EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR BIOLOGICAL, GENETIC, AND MEASUREMENT ERROR APPLICATIONS

A Dissertation

by

TANYA PAMELA GARCIA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2011

Major Subject: Statistics

EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR BIOLOGICAL, GENETIC, AND MEASUREMENT ERROR APPLICATIONS

A Dissertation

by

TANYA PAMELA GARCIA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Yanyuan Ma
Committee Members,	Raymond J. Carroll
	Qi Li
	Mohsen Pourahmadi
Head of Department,	Simon J. Sheather

August 2011

Major Subject: Statistics

ABSTRACT

Efficient Semiparametric Estimators for Biological, Genetic, and Measurement Error Applications.

(August 2011)

Tanya Pamela Garcia, B.S., University of California, Irvine;

M.S., University of California, Berkeley;

M.S., University of Western Ontario

Chair of Advisory Committee: Dr. Yanyuan Ma

Many statistical models, like measurement error models, a general class of survival models, and a mixture data model with random censoring, are semiparametric where interest lies in estimating finite-dimensional parameters in the presence of infinite-dimensional nuisance parameters. Developing efficient estimators for the parameters of interest in these models is important because such estimators provide better inferences.

For a general regression model with measurement error, we utilize semiparametric theory to develop an unprecedented estimation procedure which delivers consistent estimators even when the model error and latent variable distributions are misspecified. Until now, root-n consistent estimators for this setting were not attainable except for special cases, like a polynomial relationship between the response and mismeasured variables. Through simulation studies and a nutrition study application, we demonstrate that our method outperforms existing methods which ignore measurement error or require a correct model error distribution.

In randomized clinical trials, scientists often compare two-sample survival data with a log-rank test. The two groups typically have nonproportional hazards, however, and using

a log rank test results in substantial power loss. To ameliorate this issue and improve model efficiency, we propose a model-free strategy of incorporating auxiliary covariates in a general class of survival models. Our approach produces an unbiased, asymptotically normal estimator with significant efficiency gains over current methods.

Lastly, we apply semiparametric theory to mixture data models common in kin-cohort designs of Huntington's disease where interest lies in comparing the estimated age-at-death distributions for disease gene carriers and non-carriers. The distribution of the observed, possibly censored, outcome is a mixture of the genotype-specific distributions where the mixing proportions are computed based on the genotypes which are independent of the trait outcomes. Current methods for such data include a Cox proportional hazards model which is susceptible to model misspecification, and two types of nonparametric maximum likelihood estimators which are either inefficient or inconsistent. Using semiparametric theory, we propose an inverse probability weighting estimator (IPW), a nonparametrically imputed estimator and an optimal augmented IPW estimator which provide more reasonable estimates for the age-at-death distributions, and are not susceptible to model misspecification nor poor efficiencies.

To my Mom, for her unconditional love and encouragement throughout this journey.

ACKNOWLEDGEMENTS

I cannot thank Yanyuan Ma, my advisor, mentor, and friend, enough for persevering with me throughout the time it took me to complete this research. Her enthusiasm, patience, kindness and faith in my abilities inspired me to better myself as a researcher and kept me motivated through the more difficult times of this process. For the opportunities she bestowed on me, her willingness to help me at nearly any hour, her wonderful sense of humor, and her friendship, I am entirely grateful.

I am also grateful to the members of my dissertation committee, Raymond J. Carroll, Qi Li, and Mohsen Pourahmadi, for generously giving their time and expertise to improve my work.

Raymond Carroll went out of his way to provide careful insight about the content and presentation of Chapter II. His guidance led to this Chapter being acceptable by peer-reviewers, and I thank him for saving me from another rejection.

Although they are not members of my dissertation committee, Guosheng Yin and Yuanjia Wang helped me considerably in the preparation of this work. Guosheng and Yuanjia boosted my confidence by granting me the opportunity to work with them on the material that developed into Chapters III and IV. Writing with them and understanding their methods for approaching difficult problems has been one of the most rewarding and enjoyable experiences of my graduate career.

I am especially thankful to the following organizations for their belief in my potential and generously funding my studies: National Science Foundation's Texas A&M University System Louis Stokes Alliance for Minority Participation Bridge to the Doctorate Fellowship, the National Consortium for Graduate Degrees for Minorities in Engineering and Science, Inc. (GEM) Ph.D. Fellowship, Philanthropic Educational Organization (P.E.O.)

Scholar Award, and Raymond Carroll for his summer support.

I thank Mark J. DeBonis for inspiring me to study Mathematics. Had I not taken his course during my first undergraduate year, I would have surely switched majors. His excitement about mathematics and compassion for students has instilled in me the desire to inspire others as he has me.

To the girls of the Statistics Department, Hsiang-chun Chen, Anh Dao, Priya Kohli, Jenn Rolfes, Robyn Ball, and Trijya Singh: thank you so much for the laughs, the lunches, the pep-talks, and improving the Blocker ambiance. All of you hold a dear place in my heart.

This dissertation would never have been accomplished without the love, patience, and support of my family, Lita, Julio, Liz, and Pepper Garcia. My mom never faltered in telling me exactly what I needed to hear to overcome my moments of doubt; my sister could always make me laugh; and my dad instilled in me the belief that I could overcome any challenge. My achievements are results of their compassion and encouragement.

Finally, I would like to thank my darling husband, Sylvain Marcel, and my little dog, Mishka, for their love and sacrifice. Without complaint, my husband kindly supported my late nights and weekends of work, the extra responsibilities in our home, and he changed his work schedule to accommodate mine. I could not have asked for a better husband and friend.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS viii

LIST OF TABLES x

LIST OF FIGURES xii

CHAPTER

I INTRODUCTION 1

II SEMIPARAMETRIC ESTIMATORS FOR RESTRICTED MOMENT MODELS WITH MEASUREMENT ERROR 5

2.1 Introduction 5
2.2 Main results 7
2.3 Extensions for the measurement error distribution 13
2.4 Simulations 15
2.5 A case study in nutrition 23
2.6 Discussion 25

III EFFICIENCY IMPROVEMENT IN A CLASS OF SURVIVAL MODELS THROUGH MODEL-FREE COVARIATE INCORPORATION 27

3.1 Introduction 27
3.2 Notation and semiparametric models 29
3.3 Improving efficiency through covariate augmentation 33
3.4 Simulations 36
3.5 Application to leukemia study 40
3.6 Discussion 43

IV SEMIPARAMETRIC ESTIMATION FOR CENSORED MIXTURE DATA WITH APPLICATION TO THE COOPERATIVE HUNTINGTON'S OBSERVATIONAL RESEARCH TRIAL 44

CHAPTER Page

4.1 Introduction 44
4.2 Existing estimators for censored mixture data 48
4.3 Proposed nonparametric estimators for censored mixture data . 53
4.4 Simulations 57
4.5 Analysis of the COHORT data 65
4.6 Discussion 69

V CONCLUSION 71

REFERENCES 74

APPENDIX A 82

APPENDIX B 91

APPENDIX C 92

VITA 100

LIST OF TABLES

TABLE Page

1 Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter β in Model 1 and under the two proposed situations. The true parameter is $\beta_{\text{true}} = 0.5$. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar). 19

2 Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter $\hat{\beta}$ in Model 2 and under Situation 1. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). The true parameter is $\beta_{\text{true}} = (0.5, 0.7)^T$. Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar). 20

3 Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter β in Model 2 and under Situation 2. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). The true parameter is $\beta_{\text{true}} = (0.5, 0.7)^T$. Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar). 21

TABLE Page

4 Simulation with $\beta = (0, 0)'$ and different sample sizes n. The conditional correlation between T and W given Z is 0.15 ($\rho = 0.4$) and 0.27 ($\rho = 0.7$). Bias, sample standard error (se), median of estimated standard errors ($\hat{\text{se}}$), and the 95% coverage probability (CI) are given for unadjusted estimator $\hat{\beta}$ and adjusted estimator $\hat{\beta}_{\text{AUG}}$, respectively. . . . 38

5 Simulation with $\beta = (0.5, -0.5)'$ and different sample sizes n. The conditional correlation between T and W given Z is 0.11 ($\rho = 0.4$) and 0.19 ($\rho = 0.7$). Bias, sample standard error (se), median of estimated standard errors ($\hat{\text{se}}$), and the 95% coverage probability (CI) are given for unadjusted estimator $\hat{\beta}$ and adjusted estimator $\hat{\beta}_{\text{AUG}}$, respectively. 39

6 Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the first simulation with proportional hazards generated data, sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations. 59

7 Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the second simulation with non-proportional hazards generated data, sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations. 60

8 Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the third simulation replicating the COHORT data, sample size $n = 1000$, 65% censoring rate, 1000 simulations. 64

9 Estimated survival rates and 95% confidence intervals (in parentheses) for carrier (C) and non-carrier (NC) groups in COHORT data. 67

LIST OF FIGURES

FIGURE Page

1 Quantile-quantile plots of the measurement error U for the original first and third readings of the 24 hour recall surveys (top) and after the logarithm transform (bottom). 24

2 Crossing survival curves when $R(t) = t$, with (a) $\gamma_1 = 0.6$ and $\gamma_2 = 2$; (b) $\gamma_1 = 2$ and $\gamma_2 = 0.6$. 32

3 Kaplan-Meier survival curves under different treatments in the leukemia study. 41

4 Kaplan-Meier survival curves for (a) males; (b) females; (c) individuals with ages less than 60 years; (d) individuals with ages over 60 years. Survival curves in each case are stratified by treatment where the solid line corresponds to chemoimmunotherapy with rituximab, and the dashed line corresponds to chemotherapy. 42

5 Simulation 3. True survival curve (solid) and the mean of 1000 simulations at each time point (short-dashed for carrier group, long-dashed for non-carrier group), 95% pointwise confidence band (upper band dotted, lower band dash-dotted) of the estimated survival curves. The mean and true survival curves are indistinguishable in EFFIMP and EFFAIPW estimators. Sample size is 1000, censoring rate is 65%. 63

6 Estimated survival curves and 95% point-wise confidence bands (upper band dotted, lower band dash-dotted) for Huntington disease data using the efficient complete data influence function based optimal AIPW and IMP, the Cox-based, and the type I NPMLE estimator. The short-dashed curve denotes survival rates for persons possessing a carrier gene, the long-dashed curve for persons possessing a non-carrier gene, and the solid curve corresponds to survival rates for general US population in 2003. 66

CHAPTER I

INTRODUCTION

A vast majority of statistical models are semiparametric, where interest lies in estimating finite-dimensional parameters of interest in the presence of infinite-dimensional nuisance parameters. Examples of infinite-dimensional nuisance parameters are unknown error distributions in regression and unknown baseline hazard functions in survival analysis. Leaving these error distributions, baseline hazard functions, and other nuisance parameters unspecified leads to more general models. Consequently, results based on these general models are more appealing as they tend to be robust and have greater applicability.

To describe a semiparametric model in general, consider independent, and identically distributed random vectors $X_1, \ldots, X_n$ which are drawn from a class of densities indexed by a parameter θ,

$$\{\mathcal{P}_\theta : \theta \in \Theta\}.$$

In the semiparametric case, θ is composed of a p-dimensional parameter β, and an infinite-dimensional parameter η. In some instances, θ is conveniently written as (β, η), and in others, we write β as a function, $\beta(\theta)$. A main objective for these models is to prove general estimation procedures for β and determine the efficient estimator within a class of regular, semiparametric estimators. We restrict our attention to regular estimators to avoid estimators with unfavorable local properties (Newey, 1990).

To derive consistent, regular asymptotically linear (RAL) and locally efficient estimators under the semiparametric theory framework, we use so-called influence functions. An

---

The format and style follow that of *Biometrics*.

influence function uniquely characterizes the RAL estimator $\hat{\beta}_n$ for β based on the observed random vectors X_i through

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(X_i) + o_p(1),$$

where the influence functions $\varphi(X_i)$ are independent, identically distributed, mean zero random vectors of length p. In the above, β_0 is the true parameter value and $o_p(1)$ is a term converging in probability to zero as n tends to infinity. The asymptotic variance of $\hat{\beta}_n$ equals the variance of φ. As the influence function with the smallest variance yields the most efficient RAL estimator, the search for the efficient RAL estimator is equivalent to determining the most efficient influence function.

To facilitate characterizing influence functions, we take a geometric approach so that influence functions are viewed as elements of a Hilbert space $\mathcal{H}$ composed of mean zero, finite variance functions. All influence functions of interest are in fact orthogonal to the so-called tangent space in $\mathcal{H}$. The general method for determining (efficient) semiparametric estimators is thus to specify the Hilbert space $\mathcal{H}$, determine the tangent space and its orthogonal complement, and, within the latter space, identify the influence function with smallest variance, i.e., the most efficient RAL estimator. Further details about this procedure are available in Tsiatis (2006).

In this dissertation, we apply this methodology to three specific semiparametric models: a measurement error model, a general class of survival models, and a mixture data model with random censoring. In the chapters that follow, we explore each model, provide a general class of semiparametric estimators and characterize the properties of the optimal estimator therein. The consequences of this work are outlined below.

In Chapter II, under the semiparametric framework, we construct root-n consistent, asymptotically normal and locally efficient estimators for regression with errors in covariates and an unspecified model error distribution. Until now, root-n consistent estimators

for this setting were not attainable except for special cases, such as a polynomial relationship between the response and mismeasured variables. The proposed method is the first to deliver root-n consistent estimators when the distributions for both the model error and the mismeasured variable are unknown and can be misspecified. The estimators are based on the semiparametric efficient score which is calculated under several possibly incorrect distribution assumptions resulting from the misspecified model error distribution, from the misspecified error-prone covariates' distribution, or from both. A simulation study demonstrates that the method is robust and outperforms methods which either ignore measurement error, or allow measurement error but require a correctly specified model error distribution. A real data example illustrates the performance of our method.

In Chapter III, we consider randomized clinical trials, where we are often concerned with comparing two-sample survival data. Although the log-rank test is usually suitable for this purpose, it may result in substantial power loss when the two groups have nonproportional hazards. In a more general class of survival models of Yang and Prentice (2005), which includes the log-rank test as a special case, we improve model efficiency by incorporating auxiliary covariates that are correlated with the survival times. In a model-free form, we augment the estimating equation with auxiliary covariates, and establish the efficiency improvement using the semiparametric theories in Zhang et al. (2008) and Lu and Tsiatis (2008). Under minimal assumptions, our approach produces an unbiased, asymptotically normal estimator with additional efficiency gain. Simulation studies and an application to a leukemia study show the satisfactory performance of the proposed method.

Lastly, we apply semiparametric theory to mixture data models common in kin-cohort designs (Struewing et al., 1997; Wacholder et al., 1998) and interval mapping of quantitative traits (QTL, Lander and Botstein, 1989). In these studies, the distribution of the observed, possibly censored, outcome is a mixture of the genotype-specific distributions where the mixing proportions are computed based on the genotypes which do not depend

on the trait outcomes. In this work, we examine estimators for a kin-cohort study of Huntington's disease where interest lies in comparing the estimated age-at-death distributions for disease gene carriers and non-carriers. Current literature on statistical methods for such data include a Cox proportional hazards based approach (Diao and Lin, 2005) and other parametric approaches (Moore et al., 2001) which are too restrictive and susceptible to model mis-specification. Current nonparametric approaches include two types of nonparametric maximum likelihood estimators (NPMLEs, Chatterjee and Wacholder, 2001; Wacholder et al., 1998), but we demonstrate one is not efficient while the other is not even consistent. Using a semiparametric approach, we propose several estimators including an inverse probability weighting (IPW) estimator, a nonparametrically imputed (IMP) estimator and an optimal augmented IPW (AIPW) estimator. We show the validity of these estimators and derive their asymptotic properties. Through simulation experiments and an application to the study data for Huntington's disease, we demonstrate that the proposed estimators lead to more reasonable estimates for the age-at-death distributions, and are not susceptible to model mis-specification nor poor efficiencies.

Finally, we conclude with summary remarks in Chapter V about the three models explored in this work.

CHAPTER II

SEMIPARAMETRIC ESTIMATORS FOR RESTRICTED MOMENT MODELS WITH MEASUREMENT ERROR

2.1 Introduction

Regression is arguably the most familiar topic in statistics and has motivated a vast amount of literature. Yet consistent estimation in regression with classic measurement error is only resolved for certain models, e.g. when the mean model is a polynomial or the model error distribution is specified. Our work develops the first consistent, regular asymptotically linear estimator (Newey, 1990) available for regression with errors-in-covariates in a general setting.

A general regression model characterizes the relationship between a response variable Y and covariates (X, Z) under minimal model assumptions. Regression relates Y to (X, Z) through

$$Y = m(X, Z; \beta) + \epsilon,$$

where m is decided up to the p-dimensional parameter β and the model error ϵ is only assumed to satisfy $E(\epsilon|X, Z) = 0$. Without distributional assumptions imposed on ϵ, this general regression model is also known as a restricted moment model (RMM). The RMM is a semiparametric statistical model with consistent estimators being the solutions to the linear estimating equation

$$\sum_{i=1}^{n} A(X_i, Z_i)\{Y_i - m(X_i, Z_i; \beta)\} = 0,$$

where $A(X, Z)$ is an arbitrary function in $R^{p\times 1}$ which does not cause the above estimating equation to degenerate. The efficient estimator is obtained when $A(X_i, Z_i) =$

$\partial m(X_i, Z_i; \beta)/\partial\beta E(\epsilon_i^2|X_i, Z_i)^{-1}$, commonly known as the optimal generalized estimating equation (Liang and Zeger, 1986).

We consider the situation when part of the covariates, say Z is precisely measured, while the remaining covariates, say X, are measured with error. In place of X, a surrogate variable W is observed. Surrogacy means Y and W are conditionally independent given (X, Z). We adopt here a modern functional measurement error model framework (Carroll et al., 2006, Chapter 7.2), which assumes that the conditional distribution of X given Z and the distribution of Z are completely unspecified. Our assumptions are motivated by the fact that many common regression models have measurement error along with an unknown model error distribution. Compared to the models considered in Tsiatis and Ma (2004), our model is less stringent because it allows an unspecified model error distribution and unspecified covariate distribution, not just the latter.

With an unspecified model error distribution, the RMM with measurement error is a very different problem compared to the model considered in Tsiatis and Ma (2004), where the model error distribution has a known parametric form. Consequently, the semiparametric treatment here is also drastically different. Our problem is also much more difficult than the model of Tsiatis and Ma (2004) both in terms of mathematical derivation and numerical computation. The RMM subject to measurement error possesses three unknown distributions: the conditional distribution of X given Z, $p_{X|Z}(x|z)$, the model error distribution of ϵ given (X, Z), $p_{\epsilon|X,Z}(\epsilon|x, z)$, and the unknown density for the observed covariates $p_Z(z)$. The first two distributions become nuisance parameters of infinite dimension that cannot be ignored and are difficult to model. Arbitrarily adopting a distribution for ϵ or for X given Z may cause bias, and estimating them is difficult. First of all, $p_{X|Z}(x|z)$ is a model for unobservable variables. Its estimation would thus involve deconvolution (Stefanski and Carroll, 1990) which results in a very slow rate (Carroll and Hall, 1988; Fan, 1991). When the measurement error is not additive or is correlated with X given Z, $p_{X|Z}(x|z)$ may not even

be identifiable. The estimation of $p_{\epsilon|X,Z}(\epsilon|x,z)$ is equally challenging because residuals are not obtainable in measurement error models, even if model parameters were known.

Possibly due to these difficulties, aside from special m functions such as polynomials (Chan and Mak, 1985; Cheng and Schneeweiss, 1998; Cheng et al., 2000), general regression with errors in covariates has not been well studied. No consistent estimator is known in the literature, even though measurement error models have received extensive attention in recent years. Chen et al. (2009) use a Sieve approach to treat an error-in-variables problem under unknown error distribution, but requires the covariates to be discrete. Hu and Schennach (2006) treated a similar problem via a deconvolution approach. Their major focus is on identifiability while here we focus on estimation and inference.

Although regression with errors in covariates is semiparametric, it is different from the models of Liang et al. (1999) or Liang and Li (2009). There, the semiparametric nature arises from an unknown smooth function of an error free variable. For an overview and recent developments on measurement error models, see Fuller (1987) for earlier results in linear models and Carroll et al. (2006) for modern approaches in linear and nonlinear models.

Until now, for an RMM with measurement error, no existing method allows an unspecified model error distribution and still guarantees consistency. Our proposed method is the first to give consistency while allowing misspecification in both the model error distribution $p_{\epsilon|X,Z}(\epsilon|x,z)$ and the conditional latent variable distribution $p_{X|Z}(x|z)$.

2.2 Main results

2.2.1 Semiparametric estimation

Our method for deriving consistent, regular asymptotically linear (RAL) and locally efficient estimators relies on semiparametric theory (Bickel et al., 1993). From a semiparametric perspective, efficient estimators correspond to efficient influence functions which

are viewed as elements of a Hilbert space $\mathcal{H}$ composed of mean zero, finite variance functions. The efficient influence function is a normalized element orthogonal to the so-called nuisance tangent space in $\mathcal{H}$. Our approach thus requires specifying the Hilbert space, characterizing the nuisance tangent space and its orthogonal complement, and, within the latter space, identifying the influence function with smallest variance. See Appendix A for a more detailed description and Tsiatis (2006) for a thorough introduction of this method. While our method uses the fundamentals of Bickel et al. (1993) and Tsiatis (2006), it is not a simple application of their results. As shown below, the efficient influence function may not be solved for directly and requires a particular mapping between the space of observed and unobservable variables so as to utilize the properties of the RMM without measurement error.

To help the readers focus on the core methodology, we assume, for now, the conditional density of $p_{W|X,Z}$ contains no additional unknown parameters. We explain how to handle additional parameters in $p_{W|X,Z}$ in Section 2.3. To place the RMM with measurement error in the semiparametric model framework, we write the probability density function of the observable random variables (W, Y, Z) as $p_{W,Y,Z}(w, y, z; \beta, \eta_1, \eta_2, \eta_3)$ which equals

$$\int p_{W|X,Z}(w|x, z)\eta_1(x, z)\eta_2\{y - m(x, z; \beta), x, z\}\eta_3(z)dx, \tag{2.1}$$

where β is the finite p-dimensional parameter of interest, $\eta_1(x, z) \equiv p_{X|Z}(x|z)$, $\eta_2(\epsilon, x, z) \equiv p_{\epsilon|X,Z}(\epsilon|x, z)$, and $\eta_3(z) \equiv p_Z(z)$ are infinite dimensional nuisance parameters. Doing so, we see that $p_{W,Y,Z}$, the RMM with measurement error, is tightly linked to the RMM without measurement error with probability density expressed as $p_{X,Y,Z} \equiv \eta_1(x, z)\eta_2\{y - m(x, z; \beta), x, z\}\eta_3(z)$. The lack of measurement error in the latter model implies that X, in addition to Z, is precisely observed there.

Proposition 1 in Appendix A describes the Hilbert space, nuisance tangent space and

the orthogonal complement for the RMM without measurement error. Based on (2.1) and the results in Proposition 1, we characterize the corresponding spaces for the measurement error model as shown below.

Theorem 1. For the RMM with measurement error, the Hilbert space is equal to $\mathcal{H} = \{f(W,Y,Z) : E(f) = 0, \text{var}(f) < \infty\}$; the nuisance tangent space is defined as $\Lambda = [E\{f(X,Y,Z)|W,Y,Z\} : E(f\epsilon|X,Z) = 0, E(f) = 0, \text{var}(f) < \infty]$; the nuisance tangent space orthogonal complement is

$$\Lambda^{\perp} = [f(W,Y,Z) : E\{f(W,Y,Z)|X,Y,Z\} = g(X,Z)\epsilon],$$

where g is an arbitrary function of (X,Z) with finite variance; the score vector with respect to β is $S_{\beta} = -E\{m'_{\beta}(X,Z;\beta)\partial \text{log}\eta_2(\epsilon,X,Z)/\partial\epsilon|W,Y,Z\}$, where $m'_{\beta}(X,Z;\beta)$ denotes $\partial m(X,Z;\beta)/\partial\beta$.

The proof of Theorem 1 is in Appendix A It is worth emphasizing that the description of the orthogonal complement $\Lambda^{\perp}$ above only requires the distribution $p_{W|X,Z}$. In other words, $\Lambda^{\perp}$ is invariant to misspecification of both nuisance parameters η_1 and η_2. This result is critical because it means even with an incorrectly specified η_1 and η_2, and consequently an incorrect $\mathcal{H}$ and Λ, one will still obtain a correct $\Lambda^{\perp}$ and, hence, consistent semiparametric RAL estimators for β.

Since the orthogonal complement $\Lambda^{\perp}$ contains all possible influence functions of semiparametric estimators, any function f satisfying the requirement of $\Lambda^{\perp}$ leads to a semiparametric estimator for β. Theorem 1 thus implies that if an arbitrary f satisfies $E(f|X,Y,Z) = g(X,Z)\epsilon$ for some $g(X,Z)$, then it would yield a semiparametric estimator for β as the solution to the estimating equation $\sum_{i=1}^{n} f(W_i,Y_i,Z_i;\beta) = 0$. In particular, the efficient score vector $S_{\text{eff}}(W,Y,Z)$ lies in $\Lambda^{\perp}$ and satisfies $E\{S_{\text{eff}}(W,Y,Z)|X,Y,Z\} = g(X,Z)\epsilon$ for some $g(X,Z)$. Consequently, the efficient score is used to construct estimating equations whose solution will be a semiparametric efficient estimator for β.

To this end, we express $S_{\text{eff}}(W, Y, Z)$ as a function of elements from the RMM without measurement error model so as to exploit the properties already established in Theorem 1 and Proposition 1. Specifically, first define conjugate linear operators $\mathcal{K}$ and $\mathcal{K}^*$ as $\mathcal{K}\{f(X, Y, Z)\} = E\{f(X, Y, Z)|W, Y, Z\}$ and $\mathcal{K}^*\{f(W, Y, Z)\} = E\{f(W, Y, Z)|X, Y, Z\}$. Through a careful analytic derivation using conjugacy (see Appendix A), we show that $S_{\text{eff}}(W, Y, Z) = \mathcal{K}\{d(X, Y, Z)\}$, where $d(X, Y, Z)$ satisfies

$$\epsilon E(d\epsilon|X, Z) + \mathcal{K}^* \circ \mathcal{K}(d)E(\epsilon^2|X, Z) - \epsilon E\{\mathcal{K}^* \circ \mathcal{K}(d)\epsilon|X, Z\} = m'_\beta(X, Z; \beta)\epsilon. \quad (2.2)$$

Here $\circ$ means the composite operation. Certainly, $\mathcal{K}^*$ can be calculated. If η_1, η_2 were known, we could also calculate $\mathcal{K}$ and the operation $E(\cdot|X, Z)$. In practice however, η_1, η_2 are both unknown. To circumvent this difficulty, we recommend proposing arbitrary models for η_1 and η_2, and carrying out the two operations $\mathcal{K}$ and $E(\cdot|X, Z)$ under the proposed models. Our proposed estimation procedure is thus as follows.

Procedure for estimating β:

1. Posit arbitrary models for η_1 and η_2 that follow the usual distribution requirements and such that η_2 satisfies $E(\epsilon|X, Z) = 0$.

2. Perform $\mathcal{K}, \mathcal{K}^*, E(\cdot|X, Z)$ under the known $p_{W|X,Z}$ and the proposed η_1, η_2 models, then solve for $d(X, Y, Z)$ from (2.2).

3. Form the score vector $S(W, Y, Z; \beta, \eta_1, \eta_2) = \mathcal{K}(d)$ by calculating $\mathcal{K}$ under the proposed η_1 model.

4. Solve the estimating equation $\sum_{i=1}^n S(W_i, Y_i, Z_i; \beta, \eta_1, \eta_2) = 0$ for the estimator $\widehat{\beta}$.

Having $\Lambda^\perp$ invariant to misspecification of η_1, η_2 ensures several robust consistency properties as we state in the following remarks.

Remark 1. When the covariate distribution η_1 is misspecified, the above algorithm still yields a consistent estimator.

Remark 2. When η_2, i.e. the conditional distribution of the model error ϵ on (X, Z), is misspecified, the above algorithm still yields a consistent estimator. This robust property is especially useful since it allows misspecification of the variance-covariance structure of η_2. This is in contrast to all existing methods in the literature, including that from Tsiatis and Ma (2004), which are extremely sensitive to the variance-covariance misspecification.

Remark 3. When both η_1, η_2 are misspecified, the algorithm still provides a consistent estimator.

Remark 4. Finally, for correctly specified nuisance parameters η_1 and η_2, this procedure gives the optimal estimator, in that its estimation variance achieves the semiparametric efficiency bound. Such results follow because, in this case, the resulting estimator indeed solves the true efficient score estimating equation defined as $\sum_{i=1}^{n} S_{\text{eff}}(W_i, Y_i, Z_i; \beta) = 0$.

A proof of Remark 3, which encompasses Remarks 1 and 2, is given in Appendix A.

2.2.2 Theoretical properties

We now establish the theoretical properties of the proposed estimation procedure in Theorem 2, and provide an outline of its proof in Appendix A.

Theorem 2. For any function $f(W, Y, Z; \beta)$ such that $E(f|X, Y, Z) = g(X, Z)\{Y - m(X, Z; \beta)\}$ for some function $g(X, Z)$, under suitable regularity conditions, the root of the estimating equation $\sum_{i=1}^{n} f(W_i, Y_i, Z_i; \beta) = 0$, denoted $\widehat{\beta}$, is an RAL estimator with influence function $E\{f(W, Y, Z)S_\beta^T(W, Y, Z)\}^{-1} f(W, Y, Z; \beta_0)$. Therefore,

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to N\{0, A^{-1}B(A^{-1})^\top\}$$

in distribution when $n \to \infty$, where the terms $A = E\{\partial f(W, Y, Z; \beta)/\partial\beta^\top|_{\beta_0}\}$ and $B = \text{var}\{f(W, Y, Z; \beta_0)\}$.

Remark 5. Our algorithm in Section 2.2.1 ultimately solves for $\hat{\beta}$ from an estimating equation of the form $\sum_{i=1}^{n} S(W_i, Y_i, Z_i; \beta, \eta_1, \eta_2) = 0$ where $S(W, Y, Z; \beta, \eta_1, \eta_2) = \mathcal{K}(d)$.

$\mathcal{K}(d)$ as a function lies in $\Lambda^{\perp}$ and satisfies $E\{\mathcal{K}(d)|X,Y,Z\} = g(X,Z)\epsilon$ for some $g(X,Z)$, which is required in Theorem 2. Thus, from Theorem 2, the resulting $\hat{\beta}$ from our proposed estimating procedure is indeed a consistent RAL estimator.

Not only does our algorithm lead to consistent estimators, but under certain conditions, it also leads to estimators with the same asymptotic efficiency as that of the true model. To justify this explicitly, suppose we posit parametric models for $\eta_1(x,z)$ and $\eta_2(\epsilon,x,z)$ denoted by $\eta_1(x,z;\gamma_1)$ and $\eta_2(\epsilon,x,z;\gamma_2)$, respectively, where γ_1 and γ_2 are finite dimensional parameters. The truth is denoted by $\eta_{10}(x,z) = \eta_1(x,z;\gamma_{10})$ and $\eta_{20}(\epsilon,x,z) = \eta_2(\epsilon,x,z;\gamma_{20})$. Let $\gamma = (\gamma_1,\gamma_2)^{\top}$, γ_0 be the true value of γ, and $\widehat{\gamma}$ be a root-n consistent estimator.

Assume $\widehat{\beta}$ solves $\sum_{i=1}^{n} f(W_i,Y_i,Z_i;\beta,\widehat{\gamma}) = 0$, and, for f satisfying the conditions in Theorem 2, $\widetilde{\beta}$ solves $\sum_{i=1}^{n} f(W_i,Y_i,Z_i;\beta,\gamma_0) = 0$. Our previous analysis has warranted that both $\widehat{\beta}$ and $\widetilde{\beta}$ are root-n consistent estimators. A stronger result here is that $\widehat{\beta}$ and $\widetilde{\beta}$ also have the same asymptotic efficiency, even though the former is derived from an estimating equation involving the estimated $\widehat{\gamma}$, and the latter, the true value γ_0.

Theorem 3. Consider parametric submodels for $\eta_1(x,z)$ and $\eta_2(\epsilon,x,z)$ denoted through $\eta_1(X,Z;\gamma_1)$ and $\eta_2(\epsilon,X,Z;\gamma_2)$, respectively, with the truth defined at $\gamma_0 = (\gamma_{10},\gamma_{20})^{\top}$. Assume $\widehat{\gamma}$ is such that $n^{1/2}(\widehat{\gamma}-\gamma_0)$ is bounded in probability. Let f be such that it satisfies $E(f|X,Y,Z) = g(X,Z)\epsilon$ for some function $g(X,Z)$. The efficiency of the estimator $\widehat{\beta}$ obtained as the root of $\sum_{i=1}^{n} f(W_i,Y_i,Z_i;\beta,\widehat{\gamma}) = 0$ is asymptotically equivalent to the estimator obtained from solving $\sum_{i=1}^{n} f(W_i,Y_i,Z_i;\beta,\gamma_0) = 0$. Both $n^{1/2}(\widehat{\beta}-\beta_0)$ and $n^{1/2}(\widetilde{\beta}-\beta_0)$ are asymptotically normal with mean zero and covariance

$$V = C^{-1}\text{var}\{f(W,Y,Z;\beta_0,\gamma_0)\}(C^{-1})^{\top},$$

where $C = E[\partial/\partial\beta^{\top}\{f(W,Y,Z;\beta_0,\gamma_0)\}]$.

Details of the proof of Theorem 3 are provided in Appendix A.

Remark 6. A particularly interesting case is when f is the efficient score S_{eff}. Since $S_{\text{eff}} \in \Lambda^{\perp}$, Theorem 3 tells us that if correct parametric models are used for $\eta_1(x,z), \eta_2(\epsilon,x,z)$, and root-$n$ estimators can be found for the nuisance parameters, then it is as if $\eta_1(x,z)$, $\eta_2(\epsilon,x,z)$ are known precisely. In this case, we achieve the optimal semiparametric efficiency. This is a stronger statement than that in Remark 4.

In practice, a correct parametric model is certainly not easy to obtain. We may be obliged to estimate $\eta_1(x,z)$ and $\eta_2(\epsilon,x,z)$, both of which depend on estimated densities of X, W, Z, and Y. Doing so can lead to a complicated procedure which is sensitive to numerical procedures. If efficiency is an important issue, we suggest proposing a relatively large model for η_1 and η_2, and proceeding with the locally efficient estimator.

2.3 Extensions for the measurement error distribution

We now extend our method to the case where the conditional probability density $p_{W|X,Z}$ contains an additional unknown parameter α, denoted by $p_{W|X,Z}(W|X,Z;\alpha)$. Estimating α typically involves either using additional information, or resorting to more sophisticated methods when no additional information is available. We now discuss both in detail.

2.3.1 With additional information

Additional information for estimating α include results from an outside experiment or from additional information such as replicated W values. Following Carroll et al. (2006), this outside information is used in forming the estimating equation

$$\sum_{j=1}^{m} \varphi(W_j, Z_j; \alpha) = 0,$$

from which consistent estimators of α can be obtained. Here $\varphi(W,Z;\alpha)$ is an appropriate estimating function for α, and is based on independent, identically distributed data (W_j, Z_j) different from the data used to estimate β.

Under this modification, Step 2 of our "Procedure for Estimating β" would first involve obtaining the consistent estimator $\hat{\alpha}$ as the root of the above estimating equation, and then calculating the relevant components of (2.2) under $p_{W|X,Z}(w|x,z;\hat{\alpha})$. That is, $\mathcal{K}$, $\mathcal{K}^*$ and $E(\cdot|X,Z)$ are calculated under $p_{W|X,Z}(w|x,z;\hat{\alpha})$ and the proposed η_1, η_2 models. The subsequent steps in the procedure would follow as given. Letting $\hat{\beta}_n(\hat{\alpha})$ denote the resulting estimator under the calculated $\hat{\alpha}$, we demonstrate in Appendix A that $\hat{\beta}_n(\hat{\alpha})$ retains the consistency and robustness properties under misspecified η_1, η_2. The estimation variance is also shown to be $V_\beta(\alpha_0) + \beta'(\alpha_0)^2 V_\alpha(\alpha_0)$ where $V_\beta(\alpha_0)$ is the estimation variance under the known α given in Theorem 2 evaluated at α_0, and $V_\alpha(\alpha_0)$ is the estimation variance for α evaluated at α_0. In contrast to the case of known α, the estimation variance for $\hat{\beta}_n(\hat{\alpha})$ is larger due to the extra variability incurred by having to estimate α.

2.3.2 Without additional information

Sometimes, a problem is still identifiable even when no additional information is available. In this situation, we modify the procedure discussed in Section 2.2.1 to estimate α along with β. To be specific, for $\theta = (\beta^T, \alpha^T)^T$, the probability density of (Y, W, Z) now equals

$$p_{W,Y,Z}(w,y,z;\theta,\eta_1,\eta_2,\eta_3) = \int p_{W|X,Z}(w|x,z;\alpha)\eta_1(x,z)\eta_2\{y - m(x,z;\beta),x,z\}\eta_3(z)dx,$$

and the score vector with respect to θ is $S_\theta = [S_{\theta,1}^T, S_{\theta,2}^T]^T$, where

$$\begin{aligned} S_{\theta,1} &= -E\{m'_\beta(X,Z;\beta)\partial \log \eta_2(\epsilon,X,Z)/\partial\epsilon | W,Y,Z\}, \\ S_{\theta,2} &= E\{\partial \log p_{W|X,Z}(w|x,z;\alpha)/\partial\alpha^T | W,Y,Z\}. \end{aligned}$$

For this new model, the nuisance tangent space Λ and its orthogonal complement $\Lambda^\perp$ have the same forms as given in Theorem 1, where β is replaced by θ. Consequently, θ can be estimated using the same procedure described in Section 2.2.1 and consistency results for the estimator $\hat{\theta}$ still hold even when η_1, η_2, or both are misspecified. A proof of this consistency result is identical to the one provided in Appendix A, where everywhere β is replaced

by θ and the expectations are calculated under the forms of η_1, η_2 and $p_{W|X,Z}(w|x, z; \alpha)$. Likewise, the results on asymptotic normality and estimation variance and efficiency established in Theorems 2 and 3 continue to hold in the same form with β replaced by θ.

2.4 Simulations

2.4.1 Implementation of the proposed method

Constructing the semiparametric estimator for β involves solving the integral equation in (2.2) for $d(X, Y, Z)$. Our general idea for implementation is to approximate $d(X, Y, Z_i)$ at each observed Z_i, $i = 1, \ldots, n$, by a linear combination of basis functions and then solve for the coefficients.

To make explicit our method, let $h_1, \ldots, h_q$ and $g_1, \ldots, g_q$ denote sets of real-valued basis functions, where the chosen number of bases q gives an accurate approximation and permits fast computation. The actual basis functions are also chosen to minimize the error between d and its summand approximation; typical basis functions that meet this criteria include Hermite polynomials, Chebychev polynomials, Fourier series, and Legendre polynomials. We express d as

$$d(X, Y, Z) = \sum_{j,k=1}^{q} c_{jk,Z} h_j(X) g_k(Y),$$

where each $c_{jk,Z}$ is a p-dimensional vector of unknown coefficients for $j, k = 1, \ldots, q$.

The operators in (2.2) involve computing expectations under posited, unknown distributions $\eta_1(x, z)$ and $\eta_2(\epsilon, x, z)$. To handle the unknown η_1, we discretize the posited density for $\eta_1(x, z)$ at r points $x_1, \ldots, x_r$ across the support of X with weights given by $\eta_1(x, z)$=$\sum_{s=1}^{r} p_s(z) I(x = x_s)$ and $\sum_{s=1}^{r} p_s(z) = 1$ for all z in the support of Z.

Under this setup, the linear integral equation (2.2) can be written as

$$\sum_{j,k=1}^{q} c_{jk,Z} \left\{ a_{jk,Z}(X, Y) \sigma^2(X, Z) + b_{jk,Z}(X, Y) \right\}$$

where $\sigma^2(X,Z) = \int \epsilon^2 \eta_2(\epsilon, X, Z) d\epsilon$, and

$$\begin{aligned} & a_{jk,Z}(X,Y) \equiv \mathcal{K}^* \circ \mathcal{K}\{h_j(X) g_k(Y)\} \\ &= \int \frac{\sum_{s=1}^r h_j(x_s) g_k(Y) p_{W|X,Z}(w|x_s, Z) \eta_2(\epsilon, x_s, Z) p_s(Z)}{\sum_{s=1}^r p_{W|X,Z}(w|x_s, Z) \eta_2(\epsilon, x_s, Z) p_s(Z)} p_{W|X,Z}(w|X,Z) d\mu(w), \end{aligned}$$

and

$$\begin{aligned} b_{jk,Z}(X,Y) &\equiv \epsilon E\{h_j(X) g_k(Y) \epsilon | X, Z\} - \epsilon E\{\mathcal{K}^* \circ \mathcal{K}\{h_j(X) g_k(Y)\} \epsilon | X, Z\} \\ &= \{Y - m(X,Z;\beta)\} \int a_{jk}(X,Y,Z) \epsilon \eta_2(\epsilon, X, Z) d\epsilon \\ &\quad - \{Y - m(X,Z;\beta)\} \int h_j(X) g_k\{m(X,Z;\beta) + \epsilon\} \epsilon \eta_2(\epsilon, X, Z) d\epsilon. \end{aligned}$$

To ultimately solve for the coefficients defining $d(X,Y,Z_i)$ at each observed Z_i, $i = 1, \ldots, n$, the functions $a_{jk,Z_i}(X,Y)$, $b_{jk,Z_i}(X,Y)$, and $m'_\beta(X, Z_i; \beta)\{Y - m(X, Z_i; \beta)\}$ are evaluated at q^2 points (x_ℓ, y_m) for $\ell, m = 1, \ldots, q$. Doing so leads to p linear systems of size $q^2 \times q^2$, from which we may solve for c_{jk,Z_i}'s. The process is repeated for each Z_i, $i = 1, \ldots, n$.

Upon solving for the coefficients, we invoke the relation $\mathcal{K}(d) = S(W,Y,Z;\beta,\eta_1,\eta_2)$ to form $\sum_{i=1}^n S(W_i, Y_i, Z_i; \beta, \eta_1, \eta_2) = 0$ whose root, $\widehat{\beta}$, is the desired estimate. Specifically, $S(W,Y,Z;\beta,\eta_1,\eta_2) = \mathcal{K}\{d(X,Y,Z)\}$ equals

$$\frac{\sum_{s=1}^r \sum_{j,k} c_{jk,Z} h_j(x_s) g_k(Y) p_{W|X,Z}(W|x_s, Z) \eta_2(\epsilon, x_s, Z) p_s(Z)}{\sum_{s=1}^r p_{W|X,Z}(W|x_s, Z) \eta_2(\epsilon, x_s, Z) p_s(Z)},$$

and is evaluated at the observed data (W_i, Y_i, Z_i), $i = 1, \ldots, n$.

2.4.2 Simulation examples

To illustrate the performance of our method, we consider two simulated examples where Y is related to unobserved covariates X through a nonlinear function $m(X;\beta)$, and the measurement error is normal additive.

Our first model is

$$Y = 0.7 \exp(-\beta X^2) + \epsilon, \quad W = X + U,$$

where $\beta = 0.5$, U is normally distributed with mean 0 and variance 0.1, and the covariate X has a uniform distribution on $[1.1 - \sqrt{0.9}, 1.1 + \sqrt{0.9}]$. To demonstrate the robustness of our method, we considered two situations for the nuisance parameters where the posited densities for η_1 and η_2 were different from the true densities.

Situation 1: The true distribution for model error ϵ is a t-distribution with 5 degrees of freedom, whereas the posited density η_2 is $N(0, 0.4^2)$. To contrast from the uniform distribution of the covariate X, we posited a misspecified η_1 as $N(1.1, 0.1)$.

Situation 2: Secondly, we considered the true distribution for ϵ as a mixture of normals, $N(0.5, 0.4^2)$ and $N(-0.5, 0.4^2)$, with equal weights. In comparison, we posited a misspecified η_2 as $N(0, 0.29)$. As in Situation 1, the posited and misspecified distribution for η_1 is $N(1.1, 0.1)$.

As an extension to the first model, the second model under consideration is

$$\begin{aligned} Y &= \beta_2 \exp(-\beta_1 X^2) + \epsilon, \\ W &= X + U, \end{aligned}$$

where $\beta = (0.5, 0.7)^T$. The rest of the simulation set-up is identical to the first model.

Following the procedure described in Section 2.4.1, $d(X, Y)$ in (2.2) was approximated using five basis functions, with Hermite polynomials being the sets of real-valued basis functions $h_1, \ldots, h_q$ and $g_1, \ldots, g_q$ approximating $d(X, Y)$. The posited density $\eta_1(x)$, was discretized at $r = 200$ grid points allocated evenly across the range of $\mu_X \pm 3\sigma_X$ where μ_X and σ_X represent the mean and standard deviation, respectively, for $\eta_1(x)$; that is, $\mu_X = 1.1$ and $\sigma_X = \sqrt{0.1}$. The values for q and r were selected empirically, and provided numerically accurate and stable results for all situations considered. The functions $a_{jk}(X, Y), b_{jk}(X, Y)$, and $m'_\beta(X; \beta)\{Y - m(X; \beta)\}$ described in Section 2.4.1 were evaluated at q^2 Hermite quadrature points (x_ℓ, y_m) for $\ell, m = 1, \ldots, q$. All integrals were

calculated using Gauss quadrature approximation. Finally, the estimator $\hat{\beta}$ was obtained using a modification of the Powell hybrid method (Moré et al., 1984).

To illustrate the performance of our method we compared it with three other candidate methods. First, to demonstrate that the measurement error is not ignorable, we considered the "naive" estimator which employs least squares and assumes X and W are the same. A possible competing method is the Tsiatis and Ma (2004) method (TM) which accounts for misspecification in the covariate distribution, but requires a correctly specified model error distribution. In particular, to demonstrate the sensitivity of the variance-covariance structure of posited η_2 in the TM method, we specified two variance structures of η_2 for each situation described above: (1) homoscedastic variance, σ_ϵ^2 and (2) heteroscedastic variance, $\sigma_\epsilon^2 x^2$. In each simulation, we compared our method to the Tsiatis-Ma estimator assuming homoscedastic model errors (TM-Hom), and the Tsiatis-Ma estimator assuming heteroscedastic model errors (TM-Het). Because our semiparametric estimator is free of assumptions about the variance-covariance structure of the model error, we expect our method (Semipar) to be robust and outperform the TM method when the variance structure is misspecified.

For each scenario described, 1000 simulations with sample size $n = 500$ were conducted. The variance-covariance matrix for the estimator was estimated using the empirical version of the variance described in Theorem 2, and the coverage rate of the estimated 95%-confidence interval was also calculated. Results are summarized in Tables 1, 2 and 3.

Table 1: Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter β in Model 1 and under the two proposed situations. The true parameter is $\beta_{\text{true}} = 0.5$. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar).

	Situation 1			
	Naive	TM-Hom	TM-Het	Semipar
Gen. Hom. Error				
bias	-0.026	0.002	-0.057	0.001
var	0.004	0.004	0.003	0.004
$\widehat{\text{var}}$	0.005	0.004	0.004	0.004
CI (%)	88.3	94.1	78.3	94.1
Gen. Het. Error				
bias	-0.017	-0.036	-0.019	0.008
var	0.007	0.004	0.006	0.007
$\widehat{\text{var}}$	0.008	0.004	0.006	0.007
CI (%)	93.0	87.6	91.9	94.5

	Situation 2			
	Naive	TM-Hom	TM-Het	Semipar
Gen. Hom. Error				
bias	-0.022	0.007	0.025	0.003
var	0.012	0.006	0.016	0.007
$\widehat{\text{var}}$	0.013	0.007	0.017	0.007
CI (%)	91.4	96.0	95.5	94.8
Gen. Het. Error				
bias	-0.013	-0.018	-0.004	0.004
var	0.008	0.0102	0.008	0.008
$\widehat{\text{var}}$	0.008	0.011	0.008	0.009
CI (%)	94.9	96.8	95.1	95.5

Table 2: Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter $\hat{\beta}$ in Model 2 and under Situation 1. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). The true parameter is $\beta_{\text{true}} = (0.5, 0.7)^T$. Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar).

Situation 1	Naive		TM-Hom	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Gen. Hom. Error				
bias	-0.088	-0.054	0.004	0.004
var	0.005	0.002	0.009	0.003
$\widehat{\text{var}}$	0.009	0.003	0.009	0.003
CI (%)	91.0	90.4	95.6	95.0
Gen. Het. Error				
bias	0.074	-0.049	-0.042	-0.026
var	0.021	0.002	0.009	0.002
$\widehat{\text{var}}$	0.021	0.003	0.009	0.002
CI (%)	68.0	93.9	91.0	90.4

	TM-Het		Semipar	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Gen. Hom. Error				
bias	0.251	0.187	-0.022	-0.015
var	0.064	0.023	0.007	0.003
$\widehat{\text{var}}$	0.065	0.023	0.009	0.002
CI (%)	63.7	48.0	95.6	95.1
Gen. Het. Error				
bias	-0.035	-0.003	-0.038	0.023
var	0.005	0.0004	0.024	0.010
$\widehat{\text{var}}$	0.006	0.0005	0.025	0.011
CI (%)	91.4	96.6	97.7	94.7

Table 3: Bias of the parameter estimates (bias), sample variances (var), the mean of estimated variances ($\widehat{\text{var}}$), and estimated 95% confidence interval coverage probabilities (CI) for the parameter β in Model 2 and under Situation 2. Data is generated with homoscedastic model errors (Gen. Hom. Error) and with heteroscedastic model errors (Gen. Het. Error). The true parameter is $\beta_{\text{true}} = (0.5, 0.7)^T$. Methods reported include the naive method (Naive), Tsiatis-Ma method for homoscedastic model errors in estimation procedure (TM-Hom) and for heteroscedastic model errors in estimation procedure (TM-Het), and our proposed method (Semipar).

Situation 2	Naive		TM-Hom	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Gen. Hom. Error				
bias	-0.085	-0.054	0.010	0.003
var	0.009	0.003	0.015	0.005
$\widehat{\text{var}}$	0.010	0.004	0.015	0.005
CI (%)	78.3	83.6	94.0	94.3
Gen. Het. Error				
bias	-0.081	-0.055	-0.028	-0.018
var	0.012	0.003	0.014	0.004
$\widehat{\text{var}}$	0.014	0.004	0.013	0.003
CI (%)	80.3	84.2	90.4	91.9

	TM-Het		Semipar	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Gen. Hom. Error				
bias	0.035	0.022	-0.017	-0.011
var	0.041	0.010	0.010	0.004
$\widehat{\text{var}}$	0.041	0.010	0.011	0.004
CI (%)	93.2	94.3	95.4	95.4
Gen. Het. Error				
bias	-0.026	0.001	-0.025	-0.014
var	0.013	0.003	0.017	0.005
$\widehat{\text{var}}$	0.014	0.004	0.018	0.005
CI (%)	91.4	95.0	95.4	93.9

Parameter estimates from the naive method are largely biased, indicating that the measurement error is significant enough and cannot be ignored. Furthermore, the naive method has the worst coverage probabilities compared to the other methods, with coverage probabilities tending to be less than the nominal 95%. The TM method improves over the naive method when the model error variance structure is correctly specified. Under this scenario, the TM method has one less nonparametric term than the Semipar method and thus is more precise. In these cases, the TM method has little bias and nearly perfect nominal coverage probabilities. However, the TM method heavily relies on the correctness of the model error variance assumption. When the variance structure is misspecified, the TM method performs poorly compared to the Semipar method. The large bias and smaller coverage probabilities are most notable for the results of Situation 1 in Table 1 both when the true variance structure is homoscedastic while we employ TM-Het, and when the true variance structure is heteroscedastic while we employ TM-Hom. Even worse, in Table 2 for Situation 1 when the true variance structure is homoscedastic and we use TM-Het, we see very large bias and coverage probabilities less than 65%. In these notable cases and throughout, the Semipar method consistently had little bias, the average of the estimated variances closely approximates the sample variances, and the estimated 95% confidence interval coverage probabilities were close to nominal.

These results demonstrate that measurement error cannot simply be ignored as the naive method certainly leads to unreliable parameter estimates. Even after accounting for measurement error, our results show that using a method (i.e., TM) which relies on a correctly specified variance structure for the model error also gives incorrect results when the assumption does not hold. In practice, specifying the model error variance structure correctly is almost impossible since residuals are not obtainable in the measurement error models. In summary, the Semipar method is a significant improvement over the TM method, giving unbiased parameter estimates with actual coverages close to nominal for

the general regression with measurement error when the model error *and* latent variable distributions are misspecified.

2.5 A case study in nutrition

Flag et al. (2000) carried out a study to evaluate the validity of a Nutrition Survey conducted by the American Cancer Society in 1992-1993. Four-hundred forty one participants completed four 24 hour dietary recall interviews given over a one-year period, as well as a second FFQ survey similar to that from the original study. The data consist of estimates for energy, calorie percentages from fat intake, and estimates of saturated fat intake. Interest lies in understanding the relationship between Percent Calories from Fat (Y) and Saturated Fat intake. Because Saturated Fat intake was calculated through repeated measurements, only an approximation is available.

In our analysis, we considered the male subgroup of respondents which consisted of 317 individuals, each with two repeated measurements of Saturated Fat intake. With a log transformation, the difference of the two measurements is acceptably normally distributed; see Figure 1 for the qqplot of the difference of these measurements before and after the log transformation. This condition was further evaluated through a Pearson Chi-squared test where we used 10 to 20 bins for testing and obtained a p-value at least as large as 0.63, thus assuring the normality assumption. We denote the log transformation of Saturated Fat intake X, and the corresponding average of the two measurements W. Our analysis warrants assuming $W = X + U$, where U is a mean zero normal random variable with variance 0.332^2. Because nutrition models usually assume Percent Calories from Fat is related to Saturated Fat intake through a linear regression, we have

$$Y = \beta_1 \exp(X) + \beta_2 + \epsilon, \quad W = X + U, \quad E(\epsilon|X) = 0, \quad U \sim N(0, 0.332^2).$$

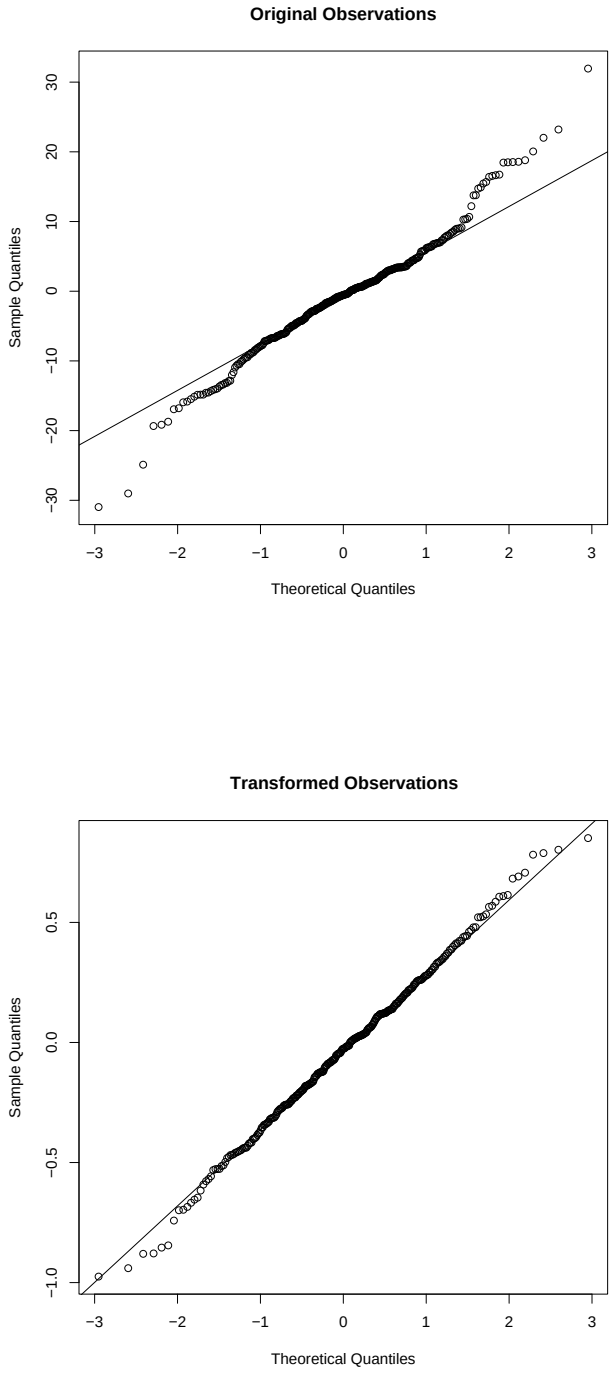

Figure 1: Quantile-quantile plots of the measurement error U for the original first and third readings of the 24 hour recall surveys (top) and after the logarithm transform (bottom).

To estimate β_1 and β_2 using our method, we posited a normal distribution with mean zero and variance 0.316 for the covariate distribution $\eta_1(x)$, and a normal model with mean zero and variance 0.29 for the density of the model error given X, $\eta_2(\epsilon, x)$. Furthermore, we calculated the corresponding estimated variances using the sandwich estimator described in Theorem 2. For comparison, we also included the naive estimate calculated as the maximum likelihood estimator assuming X and W are the same.

The naive estimate yields $\widetilde{\beta} = (0.709, -0.819)^T$ with estimated variances var$(\widetilde{\beta}_1) = 0.0064$ and var$(\widetilde{\beta}_2) = 0.0117$. In contrast, the proposed method yields estimates $\widehat{\beta} = (1.611, -1.768)^T$ with estimated variances var$(\widehat{\beta}_1) = 0.0164$ and var$(\widehat{\beta}_2) = 0.0243$. The stark contrast between the estimates obtained from the naive approach and from the semiparametric method approach indicates that the measurement error here needs to be taken into account. If either the distribution of X or ϵ given X is misspecified, then the classical maximum likelihood estimator, even taking into account the measurement error, would be inconsistent. Our method, however, allows for both distributions to be misspecified and gives a more reasonable relationship. Results from our method imply that a one unit increase in Saturated Fat (X) is associated with an estimated increase of 1.611 units in the mean of Percent Calories (Y), more than twice as large as the naive estimates would conclude. Hence, ignoring the measurement error gravely underestimates how much Saturated Fat intake affects an individual's Calorie Percentage.

2.6 Discussion

For the measurement error problem in restricted moment models, identifiability of β is relatively easy to analyze. The proof of identifiability usually resorts to deconvolution and invokes basic results from Fourier inversion. More formal justifications of identifiability are discussed in Hu and Schennach (2006) and Chen et al. (2009). Based on the established identifiability, our focus here is on statistical methodology which provides consistent esti-

mators when both the measurement error and latent variable distributions are unknown, and on valid inference tools. The proposed estimator is derived via a semiparametric procedure different from that in Tsiatis and Ma (2004), and is the first known in its generality that is robust to various distribution mis-specifications.

CHAPTER III

EFFICIENCY IMPROVEMENT IN A CLASS OF SURVIVAL MODELS THROUGH MODEL-FREE COVARIATE INCORPORATION*

3.1 Introduction

In randomized two-arm clinical trials, we are often interested in comparing patient survival between treatment and control groups. Typically, the log-rank test is used to examine the survival differences, and under a proportional hazards assumption (Cox, 1972), the log-rank test is known to be optimal. However, under nonproportional hazards, especially when the two survival curves cross, the log-rank test may incur substantial power loss. Extensive research has been carried out to extend the scope of the proportional hazards model. For example, Hess (1994) explored nonparametric modifications of the Cox model, and Verweij and Van Houlwelingen (1995) studied time-varying coefficients in the regression model. Hsieh (2001), Bagdonavicius et al. (2004), and Zeng and Lin (2007) studied more general classes of hazard regressions. To accommodate nonproportional hazards, Yang and Prentice (2005) proposed a novel class of survival models, which includes the proportional hazards and the proportional odds structures (Bennet, 1983), and other flexible modeling structures in between. In this paper, we demonstrate that incorporating auxiliary covariates into the general class of survival models by Yang and Prentice (2005) improves estimation efficiency and provides a better approach to comparing survival curves.

In regression models, covariates are generally included to account for their effects on

---

* Reprinted with kind permission from Springer Science+Business Media: *Lifetime Data Analysis*, "Efficiency improvement in a class of survival models through model through model-free covariate incorporation", **18**, 2011, p. 1-14, by Tanya P. Garcia, Yanyuan Ma, and Guosheng Yin.

treatment for post-randomization adjustment. For an extensive overview of covariate analysis, see Senn (1989), Hauck et al. (1998), Koch et al. (1998), Tangen and Koch (1999), Lesaffre and Senn (2003), Grouin et al. (2004), and Zhang et al. (2008). Although including covariates such as patients' health conditions and demographic information can generally lead to improved parameter estimates, this efficiency gain is not always guaranteed. For logistic regression, Robinson and Jewell (1991) demonstrated that directly modeling non-confounding predictive covariates using nonlinear regression may actually lead to a loss of precision. The interpretations of the parameters in the unadjusted and covariate-adjusted logistics models are different; in the former, the parameters characterize the unconditional odds ratios, and in the latter, they refer to the conditional odds ratios. Likewise, Wickramaratne and Holford (1989) provided an example with $2 \times 2 \times 2$ contingency tables where the variance of the stratified estimate is higher than that of the estimate using the pooled (log) odds ratio. Therefore, one needs to be cautious when incorporating covariates into a model to avoid worsened precision and conflicting parameter interpretations.

In the analysis of time-to-event data, the aforementioned general classes of survival models typically handle auxiliary covariates in the usual regression setting. On the other hand, Lu and Tsiatis (2008) proposed incorporating covariates through augmented estimating equations in the log-rank test. Their method does not impose extra modeling assumptions and may gain substantial efficiency. To improve modeling efficiency and robustness in the flexible class of Yang and Prentice (2005), we propose incorporating auxiliary covariates via the semiparametric principles in Zhang et al. (2008) and Lu and Tsiatis (2008). Accordingly, we take an unbiased estimating equation and augment it with auxiliary covariates that carry extra information about the times to events beyond the treatments. For an appropriate, model-free choice of the augmentation term, we can produce a consistent, more efficient estimator, and avoid model misspecification. The resulting model is also a generalization of the results by Lu and Tsiatis (2008) which only handles the proportional

hazards setting.

The rest of this chapter is organized as follows. Section 3.2 introduces the notation and model assumptions needed for improving efficiency through covariate augmentation. We also briefly describe the class of survival models of Yang and Prentice (2005) along with asymptotic results. We propose the augmented estimating equations with auxiliary covariates, and develop the asymptotic properties using semiparametric theory in Section 3.3. Simulation studies in Section 3.4 and an application to the leukemia study in Section 3.5 demonstrate the satisfactory performance of the proposed method.

3.2 Notation and semiparametric models

In a typical randomized two-arm study, we observe independent and identically distributed data $(Y_i, \delta_i, Z_i, W_i)$ for $i = 1, \ldots, n$. Here, $Y_i = \min(T_i, C_i)$ denotes an individual's event time, where T_i is the survival time and C_i is the censoring time; $\delta_i = I(T_i \leq C_i)$ is the censoring indicator; the treatment indicator $Z_i = 0$ if the subject is in the control group and 1 otherwise; and W_i is a vector of auxiliary covariates that are correlated with T_i, such as a patient's age, health condition and demographic information.

Intuitively, when T and W are correlated conditional on Z, i.e., W contains extra information about T given Z, then inclusion of W in a regression model would improve efficiency. To ensure consistency and asymptotic normality of the resulting improved estimator, we require the following two assumptions. First, we assume Z and W are independent,

$$Z \amalg W, \tag{3.1}$$

which is generally satisfied by randomization, where the randomization probability $P(Z = 1) = \pi$, $0 < \pi < 1$, is usually known. Condition (3.1) ensures that our model-free incorporation of W will produce an unbiased estimator based on (Y, δ, Z, W). Second, we

assume C and (T, W) are independent conditional on Z,

$$C \amalg (T, W)|Z,$$

which implies independent censoring and thus guarantees model identifiability.

Before illustrating the method of covariate augmentation, we first briefly describe the general class of survival models by Yang and Prentice (2005), in which the auxiliary covariate W is not considered. For $j = 0, 1$, let $\lambda_j(t)$ denote the hazard function under treatment j and $S_j(t) = \exp\{-\int_0^t \lambda_j(s)ds\}$ denote the survival function. Yang and Prentice (2005) proposed the following general class of survival models:

$$\lambda_1(t) = \frac{\gamma_1\gamma_2}{\gamma_1 + (\gamma_2 - \gamma_1)S_0(t)}\lambda_0(t),\ 0 < t < \tau, \tag{3.2}$$

where $\tau = \sup\{t : S_0(t) > 0\}$, and $\gamma_1, \gamma_2 > 0$. Under model (3.2), the hazard ratio $\lambda_1(t)/\lambda_0(t)$ has several appealing properties: it monotonically increases when $\gamma_2 > \gamma_1$, and monotonically decreases when $\gamma_2 < \gamma_1$, leading to a broad range of nonproportional hazards models; it includes the Cox proportional hazards model when $\gamma_1 = \gamma_2$, and the proportional odds model when $\gamma_2 = 1$. In addition, as

$$\gamma_1 = \lim_{t\downarrow 0} \frac{\lambda_1(t)}{\lambda_0(t)},\ \gamma_2 = \lim_{t\uparrow \tau} \frac{\lambda_1(t)}{\lambda_0(t)},$$

γ_1 and γ_2 are naturally interpreted as the short- and long-term hazard ratios, respectively.

In two-arm clinical trials, it may happen that one group initially exhibits a higher survival rate, but later shows a lower survival rate, or vice versa. Such phenomenon may lead to nonproportional hazards, or even crossing survival curves, a feature captured by model (3.2). To see this, define the odds function of the treatment 0 group as

$$R(t) = \frac{1 - S_0(t)}{S_0(t)}.$$

Then the survival function for each treatment satisfies

$$S_0(t) = \{1 + R(t)\}^{-1},\ S_1(t) = \left\{1 + \frac{\gamma_1}{\gamma_2}R(t)\right\}^{-\gamma_2},$$

from which it can be seen that $S_0(t)$ and $S_1(t)$ would cross when $\gamma_1 < 1$ and $\gamma_2 > 1$, or $\gamma_1 > 1$ and $\gamma_2 < 1$. Figure 2 shows examples of crossed survival curves for $R(t) = t$.

To better understand treatment short- and long-term effects, Yang and Prentice (2005) developed a pseudo-likelihood procedure for estimating γ_1, γ_2, or equivalently, $\beta = (\beta_1, \beta_2)'$ where each component $\beta_j = \log\gamma_j$ for $j = 1, 2$. Define the estimated martingale as

$$\hat{M}_i(t;\beta) = \delta_i I(Y_i \le t) - \int_0^t I(Y_i \ge s) \frac{d\hat{R}(s;\beta)}{\exp(-\beta_1 Z_i) + \exp(-\beta_2 Z_i)\hat{R}(s;\beta)},$$

where the estimated version of $R(t)$ is

$$\begin{aligned} \hat{R}(t;\beta) &= \frac{1}{\prod_{s\le t}\{1 - \Delta\hat{\Psi}(s;\beta_2)\}} \\ &\quad \times \int_0^t \prod_{u \le s^-} \{1 - \Delta\hat{\Psi}(u;\beta_2)\} \frac{\sum_{i=1}^n \delta_i \exp(-\beta_1 Z_i) I(Y_i \le s)}{\sum_{i=1}^n I(Y_i \ge s)} ds, \end{aligned}$$

and $\Delta\hat{\Psi}(t;\beta_2)$ is the jump size of $\hat{\Psi}$ at t for

$$\hat{\Psi}(t;\beta_2) = \int_0^t \frac{\sum_{i=1}^n \delta_i \exp(-\beta_2 Z_i) I(Y_i \le s)}{\sum_{i=1}^n I(Y_i \ge s)} ds.$$

The estimator $\hat{\beta}$ is the zero of

$$U(\beta) = \sum_{i=1}^n \int_0^\tau g_i(t;\beta) d\hat{M}_i(t;\beta), \tag{3.3}$$

where $g_i = (g_{1i}, g_{2i})'$ with

$$\begin{aligned} g_{1i}(t;\beta) &= Z_i \frac{\exp(-\beta_1 Z_i)}{\exp(-\beta_1 Z_i) + \exp(-\beta_2 Z_i)\hat{R}(t;\beta)}, \\ g_{2i}(t;\beta) &= Z_i - g_{1i}(t;\beta). \end{aligned}$$

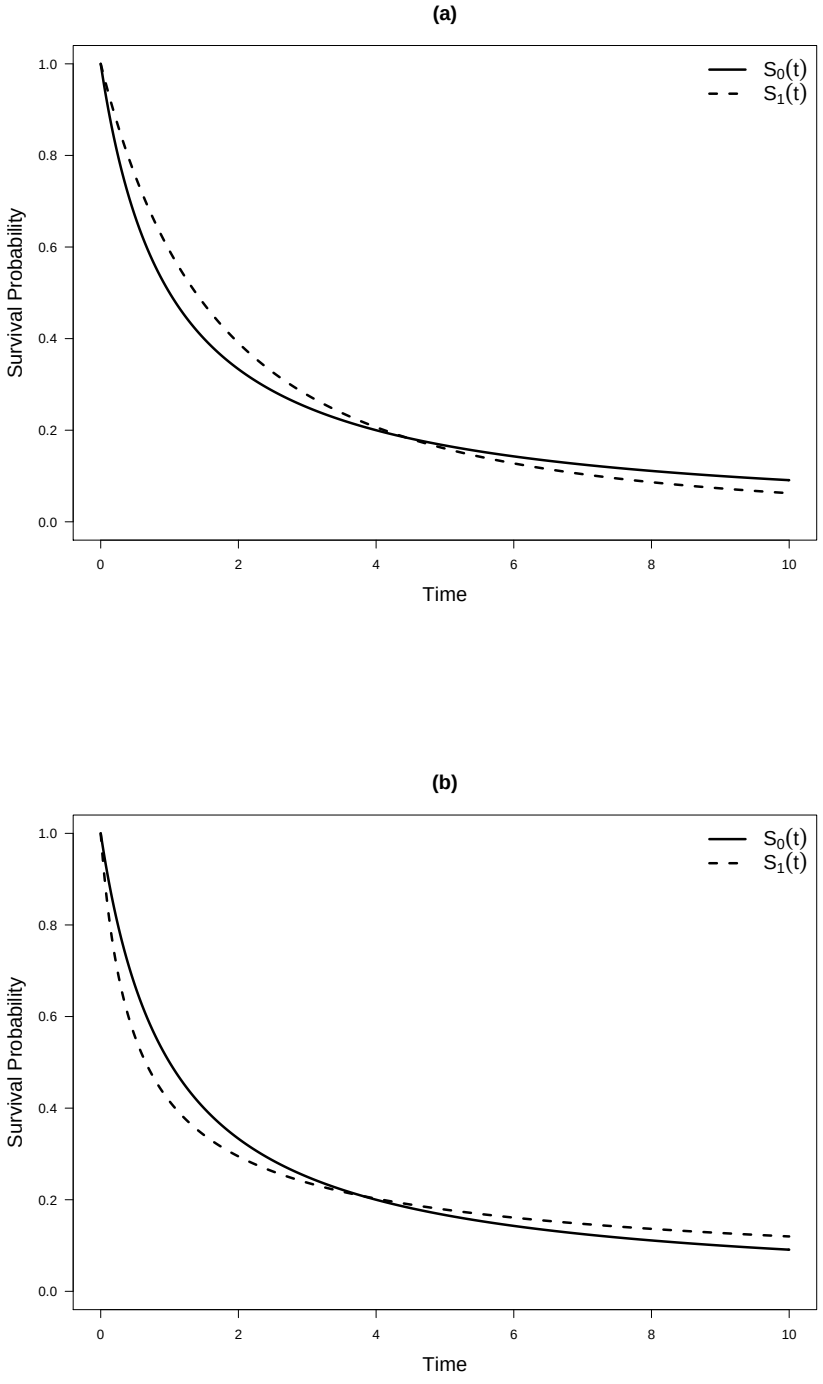

Figure 2: Crossing survival curves when $R(t) = t$, with (a) $\gamma_1 = 0.6$ and $\gamma_2 = 2$; (b) $\gamma_1 = 2$ and $\gamma_2 = 0.6$.

Yang and Prentice (2005, Thm. A.2) showed that under regularity conditions,

$$U(\beta) = \sum_{i=1}^{n} \int_0^\tau \zeta_i(t;\beta) dM_i(t;\beta)$$

with $\zeta_i = (\zeta_{1i}, \zeta_{2i})'$ defined in Appendix B and $M_i(t;\beta)$ analogous to $\hat{M}_i(t;\beta)$ except with $\hat{R}(t;\beta)$ replaced by $R(t)$. Furthermore, $\hat{\beta}$ is a consistent estimator of the true parameter β_0, and is asymptotically normal with covariance $\Gamma^{-1}\Omega(\Gamma^{-1})'$ where

$$\Gamma = E\left\{-\frac{\partial}{\partial\beta'}U(\beta)|_{\beta=\beta_0}\right\},\ \Omega = E\left[\left\{\int_0^\tau \zeta(t;\beta_0)dM(t;\beta_0)\right\}^{\otimes 2}\right],$$

with $v^{\otimes 2} = vv'$ for any vector v.

3.3 Improving efficiency through covariate augmentation

Following Zhang et al. (2008) and Lu and Tsiatis (2008), we develop a more efficient estimator for β than that obtained from (3.3) by appropriately incorporating auxiliary covariates. In this regard, under the semiparametric theory framework (Tsiatis, 2006), we construct estimating equations for β based on (Y, δ, Z, W) by augmenting (3.3) with auxiliary covariates:

$$U_{\mathrm{AUG}}(\beta) = \sum_{i=1}^{n}\left\{\int_0^\tau g_i(t;\beta)d\hat{M}_i(t;\beta) - (Z_i - \pi)\Gamma h(W_i)\right\}.$$

Here, $h(W)$ is an arbitrary function of W which could depend on some additional parameter α, in which case we write $h(W;\alpha)$. Independence of Z and W from (3.1) ensures consistency of the estimator of β regardless of the choice of h. Our augmented estimating equation above differs from that suggested in Zhang et al. (2008) in that we involve Γ, whereas they do not. We demonstrate below that involving Γ in the augmentation term will ensure improved efficiency of $\hat{\beta}_{\mathrm{AUG}}$, the root of $U_{\mathrm{AUG}}(\beta) = 0$, over $\hat{\beta}$ regardless of the posited choice of h. For optimality, with h as

$$h(W) = \frac{1}{\pi(1-\pi)}E\left\{(Z-\pi)\Gamma^{-1}\int_0^\tau g(t;\beta)d\hat{M}(t;\beta)\Big|W\right\}, \tag{3.4}$$

the resulting estimator will exploit the most information from the correlation of T and W (Zhang et al., 2008, Appendix). The derivation leading to $h(W)$ in (3.4) requires that $\int_0^\tau g(t;\beta)d\hat{M}(t;\beta) - (Z-\pi)\Gamma h(W)$ and $(Z-\pi)\Gamma h(W)$ are orthogonal to each other. This orthogonality implies that when $h(W)$ is as in (3.4), the augmented estimating equation has less variability than the unadjusted estimating equation. Consequently, if (3.4) were exactly known, the estimator $\hat{\beta}_{\mathrm{AUG}}$ would indeed be more efficient than $\hat{\beta}$.

Due to the difficulty involved in calculating the unknown conditional expectation of (3.4), we follow Zhang et al. (2008) and Lu and Tsiatis (2008) to circumvent this challenge by imposing a parametric model for $h(W)$ denoted $h(W;\alpha)$. For each component of h, we take $h_j(W;\alpha_j) = \alpha_j' q_j(W)$, $j = 1, 2$, so that $h_j(\cdot)$ is simply a linear regression model with the unknown coefficients α_j, and $q_j(W)$ is a vector of arbitrary known functions of W such as polynomials or splines. Under this parametrization, we choose the elements of $\alpha = (\alpha_1', \alpha_2')'$ to minimize the trace of

$$\mathrm{cov}\left[\hat{\Gamma}^{-1}\sum_{i=1}^{n}\left\{\int_0^\tau g_i(t;\hat{\beta})d\hat{M}_i(t;\hat{\beta}) - (Z_i-\pi)\hat{\Gamma}h(W_i;\alpha)\right\}\right],$$

where $\hat{\beta}$ is the unadjusted estimator from (3.3) and $\hat{\Gamma}$ is the sample version of Γ evaluated at $\hat{\beta}$. Simple algebra shows that

$$\begin{aligned}\hat{\alpha}_j &= \left\{\pi(1-\pi)\sum_{i=1}^{n} q_j(W_i)q_j'(W_i)\right\}^{-1} \\ &\quad \times \sum_{i=1}^{n} q_j(W_i)(Z_i-\pi)\int_0^\tau \{\hat{\Gamma}^{-1}g_i(t;\hat{\beta})\}_j d\hat{M}_i(t;\hat{\beta})\end{aligned} \tag{3.5}$$

is the desired minimizer for $j = 1, 2$, where $\{\hat{\Gamma}^{-1}g_i(t;\hat{\beta})\}_j$ corresponds to the jth row of the matrix product $\hat{\Gamma}^{-1}g_i(t;\hat{\beta})$. Thus the augmented estimator $\hat{\beta}_{\mathrm{AUG}}$ solves $U_{\mathrm{AUG}}(\beta) = 0$, where $\Gamma h(W)$ is replaced by $\hat{\Gamma}h(W;\hat{\alpha})$, and $\hat{\alpha} = (\hat{\alpha}_1', \hat{\alpha}_2')'$ is given in (3.5).

To make explicit the efficiency improvement of $\hat{\beta}_{\mathrm{AUG}}$ over $\hat{\beta}$, we first characterize the asymptotic behavior of $\hat{\beta}_{\mathrm{AUG}}$. Following similar arguments in Yang and Prentice (2005),

we have that when the augmentation term in $U_{\mathrm{AUG}}(\beta)$ is a simple linear regression as above, then

$$U_{\mathrm{AUG}}(\beta) = \sum_{i=1}^{n} \left\{ \int_0^\tau \zeta_i(t;\beta) dM_i(t;\beta) - (Z_i - \pi)\Gamma h(W_i;\alpha^*) \right\},$$

with the components of α^* defined as in (3.5) except with $\hat{\Gamma} g(t;\beta)$ replaced by $\Gamma\zeta(t;\beta)$ and everywhere evaluated at β_0 instead of $\hat{\beta}$. The estimator $\hat{\beta}_{\mathrm{AUG}}$ is consistent, asymptotically normal with covariance $\Gamma^{-1}\Omega_{\mathrm{AUG}}(\Gamma^{-1})'$, where

$$\Omega_{\mathrm{AUG}} = E\left[\left\{ \int_0^\tau \zeta(t;\beta_0) dM(t;\beta_0) - (Z - \pi)\Gamma h(W;\alpha^*) \right\}^{\otimes 2} \right].$$

In practice, estimating the covariance matrix corresponds to utilizing the sample versions $\hat{\Gamma}$ and $\hat{\Omega}_{\mathrm{AUG}}$, where $M(t;\beta_0)$, $R(t)$ and β_0 are replaced by $\hat{M}(t;\hat{\beta}_{\mathrm{AUG}})$, $\hat{R}(t;\hat{\beta}_{\mathrm{AUG}})$ and $\hat{\beta}_{\mathrm{AUG}}$, respectively.

The estimator $\hat{\beta}$ will have more variability than $\hat{\beta}_{\mathrm{AUG}}$ if the difference of their corresponding covariances results in a matrix with positive diagonal elements. Our proposed method, however, ensures this positivity by the construction of α^*. Simple algebra shows that α^* actually minimizes the diagonal element values of $\Gamma^{-1}(\Omega - \Omega_{\mathrm{AUG}})(\Gamma^{-1})'$. That is, for $j = 1, 2$, $\alpha_j^* = \{\mathrm{cov}(B_j)\}^{-1}\mathrm{cov}(A_j, B_j)$ where A_j is the jth row of the matrix product $\Gamma^{-1}\int_0^\tau \zeta(t;\beta_0) dM(t;\beta_0)$ and $B_j = (Z - \pi)q_j(W)$. With α^* as defined, the jth diagonal element of $\Gamma^{-1}(\Omega - \Omega_{\mathrm{AUG}})(\Gamma^{-1})'$ equals $\mathrm{cov}(A_j, B_j)\{\mathrm{cov}(B_j)\}^{-1}\mathrm{cov}(A_j, B_j)'$, which is certainly positive. Consequently even if the proposed form of $q_j(W)$ in $h_j(W;\alpha)$ is misspecified, our method for augmentation will lead to more efficient estimators.

For completeness, we outline the algorithm for computing $\hat{\beta}_{\mathrm{AUG}}$:

(1) Solve the unadjusted estimating equation $U(\beta) = 0$ for $\hat{\beta}$. For $j = 1, 2$, define $h_j(W;\alpha) = \alpha_j' q_j(W)$ and obtain the ordinary least square estimator $\hat{\alpha}_j$ which depends on $\hat{\beta}$ as defined in (3.5).

(2) Plug $\hat{\alpha} = (\hat{\alpha}_1', \hat{\alpha}_2')'$ in $U_{\mathrm{AUG}}(\beta) = 0$ with $\Gamma h(W)$ replaced by $\hat{\Gamma}^{-1} h(W; \hat{\alpha})$, and solve for $\hat{\beta}_{\mathrm{AUG}}$.

We implemented the two-stage algorithm in Fortran 90 by invoking a quick sorting algorithm (Singleton, 1969) and a modification of the Powell hybrid method (Moré et al., 1984) as the root-finding method. The program is available upon request.

3.4 Simulations

We conducted Monte Carlo simulation studies to compare the efficiencies of $\hat{\beta}_{\mathrm{AUG}}$ and $\hat{\beta}$. We took the odds function $R(t) \equiv t$, the identity function, which led to the hazard function for subject i as

$$\lambda(t|Z_i) = \frac{1}{\exp(-\beta_1 Z_i) + \exp(-\beta_2 Z_i)t}, \; i = 1, \ldots, n.$$

We generated independent treatment indicators Z from a Bernoulli(π) distribution; and (W, V) from a bivariate normal density with zero mean, variances 1, and correlation ρ. Setting the survival time

$$T = (\gamma_2/\gamma_1)^z \left[\{1 - \Phi(V)\}^{-1/\gamma_2^z} - 1\right],$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, simple algebra shows that the conditional distribution of T given Z is

$$F_{T|Z}(t|z; \gamma_1, \gamma_2) = 1 - \left\{\frac{1}{1 + (\gamma_1/\gamma_2)^z t}\right\}^{\gamma_2^z},$$

which coincides with the conditional distribution generated by the individual hazard function given above. To estimate the augmented expectation, we used a parametric model of the form $h_j(W; \alpha_j) = \alpha_{j0} + \alpha_{j1} W + \alpha_{j2} W^2$ for $j = 1, 2$. Finally, we generated independent censoring variables from a log-normal distribution where the normal distribution had mean c and standard deviation 0.5. Varying the value of c achieved different censoring proportions.

To exemplify the flexibility of the general class of survival curves discussed in Section 3.2, we considered two sets of β-values. First, $\beta = (0, 0)'$ yields the Cox proportional hazards model and corresponds to the situation of no treatment effect. Second, $\beta = (0.5, -0.5)'$ yields a nonproportional hazards model where the survival curves cross. In this case, the treatment effect is initially non-evident but slowly becomes positive. In all scenarios, treatment assignment probability π is 0.5, and we chose c to yield 0% and 30% censoring. We set $\rho = 0.4$ and $\rho = 0.7$ which led to a conditional correlation between T and W given Z of 0.15 and 0.27, respectively, when $\beta = (0, 0)'$. When $\beta = (0.5, -0.5)'$, $\rho = 0.4$ and $\rho = 0.7$ led to a conditional correlation of 0.11 and 0.19, respectively. Under each censoring proportion, β and ρ values, and sample sizes of 250, 350 and 400, we ran 1000 Monte Carlo simulations. Ultimately, we were interested in examining the bias of each estimator, the improvement of the sample and estimated standard errors for $\hat{\beta}_{\mathrm{AUG}}$, and the coverage probabilities.

As seen from Tables 4 and 5, in general, all the resulting estimators for $\hat{\beta}$ and $\hat{\beta}_{\mathrm{AUG}}$ are consistent; the sample standard errors and the estimated standard errors are quite close; and the coverage probabilities of the 95% confidence intervals match the nominal level. By covariate augmentation, we can generally see improvements of $\hat{\beta}_{\mathrm{AUG}}$ over $\hat{\beta}$ in terms of the standard errors. The fact that small correlations between T and W still lead to improved efficiencies demonstrates the practical usefulness of our method. As the conditional correlation between T and W given Z increases, the efficiency gain becomes more evident. Overall, our simulations demonstrate that the augmented estimating equation can produce consistent estimators with improved efficiency while not imposing additional modeling structures.

Table 4: Simulation with $\beta = (0, 0)'$ and different sample sizes n. The conditional correlation between T and W given Z is 0.15 ($\rho = 0.4$) and 0.27 ($\rho = 0.7$). Bias, sample standard error (se), median of estimated standard errors ($\hat{\text{se}}$), and the 95% coverage probability (CI) are given for unadjusted estimator $\hat{\beta}$ and adjusted estimator $\hat{\beta}_{\text{AUG}}$, respectively.

		0% Censoring				30% Censoring			
n		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{\text{AUG},1}$	$\hat{\beta}_{\text{AUG},2}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{\text{AUG},1}$	$\hat{\beta}_{\text{AUG},2}$
					$\rho = 0.4$				
250	bias	0.019	-0.003	-0.005	0.020	0.047	-0.033	0.039	-0.026
	se	0.269	0.278	0.222	0.257	0.312	0.455	0.301	0.454
	$\hat{\text{se}}$	0.262	0.264	0.211	0.236	0.303	0.462	0.286	0.450
	CI (%)	95.9	94.8	94.2	92.7	96.4	95.6	95.8	94.7
350	bias	0.012	-0.005	0.009	-0.002	0.024	0.001	0.021	0.006
	se	0.226	0.225	0.217	0.224	0.259	0.399	0.254	0.393
	$\hat{\text{se}}$	0.220	0.221	0.207	0.218	0.253	0.389	0.241	0.382
	CI (%)	94.5	94.5	93.6	94.5	95.4	95.5	94.3	94.8
400	bias	0.011	-0.003	0.009	-0.001	0.019	0.008	0.017	0.012
	se	0.212	0.213	0.203	0.213	0.235	0.365	0.228	0.363
	$\hat{\text{se}}$	0.206	0.207	0.194	0.203	0.236	0.365	0.225	0.361
	CI (%)	94.8	95.2	94.6	94.0	95.9	96.4	95.7	95.4
					$\rho = 0.7$				
250	bias	0.017	0.001	-0.006	0.023	0.050	-0.014	0.021	0.004
	se	0.274	0.282	0.225	0.261	0.307	0.463	0.262	0.448
	$\hat{\text{se}}$	0.263	0.264	0.211	0.237	0.302	0.457	0.247	0.428
	CI (%)	95.6	94.5	93.9	91.6	96.4	95.5	94.8	94.0
350	bias	0.011	-0.001	-0.003	0.015	0.024	0.010	0.007	0.026
	se	0.223	0.226	0.188	0.209	0.256	0.397	0.220	0.380
	$\hat{\text{se}}$	0.220	0.221	0.177	0.198	0.251	0.393	0.209	0.370
	CI (%)	95.0	95.0	93.5	93.7	95.8	95.6	94.7	94.6
400	bias	0.008	0.001	-0.002	0.013	0.019	0.009	0.005	0.028
	se	0.210	0.210	0.173	0.195	0.236	0.362	0.201	0.349
	$\hat{\text{se}}$	0.206	0.207	0.166	0.185	0.236	0.363	0.195	0.344
	CI (%)	95.3	95.5	94.4	94.2	96.4	96.2	94.8	95.5

Table 5: Simulation with $\beta = (0.5, -0.5)'$ and different sample sizes n. The conditional correlation between T and W given Z is 0.11 ($\rho = 0.4$) and 0.19 ($\rho = 0.7$). Bias, sample standard error (se), median of estimated standard errors ($\hat{\text{se}}$), and the 95% coverage probability (CI) are given for unadjusted estimator $\hat{\beta}$ and adjusted estimator $\hat{\beta}_{\text{AUG}}$, respectively.

		0% Censoring				30% Censoring			
n		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{\text{AUG},1}$	$\hat{\beta}_{\text{AUG},2}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{\text{AUG},1}$	$\hat{\beta}_{\text{AUG},2}$
					$\rho = 0.4$				
250	bias	0.059	-0.087	0.053	-0.081	0.059	-0.087	0.053	-0.081
	se	0.326	0.268	0.313	0.267	0.326	0.268	0.313	0.267
	$\hat{\text{se}}$	0.315	0.257	0.296	0.252	0.315	0.257	0.296	0.252
	CI (%)	96.7	94.8	95.6	94.1	96.7	94.8	95.6	94.1
350	bias	0.043	-0.083	0.041	-0.081	0.056	-0.086	0.054	-0.083
	se	0.271	0.214	0.263	0.214	0.307	0.332	0.300	0.330
	$\hat{\text{se}}$	0.262	0.215	0.248	0.211	0.297	0.334	0.286	0.327
	CI (%)	95.7	94.6	94.9	94.5	96.4	94.7	95.3	94.3
400	bias	0.039	-0.081	0.037	-0.079	0.053	-0.089	0.051	-0.087
	se	0.249	0.198	0.240	0.197	0.277	0.296	0.269	0.295
	$\hat{\text{se}}$	0.245	0.201	0.232	0.197	0.280	0.312	0.269	0.308
	CI (%)	96.1	94.5	95.9	93.7	96.5	95.7	96.7	94.5
					$\rho = 0.7$				
250	bias	0.057	-0.083	0.029	-0.060	0.065	-0.070	0.045	-0.061
	se	0.329	0.264	0.272	0.244	0.361	0.399	0.311	0.383
	$\hat{\text{se}}$	0.315	0.257	0.252	0.229	0.359	0.396	0.297	0.366
	CI (%)	96.5	94.7	94.7	93.9	96.9	95.4	95.5	94.6
350	bias	0.038	-0.077	0.021	-0.062	0.053	-0.080	0.039	-0.071
	se	0.269	0.215	0.225	0.197	0.311	0.336	0.263	0.316
	$\hat{\text{se}}$	0.260	0.214	0.209	0.191	0.297	0.334	0.248	0.311
	CI (%)	96.3	94.7	94.7	93.7	96.5	94.9	95.4	93.5
400	bias	0.035	-0.076	0.022	-0.064	0.047	-0.083	0.036	-0.073
	se	0.247	0.196	0.201	0.180	0.276	0.299	0.231	0.280
	$\hat{\text{se}}$	0.244	0.201	0.196	0.179	0.278	0.312	0.232	0.288
	CI (%)	96.5	94.9	95.5	93.5	96.6	95.5	95.3	94.9

3.5 Application to leukemia study

We applied our method to a leukemia study (Tsimberidou et al., 2006) at M.D. Anderson Cancer Center which consisted of 130 patients diagnosed with Richter's syndrome through biopsy or fine-needle aspiration. Richter's syndrome is a type of high grade non-Hodgkin's lymphoma, which usually develops in patients with chronic lymphocytic leukemia. It is a rare type of leukemia and often quickly evolves into fatal cancer. In the study, fifty-one patients were randomized into treatment 0 (chemoimmunotherapy with rituximab), and the rest, into treatment 1 (chemotherapy). With the event time being the time of death, roughly 12% of the data is censored. Gender and age for each patient were also collected for the study. Ages ranged between 29 and 77 with the median being 60 years, and the patient's gender was encoded as 1 for males. The overall objective of the study is to understand the effectiveness of each treatment and ultimately determine which treatment led to better survival rates.

Figure 3 displays the Kaplan-Meier survival curves for the treatments, and it is evident that the curves are nonproportional hazards. To capture this nonproportional hazards feature and estimate the short- and long-term effects of the treatments, we applied the flexible model of Yang and Prentice (2005). Initially disregarding the auxiliary covariate information, we first obtained the unadjusted parameter estimates $\hat{\beta}_1 = 1.1375$ and $\hat{\beta}_2 = -0.7055$ with estimated standard errors 0.7102 and 0.6272, respectively. The results imply that the estimated short-term hazard ratio is $\exp(\hat{\beta}_1) = 3.1192$ and the estimated long-term hazard ratio is $\exp(\hat{\beta}_2) = 0.4938$ when not accounting for patients' ages and gender.

To verify the need for incorporating the auxiliary covariates (i.e., determine if the covariates and event times are correlated), we plotted the Kaplan-Meier survival curves for each gender and differing age groups (using 60 years as the cutoff). As the estimated survival curves for males and females in Figure 4 (a) and (b) are distinctly different from

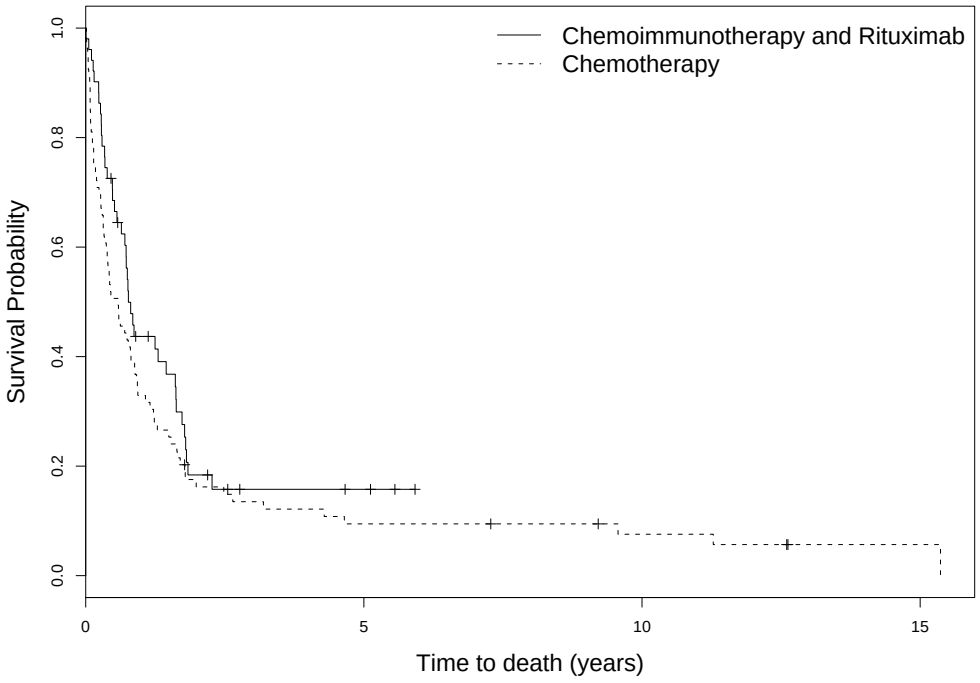

Figure 3: Kaplan-Meier survival curves under different treatments in the leukemia study.

each other, the data indicate that gender is correlated with survival times. Likewise, the different estimated survival curves based on age groups seen in Figure 4 (c) and (d) also demonstrate that age is correlated with survival times. With this evidence of correlation between the auxiliary covariates and survival times, we applied the covariate augmentation method to possibly produce more efficient estimators without imposing extra modeling structures.

In particular, we posited a linear model with an interaction for each component of the augmentation term $h(W;\alpha)$. That is, for $j = 1, 2$, we set each $h_j(W;\alpha_j) = \alpha_{j0} + \alpha_{j1}W_1 + \alpha_{j2}W_2 + \alpha_{j3}W_1W_2$, where W_1 is an indicator covariate corresponding to gender, and W_2 is the continuous covariate for age. The proposed method yielded the adjusted

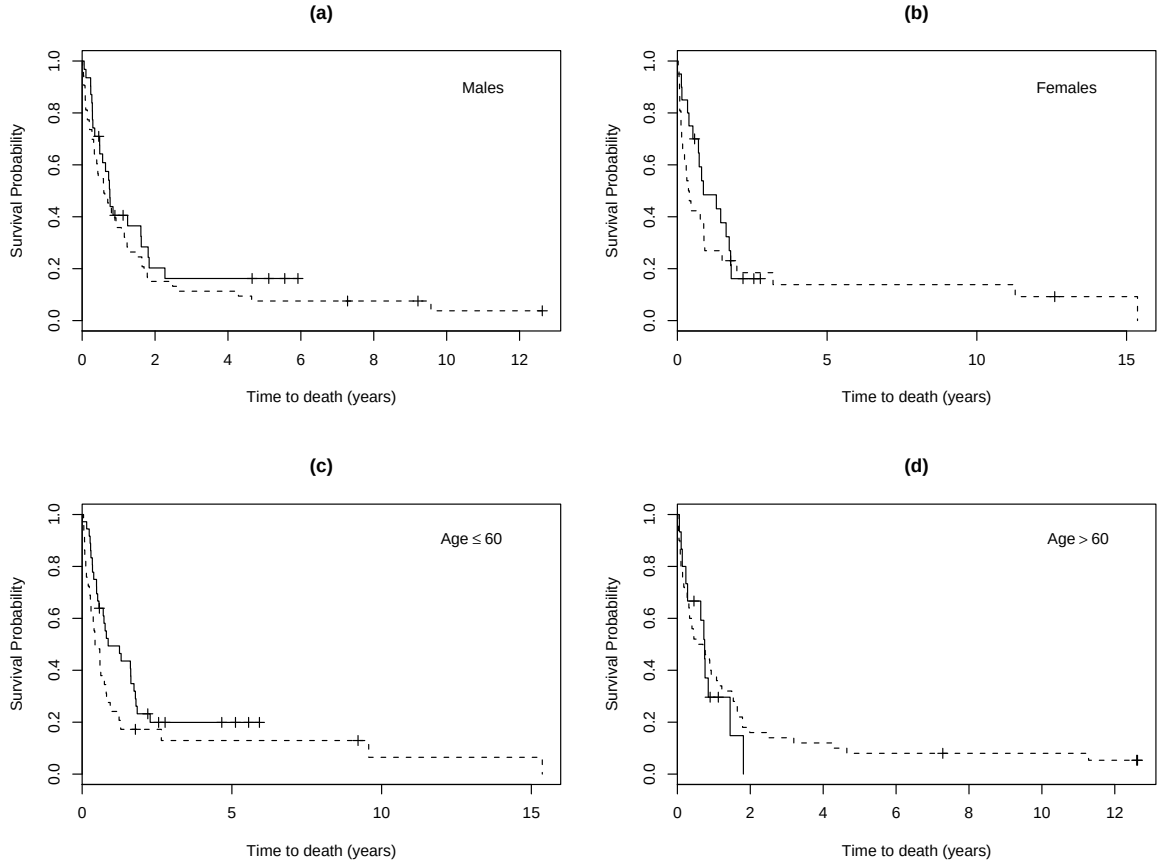

Figure 4: Kaplan-Meier survival curves for (a) males; (b) females; (c) individuals with ages less than 60 years; (d) individuals with ages over 60 years. Survival curves in each case are stratified by treatment where the solid line corresponds to chemoimmunotherapy with rituximab, and the dashed line corresponds to chemotherapy.

parameter estimates $\hat{\beta}_{\mathrm{AUG},1} = 1.0751$ and $\hat{\beta}_{\mathrm{AUG},2} = -0.7611$. The respective estimated standard errors are 0.6099 and 0.5259, which correspond to approximately 36% and 42% efficiency gains over the unadjusted estimator $\hat{\beta}$. In this case, the estimated short-term hazard ratio is 2.9303, and the estimated long-term hazard ratio is 0.4672. Incorporating the auxiliary covariates leads to reduced estimates of the short- and long-term hazard ratios, and significant efficiency improvements in the estimates.

3.6 Discussion

Our practical method for incorporating auxiliary covariates in nonproportional hazards models demonstrates improved efficiency inferences, even in the case of small correlation between the event times and covariates. Following a strategy suggested by Zhang et al. (2008) and Yang and Prentice (2005), we posit a simple linear form for the augmented term, with the actual form being motivated by scatter plots of the terms in the unadjusted estimating equation against the covariates. Even if the posited form for the augmented expectation term is incorrectly specified, the resulting $\hat{\beta}_{\mathrm{AUG}}$ displays improved efficiency over the unadjusted estimator $\hat{\beta}$. The proposed method permits the flexibility of having more efficient estimators when direct modeling of T and W is not necessary.

CHAPTER IV

SEMIPARAMETRIC ESTIMATION FOR CENSORED MIXTURE DATA WITH APPLICATION TO THE COOPERATIVE HUNTINGTON'S OBSERVATIONAL RESEARCH TRIAL

4.1 Introduction

In kin-cohort studies (Struewing et al., 1997; Wacholder et al., 1998; Gail et al., 1999) and quantitative trait locus studies (QTL, Lander and Botstein, 1989; Wu et al., 2007), a common scientific goal is to estimate the cumulative distribution function of an outcome, subject to right censoring, from mixture data of scientifically meaningful subpopulations. Current methods include parametric methods (Moore et al., 2001) which are often too restrictive and two types of nonparametric maximum likelihood estimators (NPMLEs, Chatterjee and Wacholder, 2001; Wacholder et al., 1998) which are either inefficient or inconsistent. To improve upon these methods, we propose several nonparametric estimators which are efficient, robust to model misspecification, and easy to implement.

Kin-cohort studies are recent novel designs proposed to estimate the age-specific cumulative risk of a disease in deleterious mutation carriers applicable to rare mutations. Prior to their development, population-based cohort designs or case-control family studies (Whittemore, 1995; Li et al., 1998) were used to estimate the cumulative risk, but both have disadvantages. While population-based cohort designs provide direct data on estimating the disease risk, for rare genetic exposures, they require a large number of subjects to be screened to identify the sufficient number of mutation carriers. Case-control studies allow over-sampling of subjects with rare exposures, but case-control data alone only permit estimation of relative risk or odds ratio, not absolute risk. The kin-cohort study combines prospective or case-control probands with disease histories in family members. Similar to a

case-control study, a kin-cohort study can enrich the sample of mutation carriers. However, unlike a case-control design, a kin-cohort study substantially extends our understanding of knowledge by enabling estimation of all aspects of the distribution of a disease given a genotype such as absolute cumulative risk.

In a kin-cohort study, firstly, probands possibly enriched with mutation carriers are sampled and genotyped. Next, family histories of the disease of interest in relatives of the probands are collected by administering a reliable, validated interview (Marder et al., 2003) to probands, or preferably the relatives themselves. Due to practical concerns of the cost of in-person assessments to collect blood samples, genotype information is usually not available on relatives though disease status (phenotype) information is available from systematic interview. Therefore phenotype data arise from a combination of genotype-specific subpopulations. Despite unknown genotypes in relatives, the probability of each relative having a certain genotype can be estimated from his or her relationship with the proband and the observed proband's genotype. Distributions of the observed phenotypes in the relatives are therefore a mixture of genotype-specific distributions. The missing genotype information in relatives creates complications in the analysis where the interest is on conditional distribution given the genotypes.

This mixture data structure also arises in other scientific experiments such as the interval mapping of quantitative traits (Lander and Botstein, 1989). Here, the trait distributions are mixtures of the quantitative trait locus (QTL) genotype-specific distributions where the mixing proportions are computed based on the flanking marker genotypes and recombination fractions between the marker and the putative QTL. Therefore in many controlled QTL experiments such as backcross the mixing proportions are easily obtained, and interest lies in estimating the genotype-specific distributions.

A common scientific goal of kin-cohort and QTL studies is to make inference on genotype-specific subpopulation distributions where observed data is from a mixture of

subpopulations with mixing proportions that vary across subjects but which can be obtained from available genotype data on probands or flanking markers. The focus of the current paper is to analyze such mixture data under the context of censoring. For mixture data in QTL mapping and kin-cohort studies, maximum likelihood based parametric methods are typically used (Lander and Botstein, 1989; Wu et al., 2002; Moore et al., 2001). Sometimes, the biological underpinning of the development of a disease trait suggests a suitable parametric function which offers meaningful interpretation of the biological structure (Wu et al., 2000). Still, for many situations, there may not be sufficient biological knowledge to warrant such a parametric function, and concerns of model mis-specification naturally arise. To alleviate these issues, more flexible semiparametric modeling and estimation of the distribution functions become essential (Zhao and Wu, 2008; Yu and Lin, 2008; Liu et al., 2006). Jin et al. (2007) used the exponential tilt to model the relation between different genotype-specific distributions, and provided likelihood based inference. Diao and Lin (2005) and Liu et al. (2006) proposed a Cox proportional hazards model for QTL experiments. However, a proportional hazards assumption may not be satisfied for some real data such as Huntington disease (HD) data (Langbehn et al., 2004). Ma and Wang (2011) adopted a pure nonparametric model of the conditional distributions, proposed a general class of semiparametric estimators and identified the efficient member of the class for non-censored observations.

For nonparametric time-to-event model with event time subject to censoring, Wang et al. (2007) proposed methods for kin-cohort data when the censoring times are observed for all subjects. When censoring times are random and not all observed, Wacholder et al. (1998) proposed a nonparametric maximum likelihood estimator (type I NPMLE) consisting of a combination of several NPMLEs and a linear transformation. Chatterjee and Wacholder (2001) proposed a direct maximization of the nonparametric likelihood (type II NPMLE) with respect to the conditional distributions and used an EM algorithm to find the

maximizer. Although in many situations, NPMLEs are efficient, we demonstrate here the surprising results that these two types of NPMLE have their respective limitations: the type I is inefficient and the type II is inconsistent.

Due to the shortcomings of the two NPMLEs, we take a semiparametric approach and cast this problem in a missing data framework. Given a complete data influence function (i.e., no censoring), we propose an inverse probability weighting (IPW) estimator, and then add an augmentation term to obtain the optimal estimator. We also propose an imputation (IMP) estimator which is easy to implement and does not require additional modeling assumptions for the imputation step. We demonstrate the asymptotic properties of these estimators and examine their finite sample performance through simulation studies and an application to Huntington disease data.

4.1.1 The Cooperative Huntington's Observational Research Trial (COHORT)

Huntington disease is a degenerative, genetic disorder which targets nerve cells in the brain and leads to cognitive decline, involuntary muscle spasms, and psychological problems. Affected individuals typically begin to see neurological and physical symptoms around 30-50 years of age, and eventually die from pneumonia, heart failure or other complications 10-25 years after the disease onsets. The severity of the disease has prompted the development of several organizations, like the Huntington Study Group (Huntington Study Group, 2011b), which are devoted to studying the causes, effects, and possible treatments for HD. A particular study organized by roughly 42 Huntington Study Group research centers in North America and Australia is the Cooperative Huntington's Observational Research Trial (COHORT, Huntington Study Group, 2011a). The COHORT is designed for collecting ongoing information from affected adults and at-risk family members 15 years of age and older who choose to participate.

Huntington disease is caused by unstable CAG repeats in the HD gene (Ross, 1995). In

a clinical counseling setting, CAG repeats ≥ 36 is defined as positive for Huntington gene mutation, or carrier, and CAG <36 is defined as negative, or non-carrier (Rubinsztein et al., 1996). Each year, proband participants undergo a clinical evaluation, where blood samples are genotyped for being a carrier or non-carrier of HD mutation. While the HD mutation status is ascertained in probands, high costs of in-person interviews on family members, prevents collection of blood sample in the relatives. Based on a subject's relationship with the proband and the proband's mutation status, the genotype distribution of a relative can still be obtained. Distribution of the relatives' age-at-death is therefore a mixture of the genotype-specific distributions with known mixing proportions. As the survival functions in HD mutation carriers is of great clinical interest, we apply the proposed estimators to the COHORT data to determine the cumulative risk of death from possessing the HD gene mutation.

4.2 Existing estimators for censored mixture data

Censored mixture data consist of independent, identically distributed triplets $(Q_i = q_i, X_i = x_i, \Delta_i = \delta_i)$. For the ith individual, Q_i is a p-dimensional vector of random mixture proportions with associated probability mass function p_Q which has finite support $u_1, \ldots, u_m$. Also, $X_i = \min(T_i, C_i)$ denotes a subject's event time where T_i is a continuous outcome (i.e., survival time) and C_i is a random continuous censoring time independent of T_i; and $\Delta_i = I(T_i \leq C_i)$ denotes the censoring indicator. We let $f(\cdot)$ denote the p-dimensional, unspecified probability density function of T given the mixing group membership, and $F(\cdot)$ denote its corresponding conditional cumulative distribution function. Interest lies in estimating $F(t)$ for any fixed time t, where the jth component, $F_j(t)$, $j = 1, \ldots, p$, denotes the conditional distribution of a trait given that the gene mutation status is the jth kind. For the COHORT study, $p = 2$ and $F_1(t)$ and $F_2(t)$ correspond to the age-at-death distribution for individuals with an HD carrier and non-carrier gene, respectively. Throughout, we assume

event times $x_1, \ldots, x_n$ have no ties, and that the censoring distribution is common for all p populations. Then, letting $G(\cdot)$ denote the survival function of C and $g(\cdot)$ is corresponding density, the log-likelihood of n observations is

$$\sum_{i=1}^{n} \log \left(p_Q(q_i) \left\{ q_i^T f(x_i) G(x_i) \right\}^{\delta_i} \left[\left\{ 1 - q_i^T F(x_i) \right\} g(x_i) \right\} \right]^{1-\delta_i} \Big), \tag{4.1}$$

where we use the fact that $q_i^T 1_p = 1$ with 1_p a p-dimensional vector of ones.

4.2.1 The type I NPMLE and its inefficiency

The type I NPMLE was proposed in the literature to analyze kin-cohort data (Wacholder et al., 1998). It first maximizes (4.1) with respect to $q_i^T f(x_i)$'s, then recovers $F(t)$ through a linear transformation. Although an NPMLE based estimator is usually efficient, it is not so for the mixture data context, and the magnitude of efficiency loss is non-ignorable.

To describe the type I NPMLE, we reformulate the maximization problem by evoking the assumption that Q has finite support $u_1, \ldots, u_m$ and by letting $s_j(x_k) = u_j^T f(x_k)$ and $S_j(x_k) = 1 - u_j^T F(x_k)$. The type I NPMLE then maximizes the equivalent target function

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \log \left\{ s_j(x_i)^{\delta_i} S_j(x_i)^{1-\delta_i} \right\} I(q_i = u_j) \tag{4.2}$$

with respect to $s_j(x_i)$'s and subject to $\sum_{i=1}^{n} s_j(x_i) I(q_i = u_j) \leq 1$, $s_j(x_i) \geq 0$ for $j = 1, \ldots, m$. Obviously this is equivalent to m separate maximization problems, each concerning $s_j(\cdot)$ and $S_j(\cdot)$ only, so that the maximizers are the classical Kaplan-Meier estimators. That is,

$$\widehat{S}_j(a) = \prod_{x_i \leq a, q_i = u_j} \left\{ 1 - \frac{\delta_i}{\sum_{q_k = u_j} I(x_k \geq x_i)} \right\}$$

and $s_j(a) = S_j(a^-) - S_j(a)$ for all a. Using the linear relation $u_j^T F(t) = 1 - S_j(t)$ for $j = 1, \ldots, m$, we then recover the type I NPMLE estimator as

$$\widetilde{F}(t) = \left(U^T U \right)^{-1} U^T \{ 1_m - \widehat{S}(t) \},$$

where $\widehat{S}(t) = \{\widehat{S}_1(t), \ldots, \widehat{S}_m(t)\}^T$, and $U = (u_1, \ldots, u_m)^T$. In this notation, $S(t) = 1_m - UF(t)$. The consistency of the Kaplan-Meier estimator of $S(t)$ ensures the consistency of $\widetilde{F}(t)$. The inefficiency of $\widetilde{F}(t)$, however, is evident considering that $\widetilde{F}_w(t) = (U^T\Sigma^{-1}U)^{-1}U^T\Sigma^{-1}\{1_m - \widehat{S}(t)\}$ with Σ denoting the variance-covariance matrix of $\widehat{S}(t)$ yields a more efficient estimator. In this case, because each of the m components of $\widehat{S}(t)$ is estimated using a distinct subset of the observations, Σ is a diagonal matrix. Hence, $\widetilde{F}_w(t)$ is simply a weighted version of the type I NPMLE, and this simple weighting scheme improves the estimation efficiency.

4.2.2 The type II NPMLE and its inconsistency

The type II NPMLE is considered an improvement over the type I NPMLE (Chatterjee and Wacholder, 2001). The type II NPMLE maximizes the same log likelihood (4.2), but with respect to $f(x_i)$'s in contrast to $s_j(x_i) = u_j^T f(x_i)$ as for the type I NPMLE, and subject to $\sum_{i=1}^n f(x_i) \leq 1_p$, and $f(x_i) \geq 0$ component-wise. In general, no closed form solution exists, and the Expectation-Maximization (EM) algorithm is usually implemented to obtain the $F(x_i)$'s. To be specific, we regard the genotypes $G_i = 1, \ldots, p$ as missing data, and derive the complete data log likelihood of the observations $o_i = (G_i = g_i, X_i = x_i, \Delta_i = \delta_i), i = 1, \ldots, n$, as

$$\begin{aligned} & \mathcal{L}^{\text{comp}}_{\text{type II}}\{o_1, \ldots, o_n; f(x_i), F(x_i), i = 1, \ldots, n\} \\ = & \sum_{i=1}^{n}\sum_{k=1}^{p} I(g_i = k)\log\left[f_k(x_i)^{\delta_i}\{1 - F_k(x_i)\}^{1-\delta_i}\right]. \end{aligned}$$

The EM algorithm is an iterative procedure, where at the lth iteration, the E-step is:

$$\begin{aligned} & E\left[\mathcal{L}^{\text{comp}}_{\text{type II}}\{o_1, \ldots, o_n; f(x_i), F(x_i), i = 1, \ldots, n\} | f^{(l)}(x_i), F^{(l)}(x_i), i = 1 \ldots, n\right] \\ = & \sum_{k=1}^{p}\sum_{i=1}^{n}\left[\frac{\delta_i q_{ik} f_k^{(l)}(x_i)}{\sum_{k=1}^{p} q_{ik} f_k^{(l)}(x_i)}\log f_k(x_i) + \frac{(1-\delta_i) q_{ik}\{1 - F_k^{(l)}(x_i)\}}{\sum_{k=1}^{p} q_{ik}\{1 - F_k^{(l)}(x_i)\}}\log\{1 - F_k(x_i)\}\right]. \end{aligned}$$

The M-step maximizes the above expression with respect to $f(x_i)$ and $F(x_i)$'s subject to $f(x_i) \geq 0$ and $1 \geq F(x_i) \geq 0$. To this end, let

$$c_{ik}^{(l)} = \delta_i \frac{q_{ik} f_k^{(l)}(x_i)}{\sum_{k=1}^{p} q_{ik} f_k^{(l)}(x_i)} + (1-\delta_i) \frac{q_{ik}\{1 - F_k^{(l)}(x_i)\}}{\sum_{k=1}^{p} q_{ik}\{1 - F_k^{(l)}(x_i)\}}$$

denote the known quantity based on the lth iteration. Then, the M-step reduces to p separate maximization problems of the form

$$\sum_{i=1}^{n} c_{ik}^{(l)} \left[\delta_i \log f_k(x_i) + (1-\delta_i) \log\{1 - F_k(x_i)\}\right],$$

for $k = 1, \ldots, p$. Viewing this as the log likelihood of weighted observations, where the ith observation represents $c_{ik}^{(l)}$ observations of the same value, the maximizer is a modified Kaplan-Meier estimator:

$$\begin{aligned} 1 - \check{F}_k^{(l+1)}(t) &= \prod_{x_i \leq t, \delta_i = 1} \left\{ 1 - \frac{\sum_{j=1}^{n} I(x_j = x_i, \delta_j = 1) c_{jk}^{(l)}}{\sum_{j=1}^{n} c_{jk}^{(l)} I(x_j \geq x_i)} \right\} \\ &= \prod_{x_i \leq t, \delta_i = 1} \left\{ 1 - \frac{c_{ik}^{(l)}}{\sum_{j=1}^{n} c_{jk}^{(l)} I(x_j \geq x_i)} \right\}. \end{aligned}$$

Iterating the E- and the M-steps until convergence ultimately leads to the type II estimator.

As natural as the type II NPMLE appears, we show in Appendix C the surprising result that it is an inconsistent estimator of $F(t)$. To understand this intuitively, notice that the type II NPMLE maximizes the product of m different likelihoods formed by all observations with respect to a collection of parameters, but each of these parameters should concern only one of these likelihoods formed by a subset of the observations. For example, with an uncensored observation where $Q_i = (0.5, 0.5)$ and $T_i = t_i$, the event time t_i will carry an equal weight with the type II estimator for each of the two subpopulations. However, if this observation belongs to the first subpopulation, then t_i should not be included in the support when estimating $F_2(\cdot)$. Likewise, if this observation belongs to the second subpopulation, then t_i should not be included in the support when estimating $F_1(\cdot)$.

Since we do not observe which subpopulation an observation comes from, it is difficult for the type II NPMLE to correctly identify the support for each subpopulation. In contrast, the type I NPMLE does not suffer from this difficulty: when regarding $S_j(t) = u_j^T F(t)$ as an unknown parameter, the type I correctly assigns the support of $S_j(\cdot)$ to include all observations with $q_i = u_j$. Such a computation is feasible as all q_i are observed.

4.2.3 Estimators through a Cox Proportional Hazards Model

Motivated by the work of Diao and Lin (2005) for QTL studies, we also consider the Cox proportional hazards model for genotype-specific distributions. We consider genotype G_i as missing data, and let $Z_i(G_i)$ denote a p-dimensional coding vector for the genotype effect. For simplicity, we take the G_ith component of $Z_i(G_i)$ to be one and the remaining components as zero. Other codings to indicate dominant, recessive, and additive effects can also be used.

Viewing the coding vectors $Z_i(\cdot)$ as covariates, the censored data is modeled through a proportional hazards model, where the hazard function for subject i is

$$\lambda(x|G_i = k) = \lambda_0(x) \exp\{\beta^T z_i(k)\}.$$

In the above, $\lambda_0(x)$ denotes an unspecified baseline hazard function and β is a p-dimensional parameter. Throughout, let $\Lambda_0(t)$ denote the baseline cumulative hazard function. The Cox-based estimator corresponds to replacing the $f(x_i)$'s and $F(x_i)$'s in (4.1) with their parametric forms under the proportional hazards assumption, and then maximizing it with respect to β and $\Lambda_0(t)$. The final estimates, denoted $\hat{\beta}_{\rm PH}$ and $\hat{\Lambda}_0(t)$, respectively, are then used to form the Cox-based estimator $\hat{F}_{\rm PH}(t)$, where the jth component, $j = 1, \ldots, p$, is

$$\hat{F}_{\rm PH,j}(t) = 1 - \exp\{-\exp(\hat{\beta}_{{\rm PH},j})\hat{\Lambda}_0(t)\}.$$

As this maximization has no closed form, we obtain estimates for β and $\Lambda_0(t)$ through an EM algorithm described in Appendix C.

Considering that the Cox-based estimator assumes event times follow a proportional hazards model, it is evident that this estimator will result in biased estimates when the condition is not met. For example, in the COHORT data, assuming proportional hazards is inappropriate given that the distributions for the carrier and non-carrier groups cross at around age 38.

4.3 Proposed nonparametric estimators for censored mixture data

4.3.1 The IPW and the optimal augmented IPW estimators

To compensate for the poor performance of the NPMLEs and parametric restrictions of the Cox-based estimator, we now propose a class of nonparametric estimators based on the inverse probability weighting (IPW) which are easy to implement and have satisfactory performance. Previously, Bang and Tsiatis (2000, 2002) used the IPW to estimate mean and median medical cost when the event times are subject to right censoring. For the mixture data with censoring, the IPW estimator is obtained through solving

$$n^{-1}\sum_{i=1}^{n}\frac{\delta_i\phi(q_i,x_i)}{\widehat{G}(x_i)}=0,$$

where ϕ denotes a general influence function for non-censored data (see Ma and Wang, 2011) and $\widehat{G}(t)$ is the Kaplan-Meier estimator of $G(t)$:

$$\widehat{G}(t)=\prod_{x_i\leq t}\left\{1-\frac{1-\delta_i}{\sum_{j=1}^{n}I(x_j\geq x_i)}\right\}.$$

To simplify notation in what follows, let $Y_i(u)=I(x_i\geq u)$, $Y(u)=\sum_{i=1}^{n}Y_i(u)$, $N_i^c(u)=I(X_i\leq u,\Delta_i=0)$, $\lambda^c(\cdot)$ be the hazard function for the censoring distribution,

$$M_i^c(u)=N_i^c(u)-\int_0^u I(X_i\geq s)\lambda^c(s)ds$$

denote the censoring martingale; and

$$\mathcal{B}(h,u)=E\left\{h(\cdot)|T_i\geq u\right\}=\frac{E\{h(\cdot)I(T_i\geq u)\}}{S(u)}$$

where h is any p length function.

To characterize the asymptotic behavior of the IPW estimator, we show in Appendix C that the ith influence function for the IPW estimator is

$$\phi_{\text{ipw}}(q_i, x_i, \delta_i) = \phi(q_i, t_i) - \int \frac{dM_i^c(u)}{G(u)} \left\{\phi(q_i, t_i) - \mathcal{B}\left(\phi, u\right)\right\}.$$

The two terms in ϕ_{ipw} are uncorrelated given that for filtration $\mathcal{F}(u)$, the set of σ-algebras generated by $\sigma\{q_i, I(C_i \leq r), r \leq u; I(T_i \leq x), 0 \leq x < \infty, i = 1, \ldots, n\}$, $\phi(q_i, x_i)$ is $\mathcal{F}(0)$ measurable. Hence, the estimation variance of the IPW estimator is

$$V_{\text{ipw}} = \text{cov}\{\phi(Q_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\phi^{\otimes 2}, u) - \mathcal{B}(\phi, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\},$$

and a corresponding consistent estimator is

$$\widehat{V}_{\text{ipw}} = n^{-1} \sum_{i=1}^{n} \frac{\delta_i \phi(q_i, x_i) \phi^T(q_i, x_i)}{\widehat{G}(x_i)} + n^{-1} \sum_{i=1}^{n} \int \frac{\widehat{\mathcal{B}}_1(\phi^{\otimes 2}, u) - \widehat{\mathcal{B}}_1(\phi, u)^{\otimes 2}}{\widehat{G}^2(u)} dN_i^c(u),$$

where $\widehat{\mathcal{B}}_1(h, u) = \dfrac{1}{n\widehat{S}(u)} \displaystyle\sum_{i=1}^{n} \frac{\delta_i h(q_i, x_i, \delta_i) I(x_i \geq u)}{\widehat{G}(x_i)}$ for an arbitrary function $h(q_i, x_i, \delta_i)$.

Although intuitive and easy to implement, the IPW estimator is inefficient. A modification motivated by Robins and Rotnitzky (1992), however, leads to a more efficient estimator. Using the complete data influence function ϕ, the authors provided the following general class of influence functions for censored data:

$$\phi(q_i, t_i) - \int \frac{dM_i^c(u)}{G(u)} \left\{\phi(q_i, t_i) - \mathcal{B}(\phi, u)\right\} + \int \frac{dM_i^c(u)}{G(u)} \left[h\{\bar{a}_i(u), u\} - \mathcal{B}(h, u)\right]. \quad (4.3)$$

In the mixture problem, $a_i(u) = \{q_i, I(u < T_i)\}$ and $\bar{a}_i(u)$ contains the functions $a_i(\tilde{u})$ for all $\tilde{u} \leq u$. Compared to the influence function for the IPW estimator, the estimator from (4.3) contains an augmentation term that may improve the estimation efficiency, and is thus termed the augmented IPW (AIPW) estimator. Among all the choices for h, Robins et al. (1994) and Van der Laan and Hubbard (1998) showed that

$$\begin{aligned} & h^*_{\text{eff}}\{\bar{a}_i(u), u\} = E\{\phi(Q_i, T_i) | T_i \geq u, \bar{a}_i(u)\} \\ = \;& \{I(u < X_i) + I(u = X_i, \delta_i = 0)\} E\{\phi(Q_i, T_i) | q_i, T_i \geq u\} + I(u = X_i) \delta_i \phi(q_i, u) \end{aligned}$$

with $u \leq X_i$ yields the optimal efficiency. Denoting $h_{\text{eff},i}(u) = E\{\phi(Q_i, T_i)|q_i, T_i \geq u\}$, we have that $h^*_{\text{eff}}\{\bar{a}_i(u), u\}$ and $h_{\text{eff},i}(u)$ are identical except when $u = X_i$ and $\delta_i = 1$. The functional h^*_{eff} only appears in the censoring martingale integral, so using $h_{\text{eff},i}(u)$ instead of $h^*_{\text{eff}}\{\bar{a}_i(u), u\}$ yields the same influence function.

For most problems, constructing the efficient estimator usually relies on additional model assumptions and thus prevents the estimator from achieving the efficiency bound. We now demonstrate that the AIPW estimator for the mixture data does achieve the optimal efficiency. We first note that $h_{\text{eff},i}$ can be estimated consistently using a sample version of (4.3) with IPW:

$$\widehat{h}_{\text{eff},i}(u) = \frac{\sum_{j=1}^n I(q_j = q_i)\phi(q_j, x_j)Y_j(u)\delta_j/\widehat{G}(x_j)}{\sum_{j=1}^n I(q_j = q_i)Y_j(u)\delta_j/\widehat{G}(x_j)}, \tag{4.4}$$

where we use the global estimation of $G(u)$ since we assume the censoring distribution is common for all p populations. We may relax this assumption, however, and use only observations with the same q_i values to obtain group specific censoring distributions. Furthermore, because $h_{\text{eff},i}(u)$ is not a function of C_i, the independence between the censoring and survival process gives

$$\mathcal{B}(h_{\text{eff}}, u) \;=\; \frac{E\left\{h_{\text{eff},i}(u)I(T_i \geq u)I(C_i \geq u)\right\}}{E\{I(T_i \geq u)I(C_i \geq u)\}} = \frac{E\left\{h_{\text{eff},i}(u)Y_i(u)\right\}}{E\{Y_i(u)\}}.$$

Therefore, we can approximate $\mathcal{B}(h_{\text{eff}}, u)$ with

$$\widehat{\mathcal{B}}(h_{\text{eff}}, u) = \frac{\sum_{i=1}^n h_{\text{eff},i}(u)Y_i(u)}{Y(u)},$$

which satisfies

$$\sum_{i=1}^n \int \frac{\lambda^c(u)I(x_i \geq u)}{\widehat{G}(u)} \left\{\widehat{h}_{\text{eff},i}(u) - \widehat{\mathcal{B}}(\widehat{h}_{\text{eff}}, u)\right\} du = 0.$$

This enables us to obtain the optimal estimator $\widehat{F}(t)$ practically by solving

$$\sum_{i=1}^n \frac{\delta_i \phi(q_i, x_i)}{\widehat{G}(x_i)} + \int \frac{dN_i^c(u)}{\widehat{G}(u)} \left\{\widehat{h}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{h}_{\text{eff}}, u\right)\right\} = 0. \tag{4.5}$$

The estimator is very easy to implement especially comparing to many other semiparametric problems where the efficient estimator often involves additional model assumptions (Tsiatis and Ma, 2004; Wang et al., 2010), solving integral equations (Rabinowitz, 2000) and iterative procedures (Lu and Tsiatis, 2008; Zhang et al., 2008).

In Appendix C, we demonstrate that the AIPW estimator indeed has the efficient influence function, which corresponds to replacing $h(\cdot)$ with $h_{\text{eff},i}(u)$ in (4.3). We also demonstrate the variance of the efficient estimator is

$$V_{\text{eff}} = \text{cov}\{\phi(Q_i, T_i)\} + E \int \frac{\mathcal{B}\{(\phi - h_{\text{eff}})^{\otimes 2}, u\}}{G^2(u)} \lambda^c(u) Y_i(u) du,$$

which can be estimated consistently via

$$\widehat{V}_{\text{eff}} = n^{-1} \sum_{i=1}^{n} \frac{\delta_i \phi(q_i, x_i) \phi^T(q_i, x_i)}{\widehat{G}(x_i)} + n^{-1} \sum_{i=1}^{n} \int \frac{\widehat{\mathcal{B}}_1\{(\phi - \widehat{h}_{\text{eff}})^{\otimes 2}, u\}}{\widehat{G}^2(u)} dN_i^c(u).$$

4.3.2 An imputation (IMP) estimator

Lipitz et al. (1999) proposed a conditional estimating equation for regression with missing covariates by conditioning the complete data estimating equation on the observed data. Similarly, with censored observation, we replace the unknown complete data influence function with its conditional expectation given that the event happens after the observed censoring time. Doing so yields the following imputed estimating equation:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left[\delta_i \phi(q_i, x_i) + (1 - \delta_i) E\left\{ \phi(Q_i, T_i) | T_i > x_i, q_i \right\} \right] \\ &= \sum_{i=1}^{n} \left\{ \delta_i \phi(q_i, x_i) + (1 - \delta_i) h_{\text{eff},i}(x_i) \right\}. \end{aligned}$$

In practice, we obtain the IMP estimator by solving

$$0 = \sum_{i=1}^{n} \left\{ \delta_i \phi(q_i, x_i) + (1 - \delta_i) \widehat{h}_{\text{eff},i}(x_i) \right\},$$

with $\widehat{h}_{\text{eff},i}(u)$ as in (4.4).

While in many cases the imputation method could lead to bias if the model of the missingness is mis-specified, it is straightforward to see that our proposed imputation estimator is always consistent. In practice, we often, but not always, observe that it performs competitively or even favorably in comparison with the optimal AIPW estimator. For inferences, we derive the influence function of the IMP estimator in Appendix C and find that it has a complex form containing nested conditional expectations and hence is hardly practically useful. Asymptotic analysis for imputation based estimation is often complex and can be rather involved even in parametric imputation procedures (Wang and Robins, 1998; Robins and Wang, 2000), which partially explains why the bootstrap method is usually favored in its inference.

One interesting discovery we made is that when the data arise from a single distribution (i.e., $p = 1$), the IPW, AIPW, IMP and the two NPMLEs are all equivalent to the familiar Kaplan-Meier estimator. This indicates the complexity arising from the mixture nature.

4.4 Simulations

We conducted three Monte Carlo simulation studies to illustrate the finite sample performance of five groups of estimators, yielding a total of twelve different estimators. The first group of estimators includes the IPW, optimal AIPW and IMP estimator based on the complete data ordinary least square (OLS) influence function. The second and third groups include the same three estimators based on the complete data weighted least square (WLS) and efficient (EFF) influence functions, respectively. Finally, the fourth group of estimators contains the two NPMLEs and the fifth, the Cox-based estimator.

The first two simulation studies exemplify the sensitivity of the Cox-based estimator and the robustness of the nonparametric estimators with respect to the data assumptions. To illustrate this sensitivity, we took $F(t)$ as a two-dimensional vector (i.e., $p = 2$),

and let q assume any one of 4 different possible vector values (i.e., $m = 4$). In the first simulation experiment, we generated data from a proportional hazards model; that is the two components in the true $F(t)$ have truncated exponential form where $F_1(t) = \{1-\exp(-t/4)\}/\{1-\exp(-2.5)\}$ on the interval $(0, 10)$ and $F_2(t) = F_1(t)^{0.98}$ on the interval $(0, 5)$. In the second simulation experiment, we generated data from a non-proportional hazards model where $F_1(t) = [\{1 - \exp(-t/4)\}/\{1 - \exp(-2.5)\}]^{0.5}$ on $(0, 10)$ and $F_2(t) = \{1 - \exp(-t/2)\}/\{1 - \exp(-2.5)\}$ on $(0, 5)$. For both data generation procedures, our sample size is 500 and we generated a uniform censoring distribution to achieve moderate (20%) and high (50%) censoring rates. For each scenario, we ran 1000 Monte Carlo simulations and demonstrate the performance of the twelve estimators in Tables 6 and 7.

The results in Tables 6 and 7 indicate that only when the data are generated from a proportional hazards model does the Cox-based estimator have small bias, have estimated coverage probabilities matching the nominal level, and outperforms all the nonparametric estimators. When the data follow a non-proportional hazards model, however, the Cox-based estimator deteriorates and produces highly biased estimates. These results suggest that the Cox-based estimator should not be used when the proportional hazards assumption is in doubt. It is worthy to point out that when the proportional hazards assumption does hold, the proposed AIPW estimators have similar empirical standard errors as the Cox-based estimator, indicating minimal efficiency loss of the nonparametric estimators.

Table 6: Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the first simulation with proportional hazards generated data, sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

	$F_1(t) = 0.5837$				$F_2(t) = 0.5748$			
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
Group 1: OLS based, censoring rate =20%								
IPW	0.0005	0.0430	0.0428	0.9420	0.0003	0.0418	0.0417	0.9510
AIPW	0.0006	0.0410	0.0402	0.9480	-0.0000	0.0397	0.0393	0.9490
IMP	0.0005	0.0410	0.0400	0.9360	0.0000	0.0396	0.0391	0.9500
Group 2: WLS based, censoring rate =20%								
IPW	0.0006	0.0430	0.0428	0.9410	0.0003	0.0418	0.0417	0.9510
AIPW	0.0006	0.0410	0.0402	0.9480	-0.0000	0.0397	0.0393	0.9490
IMP	0.0005	0.0410	0.0400	0.9370	0.0000	0.0396	0.0391	0.9520
Group 3: EFF based, censoring rate =20%								
IPW	0.0003	0.0432	0.0432	0.9480	0.0001	0.0429	0.0433	0.9510
AIPW	0.0007	0.0412	0.0405	0.9450	-0.0001	0.0400	0.0398	0.9430
IMP	0.0006	0.0412	0.0402	0.9410	-0.0001	0.0399	0.0393	0.9450
Group 4: NPMLE, censoring rate =20%								
type I	0.0002	0.0478	0.0468	0.9410	0.0008	0.0921	0.0891	0.9240
type II	-0.0140	-	-	-	0.0098	-	-	-
Group 5: Cox PH, censoring rate =20%								
COX	0.0002	0.0447	0.0448	0.9541	-0.0001	0.0381	0.0378	0.9431
Group 1: OLS based, censoring rate =50%								
IPW	0.0087	0.0719	0.0683	0.9260	-0.0024	0.0672	0.0666	0.9430
AIPW	0.0010	0.0469	0.0458	0.9430	-0.0006	0.0450	0.0449	0.9400
IMP	0.0052	0.0495	0.0497	0.9390	0.0018	0.0469	0.0478	0.9540
Group 2: WLS based, censoring rate =50%								
IPW	0.0087	0.0721	0.0682	0.9280	-0.0024	0.0673	0.0666	0.9430
AIPW	0.0010	0.0469	0.0458	0.9450	-0.0006	0.0450	0.0449	0.9390
IMP	0.0052	0.0495	0.0497	0.9390	0.0018	0.0469	0.0478	0.9540
Group 3: EFF based, censoring rate =50%								
IPW	0.0037	0.0705	0.0700	0.9340	-0.0036	0.0739	0.0741	0.9420
AIPW	0.0002	0.0484	0.0463	0.9410	0.0004	0.0461	0.0456	0.9520
IMP	0.0043	0.0501	0.0509	0.9470	0.0028	0.0475	0.0487	0.9590
Group 4: NPMLE, censoring rate =50%								
type I	0.0000	0.0547	0.0539	0.9460	-0.0013	0.1069	0.1024	0.9120
type II	-0.0415	-	-	-	0.0337	-	-	-
Group 5: Cox PH, censoring rate =50%								
COX	-0.0010	0.0479	0.0499	0.9620	-0.0011	0.0464	0.0470	0.9460

Table 7: Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the second simulation with non-proportional hazards generated data, sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

	$F_1(t) = 0.5063$				$F_2(t) = 0.5132$			
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
Group 1: OLS based, censoring rate =20%								
IPW	0.0071	0.0475	0.0436	0.9220	-0.0008	0.0437	0.0425	0.9410
AIPW	0.0069	0.0433	0.0391	0.9250	-0.0013	0.0402	0.0388	0.9370
IMP	0.0011	0.0408	0.0394	0.9470	-0.0009	0.0397	0.0387	0.9390
Group 2: WLS based, censoring rate =20%								
IPW	0.0071	0.0476	0.0436	0.9210	-0.0009	0.0437	0.0425	0.9410
AIPW	0.0069	0.0433	0.0391	0.9240	-0.0013	0.0402	0.0388	0.9360
IMP	0.0011	0.0408	0.0394	0.9470	-0.0009	0.0398	0.0387	0.9390
Group 3: EFF based, censoring rate =20%								
IPW	0.0064	0.0467	0.0433	0.9290	-0.0010	0.0436	0.0423	0.9400
AIPW	0.0074	0.0431	0.0391	0.9280	-0.0020	0.0405	0.0387	0.9390
IMP	0.0025	0.0409	0.0394	0.9420	-0.0023	0.0398	0.0388	0.9380
Group 4: NPMLE, censoring rate =20%								
type I	0.0007	0.0476	0.0459	0.9410	0.0002	0.0925	0.0879	0.9130
type II	-0.0227	-	-	-	0.0168	-	-	-
Group 5: Cox PH, censoring rate =20%								
COX	-0.0304	0.0520	0.0492	0.8832	0.0191	0.0334	0.0348	0.9112
Group 1: OLS based, censoring rate =50%								
IPW	0.0113	0.0847	0.0779	0.9250	-0.0047	0.0787	0.0761	0.9260
AIPW	0.0013	0.0509	0.0474	0.9340	-0.0025	0.0488	0.0487	0.9390
IMP	0.0088	0.0573	0.0567	0.9480	0.0018	0.0535	0.0551	0.9470
Group 2: WLS based, censoring rate =50%								
IPW	0.0116	0.0855	0.0778	0.9220	-0.0048	0.0790	0.0760	0.9300
AIPW	0.0013	0.0510	0.0473	0.9340	-0.0026	0.0489	0.0487	0.9390
IMP	0.0088	0.0573	0.0567	0.9490	0.0018	0.0535	0.0551	0.9470
Group 3: EFF based, censoring rate =50%								
IPW	0.0043	0.0801	0.0774	0.9290	-0.0056	0.0769	0.0758	0.9270
AIPW	-0.0008	0.0525	0.0476	0.9250	-0.0006	0.0497	0.0488	0.9410
IMP	0.0074	0.0579	0.0580	0.9550	0.0031	0.0536	0.0557	0.9520
Group 4: NPMLE, censoring rate =50%								
type I	0.0009	0.0584	0.0570	0.9360	-0.0044	0.1194	0.1080	0.8820
type II	-0.0490	-	-	-	0.0377	-	-	-
Group 5: Cox PH, censoring rate =50%								
COX	0.0273	0.0581	0.0597	0.9311	-0.0213	0.0466	0.0467	0.9261

Irrespective of the underlying data assumptions, all the nonparametric estimators we propose have ignorable finite sample bias, while the type II NPMLE is clearly biased. This inconsistency is especially evident when the censoring rate is low, and becomes somewhat masked when the censoring rate increases, especially when compared to the type I NPMLE. This may have contributed to the type II NPMLE being regarded as a consistent estimator in the literature. Moreover, this bias is not a finite sample bias since even at a sample size of $n = 2000$, the bias persists. Compared to the proposed estimators, the type I NPMLE has, for the most part, larger estimation variability, and the increased variability is rather substantial for $F_2(t)$ estimation. When the censoring rate increases, the bias of the type I NPMLE increases rather quickly, despite its asymptotic consistency. This is because in the estimation procedure, the mixture nature of the model is not taken advantage of at the maximization step. The Kaplan-Meier estimation in some subgroups could be based on very small sample sizes which can make the overall estimation unreliable.

In contrast, the three proposed nonparametric estimators have satisfactory small bias and are more efficient compared to the type I NPMLE. The optimal AIPW and IMP estimators both provide improvement over IPW in terms of estimation efficiency. When the censoring rate is moderate, IMP and AIPW perform similarly, while when the censoring rate increases, the superiority of the optimal AIPW over IMP becomes more notable. The similarity of the results in the first three groups of estimators suggests that the estimation efficiency is not sensitive to the choice of the complete data influence function ϕ. The same insensitivity of estimation efficiency to the choice of influence function is also evident in Ma and Wang (2011) for the complete data case. This phenomenon proves beneficial especially in the censoring data analysis since Robins and Rotnitzky (1992) have remarked that the best complete data influence function does not necessarily yield an optimal censoring data influence function, and finding the optimal member usually requires a computationally intensive procedure. Finally, the estimated standard error matches reasonably well with the

sample standard error, while the 95% confidence interval is close to the nominal level, with the only exception of the type I NPMLE. This is a consequence of the small subgroup sample size, and in simulation results not reported here, when we increase the sample size to 1000, the performance becomes satisfactory.

Our third simulation study mimics the COHORT study data. We generated 1000 data sets of sample size $n = 1000$ from a mixture of two distributions similar to the estimated distribution functions of the real data. According to Langbehn et al. (2004) and our own examination of the COHORT data, it is unlikely that age-at-death distribution for HD carriers and non-carriers follow a Cox model, so we did not use one. Figure 5 depicts the survival curves used where the higher curve corresponds to that for non-carriers, and the lower curve for carriers. We censored 65% of the observations with a uniformly distributed censoring process, and performed a similar analyzes as before. The results in Table 8 indicate that the estimators behave similarly as before in that all proposed nonparametric estimators have ignorable bias, and the AIPW estimator is, in general, most efficient. The type II NPMLE and Cox-based estimators have non-ignorable bias where the latter results from the true underlying model not satisfying the proportional hazards assumption.

In Figure 5, we depict the entire estimated survival curve $1 - F(t)$ using the efficiency based imputation estimator (EFFIMP) and the efficiency based AIPW estimator (EFFAIPW) as representatives of the proposed estimators, and compare them with the two NPMLEs and the Cox-based estimator. The figure displays the true and resulting mean estimated survival curves from 1000 data sets along with the 95% pointwise confidence bands. The EFFIMP and EFFAIPW estimators perform satisfactorily throughout the entire range of t, while the type I NPMLE starts to exhibit small sample estimation bias as time progresses. This confirms our observation that the type I NPMLE suffers from the small subgroup sample size difficulty and the instability of the Kaplan-Meier estimation procedure near the end of the range of the event times.

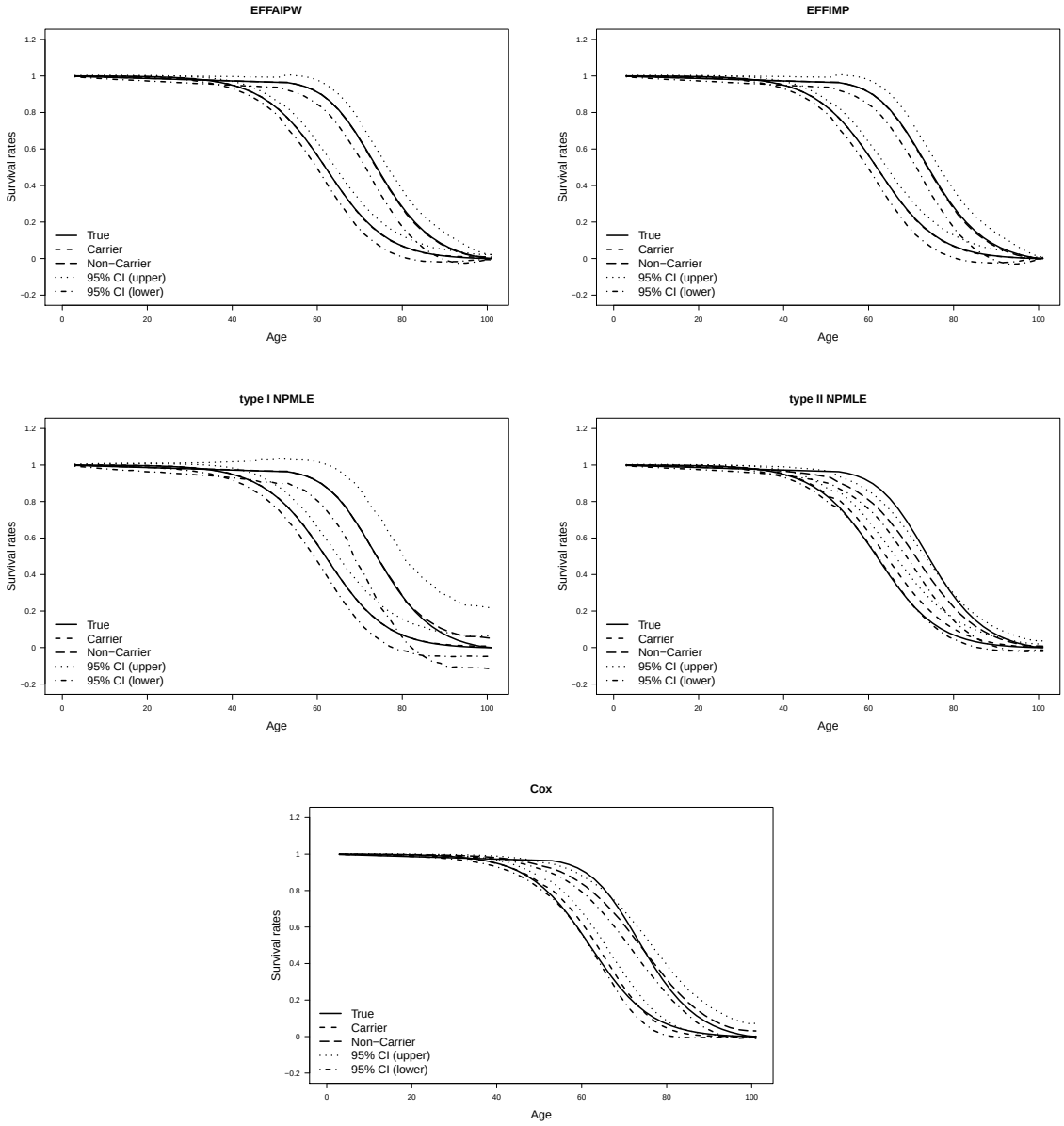

Figure 5: Simulation 3. True survival curve (solid) and the mean of 1000 simulations at each time point (short-dashed for carrier group, long-dashed for non-carrier group), 95% pointwise confidence band (upper band dotted, lower band dash-dotted) of the estimated survival curves. The mean and true survival curves are indistinguishable in EFFIMP and EFFAIPW estimators. Sample size is 1000, censoring rate is 65%.

Table 8: Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of the third simulation replicating the COHORT data, sample size $n = 1000$, 65% censoring rate, 1000 simulations.

	$F_1(t) = 0.7580$				$F_2(t) = 0.3446$			
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
Group 1: OLS based								
IPW	0.0007	0.0310	0.0304	0.9490	0.0083	0.0579	0.0544	0.9410
AIPW	0.0010	0.0281	0.0271	0.9400	0.0072	0.0550	0.0505	0.9230
IMP	0.0015	0.0289	0.0286	0.9480	0.0089	0.0557	0.0538	0.9450
Group 2: WLS based								
IPW	0.0007	0.0310	0.0304	0.9460	0.0084	0.0575	0.0543	0.9400
AIPW	0.0010	0.0280	0.0270	0.9420	0.0072	0.0547	0.0505	0.9240
IMP	0.0013	0.0288	0.0284	0.9430	0.0089	0.0555	0.0538	0.9460
Group 3: EFF based								
IPW	0.0005	0.0309	0.0303	0.9460	0.0067	0.0560	0.0537	0.9400
AIPW	-0.0010	0.0284	0.0270	0.9350	0.0095	0.0543	0.0501	0.9240
IMP	0.0005	0.0289	0.0286	0.9430	0.0097	0.0555	0.0532	0.9380
Group 4: NPMLE								
type I	-0.0005	0.0496	0.0509	0.9440	-0.0003	0.0939	0.0858	0.8820
type II	-0.0143	-	-	-	0.0925	-	-	-
Group 5: Cox PH								
COX	-0.0156	0.0376	0.0373	0.9250	0.0389	0.0455	0.0419	0.9041

The type II NPMLE and the Cox-based estimators show a non-ignorable bias for a wide range of t's. Finally, the Cox-based estimator performs poorly, both in terms of bias and its inability to capture the small crossing effect around age 38.

4.5 Analysis of the COHORT data

Data from the COHORT study consists of 4587 relatives who were assigned one of 6 different mixing proportions for being carriers or non-carriers of the HD gene. Letting p_c denote the probability of being a carrier and $(p_c, 1 - p_c)$ denote a mixture proportion, roughly 29.87% of the subjects were classified in the (0,1) proportion group, 42.93% in the (0.5,0.5) group, 22.61% in the (0.97,0.03) group, 0.044% in the (0.75,0.25) group, 3.03% in the (0.25,0.75), and 1.53% in the (1,0) group. The event time of interest is age of death, and roughly 68% of the data is censored. The overall objective of the study is to estimate the age-at-death distribution for HD gene carriers, or equivalently the corresponding survival function, and compare it with non-carriers. The severity of Huntington disease warrants that non-carriers tend to live longer, so we expect to see lower survival rates for the latter group, especially post 30-50 years old, the typical age of onset of Huntington disease. As the survival rates for non-carriers in the US should behave similarly to the general US population, we use the Kaplan-Meier estimated survival curve for the general US population in 2003 (Arias, 2006) as a base comparison.

Figure 6 displays the estimated survival curves for the carrier and non-carrier groups of the COHORT study data using the type I NPMLE estimator, the efficiency based AIPW and IMP estimators, and the Cox-based estimators. As evident in the figure, the type I NPMLE poorly estimates the survival rates, and suggests an atypical repeated crossing of the two survival curves. The contradictory behavior of this estimator is also numerically evident in Table 9 which shows the survival rates and 95% confidence intervals at different ages.

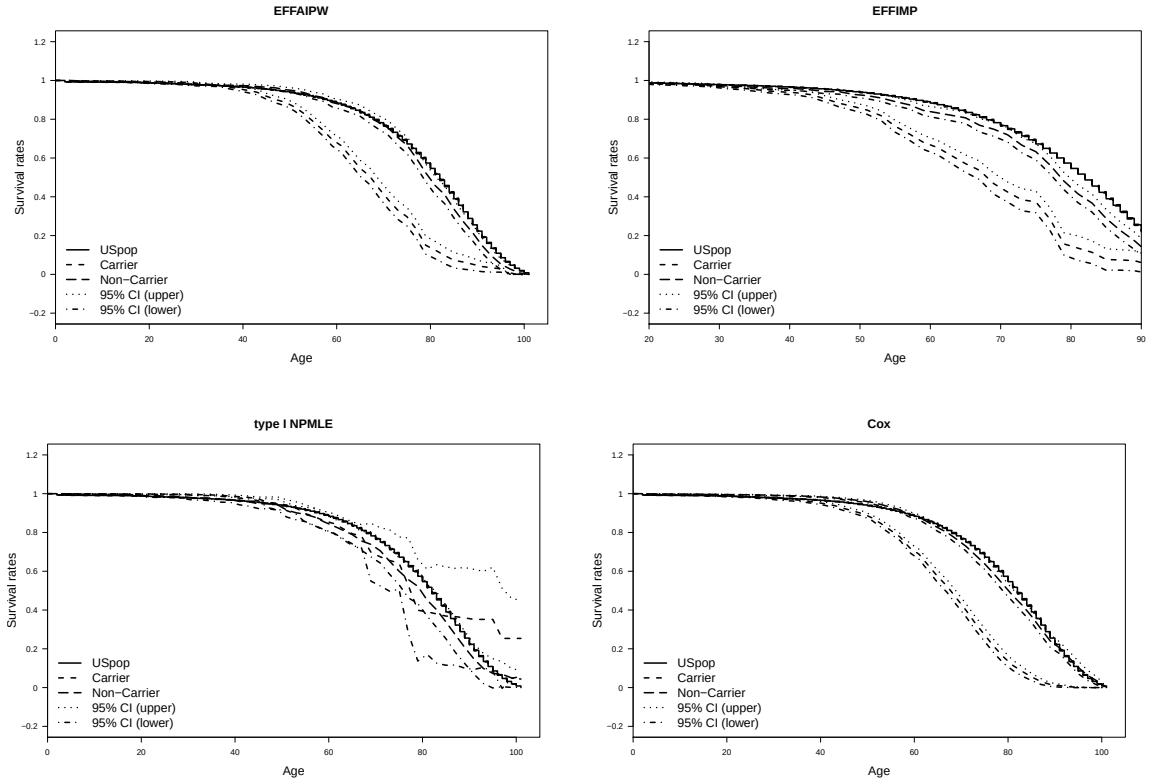

Figure 6: Estimated survival curves and 95% point-wise confidence bands (upper band dotted, lower band dash-dotted) for Huntington disease data using the efficient complete data influence function based optimal AIPW and IMP, the Cox-based, and the type I NPMLE estimator. The short-dashed curve denotes survival rates for persons possessing a carrier gene, the long-dashed curve for persons possessing a non-carrier gene, and the solid curve corresponds to survival rates for general US population in 2003.

Table 9: Estimated survival rates and 95% confidence intervals (in parentheses) for carrier (C) and non-carrier (NC) groups in COHORT data.

		Estimators		
Age	Gene	US Pop 2003	EFFAIPW	EFFIMP
40	C		0.952 (0.941, 0.963)	0.937 (0.925, 0.949)
	NC	0.965 (0.964, 0.966)	0.972 (0.964, 0.980)	0.957 (0.948, 0.967)
50	C		0.875 (0.856, 0.895)	0.859 (0.838, 0.880)
	NC	0.938 (0.936, 0.939)	0.950 (0.937, 0.963)	0.927 (0.913, 0.942)
60	C		0.675 (0.641, 0.710)	0.662 (0.623, 0.700)
	NC	0.883 (0.881, 0.885)	0.881 (0.858, 0.904)	0.836 (0.810, 0.863)
70	C		0.411 (0.365, 0.457)	0.439 (0.386, 0.492)
	NC	0.766 (0.764, 0.769)	0.762 (0.725, 0.798)	0.723 (0.684, 0.762)
80	C		0.144 (0.098, 0.189)	0.157 (0.097, 0.216)
	NC	0.546 (0.543, 0.549)	0.488 (0.441, 0.534)	0.449 (0.402, 0.496)
90	C		0.044 (0.016, 0.071)	0.051 (0.002, 0.099)
	NC	0.223 (0.220, 0.225)	0.173 (0.134, 0.213)	0.135 (0.094, 0.176)
		type I NPMLE	Cox	
40	C	0.985 (0.979, 0.992)	0.950 (0.943, 0.958)	
	NC	0.965 (0.947, 0.983)	0.982 (0.979, 0.986)	
50	C	0.945 (0.917, 0.973)	0.885 (0.872, 0.898)	
	NC	0.922 (0.890, 0.955)	0.958 (0.952, 0.965)	
60	C	0.852 (0.804, 0.900)	0.704 (0.679, 0.729)	
	NC	0.847 (0.803, 0.891)	0.886 (0.873, 0.899)	
70	C	0.686 (0.546, 0.827)	0.427 (0.396, 0.459)	
	NC	0.710 (0.640, 0.780)	0.746 (0.723, 0.770)	
80	C	0.398 (0.119, 0.676)	0.132 (0.104, 0.160)	
	NC	0.476 (0.390, 0.561)	0.498 (0.468, 0.528)	
90	C	0.352 (0.097, 0.606)	0.010 (0.003, 0.017)	
	NC	0.169 (0.087, 0.250)	0.205 (0.175, 0.234)	

Notice, in particular, the type I NPMLE suggests that being an HD carrier gives no worse chance of survival than a non-carrier, a notion contrary to the debilitating effects of Huntington disease. The wide confidence bands, especially for the carrier group, results from the inefficiency of the type I NPMLE. This poor performance is most likely a result of some proportion groups having small sample sizes; only 0.044% of patients in the study are classified in the (0.75, 0.25) group. While removing these two individuals may lead to more sensible estimates of the type I NPMLE, the overall inefficiency of this estimator may, in general lead to invalid inferences.

In contrast to the type I NPMLE, the superior performance of the AIPW suggests that carrying an HD gene mutation increases a subject's cumulative risk of death significantly in the age range 45 to 80. For example, referring to Table 9, the cumulative risk of death for carriers at age 50 is 12.4% (95%CI: 10.4%, 14.4%) compared to 5.0%(95% CI: 3.7%, 6.2%) in non-carriers. The corresponding rate at age 70 is 58.9% (95%CI: 54.2%, 63.5%) in carriers versus 23.8% (95%CI: 27.4%, 20.2%) in non-carriers. Such numerical evidence reinforces the severity of Huntington disease and the importance of groups like the Huntington Study Group for finding treatments to reduce this fatality risk.

A reason to prefer the results of the AIPW estimator, compared to the type I NPMLE say, is that the estimated cumulative risk of death in non-carriers is very similar to the US population rates estimated from the US Census data (Arias, 2006), which is expected since the risk in non-carriers reflects the general population. Likewise, the efficiency based IMP estimator provides similar estimates as AIPW, with slightly higher estimated standard errors, a consequence of the higher censoring rate. The estimation results from IMP and AIPW estimators differ only in the age range of 70 and 80, where the IMP estimator suggests a steeper decline in survival rates for patients with an HD gene mutation. Still, in general, both the AIPW and IMP estimators agree in the overall behavior of the cumulative risk of death for both the HD carrier and non-carrier groups.

Finally, as a sensitivity analysis, we also applied the Cox-based estimator to the COHORT data. In Figure 6, we see that while all other estimators considered provide evidence that the two distributions cross at age 38, the Cox-based estimator does not do so. Moreover, while, post age 40, it appears that the Cox-based estimator provides reasonable estimates for the cumulative risk of death in non-carriers as compared to the US population rates, Table 9 shows that the AIPW estimator does better. For example, at ages 40, 50, 60 and 70 (and ages in between, not shown), the AIPW estimator gives survival rates for the non-carrier group that are more similar to the US population rates than are the Cox-based estimates. This result agrees with an earlier observation by Langbehn et al. (2004) about the unlikeliness of the age-ag-death distribution for HD carriers and non-carriers to follow a Cox model.

4.6 Discussion

We propose two IPW based estimators and an IMP estimator for censored mixture data, among which the optimal AIPW achieves the optimal efficiency based on a fixed complete data influence function. These estimators are easy to compute and do not involve any iterative procedures. When the sample size is small and the censoring rate is moderate, the IMP estimator can sometimes compete or even outperform the asymptotically optimal AIPW estimator. We also point out the surprising results of the non-optimality of the type I NPMLE and the inconsistency of the type II NPMLE proposed in the literature. Our finite sample simulations suggest that the efficiency loss of the type I NPMLE and the bias of the type II estimator can be quite substantial, and the finite sample bias of the type I NPMLE can be non-ignorable when the sample size is small or the estimation region is close to the upper end of the distribution support. Through various simulation studies, we demonstrate the sensitivity of a Cox-based estimator to the underlying model assumptions. When the underlying model follows a proportional hazards model, the Cox-based estimator

outperforms all nonparametric estimators in consistency and efficiency, as expected. When these assumptions are in doubt, however, the Cox-based estimator shows a non-ignorable bias. In contrast, the proposed AIPW is robust to the underlying model assumptions and even has similar empirical standard errors to the Cox-based estimator when the proportional hazards assumption holds, thus indicating minimal efficiency loss of this nonparametric estimator.

Applying these estimators to the COHORT data, we see that the inefficiency of the type I NPMLE leads to misleading conclusions about the fatality rates of patients with and without the HD gene mutation. The optimal AIPW and IMP estimators, in contrast, show reasonable survival rate estimates as measured by the closeness of the estimates for the non-carrier group to the estimates for the general US population in 2003. Both estimators also indicate that patients with an HD gene mutation have higher mortality rates than patients without the gene between ages 45 and 80. These results are in agreement with earlier observations of Huntington disease, thus exhibiting the usefulness of our proposed estimators.

CHAPTER V

CONCLUSION

Using the results of semiparametric theory, we have derived the general class of semiparametric estimators and characterized the properties of the optimal estimator therein for parameters in (1). a restricted moment model with measurement error; (2). a general class of survival models; and (3). a model with data of censored, mixed observations, typical in kin-cohort studies. The underlying statistical basis for these models is semiparametric theory which provides general procedures to obtain practically important parameters in the presence of nuisance parameters. For each of the three models considered, nuisance parameters correspond to unknown error distributions in regression, unknown baseline hazard functions in survival analysis, and unknown conditional distributions in outcomes given group membership in a mixture data model. The versatile applications of semiparametric theory led to the following findings from this work.

Prior to this work, existing methods for a restricted moment model with measurement error did not allow an unspecified model error distribution and still guaranteed consistency. To the best of our knowledge, the estimation procedure developed in Chapter II is the first to give consistency while allowing misspecification in both the model error distribution $p_{\epsilon|X,Z}(\epsilon|x,z)$ and the conditional covariate distribution $p_{X|Z}(x|z)$. The estimators are based on the semiparametric efficient score which is calculated under several possibly incorrect distribution assumptions resulting from the misspecified model error distribution, or unobservable covariates' distribution, or both. Through various simulation studies which accounted for different variance structures in the model error distribution, we demonstrated that our method is robust and delivers impressive results with considerably less bias than a method which ignores measurement error. Moreover, our method performs considerably

better than a competing method by Tsiatis and Ma (2004) which does allow misspecification in the latent variable distribution but requires a correct model error distribution. Future work entails extending these results to having instead of just two infinite dimensional parameters corresponding to the unknown model error and latent variable distributions, we have k, say, infinite dimensional nuisance parameters. This extension incorporates many problems, including the more difficult problem of quantile regression, where $k = 1$ and the nuisance parameter corresponds to the unspecified error distribution which has conditional τth quantile as zero.

The developments in Chapter III integrate the results of Yang and Prentice (2005), Zhang et al. (2008), and Lu and Tsiatis (2008) to provide a simple and direct illustration for comparing nonproportional hazards functions with the beneficial supplement of having improved efficiency of inferences through incorporating auxiliary covariates. Although the results are intended for comparing non-proportional hazards, the flexibility of the base model from Yang and Prentice (2005) permits a wider range of comparisons, including those for proportional hazards and the proportional odds model. The proposed method takes an unbiased estimating equation without auxiliary covariates and augments it by a term adjusted for covariates. An appropriate choice of the augmentation term leads to a consistent and more efficient estimator than the corresponding unadjusted estimator. The approach of incorporating covariates does not raise the issue of model misspecification, nor does it risk any possible loss of precision. Moreover, the proposed method generalizes the results by Lu and Tsiatis (2008) which incorporates auxiliary covariate information only in a proportional hazards setting. Even when the correlation between the auxiliary covariates and the times to events is small, the method still provides more efficient estimators than the corresponding ones without covariate adjustment.

Finally, we use semiparametric theory to develop three nonparametric estimators for estimating the age-at death distributions for Huntington disease gene carriers and non-

carriers from a kin-cohort study. The three estimators are consistent, easy to compute, and are not susceptible to model misspecification nor parametric restrictions. We demonstrated that among all estimators considered, the optimal augmented inverse probability weighting (AIPW) estimator delivered the best estimates for the age-at-death distributions in the carrier and non-carrier groups. We concluded this because the estimated survival curve for the non-carrier group behaved similarly to the Kaplan-Meier estimated survival curve for the general US population in 2003. The estimated distributions from the AIPW estimator demonstrated that non-carriers tend to live longer, and carriers had lower survival rates, especially post 30-50 years old, the typical age of onset of Huntington's disease. These results agreed with previous analysis of survival rates for Huntington's disease. Although useful, the proposed estimators excluded auxiliary covariate information typically collected in clinical trials. Future work involves using the methods proposed in Chapter III to incorporate the auxiliary information in a model free manner.

REFERENCES

Arias, E. (2006). United states life tables, 2003. *National Vital Statistics Report* **54**, 1–40.

Bagdonavicius, V., Hafdi, M., and Nikulin, M. (2004). Analysis of survival data with cross-effects of survival functions. *Biostatistics* **5**, 415–425.

Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika* **87**, 329–343.

Bang, H. and Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics* **58**, 643–649.

Bennet, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273 – 277.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density. *Journal of the American Statistical Association* **83**, 1184–1186.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. London: CRC Press, 2nd edition.

Chan, L. K. and Mak, T. K. (1985). On the polynomial functional relationship. *Journal of the Royal Statistical Society B* **47**, 510–518.

Chatterjee, N. and Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics* **57**, 245–252.

Chen, X., Hu, Y., and Lewbel, A. (2009). Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica* **19**, 949–968.

Cheng, C. L. and Schneeweiss, H. (1998). Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society B* **60**, 189–199.

Cheng, C. L., Schneeweiss, H., and Thamerus, M. (2000). A small sample estimator for a polynomial regression with errors in the variables. *Journal of the Royal Statistical Society B* **62**, 699–709.

Cox, D. (1972). Regression model and life-tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187 –220.

Diao, G. and Lin, D. (2005). Semiparametric methods for mapping quantitative trait loci with censored data. *Biometrics* **61**, 789–798.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics* **19**, 1257–1272.

Flag, E., Coates, R., Calle, E., Potischman, N., and Thun, M. (2000). Validation of the American Cancer Society Cancer Prevention Study II Nutrition Survey Cohort Food Frequency Questionnaire. *Epidemiology* **11**, 462–468.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Gail, M., Pee, D., and Carroll, R. (1999). Kin-cohort designs for gene characterization. *Journal of the National Cancer Institute* **26**, 55–60.

Grouin, J. M., Day, S., and Lewis, J. (2004). Adjustment for baseline covariates: An Introductory Note. *Statistics in Medicine* **23**, 697 –699.

Hauck, W. W., Anderson, S., and Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* **19**, 249–256.

Hess, K. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* **13**, 1045 – 1062.

Hsieh, F. (2001). On heteroscedastic hazards regression models: Theory and Application. *Journal of the Royal Statistical Society, Series B* **63**, 63–79.

Hu, Y. and Schennach, S. (2006). Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments. CeMMAP working papers CWP17/06, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Huntington Study Group (2011a). Cooperative Huntington's Observational Research Trial, [ClinicalTrials.gov identifier NCT00313495]. http://www.clinicaltrials.gov. US National Institutes of Health, ClinicalTrials.gov [online].

Huntington Study Group (2011b). HSG: Huntington Study Group, Seeking Treatments that Make a Difference for Huntington Disease. http://www.huntington-study-group.org.

Jin, C., Fine, J. P., and Yandell, B. S. (2007). A unified semiparametric framework for quantitative trait loci analyses, with application to spike phenotypes. *Journal of the American Statistical Association* **102**, 56–67.

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

Lander, E. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics* **121**, 743–756.

Langbehn, D., Brinkman, R., Falush, D., Paulsen, J., and Hayden, M. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics* **65**, 267–277.

Lesaffre, E. and Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* **22**, 3586–3596.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of age of onset data from case-control family studies. *Biometrics* **54**, 1030–1039.

Liang, H., Härdle, W., and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Annals of Statistics* **27**, 1519–1536.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* **104**, 234–248.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lipitz, S., Ibrahim, J., and Zhao, L. (1999). A weighted estimation equation for missing covariate data with properties similar to maximum likelihood. *Journal of the Royal Statistical Society B* **94**, 1147–1160.

Liu, M., Lu, W., and Shao, Y. (2006). Interval mapping of quantitative trait loci for time-to-event data with the proportional hazards mixture cure model. *Biometrics* **62**, 1053–1061.

Lu, X. and Tsiatis, A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* **95**, 679–674.

Ma, Y. and Wang, Y. (2011). Efficient semiparametric estimation for mixture data. *Submitted Manuscript* .

Marder, K., Levy, G., Louis, E., Mejia-Santana, H., Cote, L., Andrews, H., Harris, J., Waters, C., Ford, B., Frucht, S., Fahn, S., and Ottman, R. (2003). Accuracy of family history data on parkinson's disease. *Neurology* **61**, 18–23.

Moore, D., Chatterjee, N., Pee, D., and M.H., G. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genetic Epedmiology* **20**, 210–227.

Moré, J., Sorensen, D., Hillstrom, K., and Garbow, B. (1984). *Sources and Development of Mathematical Software: The MINPACK Project*. New Jersey: Prentice-Hall.

Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

Rabinowitz, D. (2000). Computing the efficient score in semi-parametric problems. *Statistica Sinica* **10**, 265–280.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Robins, J. and Rotnitzky, A. (1992). *AIDS Epidemiology, Methodological Issues: Recovery of Information and Adjustment for Dependent Censoring using Surrogate Markers*. Basel: Birkhaüser.

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

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Robinson, L. and Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **59**, 227–240.

Ross, C. A. (1995). When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. *Neuron* **15**, 493–496.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., et al. (1996). Phenotypic characterization of individuals with 30-40 cag repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *American Journal of Human Genetics* **59**, 16–22.

Senn, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**, 467–475.

Singleton, R. (1969). Algorithm 347, an efficient algorithm for sorting with minimial storage. *Communications of the ACM* **12**, 185–187.

Stefanski, L. and Carroll, R. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 169–184.

Struewing, J., Hartge, P., Wacholder, S., Baker, S., Berlin, M., McAdams, M., Timmerman, M., Brody, L., and Tucker, M. (1997). The risk of cancer associated with specific mutations of brca1 and brca2 among ashkenazi jews. *New England Journal of Medicine* **336**, 1401–1408.

Tangen, C. M. and Koch, G. G. (1999). Nonparametric analysis of covariance for hypothesis testing with logrank and wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics* **9**, 307–338.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tsiatis, A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91**, 835–848.

Tsimberidou, A. M., OBrien, S., Khouri, I., Giles, F., Kantarjian, H., Champlin, R., Wen, S., Do, K., Smith, S., Lerner, S., Freireich, E., and Keating, M. J. (2006). Clinical outcomes and prognostic factors in patients with Richters Syndrome treated with chemotherapy or chemoimmunotherapy with or without stem-cell transplantation. *Journal of Clinical Oncology* **24**, 2343–2351.

Van der Laan, M. J. and Hubbard, A. E. (1998). Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. *Biometrika* **85**, 771–783.

Verweij, J. and Van Houlwelingen, H. (1995). Time dependent effects of fixed covariates in Cox regression. *Biometrics* **51**, 1550–1556.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L., and Tucker, M. (1998). The kin-cohort study for estimating penentrance. *American Journal of Epidemiology* **148**, 623–630.

Wang, L., Rotnitzky, A., and Lin, X. (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association* **105**, 1135–1146.

Wang, N. and Robins, J. M. (1998). Large sample inference in parametric multiple imputation. *Biometrika* **85**, 935–948.

Wang, Y., Clark, L. N., Marder, K., and Rabinowitz, D. (2007). Non-parametric estimation of genotype-specific age-at-onset distributions from censored kin-cohort data. *Biometrika* **94**, 403–414.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika* **82**, 57–67.

Wickramaratne, P. and Holford, T. (1989). Confounding in epidemiologic studies. response. *Biometrics* **45**, 1319–1322.

Wu, R., Ma, C., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL.* New York: Springer.

Wu, R., Ma, C., Chang, M., Littell, R., Wu, S., Yin, T., Huang, M., Wang, M., and Casella, G. (2002). A logistic mixture model for characterizing genetic determinants conserved synteny in rat and mouse for a blood pressure causing differentiation in growth trajectories. *Genetical Research* **79**, 235–245.

Wu, R., Zeng, Z., S.E., M., and O'Malley, D. (2000). The case for molecular mapping in forest tree breeding. *Plant Breeding Reviews* **19**, 41–68.

Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika* **92**, 1–17.

Yu, Z. and Lin, X. (2008). Nonparametric regression using local kernel estimating equations for correlated failure time data. *Biometrika* **95**, 123–137.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric models with censored data (with discussion). *Journal of the Royal Statistical Society, Series B* **69**, 507–564.

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

Zhao, W. and Wu, R. (2008). Wavelet-based nonparametric functional mapping of longitudinal curves. *Journal of the American Statistical Association* **103**, 714–725.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER I

Semiparametric Theory

We highlight the semiparametric theory and critical steps in this approach. First, an RAL estimator for the p-dimensional parameter β based on observed random variables (W_i, Y_i, Z_i) for $i = 1, \ldots, n$ is uniquely characterized by its influence function through

$$n^{1/2}(\widehat{\beta}_n - \beta_0) = n^{-1/2}\textstyle\sum_{i=1}^{n} \varphi(W_i, Y_i, Z_i) + o_p(1),$$

where the influence functions $\varphi(W_i, Y_i, Z_i)$, $i = 1, \ldots, n$ are independent, identically distributed, mean zero random vectors with length p. Here, $\widehat{\beta}_n$ denotes the RAL estimator, β_0 the true parameter value, and $o_p(1)$ converges in probability to zero as n tends to infinity. The asymptotic variance of $\widehat{\beta}_n$ equals the variance of φ. Hence the influence function with the smallest variance yields the most efficient RAL estimator.

From a geometric viewpoint, influence functions are elements of a Hilbert space $\mathcal{H}$ consisting of all functions $h(W, Y, Z)$ such that $E\{h(W, Y, Z)\} = 0$ and such that $E\{h(W, Y, Z)^\top h(W, Y, Z)\} < \infty$. Here and throughout, the expectation is always performed under the true distribution. Influence functions are normalized and orthogonal to a nuisance tangent space Λ, which is a subspace of $\mathcal{H}$. The general approach of deriving influence functions thus consists of deriving Λ and its orthogonal complement $\Lambda^\perp$. All the influence functions, including the efficient one, lie in $\Lambda^\perp$.

The efficient influence function, denoted $\varphi_{\text{eff}}(W, Y, Z)$, satisfies $E(\varphi S_\beta^\top) = I$, where S_β is the score vector with respect to β and I is the $p \times p$ identity matrix. The exact form of the efficient influence function involves the efficient score vector $S_{\text{eff}}(W, Y, Z)$

defined as the projection of $S_\beta(W,Y,Z)$ onto $\Lambda^\perp$, or $S_{\text{eff}}(W,Y,Z) = \Pi\{S_\beta(W,Y,Z)|\Lambda^\perp\}$. Normalizing $S_{\text{eff}}(W,Y,Z)$, we obtain the efficient influence function as $\varphi_{\text{eff}}(W,Y,Z) = [E\{S_{\text{eff}}(W,Y,Z)S_{\text{eff}}^\top(W,Y,Z)\}]^{-1}S_{\text{eff}}(W,Y,Z)$.

Results for RMM without measurement error

For the RMM without measurement error (full data model), the conditional distribution of X given Z, $\eta_1(x,z)$, the conditional distribution of ϵ given (X,Z), $\eta_2(\epsilon,x,z)$, and the distribution of Z, $\eta_3(z)$, are unknown and are nuisance parameters of infinite dimension. We emphasize that covariates (X,Z) are precisely observed in the case of no measurement error. The joint probability density function for (X,Y,Z) is $p_{X,Y,Z}(x,y,z) = \eta_1(x,z)\eta_2\{y - m(x,z;\beta),x,z\}\eta_3(z)$ such that $\int \epsilon\eta_2(\epsilon,x,z)d\epsilon = 0$ for all x,z where β is the parameter of interest. Applying semiparametric theory to the RMM without measurement error, the nuisance tangent space, its orthogonal complement and the efficient influence function are summarized in the following proposition. A detailed derivation is available in Chapter 4 of Tsiatis (2006).

Proposition 1. For the full data restricted moment model, we have that the Hilbert space $\mathcal{H}^F = \{f(X,Y,Z) : E(f) = 0, \text{var}(f) < \infty\}$. In estimating β, the nuisance tangent space is $\Lambda^F = \{f(X,Y,Z) : E(f\epsilon|X,Z) = 0, E(f) = 0, \text{var}(f) < \infty\}$; the nuisance tangent space orthogonal complement is $\Lambda^{\perp F} = \{g(X,Z)\epsilon\} = [g(X,Z)\{Y - m(X,Z;\beta)\}]$, where g is an arbitrary function of (X,Z) such that $E(g^Tg) < \infty$; the score vector with respect to β is $S_\beta^F = -m'_\beta(X,Z;\beta)\partial\text{log}\eta_2(\epsilon,X,Z)/\partial\epsilon$; the efficient score $S_{\text{eff}}^F = m'_\beta(X,Z;\beta)E(\epsilon^2|X,Z)^{-1}\epsilon$, where $m'_\beta(X,Z;\beta)$ denotes $\partial m(X,Z;\beta)/\partial\beta$; and the efficient influence function is of the form $\varphi_{\text{eff}}^F = E\{E(\epsilon^2|X,Z)^{-1}m'_\beta m_\beta'^T\}m'_\beta E(\epsilon^2|X,Z)^{-1}\epsilon$.

Proof of Theorem 1

The description of Λ follows immediately from the fact that the score vector $S_{\eta_i}(W,Y,Z)$ is equal to $E\{S_{\eta_i}(X,Y,Z)|W,Y,Z\}$ for $i = 1,2$; see (Rao, 1973, p. 330). It thus suffices to demonstrate $\Lambda^{\perp}$ is as given. We first show that any element $f(W,Y,Z)$ satisfying $E\{f(W,Y,Z)|X,Y,Z\} = g(X,Z)\epsilon$ is orthogonal to all elements of Λ. To this end, the inner product of f and an arbitrary element $E\{h(X,Y,Z)|W,Y,Z\} \in \Lambda$ satisfying $E\{h(X,Y,Z)\epsilon|X,Z\} = 0$ indeed shows orthogonality since

$$\begin{aligned} & E[E\{h^T(X,Y,Z)|W,Y,Z\}f(W,Y,Z)] = E\{h^T(X,Y,Z)f(W,Y,Z)\} \\ = \;& E[h^T(X,Y,Z)E\{f(W,Y,Z)|X,Y,Z\}] = E\{h^T(X,Y,Z)g(X,Z)\epsilon\} = 0. \end{aligned}$$

Conversely, we now demonstrate that any $f \in \Lambda^{\perp}$ must satisfy $E(f|X,Y,Z) = g(X,Z)\epsilon$. Let $k(X,Y,Z) = E(f|X,Y,Z)$ and consider

$$h(X,Y,Z) \;=\; k(X,Y,Z) - g(X,Z)\epsilon, \tag{A.1}$$

where $g(X,Z) = E(k\epsilon|X,Z)/E(\epsilon^2|X,Z)$. From $E(f) = 0$ and $E(\epsilon|X,Z) = 0$ we immediately have that $E(h\epsilon|X,Z) = 0$ and $E(h) = 0$ implying that $E(h|W,Y,Z) \in \Lambda$. Now with $f(W,Y,Z) \in \Lambda^{\perp}$ and $E(h|W,Y,Z) \in \Lambda$, we have that the inner product of these two terms is zero, and so

$$\begin{aligned} 0 \;&=\; E[f^T(W,Y,Z)E\{h(X,Y,Z)|W,Y,Z\}] = E\{f^T(W,Y,Z)h(X,Y,Z)\} \\ &=\; E[E\{f^T(W,Y,Z)|X,Y,Z\}h(X,Y,Z)] = E\{k^T(X,Y,Z)h(X,Y,Z)\} \\ &=\; E(h^Th) + E[g(X,Z)^TE\{\epsilon h(X,Y,Z)|X,Z\}] = E(h^Th), \end{aligned}$$

where the last equality holds since $E(h\epsilon|X,Z) = 0$. Thus, by properties of Hilbert spaces, whenever $E(h^Th) = 0$, we must have $h = 0$ almost surely. This and (A.1) demonstrate that $E(f|X,Y,Z) = g(X,Z)\epsilon$ almost surely, and consequently, $\Lambda^{\perp}$ is as defined.

Sufficiency and necessity of condition (2.2) for S_{eff}

First, $\mathcal{K}$ and $\mathcal{K}^*$ are conjugates of each other because $< \mathcal{K}^*\{f(W,Y,Z)\}, g(X,Y,Z) >$ equals

$$E[E\{f^T(W,Y,Z)|X,Y,Z\}g(X,Y,Z)] = E\{f^T(W,Y,Z)g(X,Y,Z)\}$$
$$= E[f^T(W,Y,Z)E\{g(X,Y,Z)|W,Y,Z\}] =< f, \mathcal{K}\{g(X,Y,Z)\} > .$$

From Proposition 1 and Theorem 1, we have the following relationships:

$$S_\beta(W,Y,Z) = \mathcal{K}\{S_\beta^F(X,Y,Z)\},\ \Lambda = \mathcal{K}(\Lambda^F),\ \mathcal{K}^*(\Lambda^\perp) \subset \Lambda^{\perp F}. \tag{A.2}$$

By definition, the efficient score vector $S_{\text{eff}}(W,Y,Z) = S_\beta(W,Y,Z) - \Pi\{S_\beta(W,Y,Z)|\Lambda\}$, where $\Pi\{S_\beta(W,Y,Z)|\Lambda\}$ denotes the projection of S_β onto Λ. Using the above relationships, we proceed to write $S_{\text{eff}}(W,Y,Z)$ as a function of elements from the full-data model, allowing us to take advantage of the properties from Proposition 1.

Since $\Pi\{S_\beta(W,Y,Z)|\Lambda\}$ is an element of Λ and $\Lambda = \mathcal{K}(\Lambda^F)$, there exists some $a^F(X,Y,Z) \in \Lambda^F$ such that $\Pi\{S_\beta(W,Y,Z)|\Lambda\} = \mathcal{K}\{a^F(X,Y,Z)\}$. Likewise, the score vector $S_\beta(W,Y,Z)$ is such that $S_\beta(W,Y,Z) = \mathcal{K}\{S_\beta^F(X,Y,Z)\} = \mathcal{K}[S_{\text{eff}}^F(X,Y,Z) + \Pi\{S_\beta^F(X,Y,Z)|\Lambda^F\}]$. Together, these results imply that the efficient score vector satisfies

$$\begin{aligned} S_{\text{eff}}(W,Y,Z) &= S_\beta(W,Y,Z) - \Pi\{S_\beta(W,Y,Z)|\Lambda\} \\ &= \mathcal{K}[S_{\text{eff}}^F(X,Y,Z) + \Pi\{S_\beta^F(X,Y,Z)|\Lambda^F\} - a^F(X,Y,Z)] \\ &= \mathcal{K}\{d(X,Y,Z)\}, \end{aligned}$$

where $d(X,Y,Z) = S_{\text{eff}}^F(X,Y,Z) - b^F(X,Y,Z)$ and we have $b^F(X,Y,Z) = a^F(X,Y,Z) - \Pi\{S_\beta^F(X,Y,Z)|\Lambda^F\}$. Now having expressed $S_{\text{eff}}(W,Y,Z)$ as $\mathcal{K}\{d(X,Y,Z)\}$, we derive the properties of $d(X,Y,Z)$ so that $d(X,Y,Z)$ may be solved explicitly.

The function $d(X,Y,Z)$ is formed by two elements, $S_{\text{eff}}^F(X,Y,Z)$ and $b^F(X,Y,Z)$

where the former lies in $\Lambda^{\perp F}$ and the latter in Λ^F. By properties of projection and orthogonality, the projection of d onto $\Lambda^{\perp F}$ gives

$$\Pi\{d(X,Y,Z)|\Lambda^{\perp F}\} \;=\; \Pi\{S^F_{\text{eff}}(X,Y,Z) - b^F(X,Y,Z)|\Lambda^{\perp F}\} = S^F_{\text{eff}}(X,Y,Z) \quad \text{(A.3)}$$

Above we showed that $\mathcal{K}\{d(X,Y,Z)\} = S_{\text{eff}}(W,Y,Z)$ which implies that $\mathcal{K}^* \circ \mathcal{K}\{d(X,Y,Z)\} = \mathcal{K}^*\{S_{\text{eff}}(W,Y,Z)\}$. Since the efficient score vector $S_{\text{eff}}(W,Y,Z)$ is an element of $\Lambda^{\perp}$, relation (A.2) implies $\mathcal{K}^* \circ \mathcal{K}\{d(X,Y,Z)\} = \mathcal{K}^*\{S_{\text{eff}}(W,Y,Z)\} \in \Lambda^{\perp F}$. Hence, orthogonality implies that $d(X,Y,Z)$ must also satisfy

$$\Pi[\mathcal{K}^* \circ \mathcal{K}\{d(X,Y,Z)\}|\Lambda^F] = 0.$$

Combining the properties of $d(X,Y,Z)$ in (A.3) and in the above display, the efficient score vector $S_{\text{eff}}(W,Y,Z)$ is such that $S_{\text{eff}}(W,Y,Z) = \mathcal{K}\{d(X,Y,Z)\}$ where $d(X,Y,Z)$ satisfies

$$\Pi\{d(X,Y,Z)|\Lambda^{\perp F}\} + \Pi[\mathcal{K}^* \circ \mathcal{K}\{d(X,Y,Z)\}|\Lambda^F] = S^F_{\text{eff}}(X,Y,Z),$$

which simplifies into expression (2.2).

Up to now, we have shown the existence of a function $d(X,Y,Z)$ such that $S_{\text{eff}} = \mathcal{K}(d)$. To complete the demonstration, we show that any function $d(X,Y,Z)$ satisfying (2.2) yields the correct S_{eff}. First, supposing that d does satisfy condition (2.2), it immediately follows that

$$\Pi(d|\Lambda^{\perp F}) = S^F_{\text{eff}},\ \Pi\{\mathcal{K}^* \circ \mathcal{K}(d)|\Lambda^F\} = 0.$$

The first result and the definition of the efficient score vector implies

$$d(X,Y,Z) = S^F_{\text{eff}}(X,Y,Z) + a^F(X,Y,Z) = S^F_{\beta}(X,Y,Z) + b^F(X,Y,Z)$$

for some $a^F, b^F \in \Lambda^F$.

But this immediately implies that $\mathcal{K}\{d(X,Y,Z)\}$ equals

$$\begin{aligned}\mathcal{K}\{S_\beta^F(X,Y,Z)+b^F(X,Y,Z)\} &= S_\beta(W,Y,Z)+b(W,Y,Z)\\ &= S_{\text{eff}}(W,Y,Z)+a(W,Y,Z),\end{aligned}$$

where $a, b \in \Lambda$. The first equality above holds because $S_\beta = \mathcal{K}(S_\beta^F)$ and $\Lambda = \mathcal{K}(\Lambda^F)$, and the second holds by the definition of the efficient score vector. Note that we have so far shown that $\mathcal{K}\{d(X,Y,Z)\}$ equals $S_{\text{eff}}(W,Y,Z) + a(W,Y,Z)$. Our argument will be complete once we show $a(W,Y,Z) \in \Lambda$ is exactly zero.

To this end, recall that d satisfying (2.2) means $\Pi\{\mathcal{K}^*\circ\mathcal{K}(d)|\Lambda^F\} = 0$, and so $\mathcal{K}^*\circ\mathcal{K}(d)$ is an element of $\Lambda^{\perp F}$. More exactly, the implication yields

$$\mathcal{K}^* \circ \mathcal{K}\{d(X,Y,Z)\} = \mathcal{K}^*\{S_{\text{eff}}(W,Y,Z)\} + \mathcal{K}^*\{a(W,Y,Z)\} \in \Lambda^{\perp F}.$$

The inner product of $\mathcal{K}^*\{a(W,Y,Z)\} \in \Lambda^{\perp F}$ and any $b^F(X,Y,Z) \in \Lambda^F$ must be zero, hence

$$0 =< \mathcal{K}^*\{a(W,Y,Z)\}, b^F(X,Y,Z) >=< a(W,Y,Z), \mathcal{K}\{b^F(X,Y,Z)\} >,$$

where the latter equality holds from the conjugacy of $\mathcal{K}$ and $\mathcal{K}^*$. But $\mathcal{K}\{b^F(X,Y,Z)\} \in \Lambda$, so because the inner product of $a(W,Y,Z)$ and $\mathcal{K}\{b^F(X,Y,Z)\}$ is zero, we must have $a(W,Y,Z) \in \Lambda^\perp$. Consequently, $a(W,Y,Z) \in \Lambda^\perp \cap \Lambda$, which holds only when $a(W,Y,Z) = 0$. Therefore, d satisfying (2.2) requires $\mathcal{K}\{d(X,Y,Z)\} = S_{\text{eff}}(W,Y,Z)$.

Proof of consistency even with misspecified η_1, η_2

Consistency follows upon showing that the score vector $S(W,Y,Z;\beta,\eta_1,\eta_2)$ from Step 3 in our algorithm is an element of $\Lambda^\perp$. For any, perhaps misspecified, η_1 and η_2, our algorithm implies that the score vector satisfies $S(W,Y,Z;\beta,\eta_1,\eta_2) = \mathcal{K}(d)$ where $d(X,Y,Z)$ satisfies (2.2). Immediately, then, $E\{S(W,Y,Z;\beta,\eta_1,\eta_2)|X,Y,Z\} = \mathcal{K}^* \circ \mathcal{K}(d)$. A simple rearrangement of (2.2) shows $E\{S(W,Y,Z;\beta,\eta_1,\eta_2)|X,Y,Z\} = \mathcal{K}^* \circ \mathcal{K}(d)$ which

equals $(m'_\beta(X,Z;\beta) - E[\{d - \mathcal{K}^* \circ \mathcal{K}(d)\}\epsilon|X,Z])E(\epsilon|X,Z)^{-1}\epsilon$. Hence we have that $E\{S(W,Y,Z;\beta,\eta_1,\eta_2)|X,Y,Z\} = g(X,Z)\epsilon$ with $g(X,Z) = m'_\beta(X,Z;\beta) - E[\{d - \mathcal{K}^* \circ \mathcal{K}(d)\}\epsilon|X,Z])E(\epsilon|X,Z)^{-1}$ and so $S(W,Y,Z;\beta,\eta_1,\eta_2) \in \Lambda^\perp$. Given that $E(\epsilon|X,Z) = 0$, we have that

$$E\{S(W,Y,Z;\beta,\eta_1,\eta_2)\} = E\{g(X,Z)\epsilon\} = E\{g(X,Z)E(\epsilon|X,Z)\} = 0.$$

Together, the above results imply that the estimator from $\sum_{i=1}^n S(W_i,Y_i,Z_i;\beta,\eta_1,\eta_2) = 0$ is consistent, even for misspecified η_1 and η_2.

Outline of the proof of Theorem 2

Given the form of the nuisance tangent space orthogonal complement $\Lambda^\perp$, it is obvious that $f \in \Lambda^\perp$. Hence, its normalized correspondence is an influence function. Under regularity conditions specified in Newey (1990), the $\widehat{\beta}$ that solves

$$\textstyle\sum_{i=1}^n f(W_i,Y_i,Z_i;\beta) = 0$$

is an RAL estimator and satisfies

$$n^{1/2}(\widehat{\beta} - \beta_0) = n^{-1/2}\textstyle\sum_{i=1}^n E\{f(W,Y,Z)S_\beta^T(W,Y,Z)\}^{-1}f(W,Y,Z;\beta_0) + o_p(1).$$

Therefore, $n^{1/2}(\widehat{\beta} - \beta_0) \to N[0, E(fS_\beta^T)^{-1}\text{var}(f)\{E(fS_\beta^T)^{-1}\}^T]$. Obviously, $E(fS_\beta^T) = A$, hence the results hold.

Proof of Theorem 3

To see that the resulting estimators from solving $\sum_{i=1}^n f(W_i,Y_i,Z_i;\beta,\hat{\gamma}) = 0$ and from solving $\sum_{i=1}^n f(W_i,Y_i,Z_i;\beta,\gamma_0) = 0$ have the same asymptotic efficiency, we show that

their first order expansions are the same. To this end, we analyze $\hat{\beta}$ first.

$$\begin{aligned}
0 &= n^{-1/2}\textstyle\sum_{i=1}^{n} f(W_i, Y_i, Z_i; \widehat{\beta}, \widehat{\gamma}) \\
&= n^{-1/2}\textstyle\sum_{i=1}^{n} f(W_i, Y_i, Z_i; \beta_0, \gamma_0) + n^{-1/2}\textstyle\sum_{i=1}^{n} \frac{\partial f(W_i, Y_i, Z_i; \beta^*, \gamma^*)}{\partial \beta^\top}(\widehat{\beta} - \beta_0) \\
&\quad + n^{-1/2}\textstyle\sum_{i=1}^{n} \frac{\partial f(W_i, Y_i, Z_i; \beta^*, \gamma^*)}{\partial \gamma^\top}(\widehat{\gamma} - \gamma_0) \\
&= n^{-1/2}\textstyle\sum_{i=1}^{n} f(W_i, Y_i, Z_i; \beta_0, \gamma_0) \\
&\quad + n^{1/2}\left[E\left\{\frac{\partial f(W_1, Y_1, Z_1; \beta_0, \gamma_0)}{\partial \beta^\top}\right\} + o_p(1)\right](\widehat{\beta} - \beta_0) \\
&\quad + n^{1/2}\left[E\left\{\frac{\partial f(W_1, Y_1, Z_1; \beta_0, \gamma_0)}{\partial \gamma^\top}\right\} + o_p(1)\right](\widehat{\gamma} - \gamma_0), \qquad \text{(A.4)}
\end{aligned}$$

where β^* lies on the line connecting $\hat{\beta}$ and β_0 and γ^* on the line connecting $\hat{\gamma}$ and γ_0.

We now demonstrate that the last term, (A.4), is zero and thus the result holds. Our construction ensures that $f(W, Y, Z; \beta_0, \gamma) \in \Lambda^\perp$ under all γ, so

$$\int f(W, Y, Z; \beta_0, \gamma) p_{W,Y,Z}(W, Y, Z; \beta_0, \gamma) d\mu(W, Y, Z) = 0$$

for all γ. Taking derivative with respect to γ, we obtain

$$\begin{aligned}
0 &= \int \frac{\partial f(W, Y, Z; \beta_0, \gamma)}{\partial \gamma^\top} p_{W,Y,Z}(W, Y, Z; \beta_0, \gamma) d\mu(W, Y, Z) \\
&\quad + \int f(W, Y, Z; \beta_0, \gamma) S_\gamma^\top(W, Y, Z; \beta_0, \gamma) p_{W,Y,Z}(W, Y, Z; \beta_0, \gamma) d\mu(W, Y, Z),
\end{aligned}$$

for all γ, where S_γ is the score vector with respect to γ. Evaluating the above equation at γ_0 gives

$$E\left\{\frac{\partial f(W, Y, Z; \beta_0, \gamma_0)}{\partial \gamma^\top}\right\} + E\left\{f(W, Y, Z; \beta_0, \gamma_0) S_\gamma^\top(W, Y, Z; \beta_0, \gamma_0)\right\} = 0.$$

However, by construction $f \in \Lambda^\perp$ while $S_\gamma \in \Lambda$, hence $E(f S_\gamma^\top) = 0$. This implies that $E\left\{\frac{\partial f(W, Y, Z; \beta_0, \gamma_0)}{\partial \gamma^\top}\right\} = 0$. Equation (A.4) therefore reduces to

$$0 = n^{-1/2}\textstyle\sum_{i=1}^{n} f(W_i, Y_i, Z_i; \beta_0, \gamma_0) + n^{1/2} E\left\{\frac{\partial f(W_1, Y_1, Z_1; \beta_0, \gamma_0)}{\partial \beta^\top}\right\}(\widehat{\beta} - \beta_0) + o_p(1),$$

which is exactly the same first order expansion for $\widetilde{\beta}$.

Estimator properties under unknown α

Let $\beta^*(\hat{\alpha}) = \lim_{n\to\infty} \hat{\beta}_n(\hat{\alpha})$. Then, given $\hat{\alpha}$, $n^{1/2}\{\hat{\beta}(\hat{\alpha}) - \beta^*(\hat{\alpha})\}$ achieves the asymptotic normality results of Theorem 2. Based on this result, we now investigate the properties of $n^{1/2}\{\hat{\beta}(\hat{\alpha}) - \beta_0\}$.

With $n^{1/2}(\hat{\beta} - \beta_0) = n^{1/2}(\hat{\beta} - \beta^*) + n^{1/2}(\beta^* - \beta_0)$, we have that $E\{n^{1/2}(\hat{\beta} - \beta_0)\} = E[E\{n^{1/2}(\hat{\beta}-\beta^*)|\hat{\alpha}\}]+E\{n^{1/2}(\beta^*-\beta_0)\} = E\{n^{1/2}\beta_0'(\alpha)(\hat{\alpha}-\alpha_0)\} = 0$, where the second equality follows from the results of Theorem 2 and using Taylor's expansion of $\beta^*(\hat{\alpha})$ about α_0, and the last equality follows from the consistency of $\hat{\alpha}$. Similarly, var$\{n^{1/2}(\hat{\beta} - \beta_0)\}$ equals

$$
\begin{aligned}
&E[\text{var}\{n^{1/2}(\hat{\beta} - \beta_0)|\hat{\alpha}\}] + \text{var}[E\{n^{1/2}(\hat{\beta} - \beta_0)|\hat{\alpha}\}] = E[\text{var}\{n^{1/2}(\hat{\beta} - \beta^*)|\hat{\alpha}\}] \\
&+\text{var}\{n^{1/2}(\beta^* - \beta_0)\} = E\{V_\beta(\hat{\alpha})\} + \text{var}\{n^{1/2}\beta_0'(\alpha)(\hat{\alpha} - \alpha_0)\} \\
&\approx V_\beta(\alpha_0) + \beta_0'(\alpha_0)^2 V_\alpha(\alpha_0),
\end{aligned}
$$

where $V_\beta(\alpha_0)$ denotes the estimation variance under the known α given in Theorem 2 evaluated at α_0, and $V_\alpha(\alpha_0) = \text{var}\{n^{1/2}(\hat{\alpha} - \alpha_0)\}$.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER II

Definition of ζ

For $\zeta_i = (\zeta_{1i}, \zeta_{2i})'$,

$$\zeta_{ji}(t;\beta) = g_{ji}(t;\beta) - \frac{\exp(-\beta_1 Z_i) + \exp(-\beta_2 Z_i)\hat{R}(t;\beta)}{\sum_{i=1}^{n} I(Y_i \geq t)} \sum_{k=1}^{n} G_{jk}(t)$$
$$+ \frac{\{\exp(-\beta_1 Z_i) + \exp(-\beta_2 Z_i)\hat{R}(t;\beta)\} \prod_{s \leq t^-} \{1 - \Delta\hat{\Psi}(s;\beta_2)\}}{\sum_{i=1}^{n} I(Y_i \geq t)}$$
$$\times \int_t^\tau \left\{ \sum_{k=1}^{n} G_{jk}(s) V_k(s) - \frac{\sum_{k=1}^{n} G_{jk}(s) \sum_{k=1}^{n} V_k(s)}{\sum_{i=1}^{n} I(Y_i \geq s)} \right\} \frac{d\hat{R}(s;\beta)}{\prod_{u \leq s}\{1 - \Delta\hat{\Psi}(s;\beta_2)\}}$$

with

$$G_{jk}(t) = \frac{g_{jk}(t;\beta) I(Y_k \geq t)}{\exp(-\beta_1 Z_k) + \exp(-\beta_2 Z_k)\hat{R}(t;\beta)},$$
$$V_k(t) = \frac{\exp(-\beta_2 Z_k) I(Y_k \geq t)}{\exp(-\beta_1 Z_k) + \exp(-\beta_2 Z_k)\hat{R}(t;\beta)}.$$

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER III

Inconsistency of the type II NPMLE

From section 4.2.2, it is easy to see that the type II optimization should only assign positive weights to the observed event times, and the estimating procedure proceeds to form

$$\check{F}(t) = \sum_{i=1}^{n} \delta_i I(x_i < t) \check{f}(x_i).$$

If a set of $f(x_i)$'s is the maximizer of (4.2) under the type II NPMLE constraints, then $q_i^T f(x_i)$'s are all non-negative and satisfy $0 \leq q_i^T F(x_i) \leq 1$. Thus, if we denote by $H(x_i; q_1, \ldots, q_n) = q_i^T F(x_i)$ and $h(x_i; q_1, \ldots, q_n) = q_i^T f(x_i)$ for all $i = 1, \ldots, n$, then the maximization is obtained at the MLE estimator of the hypothetical H. Note that H depends on q_i's which take m different values, and we view these m values as known parameters. For notational simplicity, we write $H(x; q_1, \ldots, q_n)$ as $H(x)$. Thus, if one thinks of $u_j^T F(x)$ as a particular mixture of $F(x)$, then H contains all m mixtures. In addition, H is not necessarily a valid distribution function since it may not be monotone. Denote the maximum of H as $\check{H}$. One can then recover the $F(x_i)$'s through $\check{F}(x_i) = (q_i^T r_i)^{-1} r_i \check{H}_i$ for some length p vector r_i (i.e., $r_i = F(x_i)$). If a suitable selection of r_i's exists to ensure $0 \leq (q_i^T r_i)^{-1} r_i H_i \leq 1$ and monotonicity, then we have a solution of the type II optimization problem.

Assuming to the contrary this solution is consistent, then the resulting $\check{F}(t)$ would satisfy

$$u_j^T \check{F}(t) = 1 - \check{S}_j(t)$$

for all $j = 1, \ldots, m$, where $\check{S}_j(t)$ is a consistent estimator of $S_j(t)$. Here $S_j(t)$ has the

same definition as in the type I NPMLE and is the survival function of the observations that have the common mixing proportion, u_j. Obviously, $\check{F}(t) = \check{F}(x_i)$, where x_i is the largest x value that satisfies $x_i \leq t$ and $\delta_i = 1$. Without confusion, we assume the corresponding $q_i = u_j$, thus we have

$$u_j^T \check{F}(t) = \check{H}_i = 1 - \check{S}_j(t).$$

Because $\check{S}_j$ is a consistent estimator of S_j, we must have $\check{H}_i$ converges to $1 - S_j(t)$. However, this leads to a contradiction, since $1 - S_j(t) = u_j^T F(t)$ is the distribution function corresponding to the jth mixing vector u_j, while $\check{H}(t)$ aims at estimating the hypothetic H function which contains all m different mixtures.

It is not always possible or easy to find the $\check{H}_i$'s or to identify the r_i's. For this reason, the type II NPMLE is hardly ever solved through obtaining $\check{H}_i$'s and r_i's. Instead, the EM algorithm introduced earlier is used to obtain the $f(x_i)$'s. However, the above explanation reveals the underlying reason why the type II NPMLE fails.

Conceptually, identifying the $F_1, \ldots F_p$ functions is equivalent to identifying the functions $S_1, \ldots, S_m$. The type II NPMLE maximizes the product of m different likelihoods formed by all observations with respect to a collection of parameters, but each of these parameters should concern only one of these likelihoods formed by a subset of the observations. To help further understand the mistreatment of different conditional likelihoods of the type II NPMLE, one may consider maximizing a marginal likelihood. Denoting $n_j = \sum_{i=1}^n I(q_i = u_j)$, $j = 1, \ldots, m$, this would correspond to maximizing

$$\sum_{i=1}^{n} \log \sum_{j=1}^{m} n_j \left\{u_j^T f(x_i)\right\}^{\delta_i} \left\{1 - u_j^T F(x_i)\right\}^{1-\delta_i}$$

with respect to $f(x_i)$'s. This estimator uses all observed x_i's to form a common marginal likelihood, hence is consistent. However it completely ignores the pair information provided in the data hence could be highly inefficient.

EM algorithm for Cox proportional hazards model

The complete data log likelihood of the observations $o_i = (G_i = g_i, X_i = x_i, \Delta_i = \delta_i)$ is

$$\mathcal{L}^{\text{comp}}_{\text{COX}}\{o_1, \ldots, o_n; \beta, \Lambda_0(\cdot)\} = \sum_{i=1}^{n}\sum_{k=1}^{p} I(g_i = k)\text{log}(\lambda^{\delta_i}(x_i|k)\exp[-\Lambda_0(x_i)\exp\{\beta^T z(k)\}]).$$

At the lth iteration of the EM algorithm, with current values for $\hat{\beta}_{\text{PH}}$ and $\hat{\Lambda}_0(\cdot)$ denoted by $\hat{\beta}^{(l)}_{\text{PH}}$ and $\hat{\Lambda}^{(l)}_0(\cdot)$, respectively, the E-step is

$$\begin{aligned} & E[\mathcal{L}^{\text{comp}}_{\text{COX}}\{o_1, \ldots, o_n; \beta, \Lambda_0(\cdot)\}|x_i, \delta_i, i = 1, \ldots, n] \\ = & \sum_{i=1}^{n}\sum_{k=1}^{p} p^{(l)}_{ik}(x_i, \delta_i)[\delta_i\{\text{log}\lambda_0(x_i) + \beta^T z_i(k)\} - \Lambda_0(x_i)\exp\{\beta^T z_i(k)\}], \end{aligned}$$

where

$$p^{(l)}_{ik}(x_i, \delta_i) = \frac{q_{ik}\exp[\delta_i\{\beta^{(l)}\}^T z_i(k)]\exp(-\Lambda^{(l)}_0(x_i)\exp[\{\beta^{(l)}\}^T z_i(k)])}{\sum_{k=1}^{p} q_{ik}\exp[\delta_i\{\beta^{(l)}\}^T z_i(k)]\exp(-\Lambda^{(l)}_0(x_i)\exp[\{\beta^{(l)}\}^T z_i(k)])}$$

denotes the conditional distribution of the missing genotype $G_i = k$ given (x_i, δ_i). The M-step corresponds to using a Nelson Aalen type estimator for $\Lambda_0(t)$ and maximizing the log likelihood in the E-step with respect to β. Doing so leads to a new estimate for $\hat{\beta}_{\text{PH}}$, denoted $\hat{\beta}^{(l+1)}_{\text{PH}}$, as the root of the estimating equation

$$\sum_{i=1}^{n}\delta_i\left[\sum_{k=1}^{p} p^{(l)}_{ik}(x_i, \delta_i)z_i(k) - \frac{\sum_{j=1}^{n} I(x_j \geq x_i)\sum_{k=1}^{p} p^{(l)}_{jk}(x_j, \delta_j)\exp\{\beta^T z_j(k)\}z_j(k)}{\sum_{j=1}^{n} I(x_j \geq x_i)\sum_{k=1}^{p} p^{(l)}_{jk}(x_j, \delta_j)\exp\{\beta^T z_j(k)\}}\right].$$

Lastly, the new estimated baseline hazard function is

$$\hat{\Lambda}^{(l+1)}_0(t) = \sum_{i=1}^{n}\frac{\delta_i I(x_i \leq t)}{\sum_j I(x_j \geq x_i)\sum_{k=1}^{p} p^{(l)}_{jk}(x_j, \delta_j)\exp[\{\hat{\beta}^{(l+1)}_{\text{PH}}\}^T z_j(k)]}.$$

These E- and M-steps are repeated until both $\hat{\beta}_{\text{PH}}$ and $\hat{\Lambda}_0(t)$ converge, and final estimates are used to form the Cox-based estimator $\hat{F}_{\text{PH}}(t)$.

Influence function of the IPW estimator

To derive the influence function of the IPW estimator, we first note the following several useful facts (Bang and Tsiatis, 2000; Robins and Rotnitzky, 1992)

$$\begin{aligned} Y(t) &= n\widehat{G}(t^-)\widehat{S}(t^-) \\ \frac{\widehat{G}(t)-G(t)}{G(t)} &= -\int_0^t \frac{\widehat{G}(u^-)}{G(u)}\frac{dM^c(u)}{Y(u)} \\ \frac{\delta_i}{G(x_i)} &= 1-\int \frac{dM_i^c(u)}{G(u)}. \end{aligned}$$

Using the above relations, we expand the IPW estimator as

$$\begin{aligned} n^{-1/2}\sum_{i=1}^n \frac{\delta_i\phi(q_i,x_i)}{\widehat{G}(x_i)} &= n^{-1/2}\sum_{i=1}^n \frac{\delta_i\phi(q_i,t_i)}{G(x_i)} + n^{-1/2}\sum_{i=1}^n \frac{\delta_i\phi(q_i,t_i)}{\widehat{G}(x_i)}\int_0^{x_i}\frac{\widehat{G}(u^-)}{G(u)}\frac{dM^c(u)}{Y(u)} \\ &= n^{-1/2}\sum_{i=1}^n \frac{\delta_i\phi(q_i,t_i)}{G(x_i)} + n^{-1/2}\sum_{i=1}^n \int \frac{\widehat{\mathcal{B}}_1\{\phi,u\}\widehat{S}(u)dM_i^c(u)}{G(u)\widehat{S}(u^-)} \\ &= n^{-1/2}\sum_{i=1}^n \phi(q_i,t_i) - n^{-1/2}\sum_{i=1}^n \int \frac{\{\phi(q_i,t_i)-\mathcal{B}(\phi,u)\}\,dM_i^c(u)}{G(u)} \\ &\quad +o_p(1). \end{aligned}$$

Because the complete data influence function $\phi(q_i,t_i)$ has the general form of $d(q_i,t_i)-F(t)$ (Ma and Wang, 2011), we have that $\partial\phi/\partial F^T(t) = -I_p$. This, in combination with exchanging integration and differentiation of the above expansion, implies that the ith influence function for the IPW is ϕ_{ipw} as stated. The two terms in ϕ_{ipw} are uncorrelated because $\phi(q_i,t_i)$ are $\mathcal{F}(0)$ measurable. Therefore we can compute the variance of the IPW

estimator as

$$
\begin{aligned}
V_{\text{ipw}} &= \text{cov}\{\phi(Q_i, T_i)\} + E\left[\int \frac{\{\phi(Q_i, T_i) - \mathcal{B}(\phi, u)\}^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right] \\
&= \text{cov}\{\phi(Q_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\phi, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\} \\
&+ E\left[\int \frac{E\left[\left\{\phi(Q_i, T_i)^{\otimes 2} - 2\phi(Q_i, T_i)\mathcal{B}(\phi, u)^T\right\} I(T_i \geq u) | C_i\right] I(C_i \geq u)}{G^2(u)} \lambda^c(u) du\right] \\
&= \text{cov}\{\phi(Q_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\phi^{\otimes 2}, u) - \mathcal{B}(\phi, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\}.
\end{aligned}
$$

Influence function of the AIPW estimator

From (4.5), we have

$$
\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i \phi(q_i, x_i)}{\widehat{G}(x_i)} + n^{-1/2} \sum_{i=1}^{n} \int \frac{dN_i^c(u)}{\widehat{G}(u)} \left\{\widehat{h}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{h}_{\text{eff}}, u\right)\right\} \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i \phi(q_i, x_i)}{\widehat{G}(x_i)} + n^{-1/2} \sum_{i=1}^{n} \int \frac{dM_i^c(u)}{\widehat{G}(u)} \left\{\widehat{h}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{h}_{\text{eff}}, u\right)\right\} \\
&= n^{-1/2} \sum_{i=1}^{n} \phi(q_i, t_i) - n^{-1/2} \sum_{i=1}^{n} \int \frac{\left\{\phi(q_i, t_i) - h_{\text{eff},i}(u)\right\} dM_i^c(u)}{G(u)} + o_p(1),
\end{aligned}
$$

where the last equality follows from $\mathcal{B}(\phi - h_{\text{eff}}, u) = 0$. Similar to the IPW case, the two terms in the influence function are uncorrelated which suggests that we can compute the variance of the efficient estimator as

$$
\begin{aligned}
V_{\text{eff}} &= \text{cov}\{\phi(Q_i, T_i)\} + E\left[\int \frac{\left\{\phi(Q_i, T_i) - h_{\text{eff},i}(u)\right\}^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right] \\
&= \text{cov}\{\phi(Q_i, T_i)\} + E \int \frac{\mathcal{B}\{(\phi - h_{\text{eff}})^{\otimes 2}, u\}}{G^2(u)} \lambda^c(u) Y_i(u) du.
\end{aligned}
$$

Influence function of the imputation estimator

We now analyze the asymptotic properties of the imputation estimator.

$$\begin{aligned} 0 &= n^{-1/2}\sum_{i=1}^{n}\left\{\delta_i\phi(q_i,x_i)+(1-\delta_i)\widehat{h}_{\text{eff},i}(x_i)\right\} \\ &= n^{-1/2}\sum_{i=1}^{n}\left\{\delta_i\phi(q_i,x_i)+(1-\delta_i)h_{\text{eff},i}(x_i)\right\} \\ &\quad +n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{h}_{\text{eff},i}(x_i)-h_{\text{eff},i}(x_i)\right\}. \end{aligned}$$

We now inspect the last term. In our approximation in (4.4), $\widehat{h}_{\text{eff},i}(u)$ is estimated using weighted sample averages, with the subset of data that have a common q_i value. We now analyze $\widehat{h}_{\text{eff},i}(u)$ in the kth subsample. For notational simplicity, we assume the first n_k observations have the common q value. We have

$$\begin{aligned} &\widehat{h}_{\text{eff},i}(x_i)-h_{\text{eff},i}(x_i) \\ = &\frac{n_k^{-1}\sum_{j=1}^{n_k}\phi(q_j,x_j)I(x_j\geq x_i)\delta_j/\widehat{G}(x_j)}{n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geq x_i)\delta_j/\widehat{G}(x_j)}-\frac{E\{\phi(q_i,T)I(T>x_i)|q_i,x_i\}}{E\{I(T>x_i)|q_i,x_i\}} \\ = &\frac{n_k^{-1}\sum_{j=1}^{n_k}\phi(q_j,x_j)I(x_j\geq x_i)\delta_j/\widehat{G}(x_j)-E\{\phi(q_i,T)I(T>x_i)|q_i,x_i\}}{E\{I(T>x_i)|q_i,x_i\}} \\ &-h_{\text{eff},i}(x_i)\frac{n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geq x_i)\delta_j/\widehat{G}(x_j)-E\{I(T>x_i)|q_i,x_i\}}{E\{I(T>x_i)|q_i,x_i\}}+o_p(n_k^{-1/2}). \end{aligned}$$

Using derivations similar to the IPW analysis, we first get some basic facts. For any function $f(q_i,x_i)$, we have

$$\begin{aligned} \sum_{j=1}^{n_k}\frac{\delta_j f(q_j,x_j)}{\widehat{G}(x_j)} &= \sum_{j=1}^{n_k}\frac{\delta_j f(q_j,t_j)}{G(x_j)}+n^{-1}\sum_{j=1}^{n_k}\frac{\delta_j f(q_j,x_j)}{\widehat{G}(x_j)}\int\frac{Y_j(u)}{G(u)}\frac{dM^c(u)}{\widehat{S}(u^-)} \\ &= \sum_{j=1}^{n_k}\frac{\delta_j f(q_j,t_j)}{G(x_j)}+\frac{n_k}{n}\int\frac{E\{f(\tilde{q}_k,T)I(T\geq u)|\tilde{q}_k\}dM^c(u)}{G(u)S(u)}+o_p(n_k^{1/2}). \end{aligned}$$

Here we use $\tilde{q}_k$ to represent the common q_i value in the kth group. Using the above result, we have

$$
\begin{aligned}
& n_k^{-1}\sum_{j=1}^{n_k}\phi(q_j,x_j)I(x_j\geq x_i)\delta_j/\widehat{G}(x_j) - E\{\phi(q_i,T)I(T>x_i)|q_i,x_i\} \\
= \; & n_k^{-1}\sum_{j=1}^{n_k}\frac{\delta_j\phi(q_j,t_j)I(t_j\geq x_i)}{G(x_j)} - E\{\phi(q_i,T)I(T>x_i)|q_i,x_i\} \\
& +\frac{1}{n}\int\frac{E\{\phi(q_i,T)I(T\geq x_i)I(T\geq u)|q_i,x_i\}dM^c(u)}{G(u)S(u)} + o_p(n_k^{-1/2}), \\
& n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geq x_i)\delta_j/\widehat{G}(x_j) - E\{I(T>x_i)|q_i,x_i\} \\
= \; & n_k^{-1}\sum_{j=1}^{n_k}\frac{\delta_j I(t_j\geq x_i)}{G(x_j)} - E\{I(T>x_i)|q_i,x_i\} \\
& +\frac{1}{n}\int\frac{E\{I(T\geq x_i)I(T\geq u)|q_i,x_i\}dM^c(u)}{G(u)S(u)} + o_p(n_k^{-1/2}).
\end{aligned}
$$

Inserting these forms, we have

$$
\begin{aligned}
& \widehat{h}_{\text{eff},i}(x_i) - h_{\text{eff},i}(x_i) \\
= \; & \frac{n_k^{-1}\sum_{j=1}^{n_k}\delta_j\phi(q_j,t_j)I(t_j\geq x_i)/G(x_j)}{E\{I(T>x_i)|q_i,x_i\}} - \frac{n_k^{-1}\sum_{j=1}^{n_k}h_{\text{eff},i}(x_i)\delta_j I(t_j\geq x_i)/G(x_j)}{E\{I(T>x_i)|q_i,x_i\}} \\
& +\frac{1}{E\{I(T>x_i)|q_i,x_i\}}\frac{1}{n}\int\frac{E\{\phi(q_i,T)I(T\geq x_i)I(T\geq u)|q_i,x_i\}dM^c(u)}{G(u)S(u)} \\
& -\frac{h_{\text{eff},i}(x_i)}{E\{I(T>x_i)|q_i,x_i\}}\frac{1}{n}\int\frac{E\{I(T\geq x_i)I(T\geq u)|q_i,x_i\}dM^c(u)}{G(u)S(u)} + o_p(n_k^{-1/2}).
\end{aligned}
$$

Summing up the n_k such terms in the kth group, exchanging the summation on i and j, and writing $a(q_i,t_i) = E\{h_{\text{eff},i}(C)I(C\leq t_i)|q_i\}$, we obtain

$$
\begin{aligned}
& \sum_{i=1}^{n_k}(1-\delta_i)\left\{\widehat{h}_{\text{eff},i}(x_i) - h_{\text{eff},i}(x_i)\right\} \\
& = \sum_{i=1}^{n_k}\frac{\delta_i\phi(q_i,t_i)}{G(x_i)}\{1-G(t_i)\} - \sum_{i=1}^{n_k}\frac{\delta_i}{G(x_i)}a(q_i,t_i) \\
& +\frac{n_k}{n}\int E[\phi(q_i,T)\{1-G(T)\}I(T\geq u)|\tilde{q}_k]\frac{dM^c(u)}{G(u)S(u)} \\
& -\frac{n_k}{n}\int E\left\{a(\tilde{q}_k,T)I(T\geq u)|\tilde{q}_k\right\}\frac{dM^c(u)}{G(u)S(u)} + o_p(n_k^{1/2}).
\end{aligned}
$$

In the above derivation, we used the fact that the censoring survival function in the group is the same as the global survival function $G(t)$. Now summing up all the m groups, we have

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{h}_{\text{eff},i}(x_i)-h_{\text{eff},i}(x_i)\right\} \\
= \;& n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i\phi(q_i,t_i)}{G(x_i)}-n^{-1/2}\sum_{i=1}^{n}\delta_i\phi(q_i,t_i)-n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i}{G(x_i)}a(q_i,t_i) \\
& +n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left(\phi,u\right)}{G(u)}dM_i^c(u)-n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{\phi(q,t)G(t),u\right\}}{G(u)}dM_i^c(u) \\
& -n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{a(q,t),u\right\}}{G(u)}dM_i^c(u)+o_p(1).
\end{aligned}$$

Thus, we have obtained

$$\begin{aligned}
0 \;=\; & n^{-1/2}\sum_{i=1}^{n}\left\{\phi(q_i,t_i)-a(q_i,t_i)\right\} \\
& -n^{-1/2}\sum_{i=1}^{n}\int\frac{\left\{\phi(q_i,t_i)-a(q_i,t_i)-\mathcal{B}(\phi-a,u)\right\}dM_i^c(u)}{G(u)} \\
& +n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)h_{\text{eff},i}(x_i)-n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{\phi(q,t)G(t),u\right\}}{G(u)}dM_i^c(u)+o_p(1).
\end{aligned}$$

Using similar arguments as in the IPW and AIPW cases,

$$\begin{aligned}
& \left\{\phi(q_i,t_i)-a(q_i,t_i)\right\}-\int\frac{\left\{\phi(q_i,t_i)-a(q_i,t_i)-\mathcal{B}(\phi-a,u)\right\}dM_i^c(u)}{G(u)} \\
& +(1-\delta_i)h_{\text{eff},i}(x_i)-\int\frac{\mathcal{B}\left\{\phi(q,t)G(t),u\right\}}{G(u)}dM_i^c(u)
\end{aligned}$$

is the ith influence function of the imputation estimator.

VITA

Tanya Pamela Garcia was born in Lima, Peru. In August of 2000, she entered the undergraduate program in Mathematics at the University of California, Irvine and graduated summa cum laude with a Bachelor of Science degree in June 2003. She continued on to obtain a Master of Science degree in industrial engineering and operations research from the University of California, Berkeley in December 2004, and a second Master of Science degree in statistics from the University of Western Ontario in London, Canada in August 2006. In August 2008, she began the doctoral program in statistics at Texas A&M University and in August 2011, she graduated with her Ph.D. under the advisement of Dr. Yanyuan Ma. Her mailing address is:

Department of Statistics

Texas A&M University

College Station, TX 77845-3143.