 | 0.588 | 0.595 | 0.481 | 0.570 | 0.594 | 0.827 | 0.961 | 0.993 | 0.668 | 0.912 | 0.984 |
| $k = 100$ | 0.541 | 0.583 | 0.593 | 0.478 | 0.565 | 0.590 | 0.823 | 0.953 | 0.988 | 0.678 | 0.904 | 0.974 |
| $k = 200$ | 0.540 | 0.579 | 0.587 | 0.477 | 0.556 | 0.569 | 0.817 | 0.945 | 0.976 | 0.679 | 0.886 | 0.936 |

time. But in this simulated example $k = 10$ gives us the best overall performance among the three. Under $H_2$, the power of our method (see Table 3.3) are generally higher than those reported in Table 2 (between 0.5 and 0.65) in Hall and Tajvidi (2002).

### 3.3.2 Random weighting

To validate the large sample properties of the proposed tests, we generate 100 samples for each group. For each Monte Carlo data set, the standard normal variables $\{V_i, i = 1, \ldots, n\}$ are generated $10^5$ times to get the cutoff value. Similarly, we consider alternatives with location shift and scale difference:

$$F_1 = N\Big(\binom{0}{0}, \binom{1}{0.5}\ \binom{0.5}{1}\Big) \quad H_1 : F_2 = N\Big(\binom{0.6}{0.6}, \binom{1}{0.5}\ \binom{0.5}{1}\Big) \quad H_2 : F_2 = N\Big(\binom{0}{0}, \binom{2}{0.5}\ \binom{0.5}{2}\Big)$$

From Table 3.1, the type-I error rate is around 0.05 for both fixed and data dependent $D$. When the sample size is relatively large, by virtue of the law of large numbers, the impact of random $D$ on the test statistics becomes negligible. Here, although the data dependent $D$ contains less points than the fixed $D$, which saves computational time, the power associated with the data dependent $D$ is generally much higher than that associated with the fixed $D$ (see Table 3.2 and 3.3). The higher power with the data dependent $D$ is due to its more advantageous and more efficient position in detecting the location shift or scale difference compared to the fixed $D$. For the same $k$ or $\rho$, the power pattern is similar to that in the small sample case.

Next, we investigate the power properties of the two examples given at the end of §3.2.3. To avoid the complexity caused by the varying values of $x$ and $t$, we only take one spatial point $x$ and let $t$ be large enough so that no censoring

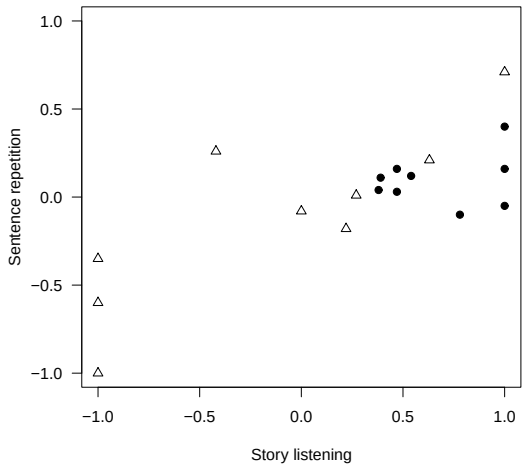



Figure 3.2: A scatter plot showing the laterality index for story listening and sentence repetition for patients(△) and controls(●).

is involved. We simulate 100 bivariate normal samples for each group. Since we know the truth in simulations, we can construct a likelihood ratio test of a simple alternative against a simple null, which is the optimal test by Neyman-Pearson lemma. We compare the power performance of the likelihood ratio test, the log-rank test ($\rho = 0$) and the Wilcoxon test ($\rho = 1$) in all the scenarios. For Example 1, we first fix $x$ at $(0, 0)$. In Figure 3.3 (a), the first sample is generated from a mean-zero normal and the mean of the second sample gradually deviates from zero. In Figure 3.3 (b), the mean of the first population is $(0.4, 0.4)$ while that of the second also gradually deviates from $(0.4, 0.4)$. Comparing the power curves in these two figures, we observe that the power of the likelihood ratio test remains the same as expected since the alternative differs from the null by the same amount in these two scenarios. However, the power of the log-rank or Wilcoxon tests are significantly higher in Figure 3.3 (b) then in Figure 3.3 (a). This matches the calibration in §3.2.3 that when the mean and $x$ are both zero, the rank based tests have no power.

If we let the respective mean of the first and the second population be $(0, 0)$ and $-(0.4, 0.4)$ while moving $x$ along $(0.2i, 0.2i)$, $i = 0, 1, \ldots, 5$, then Figure 3.3 (c) shows that the power of the proposed tests increases with $i$. Likewise, if we let the variance of the bivariate normal be $I$ and $2I$ for the first and second population, respectively and set $x$ to be $(0, i), i = 0, 1, \ldots, 5$, then Figure 3.3(d) shows that the power of the proposed tests decreases with $i$ . These two examples show that the choice of $x$ will have impact on the power of the proposed tests.

## 3.4 Applications

### 3.4.1 Application to functional magnetic resonance imaging of brain activity data

The functional magnetic resonance imaging (fMRI) of brain activity data set, previously analysed by Rosenbaum (2005) using a minimum distance pairing approach, consists of two measurements on each of the 18 subjects. All eighteen subjects are right handed. Half of them have arteriovenous malformations in the left hemisphere and the other half are normal controls. The two measurements are laterality indices calculated for listening to a story and repeating a sentence mentally after listening to it, respectively. When the subject is listening to a story or repeating a sentence while undergoing fMRI, the number of activated pixels in both the left and the right hemisphere's temporal lobe, $L$ and $R$, are recorded. For each task, the laterality index is calculated as $(L - R)/(L + R)$; see Lehéricy et al. (2002) for details. The laterality index is a continuous variable measuring the relative activation of the left and the right hemispheres during the language tasks. In some extreme cases, the laterality index is 1 if all the increased activation is on the left,

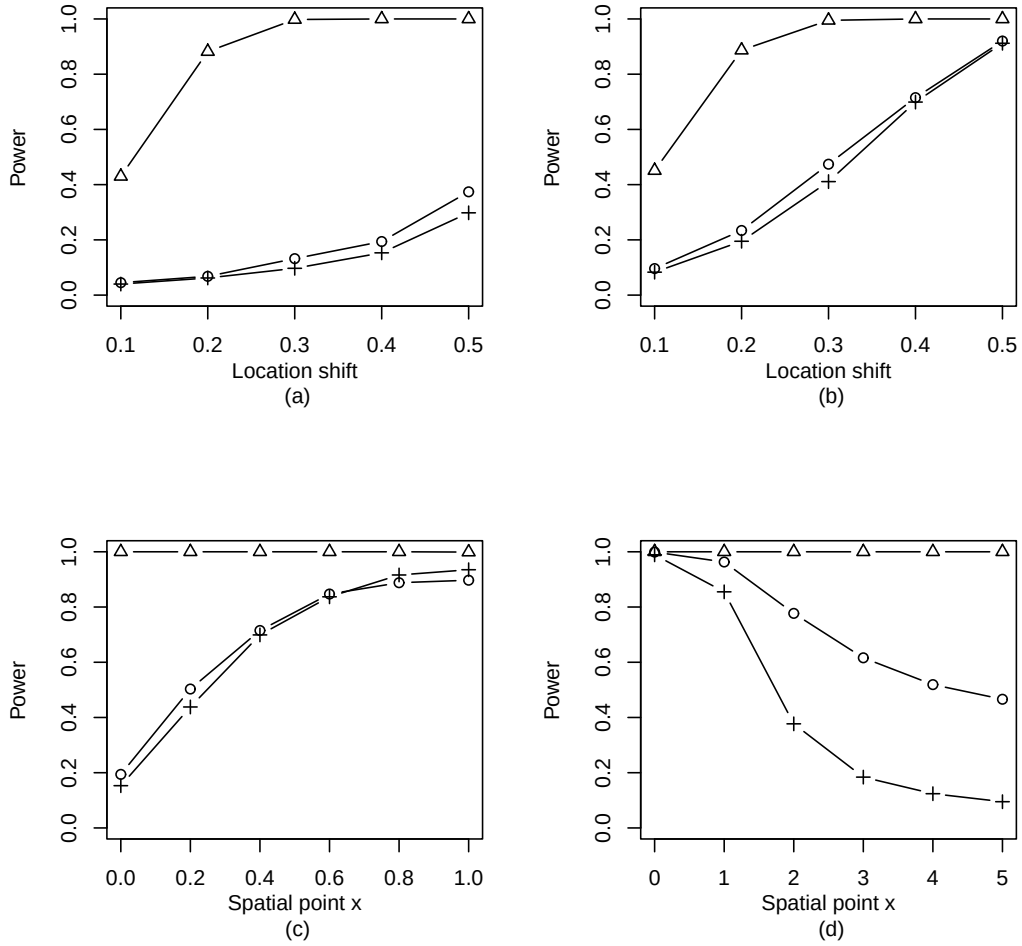



Figure 3.3: Graphs showing power curves for (a) bivariate normal with location shift and the mean under the null is $(0, 0)$, (b) bivariate normal with location shift and the mean under the null is $(0.4, 0.4)$ (c) bivariate normal with mean $(0, 0)$ against $(-0.4, -0.4)$ while $x$ deviates from the origin and (d) bivariate normal with variance $I$ against $2I$ while $x$ deviates from the origin ($\triangle$: likelihood ratio test, $\circ$: log-rank test, $+$:Wilcoxon test).

Table 3.4: $p$-values for FMRI data

| | Integral | | | Sup | | |
|---|---|---|---|---|---|---|
| | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 0$ | $\rho = 1$ | $\rho = 2$ |
| $k = 9$ | 0.0196 | 0.0230 | 0.0258 | 0.0292 | 0.0292 | 0.0384 |
| $k = 18$ | 0.0164 | 0.0172 | 0.0204 | 0.0050 | 0.0102 | 0.0266 |

$-1$ if all is on the right and 0 if the activation on both sides is the same.

Figure. 3.2 is a scatter plot of the laterality index for sentence repetition against the laterality index for story listening, with patients indicated by a triangle and controls indicated by a bullet. The primary goal of the study is to examine whether the impaired left hemisphere affects the performance of language tasks. To this end, we implement the weighted log-rank approach to compare two groups in the two-dimensional space. Due to the small sample size, we use the fixed $D$ constructed as in the previous section and 5000 permutations to obtain the cutoff value.

Table 3.4 displays the p-values corresponding to the weighted log-rank approach with variable combinations of the censoring time $k$ and weight $\rho$. All p-values are less than 0.05, so the null hypothesis that the distributions of the laterality indices for two groups are the same, is implausible. This conclusion agrees with that in Rosenbaum (2005), where the significance level is 0.0259.

### 3.4.2 Application to prostate cancer DNA microarray data

Worldwide, prostate cancer is the third most common cancer and the cause of 6% of cancer deaths in men (Parkin et al. 2001). In the United States, it is the most frequently diagnosed and the second leading cause of cancer death in men (Jemal 2003). The first step to better understanding prostate cancer is to test whether genes are expressed differentially in tumor compared to nontu-

mor samples. To explore potential molecular variation, DNA microarray profiling studies are conducted. As a new multiplex technology, DNA microarray can contain tens of thousands of probes, leading to a high dimensional problem. To illustrate the proposed method, we study a case-control prostate cancer DNA microarray data set (LaPointe 2004). The data set can be downloaded at `http://microarray-pubs-stanford.edu/prostateCA/`. The prostate cancer data set consists of 62 primary prostate tumors (61 adenocarcinomas and one adenoid cystic tumor) and 41 matched normal prostate tissues (from the noncancerous region of the prostate). For each tissue, the expressions of 5153 genes were measured. So we have $62 + 41 = 103$ samples in a 5153-dimensional space. The upper part of Figure. 3.4 displays the grayscale maps of mean expression value of each gene for tumor and nontumor groups, respectively. We realign the 5153 genes in a $72 \times 72$ matrix by row according to the percentage of missing data, from low to high, i.e., each row from left to right, the percentage of missing data is nondecreasing. The last 31 elements in the last row of the matrix are set to zero just to make up a square matrix. Throughout the following analysis, we use the permutation tests to obtain the cutoff values.

First we take out the 450 genes without any missing data and apply the weighted log-rank approach in a 450-dimensional space with $D$ composed of two 450-vectors, each being the mean expression values within the respective group. The null hypothesis of no difference in distribution is rejected at levels well below $10^{-3}$ for all reasonable combinations of $\rho$ and $k$. Then we narrow down the attention to shorter segments of genes. Specifically, we reshape the 450 genes into a $45 \times 10$ matrix (see the bottom part of Figure 3.4) and each time we test for group difference for one row, i.e., in a 10-dimensional space. Among the 45 rows, there are 29 rows with extremely small p-values and thus in Figure. 3.5, we only display

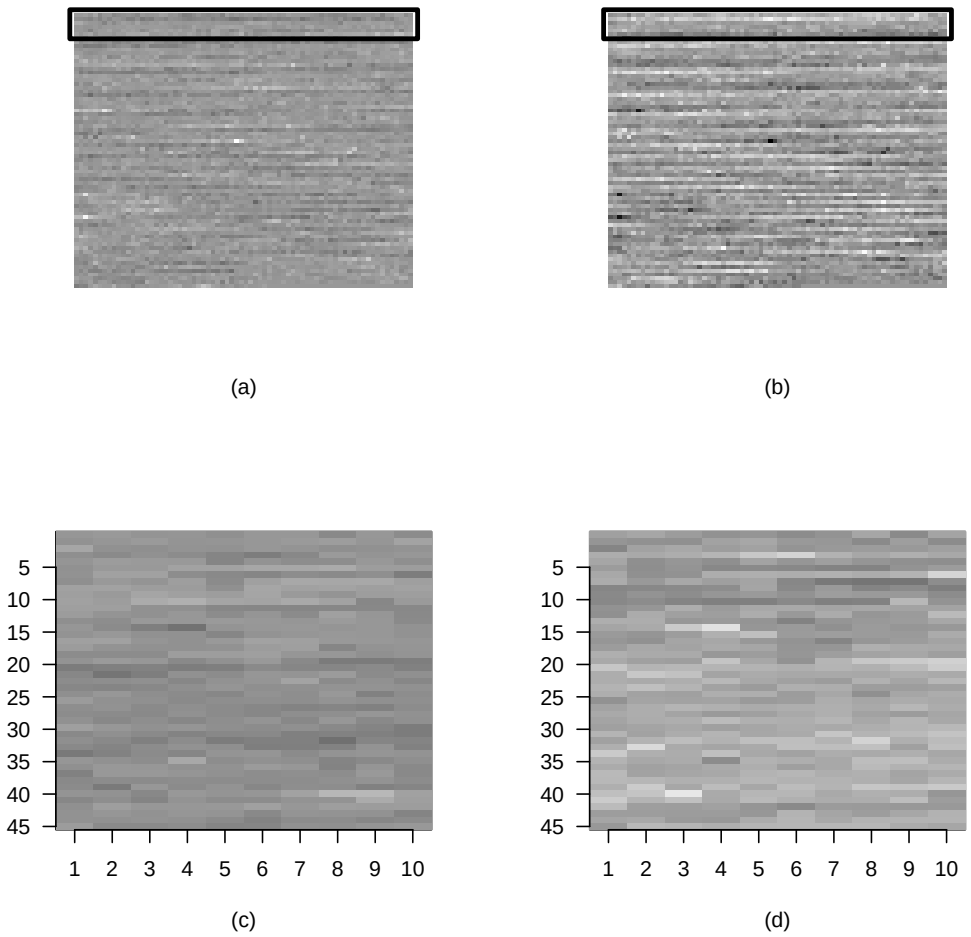



Figure 3.4: Graphs showing grayscale maps of mean gene expressions for (a) 5153 genes in cancer population, (b) 5153 genes in control population, (c) 450 genes in cancer population, (d) 450 genes in control population.

the $p$-value for the rest 16 rows $(1, 2, 4, 5, 7, 15, 17, 18, 22, 24, 29, 32, 33, 34, 35, 37)$.

Note that the patterns of the four graphs are quite similar, which shows the consistency of the procedure. For some rows, the $p$-value can drop tremendously with more samples. For row 29, censoring at the 20th observation from the mean expression value can not reject the null hypothesis at 5% level while including 50 or 103 samples can detect the difference. This might suggest the distributions for those rows are quite similar around the mean expression values but differ from each other over the places far away from the means in a 10-dimensional Euclidean space. For some other rows, e.g., row 4, censoring at the 20th observation may give the smallest $p$-value, which might suggest a discrepancy in a local neighborhood of the mean expression values. In contrast, the three $k$ values give similar $p$-values for some rows, for instance, row 37. This may imply a more uniform distribution over the 10-dimensional space.

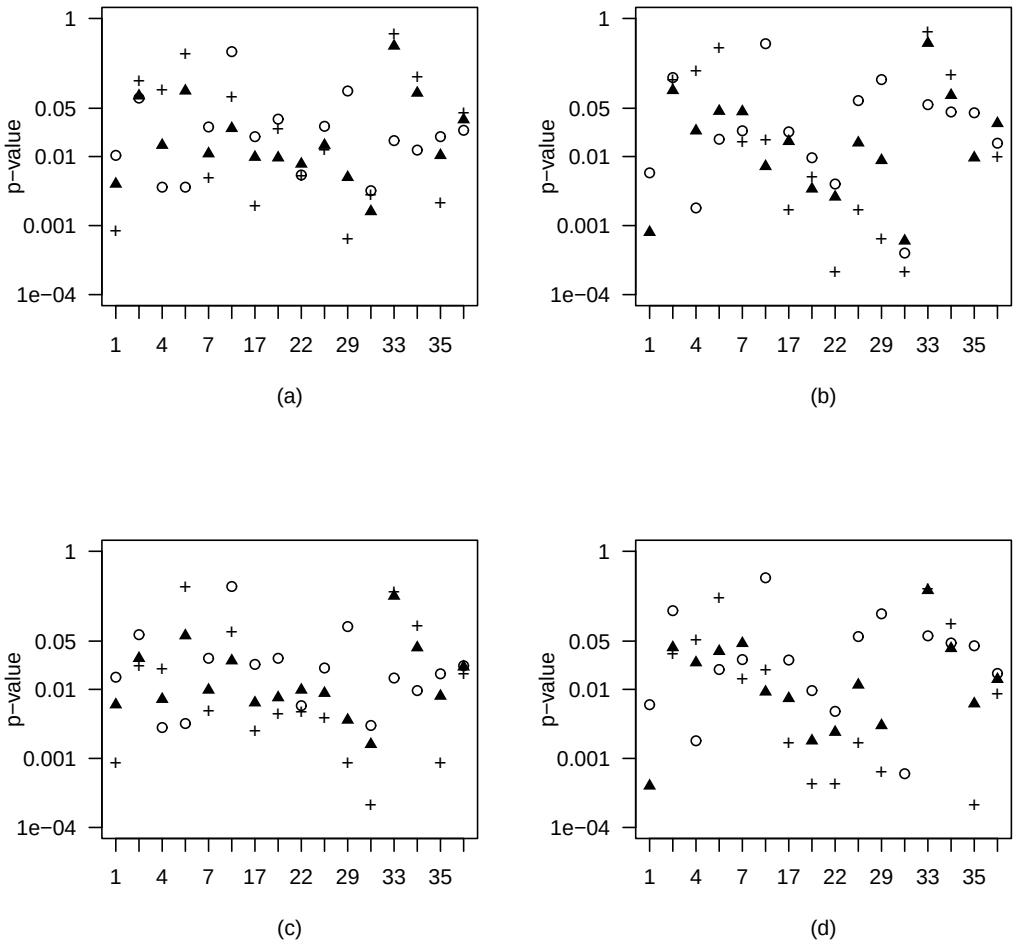



Figure 3.5: Graphs showing $p$-values for (a) the integral test when $\rho = 0$, (b) the sup test when $\rho = 0$, (c) the integral test when $\rho = 1$ and (d) the sup test when $\rho = 1$ ($\circ : k = 20$, $\blacktriangle : k = 50$, $+ : k = 103$). The p-values are plotted after a $\log_{10}$ transformation.

# Chapter 4

# Discussion

Nonparametric covariate adjustment is of importance in analysis of randomized clinical trial data. When properly done, it can result in efficiency improvement while maintaining the nonparametric nature of the usual tests. Empirical likelihood approach is nonparametric, constraint based and efficient in extracting information from data.

For randomized clinical trials, covariate information with no model assumption can be extracted from certain type of constraints or estimating equations. In this dissertation, we propose an empirical likelihood based approach for covariate adjustment. The resulting likelihood ratio test is shown to have the usual Wilks type $\chi^2$ approximation, with increased power as the number of constraints increases. The corresponding maximum empirical likelihood estimate also enjoys similar asymptotic properties. We demonstrate that the $\chi^2$ and normal approximations continue to hold as the number of constraints grows with sample size. We further show that in doing so the semiparametric efficiency can be achieved.

One of the practical issues is how to select basis functions in the constraints. From our experiences with simulations and real data analysis, it appears that

there is no universal way to deal with this issue. A related issue is how many basis functions should be used. One ad hoc way to do that is to consider variance reduction when additional constraints are added. We believe that if initial basis functions are properly chosen, then only a very small number of constraints will be needed.

It will be of interest to extend this empirical likelihood based nonparametric covariate adjustment to other situations, including observational studies. Of particular importance are survival and longitudinal studies where the response variables may be dependent or causal. For survival data, Lu and Tsiatis (2008) have introduced a general model framework for covariate adjustment and derived a semiparametric efficient score. We believe a similar approach, which makes use of suitable covariate based constraints and achieves the asymptotic efficiency, can be developed.

The second part of the thesis utilizes the powerful class of nonparametric weighted log-rank tests to test the difference between multivariate distributions. This is done by converting the original observations, possibly in a high dimensional space, into "survival times". Such an approach reduces greatly the complexity of the original problem. It also allows the application of available tools, including software packages for implementation and large sample properties for theoretical justification. Because the choice of the weight function is intuitive and well understood in survival analysis, it may be possible to use the intuition from survival analysis to gain insights into which weight function or functions should be used in the original multivariate data.

As a simple example, we take type-II censoring to illustrate the flexibility of the proposed method. This type of censoring may be viewed as a special kind of weight functions. By varying the number of failure times to be included, this type of

censoring magnifies "local" group differences as opposed to more global differences. In this connection, general weight functions provide even greater flexibility.

The usual asymptotic properties are established for the proposed tests. In particular, we establish the weak convergence of the basic processes under the null and contiguous alternative hypotheses. The weak convergence is then used to derive limiting distributions for the test statistics which are functionals of the basic processes. The limiting distributions under contiguous alternatives may shed some light on the asymptotic efficiency of the proposed tests in some simple cases.

Converting spatial points to distances is a crucial component of the proposed method. Instead of Euclidean space, one may consider a general metric space. In particular, the method is readily applicable to functional data when a suitable metric can be introduced on the corresponding function space. Modern empirical process theory is also applicable to deriving asymptotic properties.

The use of the proposed approach for high dimensional problems is of particular importance. We believe that asymptotic properties can be extended to very high dimensional problems, including those with $p$ being larger than $n$. We further note that the approach may also be extended to other kinds of high dimensional learning problems. Investigating the use of the survival analysis approach in classification problems in high dimensional spaces is worth future effort.

# Bibliography

[1] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.

[2] P. Armitage. Importance of prognostic factors in the analysis of data from clinical trials. *Controlled Clinical Trials*, 1:347–353, 1981.

[3] S.F. Assmann, S.J. Pocock, L.E. Enos, and L.E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*, 355:1064–1069, 2000.

[4] P. Billingsley. *Probability and Measure*. Wiley, New York, 1986.

[5] J.F. Box. R.A. Fisher and the design of experiments, 1922-1926. *American Statistician*, 34:1–7, 1980.

[6] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research, 2: The Design and Analysis of Cohort Studies*. IARC Scientific Publications, Lyon, France, 1987.

[7] D.P. Byar, R.M. Simon, W.T. Friedewald, J.J. Schlesselman, D.L. DeMets, J.H. Ellenberg, M.H. Gail, and J.H. Ware. Randomized clinical trials. *New England Journal of Medicine*, 295:74–80, 1976.

[8] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. Chapman and Hall/CRC, London, 1974.

[9] M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion). *Statistical Science*, 20:261–301, 2005.

[10] D.L. Donoho and M. Gasko. Breakdown properties of location estimates based on half space depth and projected outlyingness. *Annals of Statistics*, 20:1803–1827, 1992.

[11] L. Dümbgen. Limit theorems for the simplicial depth. *Statistics and Probability Letters*, 14:119–128, 1992.

[12] B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58:403–417, 1971.

[13] R.A. Fisher. *Statistical Methods For Research Workers*. Oliver and Boyd, Edinburgh, 1932.

[14] R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1966.

[15] T.R. Flemming and D.P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, New Jersey, 2005.

[16] J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

[17] L. M. Friedman, C. D. Furberg, and D. L. DeMets. *Fundamentals of Clinical Trials*. Springer-Verlag, New York, 1998.

[18] M. H. Gail, S. Wieand, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–444, 1984.

[19] E.A. Gehan. A generalized wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, 52:203–23, 1965.

[20] R.D. Gill. *Censoring and stochastic integrals*. Mathematisch Centrum 124, Amsterdam, 1980.

[21] V.P. Goldambe and C.C. Heyde. Quasi-likelihood and optimal estimation. *International Statistical Review*, 55:231–244, 1987.

[22] J.M. Grouin, S. Day, and J. Lewis. Adjustment for baseline covariates: an introductory note. *Statistics in Medicine*, 23:697–699, 2004.

[23] J. Hájek and Z. Šidák. *Theory of Rank Tests*. Academic Press, New York, 1967.

[24] J. Hájek, Z. Šidák, and P. Sen. *Theory of Rank Tests*. Academic Press, San Diego, Calif., 1999.

[25] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89:359–374, 2002.

[26] D.P. Harrington and T.R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566, 1982.

[27] W.W. Hauck, S. Anderson, and S.M. Marcus. Should we adjust for covariates in non-linear regression analyses of randomised trials. *Controlled Clinical Trials*, 19:249–256, 1998.

[28] N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Annals of Statistics*, 16:772–783, 1988.

[29] A.B. Hill. *Controlled Clinical Trials*. Blackwell, Oxford, 1960.

[30] N.L. Hjort, I.W. McKeague, and I.V. Keilegom. Extending the scope of empirical likelihood. *The Annals of Statistics*, 37:1079–1111, 2009.

[31] M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods*. Wiley, New York, 1973.

[32] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[33] Lewis J.A. Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in Medicine*, 18:1903–1904, 1999.

[34] A. Jemal, T. Murray, A. Samuels, A. Ghafoor, E. Ward, and M.J. Thun. Cancer statistics, 2003. *CA Cancer J. Clin.*, 53:5–26, 2003.

[35] O. Kempthorne. Why randomize? *Journal of Statistical Planning and Inference*, 1:1–25, 1977.

[36] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:887–906, 1956.

[37] G.G. Koch, C.M. Tangen, J.W. Jung, and I.A. Amara. Issues for covariate analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, 17:1863–1892, 1998.

[38] S.W. Lagakos and D.A. Schoenfeld. Properties of proportional hazards score tests under misspecified regression models. *Biometrics*, 40:1037–1048, 1984.

[39] S. Lang. *Calculus of several variables*. Springer-Verlag, New York, 1987.

[40] J. Lapointe, C. Li, and J.P. et al. Higgins. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences USA*, 101:811–816, 2004.

[41] Y. Le Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel.

[42] S. Lehéricy, A. Biondi, N. Sourour, M. Vlaicu, S.T. Du Montcel, L. Cohen, E. Vivas, L. Capelle, T. Faillot, A. Casasco, D. Le Bihan, and C. Marsault. 2002. *Radiology*, 223:672–82, 2002.

[43] E.L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

[44] S. Leon, A.A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Biometrics*, 59:1048–1057, 2003.

[45] D.Y. Lin, L.J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B*, 62:711–730, 2000.

[46] D.Y. Lin, L.J. Wei, and Z. Ying. Model-checking techniques based on cumulative residuals. *Biometrics*, 58:1–12, 2002.

[47] R. Liu. On a notion of simplicial depth. *Proceedings of the National Academy of Sciences USA*, 85:1732–1734, 1988.

[48] R. Liu. On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414, 1990.

[49] R. Liu and K. Singh. Ordering directional data: Concepts of data depth on circles and spheres. *Annals of Statistics*, 20:1468–1484, 1992.

[50] R. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88:252–260, 1993.

[51] X. Lu and A.A. Tsiatis. Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95:679–694, 2008.

[52] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170, 1966.

[53] N. Mantel. Confounding in epidemiologic studies. *Biometrics*, 45:1317–1318, 1989.

[54] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, 22:719–748, 1959.

[55] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[56] T.M. Morgan. Omitting covariates from the proportional hazards model. *Biometrics*, 42:993–995, 1986.

[57] A.B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

[58] A.B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18:90–120, 1990.

[59] A.B. Owen. Empirical likelihood for linear models. *The Annals of Statistics*, 19:1725–1747, 1991.

[60] A.B. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, 2001.

[61] D.M. Parkin, F.I. Bray, and S.S. Devesa. Cancer burden in the year 2000. the global picture. *European Journal of Cancer*, 37:4–66, 2001.

[62] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society: Series A*, 135:185–206, 1972.

[63] S. Piantadosi. *Clinical Trials: A Methodologic Perspective*. Wiley, New Jersey, 2005.

[64] S.J. Pocock. *Clinical Trials: A practical Approach*. Wiley, Wiley, 1983.

[65] S.J. Pocock, S.E. Assmann, L.E. Enos, and L.E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, 21:2917–2930, 2002.

[66] D. Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward, CA, 1990.

[67] S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16:356–366, 1988.

[68] R.L. Prentice. Linear rank tests with right censored data. *Biometrika*, 65:167–179, 1978.

[69] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325, 1994.

[70] G.M. Raab, S. Day, and J. Sales. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*, 21:330–342, 2000.

[71] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

[72] L.D. Robinson and N.P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 58:227–240, 1991.

[73] P.R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B*, 67:515–530, 2005.

[74] C.M. Rutter and R.M. Elashoff. Analysis of longitudinal data: Random coefficient regression modeling. *Statistics in Medicine*, 13:1211–1231, 1994.

[75] H. Scheffe. *The Analysis of Variance*. John Wiley and Sons, Inc, New York, 1959.

[76] S. Senn. Consensus and controversy in pharmaceutical statistics. *The Statistician*, 49:135–176, 2000.

[77] R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 2002.

[78] R. Simon. Use of regression models: Statistical aspects. In M.J. Staquet and R.J. Sylvester, editors, *Cancer Clinical Trials*. Oxford University Press, Oxford, 1984.

[79] E.W. Steyerberg, P.M.M. Bossuyt, and K.L. Lee. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? *American Heart Journal*, 139:745–751, 2000.

[80] A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principle yet flexible approach. *Statistics in Medicine*, 27:4658–4677, 2008.

[81] J.W. Tukey. Mathematics and the picture of data. *Proceedings of the International Congress Mathematicians*, 2:523–531, 1975.

[82] A.W. Van Der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, 2000.

[83] C.H. Wang and S.N. Srihari. A framework for object recognition in a visually complex environment and its application to locating address blocks on mail pieces. *International Journal of Computer Vision*, 2:125–511, 1988.

[84] J. Zedlewski. *Practical Empirical Likelihood Estimation with matElike*, 2008.

[85] M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64:707–715, 2008.