THE SINGLE-INDEX HAZARDS MODEL

Kai Ding

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

Chapel Hill
2010

Approved by:

Dr. Michael Kosorok, Advisor

Dr. Donglin Zeng, Advisor

Dr. Amy Herring, Reader

Dr. Jason Fine, Reader

Dr. David Richardson, Reader

Abstract

KAI DING: THE SINGLE-INDEX HAZARDS MODEL.
(Under the direction of Dr. Michael Kosorok and Dr. Donglin Zeng.)

We first propose the single-index hazards model for right censored survival data. As an extension of the Cox model, this model allows nonparametric modeling of covariate effects in a parsimonious way via a single-index. In addition, the relative importance of covariates can be assessed via this model. We consider the conventional profile-kernel method based on the local likelihood for model estimation. It is shown that this method may give consistent estimation under certain restrictive conditions, but in general it can yield biased estimation. Simulation studies are conducted to demonstrate the bias phenomena. The existence and nature of the failure of this commonly used approach is somewhat surprising.

The interpretation of covariate effects in the aforementioned single-index hazards model is difficult. Thus, we further propose the partly proportional single-index hazards model in which the effect of covariates of primary interest is represented by the regression parameter while "nuisance" covariates can have nonparametric effect on the survival time. We again consider the conventional profile-kernel method and it leads to biased estimation as well. A bias correction method is then proposed and the corrected profile local likelihood estimators are shown to be consistent, asymptotically normal and semiparametrically efficient. We evaluate the finite-sample properties of our estimators through simulation studies and illustrate the proposed model and method with an application to a dataset from the Multicenter AIDS Cohort Study (MACS).

Besides the profile-kernel method, we also study the profile stratified likelihood method based on stratification of the single-index. In the single-index hazards model, this method

may give consistent estimation under the restrictive "independent censoring" condition, but in general it can yield biased estimation. Simulation studies are conducted to demonstrate the situations in which the bias phenomena do (or do not) exist; In the partly proportional single-index hazards model, we demonstrate numerically the existence of the bias and then propose a bias correction method. The estimators from the corrected profile stratified likelihood method are shown to be consistent. Their finite-sample properties are evaluated through simulation studies. The corrected profile stratified method is applied to the aforementioned MACS study for illustration.

Acknowledgments

This dissertation could not have been written without my advisors, Dr. Michael Kosorok and Dr. Donglin Zeng, who led me to this research field and patiently guided me through the dissertation process. I wish to express my deepest appreciation to them for their support, encouragement and mentoring.

I owe many thanks to the rest of my committee members. I am deeply grateful to Dr. Amy Herring for her insightful advice, more than 3 years of financial support and being my academic advisor. My thanks also goes to Dr. Jason Fine and Dr. David Richardson for their invaluable advice and kindness throughout this research.

I would also like to thank all of my friends, particularly Yi Gong, Qianchuan He, Yijuan Hu, Zhaowei Hua, Yiyun Tang, Yingqi Zhao and Yufan Zhao, whose friendship has made this journey more enjoyable and memorable.

Finally, it's impossible to have completed this journey without the love, support and encouragement from my wife, Lan Bi. This dissertation is dedicated to her.

Table of Contents

List of Figures viii

List of Tables ix

1 Introduction 1

1.1 Motivation and Literature Review 1

1.1.1 Semiparametric and Nonparametric Regression Models for Survival Data 1

1.1.2 Single-Index Models 6

1.1.3 Partially Linear Models for Survival Data 9

1.2 Outline of Dissertation 13

2 Single-Index Hazards Model 17

2.1 Model and Data Structure 17

2.2 Profile Local Likelihood 18

2.3 Bias Analysis 19

2.4 Simulation Studies 21

2.5 Proofs of Theorems 24

3 Partly Proportional Single-Index Hazards Model 29

3.1 Model and Data Structure 29

3.2 Profile Local Likelihood 30

3.2.1 Method . . . 30
3.2.2 Bias Analysis . . . 32
3.3 Corrected Profile Local Likelihood . . . 34
3.3.1 Method . . . 34
3.3.2 Asymptotic Results . . . 38
3.3.3 Simulation Studies . . . 39
3.4 Data Application . . . 42
3.5 Proofs of Theorems . . . 44

4 Profile Stratified Likelihood **61**
4.1 Single-Index Hazards Model . . . 61
4.1.1 Method . . . 61
4.1.2 Bias Analysis . . . 62
4.2 Partly Proportional Single-Index Hazards Model . . . 66
4.2.1 Method . . . 66
4.2.2 Bias Analysis . . . 67
4.2.3 Bias Correction . . . 71
4.2.4 Data Application . . . 72
4.3 Proofs of Theorems . . . 76

5 Discussion **82**

References **86**

List of Figures

2.1 Profile local likelihood curve of γ_1 in single-index hazards model 23

3.1 Profile local likelihood curve of γ_1 in PPSIH model 36

3.2 Profile local likelihood curves (corrected and uncorrected) of γ_1 in PPSIH model 41

3.3 Cumulative baseline hazard estimator $\hat{\Lambda}(t, u)$ under PPSIH model (2). . . 45

4.1 Profile stratified likelihood curve of γ_1 in single-index hazards model . . . 65

4.2 Profile stratified likelihood curve of γ_1 in PPSIH model 70

4.3 Profile stratified likelihood curves (corrected and uncorrected) of γ_1 in PPSIH model 74

List of Tables

2.1 Simulation results of local likelihood in single-index hazards model 22

3.1 Simulation results of local likelihood in PPSIH model 35

3.2 Simulation results of corrected local likelihood in PPSIH model 40

3.3 Analysis of MACS Data under PPSIH Model (1), (2) and Cox Model . . 44

4.1 Simulation results of stratified likelihood in single-index hazards model . 64

4.2 Simulation results of stratified likelihood in PPSIH model 69

4.3 Simulation results of corrected stratified likelihood in PPSIH model . . . 73

4.4 Analysis of MACS Data under PPSIH Model (3) 75

Chapter 1

Introduction

1.1 Motivation and Literature Review

1.1.1 Semiparametric and Nonparametric Regression Models for Survival Data

In survival analysis, investigators often wish to assess the effect of covariates on the risk of the event of interest. For example, in the Multicenter AIDS Cohort Study (MACS), one important research question is to evaluate the effect of patient's baseline age, ethnicity, CD4 positive cell counts, viral loads, serum β_2-microgloburin levels and serum neopterin levels on survival time (i.e. time to death due to AIDS) among HIV positive men. The four biomarkers (CD4 positive cell counts, viral loads, serum β_2-microgloburin levels and serum neopterin levels) were identified as the most predictive prognostic factors in Mellors et al. (1997). The Cox proportional hazards model (Cox 1972) is a popular and classical choice in such scenarios due to its nice interpretation of regression parameters and the availability of efficient inference procedures implemented in all statistical software packages. In this model, the conditional hazard rate of failure time given covariates, denoted by W, is modeled as $h(t|W) = \lambda(t)e^{\beta^T W}$, where $\lambda(\cdot)$ is a completely unknown baseline hazard function. The regression parameters, β, can be nicely interpreted as the

log-hazard ratios of the covariates W. Cox (1975) also proposed the partial likelihood to estimate the regression parameters. The by now classical large sample properties of the partial likelihood estimators were later proved in Andersen and Gill (1982). See also Fleming and Harrington (1991) and Andersen, Borgan, Gill, and Keiding (1993) for the literature concerning this model.

An underlying assumption of the Cox model is the so-called proportional hazards assumption, that is, the hazard ratio remains constant over time or covariates have log-linear effects on the risk of the event of interest. However, in many real datasets, covariates may exhibit much more complicated effects than log-linear effects; thus the proportional hazards assumption may be violated and the Cox model may not be an appropriate choice. For example, in the aforementioned MACS data, testing for the proportional hazards assumption based on martingale residuals (Lin, Wei, and Ying 1993) reveals that the covariate viral load (after taking logarithmic transformation) does not satisfy this assumption ($p = .006$). Thus the inference based on the Cox model may not be valid due to model misspecification.

For this reason, many authors have considered alternatives or extensions of the Cox proportional hazards model. For example, the accelerated failure time model (Cox and Oakes 1984, chap. 5) is attractive due to its direct physical interpretation. This model takes the form $\log T = -\beta^T W + \epsilon$, where T denotes the survival time, ϵ is independent of W and has an unspecified distribution. Note that by assuming this model, the covariates W have effects on the survival time and so the interpretation is direct. The rank estimator was studied by Prentice (1978) and the least-squares estimator was studied by Buckley and James (1979). Neither estimator achieves the semiparametric efficiency bound defined in Bickel, Klaassen, Ritov, and Wellner (1993). Recently, Zeng and Lin (2007a) provided a computationally tractable and semiparametrically efficient estimator for the regression parameter β using a kernel approximation of the profile likelihood. Moreover, their method can handle time-dependent covariates as well.

Alternatively, instead of assuming a constant hazard ratio over time as in the Cox model, the proportional odds model (Bennett 1983; Pettitt 1984) assumes the odds ratio of survival to be constant over time. Consequently, the ratio of the hazards converges to unity as time increases. The model takes the form $-\log\{S_{T|W}(t)/(1-S_{T|W}(t))\} = G(t)+\beta^T W$, where $S_{T|W}(\cdot)$ denotes the conditional survival function of T given covariates W and $G(t) = \log\{F(t)/(1-F(t))\}$. Here $F(t) = P(T \leq t)$ is the baseline distribution function of the survival time T. The maximum likelihood estimation for this model was studied by Murphy, Rossini, and van der Vaart (1997). The profile likelihood estimator for the regression parameter was shown to be consistent, asymptotically normal and semiparametrically efficient. They also provided the profile likelihood ratio test for the regression coefficient β.

The Cox proportional hazards model and the proportional odds model are special cases of the generalized odds-rate model considered in Scharfstein, Tsiatis, and Gilbert (1998). The odds-rate model takes the form $g_\rho(S_{T|W}(t)) = \alpha(t) + \beta^T W$, where $S_{T|W}(t)$ has the same meaning as in the proportional odds model, $g_\rho(x)$ equals $\log(\rho^{-1}(x^{-\rho}-1))$ when $\rho > 0$ and equals $\log(-\log(x))$ when $\rho = 0$ and $\alpha(\cdot)$ is some arbitrary increasing function. If $\rho = 0$, this model is equivalent to the Cox proportional hazards model and if $\rho = 1$, this model reduces to the proportional odds model. Scharfstein et al. (1998) showed that the nonparametric maximum likelihood estimator for β is semiparametrically efficient.

Another general model which includes the proportional hazards model and the proportional odds model as special cases is the proportional hazards frailty regression model studied in Kosorok, Lee, and Fine (2004). In this model, the conditional hazard takes the form $h(t|W,U) = \lambda(t)e^{\beta^T W+\log(U)}$, where U is a continuous frailty with mean 1 within a known one-parameter family of distribution and $\lambda(\cdot)$ is an unspecified baseline hazard function. That is, the hazard given the covariates W and a random frailty U unique to each individual has the proportional hazards form multiplied by the frailty. A robust

nonparametric likelihood-based inference was carried out to allow for model misspecification. The profile likelihood estimators for the finite dimensional parameters were shown to be semiparametric efficient when the model is correctly specified. It was also proved in Kosorok et al. (2004) that the bootstrap gives valid inferences for all parameters, even under model misspecification.

An even more general class of models which includes the generalized odds-rate model as its special case is the class of linear transformation models (Dabrowska and Doksum 1988; Slud and Vonta 2004; Zeng and Lin 2007b) with the form $H(t|W) = G(\Lambda(t)e^{\beta^T W})$, where $H(\cdot|W)$ denotes the conditional baseline cumulative hazard function given covariates W, $\Lambda(\cdot)$ denotes the baseline cumulative hazard function and both $G(\cdot)$ and $\Lambda(\cdot)$ are unspecified. Equivalently, the linear transformation model can be written as $\Lambda(T) = -\beta^T W + \epsilon$, where $\Lambda(\cdot)$ is an unspecified increasing function and ϵ is a random error with a specified parametric distribution. The choice of the extreme value and standard logistic error distributions yield the proportional hazards and proportional odds model, respectively. In particular, Zeng and Lin (2007b) proposed a very general class of transformation models for counting processes which encompasses linear transformation models and which accommodates time-varying covariates and recurrent events and they also proved the semiparametric efficiency for the estimator of the regression parameter using nonparametric maximum likelihood estimation (NPMLE).

Among other extensions of the Cox model is the fully nonparametric model of the form $h(t|W) = \lambda(t, W)$ studied by Nielsen and Linton (1995), where the function $\lambda(\cdot, \cdot)$ is unspecified. One nice feature about this model is that the covariates do not need to satisfy the proportional hazards assumption and it provides the most flexible way to model covariate effects. A kernel estimator for the conditional hazard rate was proposed and its uniform convergence and asymptotic normality were established. The rate of convergence for their estimator is slower than $\sqrt{n}$ and decreases as the dimension of the covariates increases. Later, Nielsen, Linton, and Bickel (1998) studied a semiparametric model of

the form $h(t|W) = \lambda_\theta(t)g(W)$, where $\lambda_\theta(\cdot)$ is the parametric baseline hazard function indexed by a parameter θ and $g(\cdot)$ is completely unknown. Note that this multiplicative model is a special case of Nielsen and Linton (1995). A kernel smoothed estimator for the nonparametric function $g(\cdot)$ was proposed and the estimator for the regression parameter β was constructed based on a kernel smoothed profile likelihood function. The resulting estimator for β was shown to achieve the semiparametric efficiency bound. Although assuming a parametric baseline hazard function may seem reasonable in certain settings, it is more desirable to assume a nonparametric baseline hazard function instead so that the model is more robust to misspecification. Furthermore, all covariates in this model are required to satisfy the proportional hazards assumption. Instead of assuming a parametric baseline hazard function, Fan, Gijbels, and King (1997) focused on another multiplicative nonparametric model of the form $h(t|W) = \lambda(t)e^{\phi(W)}$, where the logarithm of the conditional hazard rate function is assumed to be the sum of an unknown function of covariates and an unknown function of the survival time. Note that this model is also a special case of the model studied in Nielsen and Linton (1995). The estimation of $\phi(\cdot)$ was based on its local approximation by a polynomial function and the estimation for β was based on a local version of the partial likelihood. The estimator for β was shown to be asymptotically normal but no results on semiparametric efficiency were reported. Similar to the model studied in Nielsen et al. (1998), proportionality is an implicit requirement for all covariates. Note that all covariates in these three models have nonparametric effects. Although this may seem flexible, the interpretation of the covariate effects is difficult and the nonparametric estimation of the unknown function in each of these models is feasible only if the dimension of W is low. That is, all these three models suffer from the so-called "curse of dimensionality".

1.1.2 Single-Index Models

One of the most convenient models for dimension reduction is the single-index model, which is commonly used in biometrics and econometrics, discussed by Härdle and Stoker (1989) and Härdle, Hall, and Ichimura (1993). The model takes the form $Y = \eta(\beta^T W)+\epsilon$, where Y denotes the response, the univariate smooth function $\eta(\cdot)$ is completely unknown, β is an unknown unit vector with one coordinate positive for identification purposes and $E(\epsilon|W) = 0$ almost surely. Note that, in contrast to a nonparametric model of the form $Y = \eta(W) + \epsilon$, the parsimonious single-index model is particularly attractive since the original multi-dimensional covariate vector W has been replaced by a 1-dimensional "single-index" (the linear combination $\beta^T W$). Through dimension reduction in this way, the nonparametric estimation of $\eta(\cdot)$ becomes feasible. Another attractive property about this single-index model is that the relative importance of the components of W can be fully characterized by the orientation vector β since the derivative of $E(Y|W)$ with respect to W_i, the i^{th} component of the covariates W, is proportional to β_i, the i^{th} component of β. Thus β_i characterizes how fast $E(Y|W)$ changes with W_i. This piece of information on the relative importance of components of W is practically useful for designing future studies. For instance, one only need to measure those important biomarkers but ignore those non-important ones that may be expensive to measure. We note that β does not in general represent the covariate effects as in a linear regression model. However, if $\eta(\cdot)$ is a monotone function, then β has the same role as "effect" parameters. Two popular methods for estimating the single-index model are the average derivative estimation method proposed by Härdle and Stoker (1989) and the method of Härdle et al. (1993), who used the kernel smoothing method to construct the estimator of the unknown function $\eta(\cdot)$ of the single-index and the estimator of the orientation vector β minimizes a modified mean square error function. Härdle et al. (1993) also suggested an empirical rule for selecting the bandwidth.

In some sense, a single-index $\beta^T W$ can be viewed as a principle component of the

covariate vector W. When the dimension of this covariate vector is high, one may wish to include multiple principle components into the model so that "enough" information is extracted from the covariates. Thus it may be attractive to consider a model of the form $Y = m(\beta_1^T W, \cdots, \beta_k^T W) + \epsilon$, where $m(\cdot)$ is an unknown $k-$variate function and $E(\epsilon|W) = 0$ almost surely. Here $k \geq 1$ is a pre-specified integer less than the dimension of the covariates W. This model has been studied extensively in the literature. Recent work includes Cook and Li (2002), Xia, Tong, Li, and Zhu (2002) and Yin and Cook (2002), among others.

Since it is likely that one of the dimension reduction components (or single-indices) $\beta_1^T W, \cdots, \beta_k^T W$ affects the response linearly and the other $k-1$ components affect the response nonlinearly, it is natural to consider a multiple-index model (Ichimura and Lee 1991; Horowitz 1998; Xia 2008; among others) of the form $Y = G(\beta_1^T W, \cdots, \beta_{k-1}^T W) + \beta_k^T W + \epsilon$, where $E(\epsilon|W) = 0$ almost surely and $G(\cdot)$ is an unknown link function. Compared to the model with nonparametric modeling of all single-indices, this partly linear model enjoys an easier interpretation and better estimation due to the further dimension reduction in the unknown nonparametric function $m(\cdot)$. When the number of single-indices k is large (although less than the dimension of W already), it may be beneficial to consider the so-called additive-index model (Chiou and Müller 2004) of the form $Y = \sum_{j=1}^k m_j(\beta_j^T W) + \epsilon$, where $m_j(\cdot)$ is an unknown univariate function, $j = 1, \cdots, k$, and again $E(\epsilon|W) = 0$ almost surely. Note that such a model replaces the unknown function of k variables in Cook and Li (2002), Xia et al. (2002) and Yin and Cook (2002) by k unknown univariate functions and thus offers a better estimation due to dimension reduction. One special case of the additive-index model is the additive single-index model studied by Naik and Tsai (2001) which takes the form $Y = m_1(\alpha^T W_1) + m_2(\gamma^T W_2) + \epsilon$, where $W^T = (W_1^T, W_2^T)^T$. This is a special case of the additive-index model of Chiou and Müller (2004) by setting $k = 2$, $\beta_1^T = (\alpha^T, 0^T)^T$ and $\beta_2^T = (0^T, \gamma^T)^T$.

In the survival analysis context, the model with the conditional hazard function specified as $h(t|W) = \lambda(t)e^{\phi(\beta^T W)}$, where both $\lambda(\cdot)$ and $\phi(\cdot)$ are unknown, was studied by Wang (2004) and Huang and Liu (2006). If $\phi(x) = x$, this model reduces to the Cox proportional hazards model. Note that this model is similar in form to the model studied by Fan et al. (1997) except that the single-index $\beta^T W$ replaces the original covariates W for dimension reduction. In Wang (2004), covariates are allowed to be time-dependent and potentially missing. When the missing covariates are present, a two stage approach was proposed to account for the missingness. In the first stage, the missing time-dependent covariates were imputed using functional data analysis methods. In the second stage, a two-step iterative algorithm was performed to estimate the unknown function $\phi(\cdot)$. Asymptotic properties were derived for the estimator of the nonparametric function when time-dependent covariates are not missing, but there are no asymptotic properties for the estimator of β presented in that paper. Later, Huang and Liu (2006) used spline smoothing techniques to approximate the unknown link function $\phi(\cdot)$ and then employed the maximum partial likelihood to estimate the regression parameter β. They also established inference procedures for the function $\phi(\cdot)$ and the index coefficient vector β, and discussed the interpretation of the regression coefficients in detail, but no results on semiparametric efficiency were presented. Furthermore, in the aforementioned two models, all of the covariates are incorporated into one single-index term, no matter whether they have linear or nonlinear effects on the hazard and thus the interpretation of covariate effects are difficult. Also, all covariates must meet the proportional hazards requirement. Recently, Xia, Zhang, and Xu (2010) studied a very general regression model of the form $T = G(B^T W, \epsilon)$, where T is the survival time, $G(\cdot, \cdot)$ is completely unknown, B is a parameter matrix with the column dimension less than the row dimension and ϵ is independent of the covariates W. Note that this model includes the transformation model (Zeng and Lin 2007b) and the accelerated failure time model (Cox and Oakes 1984, chap. 5) as its special cases. Xia et al. (2010) also proposed a novel dimension

reduction method by introducing a nominal regression model to estimate the conditional hazard function via estimation of the central subspace in the presence of censoring. Similarly, this model treats all the covariates in the same way (via nonparametric modeling of single-indices) and so the model interpretation is difficult.

It is worthwhile at this stage to point out that given a set of available covariates, one should always screen out those covariates that are inappropriate for control before model fitting, as suggested by Greenland (1989). For example, in epidemiologic studies, "it is well known that covariates influenced by the exposure or disease are inappropriate for control, since control of such covariates may lead to considerable bias" (Greenland 1989); thus, we assume throughout this dissertation that the covariates W are those remaining after the screening. We also note that "covariates" used in this dissertation could be other types of relevant quantities. For example, principle components are widely used in genetics as "covariates", as in Kong, Pu, and Park (2006), Chen, Wang, Smith, and Zhang (2008) and Ma and Kosorok (2009). In the sequel, we will use "covariates" despite the note we just made.

1.1.3 Partially Linear Models for Survival Data

In all of the aforementioned models, all components of W are treated equally in the sense that no distinction is made as to which components are more interesting to investigators than the others. In practice, covariates W can often be partitioned into two parts, say X and Z, corresponding to the covariates of primary interest and the "nuisance" covariates (potential confounders), respectively. We assume in the sequel that X is p dimensional and Z is q dimensional. For example, in the aforementioned MACS data, one might be interested in assessing the effect of patient's ethnicity, baseline age, viral loads and CD4 counts on the risk of death due to AIDS, controlling for serum β_2-microgloburin levels and serum neopterin levels. Thus patient's ethnicity, baseline age, viral loads and CD4 counts are treated as covariates of primary interest X and serum β_2-microgloburin levels

and serum neopterin levels are treated as "nuisance" covariates Z. One might also wish to assess in particular the effect of patient's ethnicity and baseline age on the risk of death due to AIDS, controlling for the remaining 4 biomarkers. Thus in this instance the covariates of primary interest are patient's ethnicity and baseline age and the remaining 4 biomarkers are "nuisance" covariates.

Since covariates of primary interest are given more priority, X is often modeled parametrically to ensure model interpretability but Z is modeled nonparametrically to allow for model flexibility. One example of such a modeling strategy is the partially linear Cox model studied by Sasieni (1992a, b). The model takes the form $h(t|X,Z) = \lambda(t)e^{\beta^T X+\phi(Z)}$, where both $\lambda(\cdot)$ and $\phi(\cdot)$ are completely unknown. Note that by assuming proportionality of X through a parametric function $\beta^T X$, the regression parameter β can now be interpreted as the log-hazard ratio for X and the "nuisance" covariates Z can have nonparametric effects on the hazard function. The estimation method for this model in Sasieni (1992a, b) was based on a spline smoothed partial likelihood. Sasieni (1992a, b) also provided the efficient score and information bound for estimating β. However, no details were provided on the asymptotic distribution of the suggested spline based estimators.

The special case when Z is 1-dimensional and $\lambda(\cdot)$ is a parametric function indexed by a finite dimensional parameter θ in the partially linear Cox model was studied by Lu, Singh, and Desmond (2001), who proposed to estimate β by maximizing a profile likelihood after profiling out $\phi(\cdot)$ estimated by using the local likelihood. The resulting estimator for (β,θ) was $\sqrt{n}-$consistent, asymptotically normal and semiparametrically efficient. Heller (2001) considered the same model as in Lu et al. (2001) except assuming a nonparametric baseline hazard function. Similarly, his estimator for β was based on a profile likelihood after profiling out the infinite dimensional parameters using kernel smoothing. The resulting estimator for β was again shown to be semiparametrically efficient. To our best knowledge, the large sample properties of estimators of the partially

linear Cox model with multi-dimensional "nuisance" covariates have not been studied. Although the partially linear Cox model is flexible in term of modeling "nuisance" covariate effects, it has two potential drawbacks. First, as in Nielsen and Linton (1995), Fan et al. (1997) and Nielsen et al. (1998), the nonparametric estimation is only practically feasible when the dimension of the "nuisance" covariates Z is low; Second, "nuisance" covariates are required to satisfy the stringent proportional hazards assumption.

To tackle the first drawback of the partially linear Cox model in Sasieni (1992a, b), Huang (1999) studied the partly linear additive Cox model by assuming $h(t|X,Z) = \lambda(t)e^{\beta^T X+\sum_{i=1}^{q}\phi_i(Z_i)}$, where $\lambda(\cdot)$ and $\phi_i(\cdot)$ are unknown functions and Z_i is the i^{th} component of the q-dimensional covariates Z, $i = 1, \cdots, q$. Thus one unknown function of q variables in Sasieni (1992a, b) has been replaced by q unknown univariate functions and so it breaks the "curse of dimensionality". Note that this model is a special case of Sasieni (1992a, b). The polynomial spline method was used to estimate the nonparametric functions $\phi_i(\cdot), i = 1, \cdots, q$ and the estimators of the regression parameters maximize the induced spline smoothed partial likelihood, which were shown to be $\sqrt{n}$-consistent, asymptotically normal and semiparametrically efficient. Note that such a model requires estimating q unknown functions and so is computationally intense. Moreover, the "nuisance" covariates are still required to satisfy the proportional hazards assumption.

The second drawback of Sasieni's (1992a, b) model can be overcome by assuming a partly proportional hazards model (Dabrowska 1997) of the form $h(t|X,Z) = \lambda(t,Z)e^{\beta^T X}$, where $\lambda(\cdot,\cdot)$ is an unknown bivariate baseline hazard function which depends on the "nuisance" covariates Z. Note that this model includes Sasieni (1992a, b) model as a special case. The parameter estimation in Dabrowska (1997) was based on a kernel smoothed partial likelihood. It was shown that when the dimension of Z is at most 3, the estimator for β is asymptotically normal at rate $\sqrt{n}$. However, the proposed estimator fails to be $\sqrt{n}-$consistent when the dimension of Z is larger than 3. A one-step estimator was then suggested to achieve the $\sqrt{n}$ rate. Therefore, as in Sasieni (1992a,

b), this model is only practically feasible when Z is low dimensional. Furthermore, there are no results on semiparametric efficiency in this instance.

In the current setting of semiparametric modeling of covariates effects, the aforementioned strategy for dimension reduction via a single-index can also be used. For example, Xia, Tong, and Li (1999) studied the partially linear single-index model of the form $Y = \beta^T X + \eta(\gamma^T Z) + \epsilon$, where $\eta(\cdot)$ and ϵ are defined in the aforementioned single-index model. Again, the kernel smoothing method was used to construct the estimator of the unknown function $\eta(\cdot)$ of the single-index. More generally, Carroll, Fan, Gijbels, and Wand (1997) proposed the generalized partially linear single-index model of the form $g(E(Y|X,Z)) = \beta^T X + \eta(\gamma^T Z)$, where $g(\cdot)$ is a known link function and $\eta(\cdot)$ is unspecified. A local quasi-likelihood was used to estimate the unknown function of the single-index. However, a $\sqrt{n}-$consistent pilot estimator for γ and under-smoothing are needed. Later, Xia and Härdle (2006) proposed the minimum average variance estimation method which does not require a $\sqrt{n}-$consistent pilot estimator and the bandwidth can be selected at the optimal smoothing rate. Besides kernel smoothing methods, other smoothing methods have been studied. For example, Yu and Ruppert (2002) considered the penalized spline method in the partially linear single-index model (Xia et al. 1999) and showed that the penalized spline method performs better than the kernel smoothing method of Carroll et al. (1997).

In the survival analysis context, Lu, Chen, Song, and Singh (2006) considered the partially linear single-index survival model of the form $h(t|X,Z) = \lambda_\theta(t)e^{\beta^T X + \eta(\gamma^T Z)}$, where the form of the baseline hazard function is known up to an Euclidean parameter θ and $\eta(\cdot)$ is unknown. The estimation of $\eta(\cdot)$ was based on a local linear fit and the estimator for (β, γ, θ) was shown to be asymptotically normal and semiparametrically efficient. Even though the parametric baseline hazard appears reasonable in many applications, it is desirable to have a more flexible nonparametric hazard instead. In that direction, Sun, Kopciuk, and Lu (2008) studied a more general partially linear proportional hazards

model of the form $h(t|X,Z) = \lambda(t)e^{\beta^T X+\eta(\gamma^T Z)}$, where $\lambda(\cdot)$ is now unspecified. Sun et al. (2008) adopted a polynomial spline smoothing technique for estimating the unknown smooth function $\eta(\cdot)$. However, no asymptotic results were presented in this instance. Furthermore, all covariates in Lu et al. (2006) and Sun et al. (2008) must satisfy the proportional hazards assumption.

1.2 Outline of Dissertation

In this dissertation, we first consider the "single-index hazards model", a modification of the model studied in Nielsen and Linton (1995), by assuming a nonparametric baseline hazard function that depends on W through a single index $\beta^T W$. Specifically, we consider a model of the form $h(t|W) = \lambda(t, \beta^T W)$, where $\lambda(\cdot,\cdot)$ is an unknown bivariate function. Note that this model includes the Cox model and all the transformation models mentioned before as special cases. In addition, the model has several nice features. First, covariates are allowed to have nonparametric effects on the hazard function. This is particularly useful if covariates W do not satisfy the proportional hazards assumption so that the Cox model may not be appropriate. Second, the relative importance of the components of W can be fully characterized by the orientation vector β since the derivative of $h(t|W)$ with respect to W_i, the i^{th} component of the covariate vector W, is proportional to β_i, thus β_i characterizes how fast $h(t|W)$ changes with W_i. Third, this single-index hazards model is more parsimonious than the model in Nielsen and Linton (1995) since the multi-dimensional vector W has been replaced by a one-dimensional single-index $\beta^T W$. The local likelihood approach is commonly used for the single-index model. Thus we adapt this approach for parameter estimation in our single-index hazards model. Surprisingly, we find, both theoretically and numerically, that this commonly used approach in general yields inconsistent estimators and it may work only under very specific conditions.

Since the aforementioned single-index hazards model cannot in general address covariate effects, especially the effect of covariates of main interest, we further propose the "partly proportional single-index hazards model" by assuming $h(t|X,Z) = \lambda(t, \gamma^T Z)e^{\beta^T X}$, where $\lambda(\cdot,\cdot)$ is an unknown function. The model has several nice features. First, by assuming proportionality of X via the linear combination $\beta^T X$, the regression parameter β can be interpreted as the log-hazard ratio of the covariates of primary interest X for any given Z, while Z is allowed to have nonparametric effects. The nonparametric modeling of Z is particularly useful if the "nuisance" covariates do not satisfy the proportional hazards assumption so that the Cox model may yield biased results. Second, this model overcomes both drawbacks associated with Sasieni's (1992a, b) model. Specifically, it is parsimonious since the q-dimensional covariates Z have been replaced by a one-dimensional single-index $\gamma^T Z$, and thus nonparametric estimation becomes feasible. Furthermore, as in Dabrowska's (1997) model, the proportional hazards assumption is relaxed for Z. Third, similar to the single-index hazards model, the relative importance of the components of Z can be fully characterized by the orientation vector γ since the derivative of $h(t|X,Z)$ with respect to Z_i, the i^{th} component of the "nuisance" covariates Z, is proportional to γ_i, the i^{th} component of γ. Thus γ_i characterizes how fast $h(t|X,Z)$ changes with Z_i, $i = 1, \cdots, q$. To estimate the regression parameters β and γ, we construct a profile likelihood after profiling out the baseline hazard function, which is estimated based on a local likelihood function. Similar to the single-index hazards model, it is shown that this conventional profile-kernel method leads to biased estimation of the regression parameters. We also believe that this bias phenomenon extends to other model settings besides our partly proportional single-index hazards model. To address the bias issue in this model, we propose a bias correction method which is shown to have nice asymptotic properties and works well in finite-sample settings.

In addition to the profile local likelihood method, we consider another popular approach for model estimation, named the profile stratified likelihood approach based on

stratification on the single-index. In the single-index hazards model, this method may give consistent estimation under the restrictive "independence censoring" condition, but in general it can yield biased estimation. Simulation studies are conducted to demonstrate the situations in which the bias phenomena do (or do not) exist; In the partly proportional single-index hazards model, we demonstrate numerically the existence of the bias and then propose a bias correction method using a similar idea for correcting the bias in the profile local likelihood method. The estimators from the corrected profile stratified likelihood method are shown to be consistent. Their finite-sample properties are evaluated through simulation studies and this bias corrected method is applied to the aforementioned MACS study for illustration.

The remainder of this dissertation is organized as follows. Chapter 2 focuses on the bias analysis of the profile local likelihood approach in the single-index hazards model. In Section 2.1, we describe the single-index hazards model and the data structure. In Section 2.2, we describe how to adapt the commonly used profile local likelihood method for parameter estimation. We then study the asymptotic bias of this approach in Section 2.3 and identify conditions under which this approach may work. In Section 2.4, we demonstrate our findings via a series of simulation studies. In Chapter 3, we focus on the partly proportional single-index hazards model. In Section 3.1, we describe the model and the data structure. In Section 3.2, we consider again the commonly used profile local likelihood method and study the estimation bias associated with this method, both theoretically and numerically. A bias correction method is then proposed and results on the asymptotic and finite-sample properties of the corrected profile local likelihood estimator are given in Section 3.3. In Section 3.4, we illustrate the proposed model and method with an application to a dataset from the MACS. Chapter 4 studies the profile stratified likelihood method. In Section 4.1, we consider this method in the single-index hazards model and its performance is studied both asymptotically and numerically. In Section 4.2, we consider this method in the partly proportional single-index hazards

model. Specifically, we demonstrate the estimation bias numerically, propose a bias correction, give some asymptotic results of the corrected stratified likelihood method and apply the partly proportional single-index hazards model to the dataset from the MACS using the bias corrected method. Finally, the dissertation is concluded with a discussion in Chapter 5. Technical proofs are given at the end of each chapter.

Chapter 2

Single-Index Hazards Model

2.1 Model and Data Structure

We assume the following single-index hazards model

$$h(t|W) = \lambda(t, \gamma^T W), \tag{2.1}$$

where $\gamma \in \mathbb{R}^q$ and $\lambda(\cdot, \cdot)$ is an unknown bivariate function. To ensure identifiability, we first impose the restriction that $\|\gamma\| = 1$ with the last component γ_q positive, that is, the γ vector is restricted to the half unit sphere. This assumption is practically reasonable when at least one covariate has a non-zero effect.

Suppose we observe a random sample of size n, $(Y_i = T_i \wedge C_i, \Delta_i, W_i), i = 1, \cdots, n$, where T is the survival time, C is the censoring time, $a \wedge b = \min(a, b)$, $\Delta = I(T \leq C)$ is the censoring indicator and W is the covariate vector. The subscript i is used to denote the i^{th} subject. The log-likelihood function is

$$\frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \log \lambda(Y_i, \gamma^T W_i) - \Lambda(Y_i, \gamma^T W_i) \right]. \tag{2.2}$$

This function has a maximum value of infinity and thus cannot be used directly for parameter estimation. In nonparametric maximum likelihood estimation (NPMLE), we

maximize

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i \log \Lambda\{Y_i, \gamma^T W_i\} - \sum_{Y_j \le Y_i} \Lambda\{Y_j, \gamma^T W_i\}\right]. \tag{2.3}$$

Here, $\Lambda\{Y_i, \gamma^T W_i\}$ is the jump size of $\Lambda(Y_i, \gamma^T W_i)$ at Y_i. However, the profile likelihood function based on (2.3), obtained by profiling out $\Lambda\{\cdot, \cdot\}$, is a constant and is thus not a valid objective function. In the next section, we consider a commonly used estimation approach for model (2.1), the local profile likelihood approach.

2.2 Profile Local Likelihood

Local likelihood has been frequently used to estimate the unknown function in a semi-parametric model. In this approach, a local likelihood is constructed to estimate the nonparametric function and then the estimated function is plugged into the likelihood (or some variant of the likelihood) to obtain the profile likelihood function. This conventional profile-kernel method was adopted, for example, by Fan et al. (1997). Carroll et al. (1997) used the same method except that a quasi-likelihood played the role of the regular likelihood function. Specifically for our likelihood (2.3) and fixed γ, we would estimate $\Lambda\{\cdot, \cdot\}$ by maximizing the following local likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i \log \Lambda\{Y_i, u\} - \sum_{Y_j \le Y_i} \Lambda\{Y_j, u\}\right] K_{a_n}(\gamma^T W_i - u),$$

where $K_{a_n}(t) = K(t/a_n)/a_n$, K is a mean zero symmetric density function and $\Lambda\{Y_i, w\}$ is the jump size of $\Lambda(Y_i, w)$ at Y_i for each w. This is the local constant fit weighted by the function $K_{a_n}(\cdot)$. The maximizer can be found as

$$\hat{\Lambda}\{Y_i, u\} = \frac{\Delta_i K_{a_n}(\gamma^T W_i - u)}{\sum_{Y_j \ge Y_i} K_{a_n}(\gamma^T W_j - u)}. \tag{2.4}$$

After plugging (2.4) into (2.3), we obtain, up to a constant, that the profile local likelihood is

$$-\frac{1}{n}\sum_{i=1}^{n}\Delta_i \log\left(\frac{1}{na_n}\sum_{Y_j\geq Y_i}K\left(\frac{\gamma^T(W_j-W_i)}{a_n}\right)\right) \\ -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{na_n}\sum_{Y_j\leq Y_i}\frac{\Delta_j K\left(\frac{\gamma^T(W_j-W_i)}{a_n}\right)}{\frac{1}{na_n}\sum_{Y_k\geq Y_j}K\left(\frac{\gamma^T(W_k-W_i)}{a_n}\right)}. \tag{2.5}$$

We will show in Lemma 2.5.1 in Section 2.5 that the second term of (2.5) equals a constant (with respect to the parameter) asymptotically. As a result, the estimator of γ is the maximizer of the local profile likelihood function $pl_n^{loc}(\gamma)$, which only includes the first term. That is,

$$pl_n^{loc}(\gamma) = -\frac{1}{n}\sum_{i=1}^{n}\Delta_i \log\left(\frac{1}{na_n}\sum_{Y_j\geq Y_i}K\left(\frac{\gamma^T(W_j-W_i)}{a_n}\right)\right).$$

Note that this function is smooth in γ. Thus numerically it can be easily maximized. For example, the quasi-Newton search algorithm can be used.

2.3 Bias Analysis

In this section, we aim to rigorously study the estimation bias based on the profile local likelihood $pl_n^{loc}(\gamma)$. We impose the following regularity conditions:

(C1) $\gamma_0 \in \Gamma$, where $\Gamma \in \mathbb{R}^q$ is compact.

(C2) The random covariate vector W has a continuous density on its support.

(C3) The non-uniform kernel function $K(\cdot)$ has zero mean with finite second moment. Moreover, $\sup_x |K'(x)|$ is finite, where $K'(x)$ denotes the derivative function of $K(x)$.

(C4) The bandwidth $a_n = n^{\nu_1}$ with $\nu_1 \in (-1/2, 0)$.

Remark 2.3.1. Many kernel functions satisfy condition (C3), for example, the standard Gaussian kernel $K(u) = 1/\sqrt{2\pi}\exp(-u^2/2)$ and the Epanechnikov kernel $K(u) = 3/4(1-u^2)I(|u| \leq 1)$.

The following theorem gives the asymptotic limit of $pl_n^{loc}(\gamma)$.

Theorem 2.3.1. *If conditions (C1)-(C4) hold, then* $\sup_\gamma |pl_n^{loc}(\gamma) - pl^{loc}(\gamma)| \to_{a.s.} 0$, *where*

$$pl^{loc}(\gamma) = -E\left[\Delta \log\left(P(Y \geq y|\gamma^T W)|_{y=Y} f_{\gamma^T W}(\gamma^T W)\right)\right].$$

Here $f_{\gamma^T W}(\cdot)$ *is the density function of* $\gamma^T W$.

Thus the local profile likelihood estimator should converge to the maximizer of $pl^{loc}(\gamma)$ almost surely by Theorem 2.12 of Kosorok (2008). Suppose the latter is the true parameter γ_0, then the derivative of $pl^{loc}(\gamma)$ with respect to γ should be proportional to γ_0 if evaluated at γ_0. This proportionality to γ_0 is due to the restriction $\|\gamma\| = 1$. However, we show in the next theorem that this may not be true under the following two regularity conditions:

(C5) Given covariates W, T and C are independent.

(C6) $P(T > \tau) < 1$, where τ denotes the end of the study.

Remark 2.3.2. Condition (C6) implies a positive probability of non-censoring so that $pl_n^{loc}(\gamma)$ is not a constant with respect to γ.

Theorem 2.3.2. *Assume conditions (C5) and (C6) hold and suppose* C *is independent of* W *and* $W \sim N(\mu, \Sigma)$ *with* Σ *positive definite, then* $\frac{\partial}{\partial\gamma}\big|_{\gamma=\gamma_0} pl^{loc}(\gamma) \propto \gamma_0$ *if and only if* $\Sigma\gamma_0 = c\gamma_0$ *for some constant* c.

Remark 2.3.3. This theorem suggests that even in the special case where the covariate vector follows a normal distribution and C is independent of W, the profile local likelihood

approach may give consistent estimation only under the restrictive condition $\Sigma\gamma_0 = c\gamma_0$. Thus, in the more general set-up, $pl^{loc}(\gamma)$ may not be maximized at γ_0, and thus the procedure may be inconsistent.

2.4 Simulation Studies

We conduct numerical studies to demonstrate the estimation bias associated with the aforementioned profile local likelihood approach. In this section, we assume that the covariate vector $W = (W_1, W_2)$ is two dimensional and is generated from a bivariate normal distribution with zero means and unit variances. The true parameter for γ is $\gamma_0 = (-1/2, \sqrt{3}/2)^T$. The following four simulation settings are considered: (i) The censoring time C is independent of W, $\lambda_0(t, u) = 0.5e^u t$, W has no correlation; (ii) C is independent of W, $\lambda_0(t, u) = 0.5e^u t$ and the covariance between W_1 and W_2 is 0.5; (iii) C is independent of W, $\lambda_0(t, u) = 0.25(t + u^2)$ and we use the same covariance matrix as in setting (ii); (iv) C and W are dependent, $\lambda_0(t, u) = 0.25(t + u^2)$ and we use the same covariance matrix as in setting (ii). In settings (i)-(iii), C is generated from the uniform $[0, \tau]$ distribution with $\tau = 10$ and $C = 4e^{W_2} \wedge \tau$ in setting (iv). Note that in setting (i) and (ii), the proportional hazards assumption is satisfied, but this assumption is violated in settings (iii) and (iv). The censoring rate ranges approximately from 20% to 28%.

We choose the kernel function to be the standard normal density and the parameter is estimated by using the quasi-Newton search algorithm in the R software package. The initial value is set to zero. The bandwidth is chosen to be $c_1 \times IQR_1 \times n^{-1/4}$, where the tuning parameter c_1 is chosen from $\{2, 1, 1/2\}$ and IQR_1 is the inter-quartile range of $\|W\|$ in each simulated data set. For each simulation setting, the tuning parameter c_1 which gives the smallest bias when $n = 10000$ is chosen and then the same parameter value is used for other sample sizes.

Table 2.1 summarizes the simulation results in setting (i)-(iv) with sample sizes 2000,

Table 2.1: Simulation results of local likelihood in single-index hazards model

			Local likelihood		Cox model	
Simulation settings	Parameter	Sample size	Bias	SE	Bias	SE
(i) $C \perp W$	γ_1	2000	.153	.421	-.001	.026
$\lambda_0(t,u) = 0.5e^u t$		5000	.018	.183	.000	.017
$cov(W) = 0$		10000	.013	.113	.000	.012
(ii) $C \perp W$	γ_1	2000	1.198	.035	-.001	.030
$\lambda_0(t,u) = 0.5e^u t$		5000	1.195	.023	.000	.019
$cov(W) = 0.5$		10000	1.195	.017	.001	.014
(iii) $C \perp W$	γ_1	2000	1.188	.037	.500	.032
$\lambda_0(t,u) = 0.25(t+u^2)$		5000	1.189	.022	.501	.020
$cov(W) = 0.5$		10000	1.185	.017	.500	.014
(iv) $C \not\perp W$	γ_1	2000	1.338	.027	.446	.033
$\lambda_0(t,u) = 0.25(t+u^2)$		5000	1.340	.016	.448	.022
$cov(W) = 0.5$		10000	1.342	.011	.447	.015

NOTE: Each entry is based on 500 replicates.

5000 and 10000, where γ_1 is the first component of the γ vector. As expected by Theorem 2.3.2, the local likelihood approach fails in settings (ii)-(iv) due to the correlation among the vector W and γ_0 not being an eigenvector of the covariance matrix of W. Theorem 2.3.2 also suggests that the local likelihood approach may work in setting (i) because the identity covariance matrix is used. We have also reported the results from the Cox proportional hazards model. The Cox model produces consistent estimators in setting (i) and (ii) since the proportional hazards assumption is satisfied, but it gives biased estimation in setting (iii) and (iv) due to the violation of this assumption.

Figure 2.1 shows the profile local likelihood function based on a simulated data set of size 10000 in each simulation setting. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The bandwidth is $1 \times n^{-1/4}$. Again, the profile local likelihood approach gives biased estimators except in setting (i).

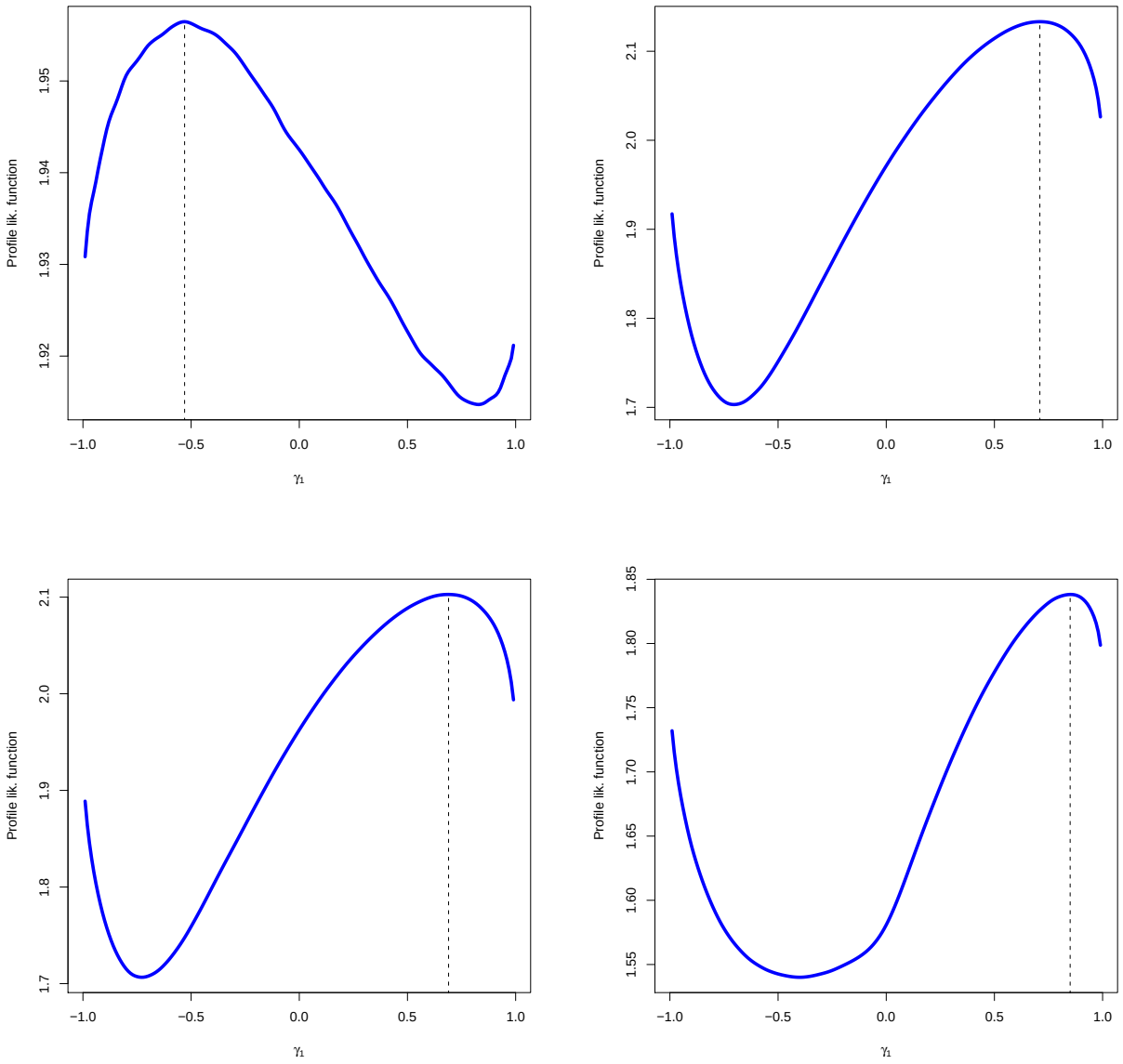

Figure 2.1: Profile local likelihood curve of γ_1 in single-index hazards model

2.5 Proofs of Theorems

We denote the second term of (2.5) by (B).

Lemma 2.5.1. *If conditions (C1)-(C4) hold, then* $\sup_\gamma |(B) - 1/n \sum_{j=1}^n \Delta_j| \to_{a.s.} 0$.

Proof of Lemma 2.5.1

We partition Γ into small cubes such that any two points in the same cube have distance no large than δ_n to be determined later. The number of partitions, denoted by m_n^*, is of order $1/\delta_n^q$. Choose one arbitrary point from each of these partitions and denote them as $\gamma^{(1)}, \ldots, \gamma^{(m_n^*)}$. For γ_1 and γ_2 in the same cube, any fixed y, w,

$$\left| \frac{1}{na_n} \sum_{Y_j \geq y} K\left(\frac{\gamma_1^T(W_j - w)}{a_n}\right) - \frac{1}{na_n} \sum_{Y_j \geq y} K\left(\frac{\gamma_2^T(W_j - w)}{a_n}\right) \right| \leq \frac{c}{a_n^2} \|\gamma_1 - \gamma_2\|, \quad \text{and}$$

$$\left| \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma_1^T(W - w)}{a_n}\right) \right] - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma_2^T(W - w)}{a_n}\right) \right] \right| \leq c_1 \|\gamma_1 - \gamma_2\|,$$

for universal constants c and c_1. If we choose $\delta_n / a_n^2 \to 0$ as $n \to \infty$, then for any $\delta > 0$,

$$\begin{aligned}
&P\left(\sup_{\gamma, y, w} \left| \frac{1}{na_n} \sum_j I(Y_j \geq y) K\left(\frac{\gamma^T(W_j - w)}{a_n}\right) - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n}\right) \right] \right| > \delta \right) \\
&\quad \leq P\left(\max_{1 \leq l \leq m_n^*} \sup_{y,z} \left| \frac{1}{na_n} \sum_j I(Y_j \geq y) K\left(\frac{\gamma^{(l)^T}(W_j - w)}{a_n}\right) \right.\right. \\
&\qquad \left.\left. - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^{(l)^T}(W - w)}{a_n}\right) \right] \right| > \frac{\delta}{2} \right) \\
&\quad \leq \sum_{l=1}^{m_n^*} P\left(\sup_{y,z} \left| \frac{1}{na_n} \sum_j I(Y_j \geq y) K\left(\frac{\gamma^{(l)^T}(W_j - w)}{a_n}\right) \right.\right. \\
&\qquad \left.\left. - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^{(l)^T}(W - w)}{a_n}\right) \right] \right| > \frac{\delta}{2} \right) \\
&\quad \leq c_0 m_n^* \exp(-c_1 n \delta^2 a_n^2),
\end{aligned}$$

where the exponential bound in the last step makes use of the result on the empirical

CDF due to Dvoretzky, Keifer and Wolfowitz (1956). Therefore,

$$\sum_{n=1}^{\infty} P\left(\sup_{\gamma,y,z} \left| \frac{1}{na_n} \sum_j I(Y_j \geq y) K\left(\frac{\gamma^T(W_j - w)}{a_n} \right) - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n} \right) \right] \right| > \delta \right)$$
$$\leq c_2 \sum_{n=1}^{\infty} \delta_n^{-q} \exp(-c_1 n \delta^2 a_n^2).$$

If we choose $\delta_n = a_n^3$, then the previous display becomes

$$c_2 \sum_{n=1}^{\infty} \frac{a_n^{-3q}}{e^{c_1 n \delta^2 a_n^2}} \leq c_3 \sum_{n=1}^{\infty} \frac{a_n^{-3q}}{(n a_n^2)^m},$$

for any positive integer m. Since $a_n = n^{\nu_1}$ with $\nu_1 \in (-1/2, 0)$, we can choose m to be larger than $(1 - 3q\nu_1)/(1 + 2\nu_1)$ such that the previous display is finite. Then, by the Borel-Cantelli lemma,

$$\sup_{\gamma,y,w} \left| \frac{1}{na_n} \sum_{Y_j \geq y} K\left(\frac{\gamma^T(W_j - w)}{a_n} \right) - \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n} \right) \right] \right| \longrightarrow_{a.s.} 0.$$

For any fixed γ, it can be shown that

$$\frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n} \right) \right] = E\left(I(Y \geq y) | \gamma^T W = \gamma^T w \right) f_{\gamma^T W}(\gamma^T w) + O(a_n^2),$$

where $f_{\gamma^T W}(\cdot)$ is the density function of $\gamma^T W$ and $O(a_n^2)$ does not depend on y and w. Hence for any given γ,

$$\sup_{y,w} \left| \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n} \right) \right] - E\left(I(Y \geq y) | \gamma^T W = \gamma^T w \right) f_{\gamma^T W}(\gamma^T w) \right| \longrightarrow 0.$$

Note that both $1/a_n E\left[I(Y \geq y) K\left(\frac{\gamma^T(W-w)}{a_n} \right) \right]$ and $E\left(I(Y \geq y) | \gamma^T W = \gamma^T w \right) f_{\gamma^T W}(\gamma^T w)$ are equi-continuous in γ. Hence,

$$\sup_{\gamma,y,w} \left| \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(W - w)}{a_n} \right) \right] - E\left(I(Y \geq y) | \gamma^T W = \gamma^T w \right) f_{\gamma^T W}(\gamma^T w) \right| \longrightarrow 0.$$

Therefore, we have proved that

$$\sup_{\gamma,y,w}\left|\frac{1}{na_n}\sum_{Y_j\geq y}K\left(\frac{\gamma^T(W_j-w)}{a_n}\right)-E\left(I(Y\geq y)|\gamma^TW=\gamma^Tw\right)f_{\gamma^TW}(\gamma^Tw)\right|\longrightarrow_{a.s.}0.$$

It then follows that

$$\sup_{\gamma}\left|(B)-\frac{1}{n}\sum_{j=1}^{n}\Delta_j\frac{1}{na_n}\sum_{i=1}^{n}\frac{I(Y_i\geq Y_j)K\left(\frac{\gamma^T(W_i-W_j)}{a_n}\right)}{E\left(I(Y\geq y)|\gamma^TW=\gamma^TW_i\right)|_{y=Y_j}f_{\gamma^TW}(\gamma^TW_i)}\right|\longrightarrow_{a.s.}0.$$

Similar arguments can be used to show that

$$\sup_{\gamma,y,w}\left|\frac{1}{na_n}\sum_{i=1}^{n}\frac{I(Y_i\geq y)K\left(\frac{\gamma^T(W_i-w)}{a_n}\right)}{E\left(I(Y\geq y)|\gamma^TW=\gamma^TW_i\right)f_{\gamma^TW}(\gamma^TW_i)}-\frac{1}{a_n}E\left[\frac{I(Y\geq y)K\left(\frac{\gamma^T(W-w)}{a_n}\right)}{E\left(I(Y\geq y)|\gamma^TW\right)f_{\gamma^TW}(\gamma^TW)}\right]\right|\longrightarrow_{a.s.}0.$$

Simple calculation shows that the second term inside the absolute value equals 1. Therefore,

$$\sup_{\gamma}\left|\frac{1}{n}\sum_{j=1}^{n}\Delta_j\frac{1}{na_n}\sum_{i=1}^{n}\frac{I(Y_i\geq Y_j)K\left(\frac{\gamma^T(W_i-W_j)}{a_n}\right)}{E\left(I(Y\geq y)|\gamma^TW=\gamma^TW_i\right)|_{y=Y_j}f_{\gamma^TW}(\gamma^TW_i)}-\frac{1}{n}\sum_{j=1}^{n}\Delta_j\right|\longrightarrow_{a.s.}0$$

and thus $\sup_\gamma|(B)-1/n\sum_{j=1}^{n}\Delta_j|\longrightarrow_{a.s.}0$.

Proof of Theorem 2.3.1

Following the proof for Lemma 2.5.1, we obtain

$$\sup_{\gamma}\left|\frac{1}{n}\sum_{i}\Delta_i\log\left(\frac{1}{na_n}\sum_{Y_j\geq Y_i}K\left(\frac{\gamma^T(W_j-W_i)}{a_n}\right)\right)-\frac{1}{n}\sum_{i}\Delta_i\log\left(E\left(I(Y\geq y)|\gamma^TW=\gamma^TW_i\right)|_{y=Y_i}f_{\gamma^TW}(\gamma^TW_i)\right)\right|\longrightarrow_{a.s.}0.$$

The second term inside the absolute value converges uniformly in γ to

$$E\left\{\Delta \log\left(E\left(I(Y \geq y)|\gamma^T W\right)|_{y=Y} f_{\gamma^T W}(\gamma^T W)\right)\right\},$$

since the involved class of functions is strong P-GC. Therefore,

$$\sup_{\gamma} \left|pl_n^{loc}(\gamma) - pl^{loc}(\gamma)\right| \to_{a.s.} 0.$$

Proof of Theorem 2.3.2

Note that $-\frac{\partial}{\partial\gamma}\big|_{\gamma=\gamma_0} pl^{loc}(\gamma)$ equals

$$\begin{aligned}
&E\left[\Delta \frac{\nabla_\gamma \left(E(I(Y \geq y)|\gamma^T W) f_{\gamma^T W}(\gamma^T W)\right)}{E(I(Y \geq y)|\gamma_0^T W) f_{\gamma_0^T W}(\gamma_0^T W)}\Big|_{y=Y}\right] \\
&\quad = E_W\left[\int \frac{\nabla_\gamma \left(E(I(Y \geq t)|\gamma^T W) f_{\gamma^T W}(\gamma^T W)\right)}{E(I(Y \geq t)|\gamma_0^T W) f_{\gamma_0^T W}(\gamma_0^T W)} \lambda_0(t, \gamma_0^T W) G_C(t) \exp\left(-\Lambda_0(t, \gamma_0^T W)\right) dt\right] \\
&\quad = E_W\left[\int \frac{\nabla_\gamma \left(E(I(Y \geq t)|\gamma^T W) f_{\gamma^T W}(\gamma^T W)\right)}{f_{\gamma_0^T W}(\gamma_0^T W)} \lambda_0(t, \gamma_0^T W) dt\right] \\
&\quad = \iint \frac{\lambda_0(y, \gamma_0^T w) f_W(w)}{f_{\gamma_0^T W}(\gamma_0^T w)} \nabla_\gamma \left(E(I(Y \geq y)|\gamma^T W = \gamma^T w) f_{\gamma^T W}(\gamma^T w)\right) dy dw,
\end{aligned}$$

where $G_C(\cdot)$ denotes the survival function of C. The quantity inside the gradient operator can be written as

$$\lim_{h\to 0} \frac{1}{h} E\left[I(Y \geq y) K\left(\frac{\gamma^T W - \gamma^T w}{h}\right)\right] = \lim_{h\to 0} \frac{1}{h} E_W\left[K\left(\frac{\gamma^T W - \gamma^T w}{h}\right) g(y, \gamma_0^T W)\right],$$

where $g(y, u) = G_C(y) \exp\left(-\Lambda_0(y, u)\right)$. Thus

$$\begin{aligned}
-\frac{\partial}{\partial\gamma}\Big|_{\gamma_0} pl^{loc}(\gamma) \;=\; & \iint \frac{\lambda_0(y, \gamma_0^T w) f_W(w)}{f_{\gamma_0^T W}(\gamma_0^T w)} \\
& \times \lim_{h\to 0} E_{\gamma_0^T W}\left[\frac{1}{h^2} K'\left(\frac{\gamma_0^T W - \gamma_0^T w}{h}\right)\left(E(W|\gamma_0^T W) - w\right) g(y, \gamma_0^T W)\right] dy dw.
\end{aligned}$$

Let $r(u) \equiv E(W|\gamma_0^T W = u)$, then the limit inside of the integral is

$$\begin{aligned}
&\lim_{h\to 0} \int \frac{1}{h^2} K'\left(\frac{u-\gamma_0^T w}{h}\right)(r(u)-w)g(y,u)f_{\gamma_0^T W}(u)du \\
&\quad = -g_2'(y,\gamma_0^T w)f_{\gamma_0^T W}(\gamma_0^T w)r(\gamma_0^T w) - g(y,\gamma_0^T w)f'_{\gamma_0^T W}(\gamma_0^T w)r(\gamma_0^T w) \\
&\qquad - g(y,\gamma_0^T w)f_{\gamma_0^T W}(\gamma_0^T w)r'(\gamma_0^T w) + g_2'(y,\gamma_0^T w)f_{\gamma_0^T W}(\gamma_0^T w)w + g(y,\gamma_0^T w)f'_{\gamma_0^T W}(\gamma_0^T w)w,
\end{aligned}$$

where $g_2'(y,u) = \frac{\partial}{\partial u} g(y,u)$. Hence, the double integral equals

$$\begin{aligned}
&\int E_W\Bigg[\lambda_0(y,\gamma_0^T W)\Bigg(- g_2'(y,\gamma_0^T W)r(\gamma_0^T W) - g(y,\gamma_0^T W)\frac{f'_{\gamma_0^T W}}{f_{\gamma_0^T W}}(\gamma_0^T W)r(\gamma_0^T W) \\
&\qquad\qquad - g(y,\gamma_0^T W)r'(\gamma_0^T W) + g_2'(y,\gamma_0^T W)W + g(y,\gamma_0^T W)\frac{f'_{\gamma_0^T W}}{f_{\gamma_0^T W}}(\gamma_0^T W)W\Bigg)\Bigg]dy \\
&\quad = -\int E_{\gamma_0^T W}\left(\lambda_0(y,\gamma_0^T W)g(y,\gamma_0^T W)r'(\gamma_0^T W)\right)dy.
\end{aligned}$$

Since $W \sim N(\mu,\Sigma)$, $r'(u) = (\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0$ for any u. Thus the last display becomes

$$-(\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0 \int E_{\gamma_0^T W}\left(\lambda_0(y,\gamma_0^T W)g(y,\gamma_0^T W)\right)dy = -E[\Delta](\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0.$$

By (C6), $E[\Delta] > 0$. Thus the display is proportional to γ_0 if and only if $\Sigma\gamma_0 \propto c\gamma_0$ for some constant c. This completes the proof.

Chapter 3

Partly Proportional Single-Index Hazards Model

3.1 Model and Data Structure

In this chapter, we assume the following partly proportional single-index hazards (PPSIH) model

$$h(t|X,Z) = \lambda(t, \gamma^T Z) e^{\beta^T X}, \tag{3.1}$$

where $\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^q$ and $\lambda(\cdot,\cdot)$ is an unknown bivariate function. When γ is a constant and Z is one dimensional, (3.1) reduces to a special case of the model studied by Dabrowska (1997). When β is zero or there is no X, (3.1) reduces to the single-index model proposed in Chapter 2. To ensure identifiability of this model, we first impose the restriction that $\|\gamma\| = 1$ with the last coordinate γ_q of γ positive, that is, γ is restricted to the half unit sphere. This assumption is practically reasonable when at least one "nuisance" covariate has a non-zero effect. Other identifiability and regularity conditions are given in Sections 3.2 and 3.3.

Suppose we observe a random sample of size n, $(Y_i = T_i \wedge C_i, \Delta_i, X_i, Z_i), i = 1, \cdots, n$, where T is the survival time, C is the censoring time, $a \wedge b = \min(a,b)$, $\Delta = I(T \leq$

C) is the censoring indicator, X is the covariate vector of main interest and Z is the "nuisance" covariate vector. The subscript i denotes the i^{th} subject. We make the standard assumptions that C is independent of T conditional on (X, Z) and that the distribution of C does not depend on the parameters (i.e. non-informative censoring). The log-likelihood function is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left(\log\lambda(Y_i,\gamma^T Z_i)+\beta^T X_i\right)-e^{\beta^T X_i}\Lambda(Y_i,\gamma^T Z_i)\right]. \tag{3.2}$$

This function has the maximum value of infinity, thus it cannot be used directly for parameter estimation. Instead, in the setting of nonparametric maximum likelihood estimation (NPMLE), we maximize

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left(\log\Lambda\{Y_i,\gamma^T Z_i\}+\beta^T X_i\right)-e^{\beta^T X_i}\sum_{Y_j\le Y_i}\Lambda\{Y_j,\gamma^T Z_i\}\right]. \tag{3.3}$$

Here, $\Lambda\{Y_i, u\}$ is the jump size of the "baseline" cumulative hazard function $\Lambda(t, u)$ at Y_i for fixed u. However, the profile likelihood function based on (3.3) obtained by profiling out $\Lambda\{\cdot,\cdot\}$ is a constant, thus it is not a valid objective function. In the next section, we consider a commonly used estimation method, namely the profile local likelihood approach, for model estimation.

3.2 Profile Local Likelihood

3.2.1 Method

In a semiparametric model, a local likelihood using kernel smoothing is frequently constructed to estimate the unspecified function and then the estimated function is plugged into the likelihood (or some variant of the likelihood) to obtain the profile likelihood function. This conventional profile-kernel method is adopted by Dabrowska (1997) and

Fan et al. (1997). Specifically for our likelihood (3.3) and fixed (β, γ), we estimate $\Lambda\{\cdot, \cdot\}$ by maximizing the following local likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left(\log\Lambda\{Y_i, u\} + \beta^T X_i\right) - e^{\beta^T X_i}\sum_{Y_j \le Y_i}\Lambda\{Y_j, u\}\right] K_{a_n}(\gamma^T Z_i - u),$$

where $K_{a_n}(t) = K(t/a_n)/a_n$, $K(\cdot)$ is a mean zero symmetric density function and $\Lambda\{Y_i, u\}$ is the jump size of $\Lambda(t, u)$ at Y_i for each u. This is the local constant fit weighted by the function $K_{a_n}(\cdot)$. The maximizer can be found as

$$\hat{\Lambda}\{Y_i, u\} = \frac{\Delta_i K_{a_n}(\gamma^T Z_i - u)}{\sum_{Y_j \ge Y_i} e^{\beta^T X_j} K_{a_n}(\gamma^T Z_j - u)}. \tag{3.4}$$

We plug $\hat{\Lambda}\{Y_i, u\}$ into (3.3) to obtain, up to a constant, the profile likelihood

$$pl_n^{loc}(\beta, \gamma) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{na_n}e^{\beta^T X_i}\sum_{Y_j \le Y_i}\frac{\Delta_j K\left(\frac{\gamma^T(Z_j - Z_i)}{a_n}\right)}{\frac{1}{na_n}\sum_{Y_k \ge Y_j} e^{\beta^T X_k} K\left(\frac{\gamma^T(Z_k - Z_i)}{a_n}\right)}, \tag{3.5}$$

where

$$pl_n^{loc}(\beta, \gamma) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i\beta^T X_i - \frac{1}{n}\sum_{i=1}^{n}\Delta_i \log\left(\frac{1}{na_n}\sum_{Y_j \ge Y_i} e^{\beta^T X_j} K\left(\frac{\gamma^T(Z_j - Z_i)}{a_n}\right)\right).$$

We will show in Lemma 3.5.1 in Section 3.5 that under some regularity conditions, the second term of (3.5) converges uniformly in β and γ to $1/n\sum_i \Delta_i$, which is a constant with respect to (β, γ). As a result, the estimator for (β, γ) is obtained by maximizing the profile local likelihood function $pl_n^{loc}(\beta, \gamma)$. Note that this function is smooth in both β and γ. Thus numerically it can be easily maximized. For example, the quasi-Newton search algorithm can be used.

3.2.2 Bias Analysis

Let (β_0, γ_0) denote the true parameter value of (β, γ). We impose the following regularity conditions:

(C1) $\beta_0 \in \mathcal{B}, \gamma_0 \in \Gamma$, where $\mathcal{B} \subset \mathbb{R}^p, \Gamma \subset \mathbb{R}^q$ are compact.

(C2) The random vector Z has a continuous positive density on its support.

(C3) The kernel function $K(\cdot)$ has mean zero with finite second moment. Moreover, $\sup_x |K'(x)|$ is finite, where $K'(x)$ denotes the derivative function of $K(x)$.

Remark 3.2.1. Many kernel functions satisfy condition (C3), for example, the standard Gaussian kernel $K(u) = 1/\sqrt{2\pi}\exp(-u^2/2)$ and the Epanechnikov kernel $K(u) = 3/4(1-u^2)I(|u| \leq 1)$.

Theorem 3.2.1. *If conditions (C1)–(C3) hold and let* $a_n = n^{\nu_1}$ *with* $\nu_1 \in (-1/2, 0)$, *then* $\sup_{\beta,\gamma} |pl_n^{loc}(\beta, \gamma) - pl^{loc}(\beta, \gamma)| \to_{a.s.} 0$, *where*

$$pl^{loc}(\beta, \gamma) = E\left[\Delta\left(\beta^T X + \log \frac{1}{E\left(I(Y \geq y)e^{\beta^T X} | \gamma^T Z\right)|_{y=Y} f_{\gamma^T Z}(\gamma^T Z)}\right)\right].$$

Here $f_{\gamma^T Z}(\cdot)$ *is the density function of* $\gamma^T Z$.

Thus the profile local likelihood estimator should converge to the maximizer of $pl^{loc}(\beta, \gamma)$ almost surely by Theorem 2.12 of Kosorok (2008). Suppose the latter is the true parameter (β_0, γ_0), then the partial derivative of $pl^{loc}(\beta, \gamma)$ with respect to β should be zero if evaluated at (β_0, γ_0) and the partial derivative of $pl^{loc}(\beta, \gamma)$ with respect to γ should be proportional to γ_0 if evaluated at (β_0, γ_0). The latter is due to the restriction $\|\gamma\| = 1$. However, we show in Theorem 3.2.2 that the true parameter (β_0, γ_0) may not maximize $pl^{loc}(\beta, \gamma)$.

Theorem 3.2.2. *Suppose* C *is independent of* (X, Z), $Z \sim N(\mu, \Sigma)$ *with* $\Sigma > 0$ *and is independent of* X, *then*

(i) $\frac{\partial}{\partial\beta}\big|_{\beta=\beta_0,\gamma=\gamma_0} pl^{loc}(\beta,\gamma)=0.$

(ii) $\frac{\partial}{\partial\gamma}\big|_{\beta=\beta_0,\gamma=\gamma_0} pl^{loc}(\beta,\gamma)\propto\gamma_0$ *if and only if* $\Sigma\gamma_0=c\gamma_0$ *for some constant* c.

Remark 3.2.2. This theorem shows that the profile local likelihood approach only works in very special cases in view of the requirement $\Sigma\gamma_0 = c\gamma_0$ and the independence assumption between X and Z. Thus, in the more general set-up, $pl^{loc}(\beta,\gamma)$ may not be maximized at (β_0,γ_0). Consequently, the profile local likelihood estimator is not in general consistent.

In addition to the theoretical investigation of the asymptotic bias, we conduct simulation studies to examine the numerical performance of the profile local likelihood estimator. For simplicity, we assume that X is one dimensional, $Z=(Z_1,Z_2)$ is two dimensional and the censoring time C is generated from the uniform $[0,10]$ distribution. Jointly, (X,Z) is generated from a multivariate normal distribution with zero means and unit variances. The true parameters are $\beta_0=1,\gamma_0=(1/2,\sqrt{3}/2)$. The following four simulation settings are considered: (i) $\lambda_0(t,u)=0.5e^u t$, X and Z are independent, $cov(Z_1,Z_2)=0$; (ii) $\lambda_0(t,u)=0.5e^u t$, X and Z are independent, $cov(Z_1,Z_2)=0.5$; (iii)$\lambda_0(t,u)=0.5e^u t$ and the covariance between any pair of $\{X,Z_1,Z_2\}$ is 0.5; (iv) $\lambda_0(t,u)=0.25e^{e^u t}$ and we use the same covariance matrix as in setting (iii). Note that in setting (i)–(iii), the proportional hazards assumption is satisfied, but this assumption is violated in setting (iv). The censoring rate ranges approximately from 19% to 25%.

We choose the kernel function to be the standard normal density. Parameters are estimated by using the quasi-Newton search algorithm in the R software package. The initial values are set to zero. The bandwidth is chosen to be $c_1\times IQR_1\times n^{\nu}$, where $c_1\in\{0.5,1,1.5,2\}$, $\nu\in\{-1/4,-1/5\}$ and IQR_1 is the inter-quartile range of $\|Z\|$ in each simulated dataset. For each simulation setting, the tuning parameters c_1 and ν which give the smallest bias when $n=400$ are chosen and then the same values are used for the cases $n=100$ and $n=200$.

Table 3.1 summarizes the simulation results in setting (i)–(iv) with sample sizes 100, 200 and 400, where γ_1 is the first coordinate of the γ vector. As expected from Theorem 3.2.2, the profile local likelihood method works in setting (i) but fails in settings (ii) due to the non-zero correlation between Z_1 and Z_2 together with the fact that γ_0 is not an eigenvector of $cov(Z)$. Our simulation results in setting (iii) and (iv) show that this method also fails when X and Z are dependent in the presence of non-zero correlation between Z_1 and Z_2. For comparison, results from the Cox model are also presented. The Cox model works very well in setting (i)–(iii) as we expect, but it gives asymptotically biased estimators in setting (iv) because of the violation of the proportional hazards assumption.

Figure 3.1 shows the profile likelihood curve of γ_1, the first coordinate of γ, in each simulation setting based on a simulated dataset with $n = 5000$. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The bandwidth is $2 \times n^{-1/4}$. It is observed again that this approach leads to asymptotically biased estimation except in case (i).

3.3 Corrected Profile Local Likelihood

3.3.1 Method

Since the profile local likelihood method generates biased results in general, this section focuses on the bias correction. The idea is to construct a "baseline hazard function" $\tilde{\lambda}(t, z; \beta, \gamma)$ such that the corrected asymptotic limit, $cpl^{loc}(\beta, \gamma)$, can be written as

$$E\left[\Delta\left(\beta^T X + \log \tilde{\lambda}(Y, Z; \beta, \gamma)\right) - e^{\beta^T X}\tilde{\Lambda}(Y, Z; \beta, \gamma)\right]$$

with $\tilde{\lambda}(t, z; \beta_0, \gamma_0) = \lambda_0(t, \gamma_0^T z)$ and $\tilde{\Lambda}(t, z; \beta, \gamma) = \int_0^t \tilde{\lambda}(v, z; \beta, \gamma)dv$. It then follows that $cpl^{loc}(\beta, \gamma)$ is maximized at the true parameter value (β_0, γ_0) by the non-negativity of

Table 3.1: Simulation results of local likelihood in PPSIH model

			Local likelihood		Cox model	
Simulation settings	Sample size	Parameters	Bias	SE	Bias	SE
(i) $X \perp Z$	100	β	-.113	.161	.030	.163
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.264	.440	.015	.134
$cov(Z) = 0$	200	β	-.091	.113	.018	.105
		γ_1	-.150	.358	.009	.090
	400	β	-.067	.077	.006	.072
		γ_1	-.045	.221	.005	.063
(ii) $X \perp Z$	100	β	-.052	.173	.028	.163
$\lambda_0(t,u) = 0.5e^u t$		γ_1	.093	.192	.014	.154
$cov(Z) = 0.5$	200	β	-.050	.110	.018	.104
		γ_1	.128	.125	.010	.103
	400	β	-.045	.076	.006	.072
		γ_1	.146	.078	.006	.072
(iii) $X \not\perp Z$	100	β	-.020	.206	.027	.193
$\lambda_0(t,u) = 0.5e^u t$		γ_1	.092	.231	.014	.164
$cov(Z) = 0.5$	200	β	-.031	.131	.019	.123
		γ_1	.144	.153	.008	.109
	400	β	-.038	.086	.006	.083
		γ_1	.167	.102	.005	.077
(iv) $X \not\perp Z$	100	β	-.045	.194	-.125	.163
$\lambda_0(t,u) = 0.25e^{e^u t}$		γ_1	.091	.204	-.048	.147
$cov(Z) = 0.5$	200	β	-.051	.123	-.139	.103
		γ_1	.137	.138	-.044	.104
	400	β	-.054	.083	-.151	.073
		γ_1	.151	.096	-.041	.070

NOTE: Each entry is based on 1000 replicates.

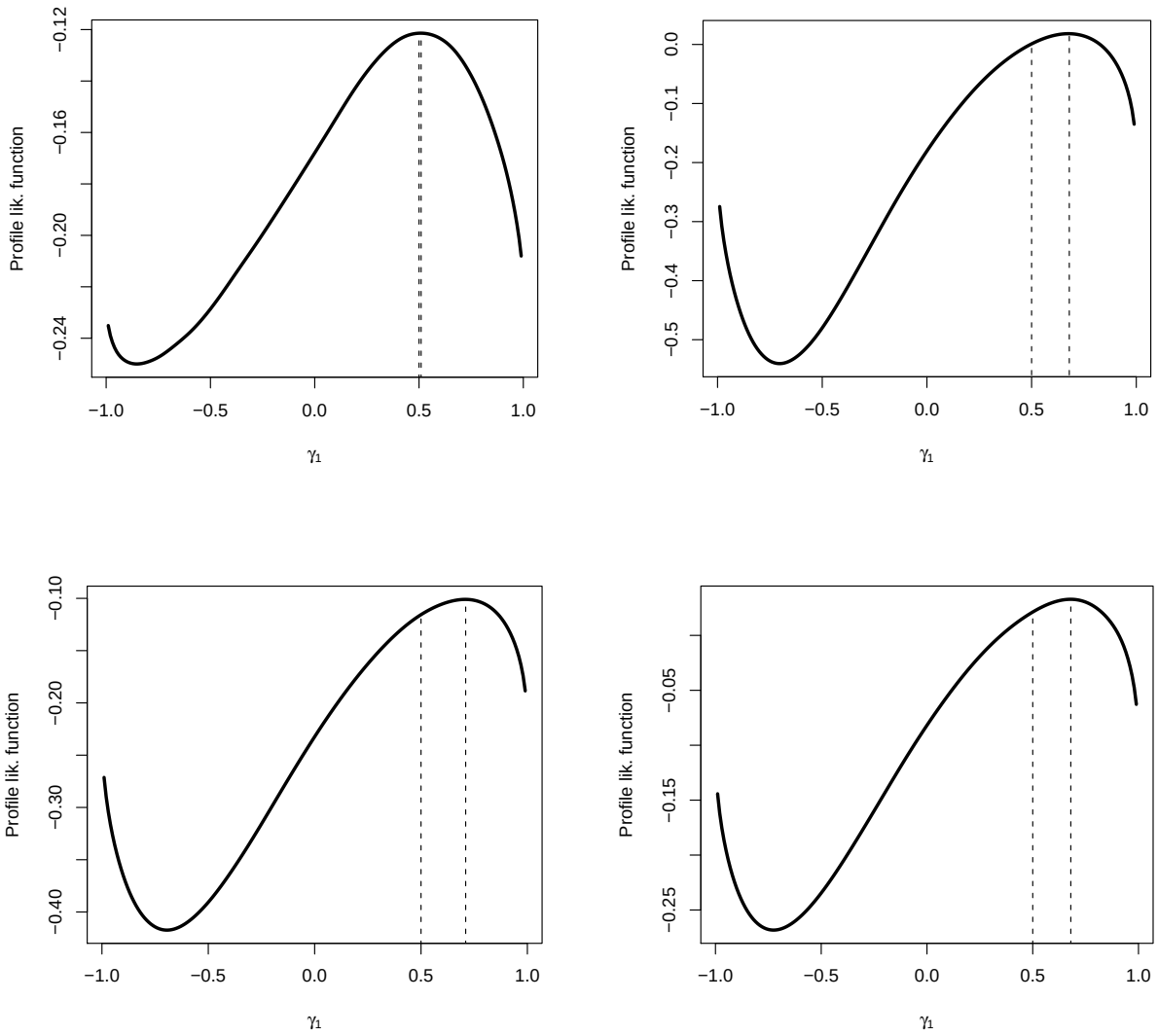

Figure 3.1: Profile local likelihood curve of γ_1 in PPSIH model

Kullback-Leibler information together with the identifiability conditions.

Consider

$$\tilde{\lambda}(t,z;\beta,\gamma) \equiv \frac{\frac{d}{dt}E\left(I(\Delta=1,Y\le t)|\gamma^T Z=\gamma^T z\right)}{E\left(I(Y\ge t)e^{\beta^T X}|\gamma^T Z=\gamma^T z\right)}.$$

With this "baseline hazard function", we will show in the proof of Theorem 3.3.1 that $\tilde{\lambda}(t,z;\beta_0,\gamma_0)=\lambda_0(t,\gamma_0^T z)$ and that $E[e^{\beta^T X}\tilde{\Lambda}(Y,Z;\beta,\gamma)]=E[\Delta]$ for any β and γ. Therefore, up to a constant ($E[\Delta]$), the corrected asymptotic limit is

$$cpl^{loc}(\beta,\gamma)=E\left[\Delta\left(\beta^T X+\log\frac{\frac{d}{dy}E\left(I(\Delta=1,Y\le y)|\gamma^T Z\right)}{E\left(I(Y\ge y)e^{\beta^T X}|\gamma^T Z\right)}\Big|_{y=Y}\right)\right].$$

Note that the difference between $cpl^{loc}(\beta,\gamma)$ and $pl^{loc}(\beta,\gamma)$ is

$$E\left[\Delta\log\left(\frac{d}{dy}\Big|_{y=Y}E\left(I(\Delta=1,Y\le y)|\gamma^T Z\right)f_{\gamma^T Z}(\gamma^T Z)\right)\right],$$

which can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i\log\left[\frac{1}{na_nb_n}\sum_{j=1}^{n}\Delta_j K\left(\frac{Y_j-Y_i}{b_n}\right)K\left(\frac{\gamma^T(Z_j-Z_i)}{a_n}\right)\right]$$

uniformly in β and γ, where a_n and b_n are bandwidth parameters. Hence, the corrected profile local likelihood function is

$$\begin{aligned}cpl_n^{loc}(\beta,\gamma)=&pl_n^{loc}(\beta,\gamma)\\&+\frac{1}{n}\sum_{i=1}^{n}\Delta_i\log\left[\frac{1}{na_nb_n}\sum_{j=1}^{n}\Delta_j K\left(\frac{Y_j-Y_i}{b_n}\right)K\left(\frac{\gamma^T(Z_j-Z_i)}{a_n}\right)\right].\end{aligned}$$

We denote its point of maximum as $(\hat{\beta}_n,\hat{\gamma}_n)$. Given $(\hat{\beta}_n,\hat{\gamma}_n)$, we propose to estimate $\Lambda(t,u)$ by

$$\hat{\Lambda}(t,u)=\sum_{Y_i\le t}\frac{\Delta_i K_{a_n}(\hat{\gamma}_n^T Z_i-u)}{\sum_{Y_j\ge Y_i}e^{\hat{\beta}_n^T X_j}K_{a_n}(\hat{\gamma}_n^T Z_j-u)}. \tag{3.6}$$

Note that this is exactly the baseline hazard estimator in Section 3.2.

3.3.2 Asymptotic Results

In addition to conditions (C1)–(C3), we further impose the following regularity conditions:

(C4) $\lambda_0(t,u)$ has a non-zero partial derivative function with respect to u for some t.

(C5) For any scalar α_1 and constant vector α_2 satisfying $\alpha_1 + \alpha_2^T X = 0$ with probability 1, we must have $\alpha_1 = 0$ and $\alpha_2 = 0$. Furthermore, the support of Z given X contains 0.

Remark 3.3.1. Conditions (C4) and (C5) are used for model identifiability. Note that $\partial/\partial u \lambda_0(t,u) = 0$ implies $\lambda_0(t,u)$ is constant in u and thus Z has no effect on the hazard function. Therefore, assuming (C4) is not unreasonable.

Theorem 3.3.1. *Under conditions (C1)–(C5), let* $a_n = n^{\nu_1}, b_n = n^{\nu_2}$ *with* $\nu_1, \nu_2 \in (-1/2, 0)$, *then* $\hat{\beta}_n \to \beta_0$ *and* $\hat{\gamma}_n \to \gamma_0$ *almost surely.*

Theorem 3.3.2. *Under conditions (C1)–(C5), let* $a_n = n^{\nu_1}, b_n = n^{\nu_2}$ *with* $\nu_1, \nu_2 \in (-1/2, -1/4)$, *then* $\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{\gamma}_{n,(-q)} - \gamma_{0,(-q)})$ *converges weakly to a zero mean Gaussian distribution, where* $\gamma_{0,(-q)}$ *is the true parameter value* γ_0 *with the last component deleted. Furthermore,* $(\hat{\beta}_n, \hat{\gamma}_n)$ *is semiparametrically efficient.*

Remark 3.3.2. The covariance matrix for $(\hat{\beta}_n, \hat{\gamma}_{n,(-q)})$ can be consistently estimated based on a plug-in estimator of the efficient score function. See the proof of Theorem 3.3.2 for details.

Theorem 3.3.3. *Under the conditions of Theorem 3.3.1,* $\sup_{t,u} |\hat{\Lambda}(t,u) - \Lambda_0(t,u)| \to_{a.s.} 0$.

Remark 3.3.3. Although plugging (3.6) into (3.3) yields biased estimation for (β, γ), it is interesting to observe that (3.6) is in fact consistent provided that $(\hat{\beta}_n, \hat{\gamma}_n)$ is consistent.

3.3.3 Simulation Studies

In this section, we consider the same simulation settings studied in Section 3.2, but now with the corrected profile local likelihood method.

We again choose the kernel function to be the standard normal density and the parameters (β, γ) are estimated by using the quasi-Newton search algorithm in the R software package. The initial values are set to zero. The results summarized in Table 3.2 are based on the bandwidths $c_i \times IQR_i \times n^{-1/3}$ for point estimation and $d_i \times IQR_i \times n^{-1/4}$ for variance estimation, where c_i and d_i are tuning parameters and IQR_1 and IQR_2 are the inter-quartile ranges of $\|Z\|$ and Y, respectively. According to our experience, selection of the tuning parameters c_i and d_i is data-adaptive and thus they can vary case by case and the bandwidths for variance estimation need to be larger than those for point estimation. The following values of (c_1, c_2, d_1, d_2) are used: In setting (i), $(c_1, c_2, d_1, d_2) = (1.5, 4, 3, 8)$; In setting (ii), $(c_1, c_2, d_1, d_2) = (1.5, 2, 3, 8)$; In setting (iii), $(c_1, c_2, d_1, d_2) = (2, 2.5, 3.5, 8)$; In setting (iv), $(c_1, c_2, d_1, d_2) = (1, 2, 3, 8)$; These simulation results suggest that the corrected profile local likelihood method works well under every simulation setting.

The profile local likelihood curve after the bias correction (dashed curve) in each simulation setting based on a simulated dataset of size 5000 is plotted in Figure 3.2. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The bandwidth is $2 \times n^{-1/4}$. In each case, the maximizer of the corrected profile local curve is around the true value 0.5 of γ_1, suggesting that the corrected method gives estimators with very little bias. For comparison, the profile local likelihood curves before the bias correction (solid curves) are also plotted.

Table 3.2: Simulation results of corrected local likelihood in PPSIH model

			Corrected Local likelihood			
Simulation settings	Sample size	Parameters	Bias	SE	SEE	CP
(i) $X \perp Z$	100	β	-.062	.156	.155	.911
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.077	.197	.208	.932
$cov(Z) = 0$	200	β	-.046	.102	.106	.929
		γ_1	-.028	.124	.125	.954
	400	β	-.037	.072	.073	.926
		γ_1	-.009	.083	.078	.937
(ii) $X \perp Z$	100	β	-.071	.156	.157	.918
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.060	.214	.255	.925
$cov(Z) = 0.5$	200	β	-.054	.103	.108	.920
		γ_1	-.017	.157	.160	.927
	400	β	-.042	.073	.075	.918
		γ_1	-.001	.108	.103	.921
(iii) $X \not\perp Z$	100	β	.025	.191	.196	.953
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.099	.288	.282	.868
$cov(Z) = 0.5$	200	β	-.006	.122	.131	.963
		γ_1	-.035	.173	.184	.938
	400	β	-.011	.087	.091	.957
		γ_1	-.017	.119	.118	.935
(iv) $X \not\perp Z$	100	β	-.030	.184	.185	.936
$\lambda_0(t,u) = 0.25e^{e^u t}$		γ_1	-.150	.255	.279	.886
$cov(Z) = 0.5$	200	β	-.038	.119	.124	.945
		γ_1	-.053	.174	.175	.927
	400	β	-.039	.083	.085	.935
		γ_1	-.005	.113	.109	.930

NOTE: Each entry is based on 1000 replicates.

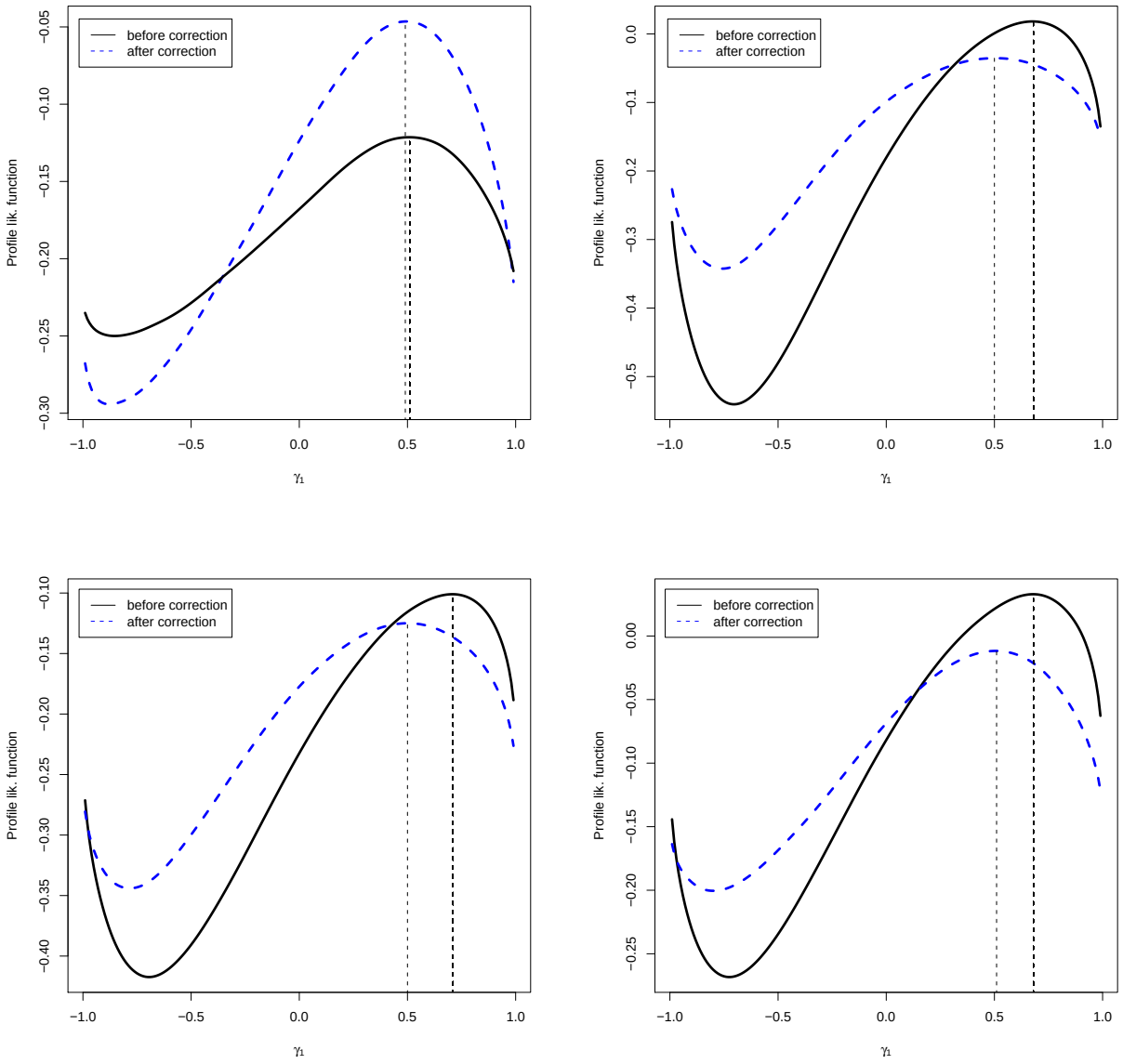

Figure 3.2: Profile local likelihood curves (corrected and uncorrected) of γ_1 in PPSIH model

3.4 Data Application

We consider a dataset from the Multicenter AIDS Cohort Study (MACS), which contains survival times and 6 covariates for 471 HIV positive men. The survival time is the time to death due to AIDS. The 6 covariates include ethnicity (white or non-white), baseline age, viral loads, CD4 positive cell counts, serum β_2-microgloburin levels and serum neopterin levels. Note that the latter 4 covariates are the most predictive prognostic factors identified in Mellors et al. (1997) and the baseline is defined to be one year after recruitment. The censoring rate in this dataset is about 27%.

In AIDS studies, one might be interested in assessing the effect of patient's ethnicity, baseline age, viral loads and CD4 counts on the risk of death due to AIDS, controlling for serum β_2-microgloburin levels and serum neopterin levels. One might also wish to assess in particular the effect of patient's ethnicity and baseline age on the risk of death due to AIDS, controlling for the remaining 4 biomarkers. Both questions can be addressed by fitting either the Cox model or our partly proportional single-index hazards (PPSIH) model. Before the model fitting, we take the logarithmic transformation of covariates viral loads and CD4 positive cell counts and then standardize all continuous covariates (log(viral load), log(CD4 counts), neopterin, microgloburin and baseline age) so that they have zero means and unit variances.

The results from fitting the Cox proportional hazards model are presented in Table 3.3. Note, however, that testing for the proportional hazards assumption based on martingale residuals (Lin et al. 1993) reveals that the covariate log(viral load) does not satisfy this assumption ($p = .006$). Thus the inference based on the Cox model may not be valid.

We now fit the proposed partly proportional single-index hazards (PPSIH) model using the corrected profile local likelihood method with the standard normal kernel function to address the first aforementioned research question. Covariates ethnicity, baseline age, viral loads and CD4 counts are treated as covariates of interest and the remaining 2 are treated as "nuisance" covariates. The bandwidths are $c_i \times IQR_i \times n^{-1/3}$ for point

estimation and $d_i \times IQR_i \times n^{-1/4}$ for variance estimation, where c_i is selected using cross-validation, $i = 1, 2$. For this dataset, we choose $c_1 = 2.5$ and $c_2 = 2.5$. For simplicity, we set $d_i = c_i$, $i = 1, 2$. The results from our model are summarized in Table 3.3 (columns labeled with "PPSIH model (1)"). Next, we investigate whether the proportional hazards assumption holds for the covariates of primary interest since this is implicitly assumed by our model. In view of the similarity of our model to a stratified Cox model, this can be done easily by fitting a stratified Cox model in SAS (version 9.2), where we use 2 strata based on the values of $\hat{\gamma}_n^T Z$. It is found that covariates log(viral load) and baseline age violate the proportional hazards assumption, with p-values equal to .017 and .019, respectively. Thus the inference based on our modified partly proportional single-index hazards model may not be valid either.

To address the second research question and the aforementioned issue using the proposed partly proportional single-index hazards model, we treat patient's baseline age and ethnicity as covariates of interest and the remaining 4 covariates as potential confounders. Again, the bandwidths are $c_i \times IQR_i \times n^{-1/3}$ for point estimation and $d_i \times IQR_i \times n^{-1/4}$ for variance estimation, where c_i is selected using cross-validation and we set $d_i = c_i$ for simplicity, $i = 1, 2$. We choose $c_1 = 2$ and $c_2 = 3$. The results are shown in Table 3.3 (columns labeled with "PPSIH model (2)"). Using the same model checking technique, we find that both covariates of interest do not seem to violate the proportional hazards assumption, with p-values equal to .119 and .607, respectively. It is interesting to note that the covariate white becomes non-significant under our model while it is significant under the Cox model. This suggests that the apparent initial significance of the white covariate may be spurious and attributable to model misspecification in this instance. The cumulative baseline hazard function plotted in Figure 3.3 also suggests that the Cox model may not be appropriate. We also conclude by using our model that the baseline age does not have any effect on the survival time and that the "nuisance" covariates sorted by relative importance are viral loads, serum β_2-microgloburin levels, CD4 positive cell

Table 3.3: Analysis of MACS Data under PPSIH Model (1), (2) and Cox Model

	PPSIH Model (1)			PPSIH Model (2)			Cox Model		
Parameter	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
age	.015	.064	.815	.011	.068	.871	.021	.053	.692
white	.596	.198	.003	.409	.263	.120	.647	.235	.006
log(CD4)	-.168	.063	.008	-.232	.119	.051	-.204	.056	<.001
log(viral)	.654	.055	<.001	.933	.041	<.001	.657	.064	<.001
microgloburin	.655	.057	<.001	.267	.115	.020	.100	.053	.059
neopterin	.756	.049	<.001	.069	.106	.515	.092	.055	.094

NOTE: "white" is an indicator for whites. Est. and SE denote the parameter estimate and (estimated) standard error, respectively.

counts and serum neopterin levels, in that order.

3.5 Proofs of Theorems

We denote the second term of the profile likelihood function (3.5) by (A).

Lemma 3.5.1. *If conditions (C1)–(C3) hold and let* $a_n = n^{\nu_1}$ *with* $\nu_1 \in (-1/2, 0)$, *then* $\sup_{\beta,\gamma} |(A) - 1/n \sum_{i=1}^{n} \Delta_i| \to_{a.s.} 0$.

Proof of Lemma 3.5.1

We partition $\mathcal{B}$ into small cubes such that any two points in the same cube have distance no large than α_n to be determined later. The number of partitions, denoted by m_n, is of order $1/\alpha_n^p$. Choose one arbitrary point from each of these partitions and denote them as $\beta^{(1)}, \ldots, \beta^{(m_n)}$. Similarly, we partition Γ into small cubes such that any two points in the same cube have distance no large than δ_n to be determined later. The number of partitions, denoted by m_n^*, is of order $1/\delta_n^q$. Choose one arbitrary point from the each of these partitions and denote them as $\gamma^{(1)}, \ldots, \gamma^{(m_n^*)}$. For β_1 and β_2 in the same

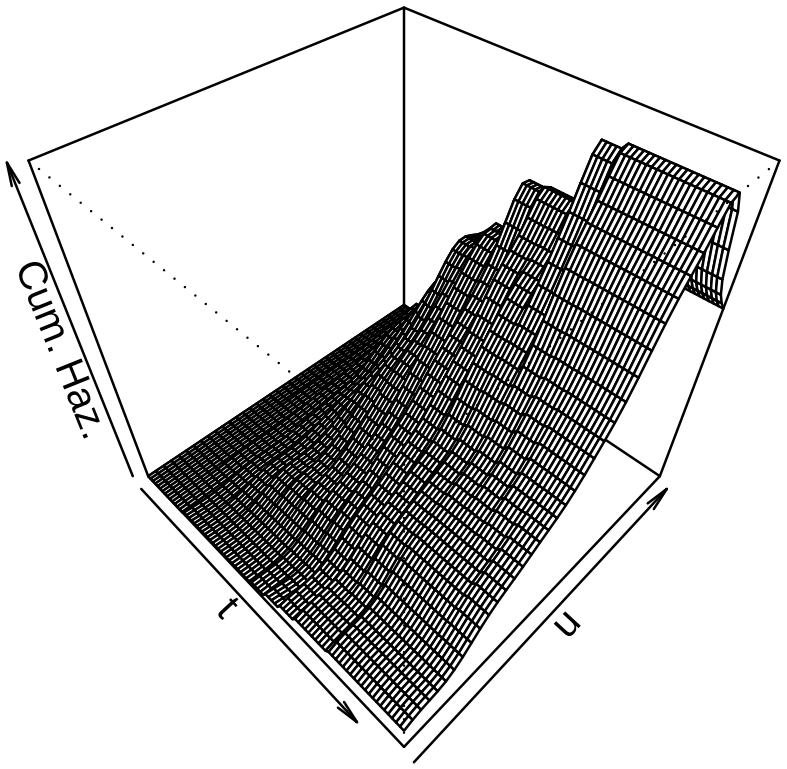

Figure 3.3: Cumulative baseline hazard estimator $\hat{\Lambda}(t, u)$ under PPSIH model (2).

cube and for any fixed y, z and $\gamma \in \Gamma$,

$$\begin{aligned}
&\left| \frac{1}{na_n} \sum_{Y_j \geq y} e^{\beta_1^T X_j} K\left(\frac{\gamma^T(Z_j - z)}{a_n}\right) - \frac{1}{na_n} \sum_{Y_j \geq y} e^{\beta_2^T X_j} K\left(\frac{\gamma^T(Z_j - z)}{a_n}\right) \right| \\
&\qquad \leq \frac{1}{na_n} \sum_j I(Y_j \geq y) K\left(\frac{\gamma^T(Z_j - z)}{a_n}\right) \left| e^{\beta_1^T X_j} - e^{\beta_2^T X_j} \right| \\
&\qquad \leq \frac{c}{a_n} \|\beta_1 - \beta_2\|
\end{aligned}$$

by condition (C3). Also,

$$\begin{aligned}
&\left| \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta_1^T X} K\left(\frac{\gamma^T(Z - z)}{a_n}\right) \right] - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta_2^T X} K\left(\frac{\gamma^T(Z - z)}{a_n}\right) \right] \right| \\
&\qquad \leq \frac{1}{a_n} E\left[I(Y \geq y) K\left(\frac{\gamma^T(Z - z)}{a_n}\right) \left| e^{\beta_1^T X} - e^{\beta_2^T X} \right| \right] \\
&\qquad \leq \frac{c_0}{a_n} \|\beta_1 - \beta_2\| E\left[K\left(\frac{\gamma^T(Z - z)}{a_n}\right) \right] \\
&\qquad \leq c\|\beta_1 - \beta_2\|.
\end{aligned}$$

Similarly, for γ_1 and γ_2 in the same cube and for any fixed y, z and $\beta \in \mathcal{B}$,

$$\begin{aligned}
&\left| \frac{1}{na_n} \sum_{Y_j \geq y} e^{\beta^T X_j} K\left(\frac{\gamma_1^T(Z_j - z)}{a_n}\right) - \frac{1}{na_n} \sum_{Y_j \geq y} e^{\beta^T X_j} K\left(\frac{\gamma_2^T(Z_j - z)}{a_n}\right) \right| \\
&\qquad \leq \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^T X_j} \left| K\left(\frac{\gamma_1^T(Z_j - z)}{a_n}\right) - K\left(\frac{\gamma_2^T(Z_j - z)}{a_n}\right) \right| \\
&\qquad \leq \frac{c}{a_n^2} \|\gamma_1 - \gamma_2\| \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
&\left| \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma_1^T(Z - z)}{a_n}\right) \right] - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma_2^T(Z - z)}{a_n}\right) \right] \right| \\
&\qquad \leq \frac{c_1}{a_n} E\left| K\left(\frac{\gamma_1^T(Z - z)}{a_n}\right) - K\left(\frac{\gamma_2^T(Z - z)}{a_n}\right) \right| \\
&\qquad \leq c\|\gamma_1 - \gamma_2\|.
\end{aligned}$$

If we choose $\alpha_n/a_n \to 0$ and $\delta_n/a_n^2 \to 0$ as $n \to \infty$, then for any $\delta > 0$,

$$
\begin{aligned}
& P\Bigg(\sup_{\beta,\gamma,y,z} \Bigg| \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^T X_j} K\left(\frac{\gamma^T (Z_j - z)}{a_n} \right) \\
& - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T (Z - z)}{a_n} \right) \right] \Bigg| > \delta \Bigg) \\
\leq \; & P\Bigg(\max_{1 \leq k \leq m_n} \sup_{\gamma,y,z} \Bigg| \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^{(k)T} X_j} K\left(\frac{\gamma^T (Z_j - z)}{a_n} \right) \\
& - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^{(k)T} X} K\left(\frac{\gamma^T (Z - z)}{a_n} \right) \right] \Bigg| > \frac{\delta}{2} \Bigg) \\
\leq \; & P\Bigg(\max_{1 \leq l \leq m_n^*} \max_{1 \leq k \leq m_n} \sup_{y,z} \Bigg| \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^{(k)T} X_j} K\left(\frac{\gamma^{(l)T} (Z_j - z)}{a_n} \right) \\
& - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^{(k)T} X} K\left(\frac{\gamma^{(l)T} (Z - z)}{a_n} \right) \right] \Bigg| > \frac{\delta}{4} \Bigg) \\
\leq \; & \sum_{k=1}^{m_n} \sum_{l=1}^{m_n^*} P\Bigg(\sup_{y,z} \Bigg| \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^{(k)T} X_j} K\left(\frac{\gamma^{(l)T} (Z_j - z)}{a_n} \right) \\
& - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^{(k)T} X} K\left(\frac{\gamma^{(l)T} (Z - z)}{a_n} \right) \right] \Bigg| > \frac{\delta}{4} \Bigg) \\
\leq \; & c_0 m_n m_n^* \exp(-c_1 n \delta^2 a_n^2),
\end{aligned}
$$

where the exponential bound in the last step makes use of the result on the empirical CDF due to Dvoretzky et al. (1956). Therefore,

$$
\begin{aligned}
& \sum_{n=1}^{\infty} P\Bigg(\sup_{\beta,\gamma,y,z} \Bigg| \frac{1}{na_n} \sum_j I(Y_j \geq y) e^{\beta^T X_j} K\left(\frac{\gamma^T (Z_j - z)}{a_n} \right) \\
& \qquad - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T (Z - z)}{a_n} \right) \right] \Bigg| > \delta \Bigg) \\
\leq & c_2 \sum_{n=1}^{\infty} \alpha_n^{-p} \delta_n^{-q} \exp(-c_1 n \delta^2 a_n^2).
\end{aligned}
$$

If we choose $\alpha_n = a_n^2, \delta_n = a_n^3$, then the previous display becomes

$$
c_2 \sum_{n=1}^{\infty} \frac{a_n^{-2p-3q}}{e^{c_1 n \delta^2 a_n^2}} \leq c_3 \sum_{n=1}^{\infty} \frac{a_n^{-2p-3q}}{(n a_n^2)^m},
$$

for any positive integer m. By assumption, $a_n = n^{\nu_1}$ with $\nu_1 \in (-1/2, 0)$, we can choose m to be larger than $(1-(2p+3q)\nu_1)/(1+2\nu_1)$ such that the previous display is finite. Then, by the Borel-Cantelli lemma,

$$\sup_{\beta,\gamma,y,z} \left| \frac{1}{na_n} \sum_{Y_j \geq y} e^{\beta^T X_j} K\left(\frac{\gamma^T(Z_j - z)}{a_n}\right) - \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T(Z-z)}{a_n}\right)\right] \right|$$

$$\longrightarrow_{a.s.} 0.$$

For any fixed β, γ, it can be shown that

$$\frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T(Z-z)}{a_n}\right)\right] = \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta^T X}\right] \Big|_{w=\gamma^T z} + O(a_n^2),$$

where $O(a_n^2)$ does not depend on y and z. Hence for any given β, γ,

$$\sup_{y,z} \left| \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T(Z-z)}{a_n}\right)\right] - \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta^T X}\right] \Big|_{w=\gamma^T z} \right|$$

converges to 0. Note that we have shown $1/a_n E[I(Y \geq y) e^{\beta^T X} K(\gamma^T(Z-z)/a_n)]$ is equi-continuous in β. Also,

$$\begin{aligned} &\left| \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta_1^T X}\right] \Big|_{w=\gamma^T z} - \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta_2^T X}\right] \Big|_{w=\gamma^T Z} \right| \\ &\quad = \left| \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w)\left(e^{\beta_1^T X} - e^{\beta_2^T X}\right)\right] \Big|_{w=\gamma^T z} \right| \\ &\quad \leq c\|\beta_1 - \beta_2\|. \end{aligned}$$

So, $d/dw E[I(Y \geq y, \gamma^T z \leq w) e^{\beta^T X}]|_{w=\gamma^T z}$ is equi-continuous in β as well. Thus,

$$\sup_{\beta,y,z} \left| \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T(Z-z)}{a_n}\right)\right] - \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta^T X}\right] \Big|_{w=\gamma^T z} \right|$$

converges to 0. Similarly, it can be shown that both terms inside of the absolute value are equi-continuous in γ, hence

$$\sup_{\beta,\gamma,y,z} \left| \frac{1}{a_n} E\left[I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T (Z - z)}{a_n} \right) \right] - \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta^T X} \right] \Big|_{w=\gamma^T z} \right|$$

converges to 0. Therefore, we have proved that

$$\sup_{\beta,\gamma,y,z} \left| \frac{1}{n a_n} \sum_{Y_j \geq y} e^{\beta^T X_j} K\left(\frac{\gamma^T (Z_j - z)}{a_n} \right) - \frac{d}{dw} E\left[I(Y \geq y, \gamma^T Z \leq w) e^{\beta^T X} \right] \Big|_{w=\gamma^T z} \right|$$

converges to 0 almost surely. Note that the second term inside the absolute value equals $E(I(Y \geq y) e^{\beta^T X} | \gamma^T Z = \gamma^T z) f_{\gamma^T Z}(\gamma^T z)$. It then follows that

$$\sup_{\beta,\gamma} \left| (A) - \frac{1}{n} \sum_{j=1}^{n} \Delta_j \frac{1}{n a_n} \sum_{i=1}^{n} \frac{I(Y_i \geq Y_j) e^{\beta^T X_i} K\left(\frac{\gamma^T (Z_i - Z_j)}{a_n} \right)}{E\left(I(Y \geq y) e^{\beta^T X} | \gamma^T Z = \gamma^T Z_i \right) |_{y=Y_j} f_{\gamma^T Z}(\gamma^T Z_i)} \right|$$

converges to 0 almost surely. Similar arguments can be used to show that

$$\sup_{\beta,\gamma,y,z} \left| \frac{1}{n a_n} \sum_{i=1}^{n} \frac{I(Y_i \geq y) e^{\beta^T X_i} K\left(\frac{\gamma^T (Z_i - z)}{a_n} \right)}{E\left(I(Y \geq y) e^{\beta^T X} | \gamma^T Z = \gamma^T Z_i \right) f_{\gamma^T Z}(\gamma^T Z_i)} - (*) \right| \longrightarrow_{a.s.} 0,$$

where $(*)$ equals

$$\frac{1}{a_n} E\left[\frac{I(Y \geq y) e^{\beta^T X} K\left(\frac{\gamma^T (Z - z)}{a_n} \right)}{E\left(I(Y \geq y) e^{\beta^T X} | \gamma^T Z \right) f_{\gamma^T Z}(\gamma^T Z)} \right].$$

Simple calculation shows that the above display equals 1. Therefore,

$$\sup_{\beta,\gamma} \left| (A) - \frac{1}{n} \sum_{j=1}^{n} \Delta_j \right| \longrightarrow_{a.s.} 0.$$

Proof of Theorem 3.2.1

$1/n \sum_{i=1}^{n} \Delta_i \beta^T X_i$ converges uniformly on a compact set of β to $E[\Delta \beta^T X]$ almost surely since $\{\Delta \beta^T X : \beta \in \mathcal{B}\}$ is a Glivenko-Cantelli class. Following the proof for

Lemma 3.5.1, we obtain

$$\sup_{\beta,\gamma}\left|\frac{1}{n}\sum_i \Delta_i \log\left(\frac{1}{na_n}\sum_{Y_j\geq Y_i} e^{\beta^T X_j} K\left(\frac{\gamma^T(Z_j-Z_i)}{a_n}\right)\right)\right.$$
$$\left. -\frac{1}{n}\sum_i \Delta_i \log\left(E\left(I(Y\geq y)e^{\beta^T X}|\gamma^T Z=\gamma^T Z_i\right)\Big|_{y=Y_i} f_{\gamma^T Z}(\gamma^T Z_i)\right)\right| \longrightarrow_{a.s.} 0.$$

The second term inside the absolute value converges uniformly in β and γ to

$$E\left\{\Delta \log\left(E\left(I(Y\geq y)e^{\beta^T X}|\gamma^T Z\right)\Big|_{y=Y} f_{\gamma^T Z}(\gamma^T Z)\right)\right\},$$

since the involved class of functions is strong P-GC. Therefore,

$$\sup_{\beta,\gamma}\left|pl_n^{loc}(\beta,\gamma)-pl^{loc}(\beta,\gamma)\right| \longrightarrow_{a.s.} 0.$$

Proof of Theorem 3.2.2

$$\frac{\partial}{\partial\beta}\Big|_{\beta=\beta_0,\gamma=\gamma_0} pl^{loc}(\beta,\gamma) = E[\Delta X] - E\left[\Delta\frac{E(I(Y\geq y)Xe^{\beta_0^T X}|\gamma_0^T Z)}{E(I(Y\geq y)e^{\beta_0^T X}|\gamma_0^T Z)}\Big|_{y=Y}\right].$$

The second term equals

$$E_{X,Z}\left[\int \frac{E(I(Y\geq t)Xe^{\beta_0^T X}|\gamma_0^T Z)}{E(I(Y\geq t)e^{\beta_0^T X}|\gamma_0^T Z)} G_{C|X,Z}(t)\lambda_0(t,\gamma_0^T Z)e^{\beta_0^T X}\exp\left(-e^{\beta_0^T X}\Lambda_0(t,\gamma_0^T Z)\right)dt\right]$$
$$= E_{\gamma_0^T Z}\left[\int E(I(Y\geq t)Xe^{\beta_0^T X}|\gamma_0^T Z)\lambda_0(t,\gamma_0^T Z)dt\right]$$
$$= E[\Delta X],$$

where $G_{C|X,Z}(t)$ is the conditional survival function of censoring time C. By assumption, $G_{C|X,Z}(t) = G_C(t)$, the marginal survival function of C. Thus

$$\frac{\partial}{\partial\beta}\Big|_{\beta=\beta_0,\gamma=\gamma_0} pl^{loc}(\beta,\gamma) = 0$$

and (i) holds. Next, by the independence assumptions, $-\partial pl^{loc}(\beta,\gamma)/\partial\gamma|_{\beta=\beta_0,\gamma=\gamma_0}$ equals

$$
\begin{aligned}
&E\left[\Delta\frac{\nabla_\gamma\left(E(I(Y\geq y)e^{\beta_0^TX}|\gamma^TZ)f_{\gamma^TZ}(\gamma^TZ)\right)}{E(I(Y\geq y)e^{\beta_0^TX}|\gamma_0^TZ)f_{\gamma_0^TZ}(\gamma_0^TZ)}\bigg|_{y=Y}\right]\\
&\quad=E_Z\Bigg[\int\frac{\nabla_\gamma\left(E(I(Y\geq t)e^{\beta_0^TX}|\gamma^TZ)f_{\gamma^TZ}(\gamma^TZ)\right)}{E(I(Y\geq t)e^{\beta_0^TX}|\gamma_0^TZ)f_{\gamma_0^TZ}(\gamma_0^TZ)}\\
&\qquad\times\lambda_0(t,\gamma_0^TZ)E_{X|Z}\left(G_C(t)e^{\beta_0^TX}\exp\left(-e^{\beta_0^TX}\Lambda_0(t,\gamma_0^TZ)\right)\right)dt\Bigg]\\
&\quad=E_Z\left[\int\frac{\nabla_\gamma\left(E(I(Y\geq t)e^{\beta_0^TX}|\gamma^TZ)f_{\gamma^TZ}(\gamma^TZ)\right)}{f_{\gamma_0^TZ}(\gamma_0^TZ)}\times\lambda_0(t,\gamma_0^TZ)dt\right]\\
&\quad=\iint\frac{\lambda_0(y,\gamma_0^Tw)f_Z(w)}{f_{\gamma_0^TZ}(\gamma_0^Tw)}\nabla_\gamma\left(E(I(Y\geq t)e^{\beta_0^TX}|\gamma^TZ=\gamma^Tw)f_{\gamma^TZ}(\gamma^Tw)\right)dydw,
\end{aligned}
$$

where ∇_γ means taking the gradient with respect to γ and then evaluating at γ_0. The quantity inside the gradient operator can be written as

$$
\begin{aligned}
&\lim_{h\to0}\frac{1}{h}E\left[I(Y\geq y)e^{\beta_0^TX}K\left(\frac{\gamma^TZ-\gamma^Tw}{h}\right)\right]\\
&\qquad=\lim_{h\to0}\frac{1}{h}E_Z\left[K\left(\frac{\gamma^TZ-\gamma^Tw}{h}\right)g(y,\gamma_0^TZ)\right],\\
&\qquad=\lim_{h\to0}\int g(y,\gamma_0^T\tilde{w})\frac{1}{h}K\left(\frac{\gamma^T\tilde{w}-\gamma^Tw}{h}\right)f_Z(\tilde{w})d\tilde{w}
\end{aligned}
$$

where

$$
g(y,\gamma_0^TZ)=E\left[G_C(y)e^{\beta_0^TX}\exp\left(-e^{\beta_0^TX}\Lambda_0(y,\gamma_0^TZ)\right)\Big|Z\right].
$$

Note the validity of this notation depends on the assumption that X is independent of Z. Thus $-\partial pl^{loc}(\beta,\gamma)/\partial\gamma|_{\beta=\beta_0,\gamma=\gamma_0}$ equals

$$
\begin{aligned}
&\iint\frac{\lambda_0(y,\gamma_0^Tw)f_Z(w)}{f_{\gamma_0^TZ}(\gamma_0^Tw)}\lim_{h\to w0}E_Z\left[\frac{1}{h^2}K'\left(\frac{\gamma_0^TZ-\gamma_0^Tw}{h}\right)(Z-w)g(y,\gamma_0^TZ)\right]dydw\\
&=\iint\frac{\lambda_0(y,\gamma_0^Tw)f_Z(w)}{f_{\gamma_0^TZ}(\gamma_0^Tw)}\lim_{h\to0}E\left[\frac{1}{h^2}K'\left(\frac{\gamma_0^TZ-\gamma_0^Tw}{h}\right)(E(Z|\gamma_0^TZ)-w)g(y,\gamma_0^TZ)\right]dydw.
\end{aligned}
$$

Let $r(u) \equiv E(Z|\gamma_0^T Z = u)$, then the limit inside of the integral is

$$\begin{aligned}
&\lim_{h\to 0} \int \frac{1}{h^2} K'\left(\frac{u-\gamma_0^T w}{h}\right)(r(u)-w)g(y,u)f_{\gamma_0^T Z}(u)du \\
&= -g_2'(y,\gamma_0^T w)f_{\gamma_0^T Z}(\gamma_0^T w)r(\gamma_0^T w) - g(y,\gamma_0^T w)f'_{\gamma_0^T Z}(\gamma_0^T w)r(\gamma_0^T w) \\
&\quad - g(y,\gamma_0^T w)f_{\gamma_0^T Z}(\gamma_0^T w)r'(\gamma_0^T w) + g_2'(y,\gamma_0^T w)f_{\gamma_0^T Z}(\gamma_0^T w)w \\
&\quad + g(y,\gamma_0^T w)f'_{\gamma_0^T Z}(\gamma_0^T w)w.
\end{aligned}$$

Hence, the double integral equals

$$\begin{aligned}
&\int E_Z\Bigg[\lambda_0(y,\gamma_0^T Z)\Bigg(- g_2'(y,\gamma_0^T Z)r(\gamma_0^T Z) - g(y,\gamma_0^T Z)\frac{f'_{\gamma_0^T Z}}{f_{\gamma_0^T Z}}(\gamma_0^T Z)r(\gamma_0^T Z) \\
&\qquad - g(y,\gamma_0^T Z)r'(\gamma_0^T Z) + g_2'(y,\gamma_0^T Z)Z + g(y,\gamma_0^T Z)\frac{f'_{\gamma_0^T Z}}{f_{\gamma_0^T Z}}(\gamma_0^T Z)Z\Bigg)\Bigg]dy \\
&= -\int E_{\gamma_0^T Z}\left(\lambda_0(y,\gamma_0^T Z)g(y,\gamma_0^T Z)r'(\gamma_0^T Z)\right)dy.
\end{aligned}$$

Since $Z \sim N(\mu,\Sigma)$, $r'(u) = (\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0$ for any u. Thus the last display becomes

$$-(\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0 \int E_{\gamma_0^T Z}\left(\lambda_0(y,\gamma_0^T Z)g(y,\gamma_0^T Z)\right)dy = -E[\Delta](\gamma_0^T\Sigma\gamma_0)^{-1}\Sigma\gamma_0.$$

Thus it is proportional to γ_0 if and only if $\Sigma\gamma_0 \propto c\gamma_0$ for some constant c.

Proof of Theorem 3.3.1

By similar arguments to the proof of Lemma 3.5.1, we can show that

$$\frac{1}{n}\sum_{i=1}^n \Delta_i \log\left[\frac{1}{na_nb_n}\sum_{j=1}^n \Delta_j K\left(\frac{Y_j - Y_i}{b_n}\right)K\left(\frac{\gamma^T(Z_j - Z_i)}{a_n}\right)\right]$$

converges uniformly in γ to

$$E\left[\Delta \log\left(\frac{d}{dy}\bigg|_{y=Y} E\left(I(\Delta=1, Y\le y)|\gamma^T Z\right)f_{\gamma^T Z}(\gamma^T Z)\right)\right]$$

almost surely. Thus $cpl_n^{loc}(\beta,\gamma)$ converges almost surely to

$$E\left[\Delta\left(\beta^T X + \log \frac{\frac{d}{dy}E\left(I(\Delta=1, Y\le y)|\gamma^T Z\right)}{E\left(I(Y\ge y)e^{\beta^T X}|\gamma^T Z\right)}\Big|_{y=Y}\right)\right]$$
$$= E\left[\Delta\left(\beta^T X + \log\tilde{\lambda}(Y,Z;\beta,\gamma)\right)\right]$$

uniformly in (β,γ). Also, we can readily prove that

$$E[e^{\beta^T X}\tilde{\Lambda}(Y,Z;\beta,\gamma)] = E[\Delta]$$

for any β,γ. In addition,

$$E\left(I(Y\ge t)e^{\beta_0^T X}|\gamma_0^T Z = \gamma_0^T z\right)$$
$$= E\left[e^{\beta_0^T X}\exp\left(-e^{\beta_0^T X}\Lambda_0(t,\gamma_0^T Z)\right)G_{C|X,Z}(t)\Big|\gamma_0^T Z = \gamma_0^T z\right],$$

where $G_{C|X,Z}(t)$ is the survival function for the censoring time C conditional on covariates (X,Z), and

$$\frac{d}{dt}E\big(I(\Delta=1, Y\le t)|\gamma_0^T Z = \gamma_0^T z\big)$$
$$= \lambda_0(t,\gamma_0^T z)E\left[e^{\beta_0^T X}\exp\left(-e^{\beta_0^T X}\Lambda_0(t,\gamma_0^T Z)\right)G_{C|X,Z}(t)\Big|\gamma_0^T Z = \gamma_0^T z\right].$$

Thus we have proved that $\tilde{\lambda}(t,z;\beta_0,\gamma_0) = \lambda_0(t,\gamma_0^T z)$. Now suppose that (β^*,γ^*) maximizes

$$E[\Delta(\beta^T X + \log\tilde{\lambda}(Y,Z;\beta,\gamma))],$$

then it should also maximize

$$E[\Delta(\beta^T X + \log\tilde{\lambda}(Y,Z;\beta,\gamma)) - e^{\beta^T X}\tilde{\Lambda}(Y,Z;\beta,\gamma)]. \quad \text{Thus}$$

$$E\Big[\Delta\big(\beta^{*T}X+\log\tilde{\lambda}(Y,Z;\beta^*,\gamma^*)\big)-e^{\beta^{*T}X}\tilde{\Lambda}(Y,Z;\beta^*,\gamma^*)\Big]$$
$$\geq E\left[\Delta\left(\beta_0^TX+\log\lambda_0(Y,\gamma_0^TZ)\right)-e^{\beta_0^TX}\Lambda_0(Y,\gamma_0^TZ)\right].$$

Since the Kullback-Leibler information is always non-negative, it then follows, with probability 1, that

$$\exp\Big\{\Delta\big(\beta^{*T}X+\log\tilde{\lambda}(Y,Z;\beta^*,\gamma^*)\big)-e^{\beta^{*T}X}\tilde{\Lambda}(Y,Z;\beta^*,\gamma^*)\Big\}$$
$$=\exp\left\{\Delta\left(\beta_0^TX+\log\lambda_0(Y,\gamma_0^TZ)\right)-e^{\beta_0^TX}\Lambda_0(Y,\gamma_0^TZ)\right\}. \quad \text{(A.1)}$$

By choosing $\Delta=0$ and $Y=\tau$ in (A.1), we obtain $\exp(-e^{\beta^{*T}X}\tilde{\Lambda}(\tau,Z;\beta^*,\gamma^*))=\exp(-e^{\beta_0^TX}\Lambda_0(\tau,\gamma_0^TZ))$. Next, we choose $\Delta=1$ and integrate both sides of (A.1) with respect to Y for y to τ, $0\leq y<\tau$. The resulting equation together with the above display implies that $e^{\beta^{*T}X}\tilde{\Lambda}(y,Z;\beta^*,\gamma^*)=e^{\beta_0^TX}\Lambda_0(y,\gamma_0^TZ)$ with probability 1. After taking the logarithmic transformation on both sides, we obtain $\beta^*=\beta_0$ by condition (C5) after setting $Z=0$. Hence, $\tilde{\lambda}(y,Z;\beta^*,\gamma^*)=\lambda_0(y,\gamma_0^TZ)$. By condition (C4), this implies $\gamma^*\propto\gamma_0$ if we take derivative w.r.t. Z on both sides. We further conclude that $\gamma^*=\gamma_0$ in view of the restrictions that γ^* and γ_0 have unit norms and positive last coordinates. Finally we obtain the consistency of $(\hat{\beta}_n,\hat{\gamma}_n)$ by Theorem 2.12 of Kosorok (2008).

Proof of Theorem 3.3.2

Let $\mathbb{P}_n$ denote the empirical measure. Define

$$g_{1n}(y,z;\beta,\gamma)=\mathbb{P}_n\left[I(Y\geq y)Xe^{\beta^TX}\frac{1}{a_n}K\left(\frac{\gamma^TZ-\gamma^Tz}{a_n}\right)\right]$$
$$g_{2n}(y,z;\beta,\gamma)=\mathbb{P}_n\left[I(Y\geq y)e^{\beta^TX}\frac{1}{a_n}K\left(\frac{\gamma^TZ-\gamma^Tz}{a_n}\right)\right]$$
$$g_{3n}(y,z;\beta,\gamma)=\mathbb{P}_n\left[I(Y\geq y)e^{\beta^TX}K^{(1)}\left(\frac{\gamma^TZ-\gamma^Tz}{a_n}\right)\frac{Z-z}{a_n^2}\right]$$

$$g_{4n}(y,z;\gamma) = \mathbb{P}_n\left[\Delta K^{(1)}\left(\frac{\gamma^T Z - \gamma^T z}{a_n}\right)\frac{1}{b_n}K\left(\frac{Y-y}{b_n}\right)\frac{Z-z}{a_n^2}\right]$$
$$g_{5n}(y,z;\gamma) = \mathbb{P}_n\left[\Delta\frac{1}{a_n}K\left(\frac{\gamma^T Z - \gamma^T z}{a_n}\right)\frac{1}{b_n}K\left(\frac{Y-y}{b_n}\right)\right].$$

Let $g_{k0}(\cdot)$ be the expectation of $g_{kn}(\cdot)$, for $k = 1, \cdots, 5$. By definition of $(\hat{\beta}_n, \hat{\gamma}_n)$, we have the score equation for β,

$$\sqrt{n}\mathbb{P}_n\left[\Delta\left(X - \frac{g_{1n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{2n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}\right)\right] = 0.$$

We denote by $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ the empirical process and let $(\tilde{Y}, \tilde{\Delta}, \tilde{Z})$ be an independent copy of (Y, Δ, Z), where P is the probability measure. The above display can be written as

$$\begin{aligned}
&\mathbb{G}_n\Bigg\{\Delta\left(X - \frac{g_{1n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{2n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}\right) - Xe^{\hat{\beta}_n^T X}\tilde{P}\left[\frac{\tilde{\Delta}I(\tilde{Y}\le Y)\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)}{g_{2n}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)}\right] \\
&\quad + e^{\hat{\beta}_n^T X}\tilde{P}\left[\frac{\tilde{\Delta}g_{10}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)I(\tilde{Y}\le Y)\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)}{g_{2n}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)g_{20}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)}\right]\Bigg\} \\
&\quad + \sqrt{n}P\left[\Delta\left(X - \frac{g_{10}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{20}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}\right)\right] = 0.
\end{aligned}$$

The first term of the above equation can be written as

$$\begin{aligned}
&\mathbb{G}_n\Bigg[\Delta\left(X - \frac{s_{10}(Y,\gamma_0^T Z)}{s_2(Y,\gamma_0^T Z)}\right) - Xe^{\beta_0^T X}\Lambda_0(Y,\gamma_0^T Z) \\
&+ e^{\beta_0^T X}\int_0^Y \frac{s_{10}(t,\gamma_0^T Z)}{s_2(t,\gamma_0^T Z)}\lambda_0(t,\gamma_0^T Z)dt\Bigg] + o_P(1) \equiv \mathbb{G}_n\left[\tilde{\ell}_1(D;\beta_0,\gamma_0)\right] + o_P(1),
\end{aligned}$$

where

$$s_{10}(y,u) = E\left[I(Y\ge y)Xe^{\beta_0^T X}|\gamma_0^T Z = u\right]$$
$$s_2(y,u) = E\left[I(Y\ge y)e^{\beta_0^T X}|\gamma_0^T Z = u\right]$$

and $D \equiv (Y, \Delta, X, Z)$ denotes the data. Next, without loss of generality, we assume the last coordinate of $\gamma_{q\times 1}$ is positive since the norm of γ is non-zero. That is, $\gamma^T = (\gamma_{(-q)}^T, \gamma_q)$, where $\gamma_{(-q)}$ is the γ vector with the last coordinate γ_q deleted and $\gamma_q > 0$. Thus the score function for γ is

$$\frac{\partial cpl_n^{loc}(\beta,\gamma)}{\partial \gamma_{(-q)}} = \left(\frac{\partial \gamma}{\partial \gamma_{(-q)}}\right)^T \frac{\partial cpl_n^{loc}(\beta,\gamma)}{\partial \gamma} \equiv M(\gamma)\frac{\partial cpl_n^{loc}(\beta,\gamma)}{\partial \gamma},$$

where

$$M(\gamma) = \left(I_{(q-1)\times(q-1)}, -\frac{\gamma_{(-q)}}{\gamma_q}\right).$$

Here, $I_{(q-1)\times(q-1)}$ is the identity matrix. Thus the score equation for γ is

$$-\sqrt{n}\mathbb{P}_n\left[\Delta M(\hat{\gamma}_n)\left(\frac{g_{3n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{2n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)} - \frac{g_{4n}(Y,Z;\hat{\gamma}_n)}{g_{5n}(Y,Z;\hat{\gamma}_n)}\right)\right] = 0.$$

It can be written as

$$\begin{aligned}
-\mathbb{G}_n\Bigg\{ M(\hat{\gamma}_n)\Bigg(&\Delta\frac{g_{3n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{2n}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)} + e^{\hat{\beta}_n^T X}\tilde{P}\left[\frac{\tilde{\Delta} I(\tilde{Y}\le Y)K^{(1)}\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)\frac{Z-\tilde{Z}}{a_n^2}}{g_{2n}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)}\right] \\
&- e^{\hat{\beta}_n^T X}\tilde{P}\left[\frac{\tilde{\Delta} g_{30}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n) I(\tilde{Y}\le Y)\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)}{g_{2n}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n) g_{20}(\tilde{Y},\tilde{Z};\hat{\beta}_n,\hat{\gamma}_n)}\right] - \Delta\frac{g_{4n}(Y,Z;\hat{\gamma}_n)}{g_{5n}(Y,Z;\hat{\gamma}_n)} \\
&- \Delta\tilde{P}\left[\frac{\tilde{\Delta}K^{(1)}\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)\frac{Z-\tilde{Z}}{a_n^2}\frac{1}{b_n}K\left(\frac{Y-\tilde{Y}}{b_n}\right)}{g_{5n}(\tilde{Y},\tilde{Z};\hat{\gamma}_n)}\right] \\
&+ \Delta\tilde{P}\left[\frac{\tilde{\Delta} g_{40}(\tilde{Y},\tilde{Z};\hat{\gamma}_n)\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T(Z-\tilde{Z})}{a_n}\right)\frac{1}{b_n}K\left(\frac{Y-\tilde{Y}}{b_n}\right)}{g_{5n}(\tilde{Y},\tilde{Z};\hat{\gamma}_n) g_{50}(\tilde{Y},\tilde{Z};\hat{\gamma}_n)}\right]\Bigg)\Bigg\} \\
&- \sqrt{n}P\left[M(\hat{\gamma}_n)\Delta\left(\frac{g_{30}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{20}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)} - \frac{g_{40}(Y,Z;\hat{\gamma}_n)}{g_{50}(Y,Z;\hat{\gamma}_n)}\right)\right] = 0.
\end{aligned}$$

Similarly, the first term of the above equation can be written as

$$\mathbb{G}_n\Bigg\{ - M(\gamma_0)\Bigg[\Delta\left(\frac{s_{01}(Y,\gamma_0^T Z)}{s_2(Y,\gamma_0^T Z)} - Z\right)\frac{\lambda_0^{(0,1)}(Y,\gamma_0^T Z)}{\lambda_0(Y,\gamma_0^T Z)} + e^{\beta_0^T X} Z\Lambda_0^{(0,1)}(Y,\gamma_0^T Z)$$
$$- e^{\beta_0^T X}\int_0^Y \frac{s_{01}(t,\gamma_0^T Z)}{s_2(t,\gamma_0^T Z)}\lambda_0^{(0,1)}(t,\gamma_0^T Z)dt\Bigg]\Bigg\} + o_P(1) \equiv \mathbb{G}_n\left[\tilde{\ell}_2(D;\beta_0,\gamma_0)\right] + o_P(1),$$

where

$$s_{01}(y,u) = E[I(Y \geq y)Ze^{\beta_0^T X}|\gamma_0^T Z = u] \quad \text{and}$$
$$\lambda_0^{(0,1)}(t,u) = \frac{\partial\lambda_0(t,u)}{\partial u}.$$

Thus combined, we have

$$o_P(1) + \mathbb{G}_n\begin{pmatrix} \tilde{\ell}_1(D;\beta_0,\gamma_0) \\ \tilde{\ell}_2(D;\beta_0,\gamma_0) \end{pmatrix} + \sqrt{n}P\begin{pmatrix} \Delta\left(X - \frac{g_{10}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{20}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}\right) \\ -M(\hat{\gamma}_n)\Delta\left(\frac{g_{30}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)}{g_{20}(Y,Z;\hat{\beta}_n,\hat{\gamma}_n)} - \frac{g_{40}(Y,Z;\hat{\gamma}_n)}{g_{50}(Y,Z;\hat{\gamma}_n)}\right) \end{pmatrix} = 0.$$

By Taylor's expansion, the third term equals

$$\sqrt{n}P\begin{pmatrix} \Delta\left(X - \frac{g_{10}(Y,Z;\beta_0,\gamma_0)}{g_{20}(Y,Z;\beta_0,\gamma_0)}\right) \\ -M(\gamma_0)\Delta\left(\frac{g_{30}(Y,Z;\beta_0,\gamma_0)}{g_{20}(Y,Z;\beta_0,\gamma_0)} - \frac{g_{40}(Y,Z;\gamma_0)}{g_{50}(Y,Z;\gamma_0)}\right) \end{pmatrix}$$
$$+ P\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\sqrt{n}\begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{\gamma}_{n,(-q)} - \gamma_{0,(-q)} \end{pmatrix}, \quad \text{where}$$

$$\Sigma_{11} = \frac{\partial}{\partial\beta}\Big|_{\beta^*,\gamma^*}\left[\Delta\left(X - \frac{g_{10}(Y,Z;\beta,\gamma)}{g_{20}(Y,Z;\beta,\gamma)}\right)\right]$$
$$\Sigma_{12} = \frac{\partial}{\partial\gamma_{(-q)}}\Big|_{\beta^*,\gamma^*}\left[\Delta\left(X - \frac{g_{10}(Y,Z;\beta,\gamma)}{g_{20}(Y,Z;\beta,\gamma)}\right)\right]$$
$$\Sigma_{21} = \frac{\partial}{\partial\beta}\Big|_{\beta^*,\gamma^*}\left[-M(\gamma)\Delta\left(\frac{g_{30}(Y,Z;\beta,\gamma)}{g_{20}(Y,Z;\beta,\gamma)} - \frac{g_{40}(Y,Z;\gamma)}{g_{50}(Y,Z;\gamma)}\right)\right]$$

$$\Sigma_{22} = \frac{\partial}{\partial \gamma_{(-q)}}\Big|_{\beta^*,\gamma^*} \left[-M(\gamma)\Delta\left(\frac{g_{30}(Y,Z;\beta,\gamma)}{g_{20}(Y,Z;\beta,\gamma)} - \frac{g_{40}(Y,Z;\gamma)}{g_{50}(Y,Z;\gamma)} \right) \right],$$

for β^* falling between $\hat{\beta}_n$ and β_0 and γ^* falling between $\hat{\gamma}_n$ and γ_0. After replacing $g_{k0}(\cdot), k = 1, \cdots 5$, by their limits together with the condition that $a_n = n^{\nu_1}$ and $b_n = n^{\nu_2}$ with $\nu_1, \nu_2 \in (-1/2, -1/4)$, we can show that

$$\sqrt{n}P \begin{pmatrix} \Delta\left(X - \frac{g_{10}(Y,Z;\beta_0,\gamma_0)}{g_{20}(Y,Z;\beta_0,\gamma_0)} \right) \\ -M(\gamma_0)\Delta\left(\frac{g_{30}(Y,Z;\beta_0,\gamma_0)}{g_{20}(Y,Z;\beta_0,\gamma_0)} - \frac{g_{40}(Y,Z;\gamma_0)}{g_{50}(Y,Z;\gamma_0)} \right) \end{pmatrix} = o_P(1),$$

$$P\Sigma_{11} \to_P P\left[\frac{\partial}{\partial \beta}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_1(D;\beta,\gamma) \right], \quad P\Sigma_{12} \to_P P\left[\frac{\partial}{\partial \gamma_{(-q)}}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_1(D;\beta,\gamma) \right] \quad \text{and}$$

$$P\Sigma_{22} \to_P P\left[\frac{\partial}{\partial \gamma_{(-q)}}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_2(D;\beta,\gamma) \right].$$

Later in this proof, we will show that $(\tilde{\ell}_1(D;\beta_0,\gamma_0)^T, \tilde{\ell}_2(D;\beta_0,\gamma_0)^T)$ corresponds to the score function of some submodel and thus by the usual equality in classical likelihood theory,

$$-P\left[\frac{\partial}{\partial \beta}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_1(D;\beta,\gamma) \right] = P\left[\tilde{\ell}_1(D;\beta_0,\gamma_0)\tilde{\ell}_1(D;\beta_0,\gamma_0)^T \right]$$

$$-P\left[\frac{\partial}{\partial \gamma_{(-q)}}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_1(D;\beta,\gamma) \right] = P\left[\tilde{\ell}_1(D;\beta_0,\gamma_0)\tilde{\ell}_2(D;\beta_0,\gamma_0)^T \right] \quad \text{and}$$

$$-P\left[\frac{\partial}{\partial \gamma_{(-q)}}\Big|_{\beta_0,\gamma_0} \tilde{\ell}_2(D;\beta,\gamma) \right] = P\left[\tilde{\ell}_2(D;\beta_0,\gamma_0)\tilde{\ell}_2(D;\beta_0,\gamma_0)^T \right].$$

Combined, we have

$$\tilde{I}_{\beta_0,\gamma_0}\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{\gamma}_{n,(-q)} - \gamma_{0,(-q)} \end{pmatrix} = \mathbb{G}_n \begin{pmatrix} \tilde{\ell}_1(D;\beta_0,\gamma_0) \\ \tilde{\ell}_2(D;\beta_0,\gamma_0) \end{pmatrix} + o_P(1), \quad \text{where}$$

$$\tilde{I}_{\beta_0,\gamma_0} = P\begin{pmatrix} \tilde{\ell}_1(D;\beta_0,\gamma_0)\tilde{\ell}_1(D;\beta_0,\gamma_0)^T & \tilde{\ell}_1(D;\beta_0,\gamma_0)\tilde{\ell}_2(D;\beta_0,\gamma_0)^T \\ \tilde{\ell}_2(D;\beta_0,\gamma_0)\tilde{\ell}_1(D;\beta_0,\gamma_0)^T & \tilde{\ell}_2(D;\beta_0,\gamma_0)\tilde{\ell}_2(D;\beta_0,\gamma_0)^T \end{pmatrix}.$$

It then follows that

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{\gamma}_{n,(-q)} - \gamma_{0,(-q)} \end{pmatrix} \rightsquigarrow N\left(0, \tilde{I}^{-1}_{\beta_0,\gamma_0}\right).$$

Next, we prove semiparametric efficiency defined in Bickel et al. (1993) by showing that the influence function for estimating $(\beta_0, \gamma_{0,(-q)})$ is in fact the efficient influence function. Note that each coordinate of the influence function for estimating $(\beta_0, \gamma_{0,(-q)})$ is some linear combination of the form

$$\begin{aligned} v_1^T \left[\Delta\left(X - \frac{s_{10}(Y,\gamma_0^T Z)}{s_2(Y,\gamma_0^T Z)}\right) - Xe^{\beta_0^T X}\Lambda_0(Y,\gamma_0^T Z) + e^{\beta_0^T X}\int_0^Y \frac{s_{10}(t,\gamma_0^T Z)}{s_2(t,\gamma_0^T Z)}\lambda_0(t,\gamma_0^T Z)dt\right] \\ + v_2^T M(\gamma_0)\left[\Delta\left(Z - \frac{s_{01}(Y,\gamma_0^T Z)}{s_2(Y,\gamma_0^T Z)}\right)\frac{\lambda_0^{(0,1)}(Y,\gamma_0^T Z)}{\lambda_0(Y,\gamma_0^T Z)}\right. \\ \left. - e^{\beta_0^T X}\left(Z\Lambda_0^{(0,1)}(Y,\gamma_0^T Z) - \int_0^Y \frac{s_{01}(t,\gamma_0^T Z)}{s_2(t,\gamma_0^T Z)}\lambda_0^{(0,1)}(t,\gamma_0^T Z)dt\right)\right], \end{aligned}$$

for some vectors v_1 and v_2. This function is exactly the score function of a submodel $(\beta_0 + \epsilon v_1, \gamma_{0,(-q)} + \epsilon v_2, \lambda_0(t,u) + \epsilon v_1^T \mu(t,u)\lambda_0(t,u) + \epsilon v_2^T M(\gamma_0)\eta(t,u)\lambda_0^{(0,1)}(t,u))$, where $\mu(t,u) = -s_{10}(t,u)/s_2(t,u)$ and $\eta(t,u) = -s_{01}(t,u)/s_2(t,u)$. Thus the influence function belongs to the tangent space of the model at $(\beta_0, \gamma_0, \lambda_0(\cdot))$ and hence is the efficient influence function.

Proof of Theorem 3.3.3

In empirical process notation, $\hat{\Lambda}(t,u)$ can be written as

$$\mathbb{P}_n\left[\frac{I(Y \le t)\Delta\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T Z - u}{a_n}\right)}{\frac{1}{n}\sum_{Y_j \ge Y} e^{\hat{\beta}_n^T X_j}\frac{1}{a_n}K\left(\frac{\hat{\gamma}_n^T Z_j - u}{a_n}\right)}\right].$$

By Theorem 3.3.1, $\hat{\beta}_n \longrightarrow_{a.s.} \beta_0$ and $\hat{\gamma}_n \longrightarrow_{a.s.} \gamma_0$. Similar arguments to those used in the proof of Lemma 3.5.1 can be used to prove that

$$\sup_{y,u} \left| \frac{1}{n} \sum_{Y_j \geq y} e^{\hat{\beta}_n^T X_j} \frac{1}{a_n} K\left(\frac{\hat{\gamma}_n^T Z_j - u}{a_n} \right) - E\left(I(Y \geq y) e^{\beta_0^T X} | \gamma_0^T Z = u \right) f_{\gamma_0^T Z}(u) \right| \longrightarrow_{a.s.} 0.$$

Using standard empirical process arguments, we have

$$\sup_{t,u} \left| \hat{\Lambda}(t,u) - E\left[\frac{\Delta I(Y \leq t) \frac{1}{a_n} K\left(\frac{\gamma_0^T Z - u}{a_n} \right)}{E\left(I(Y \geq y) e^{\beta_0^T X} | \gamma_0^T Z = u \right) |_{y=Y} f_{\gamma_0^T Z}(u)} \right] \right| \longrightarrow_{a.s.} 0.$$

Straightforward calculation reveals that the second term inside the absolute value in the above display equals $\Lambda_0(t,u)$. This completes the proof.

Chapter 4

Profile Stratified Likelihood

4.1 Single-Index Hazards Model

4.1.1 Method

For a fixed γ, model (2.1) can be viewed as a stratified hazards model with strata defined by values of $\gamma^T W$. The model is in spirit similar to the stratified Cox model, which is commonly used in epidemiologic studies, such as in Motzer et al. (1999) and Chow et al. (2006). This similarity also allows us to consider the following profile stratified likelihood approach.

For any fixed γ, we stratify the range R of $\gamma^T W_i$, $i = 1, \cdots, n$ based on the sample quantiles. Specifically, let $\min_{1\le i\le n} \gamma^T W_i = t_0 < \cdots < t_{J_n} = \max_{1\le i\le n} \gamma^T W_i$ be a partition of R into J_n subintervals $[t_{k-1}, t_k), k = 1, \cdots, J_n$ such that $t_k = \mathbb{F}_n^{-1}(k/J_n)$, where $\mathbb{F}_n(\cdot)$ is the empirical distribution of $\gamma^T W_i, i = 1, \cdots, n$. Note that t_i is a random variable depending on γ. Also, it is clear by the above construction that $I(\gamma^T W_i \in [t_{k-1}, t_k)) = I(\mathbb{F}_n(\gamma^T W_i) \in [(k-1)/J_n, k/J_n))$. Let $S_k \equiv [(k-1)/J_n, k/J_n)$. We assume for any (t, u),

$$\lambda(t, u) = \sum_{k=1}^{J_n} I(u \in [t_{k-1}, t_k))\lambda_k(t) = \sum_{k=1}^{J_n} I(\mathbb{F}_n(u) \in S_k)\lambda_k(t),$$

thus we assume the baseline hazard function takes different forms on different strata defined by the sample quantiles of $\gamma^T W$. Plug this into (2.2) and then in the setting of the NPMLE, we maximize

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i \sum_{k=1}^{J_n} \log \Lambda_k\{Y_i\} I(\mathbb{F}_n(\gamma^T W_i) \in S_k) - \sum_{k=1}^{J_n} I(\mathbb{F}_n(\gamma^T W_i) \in S_k) \sum_{Y_j \le Y_i} \Lambda_k\{Y_j\}\right]. \quad (4.1)$$

The maximizer for $\Lambda_k\{Y_i\}$ is

$$\hat{\Lambda}_k\{Y_i\} = \frac{\Delta_i I(\mathbb{F}_n(\gamma^T W_i) \in S_k)}{\sum_{Y_j \ge Y_i} I(\mathbb{F}_n(\gamma^T W_j) \in S_k)}. \quad (4.2)$$

After plugging (4.2) into (4.1), we obtain, up to a constant, that the profile stratified likelihood function is

$$pl_n^s(\gamma) \;=\; -\frac{1}{n}\sum_{i=1}^{n}\Delta_i \sum_{k=1}^{J_n}\left[I\left(\mathbb{F}_n(\gamma^T W_i) \in S_k\right) \log\left(\frac{J_n}{n}\sum_{Y_j \ge Y_i} I\left(\mathbb{F}_n(\gamma^T W_j) \in S_k\right)\right)\right].$$

Note that $pl_n^s(\gamma)$ is not smooth in γ so that it is numerically difficult to find its maximizer. In the following simulation studies, we use the grid search to find the maximum point. However, the grid search becomes infeasible when the dimension of W increases.

4.1.2 Bias Analysis

Let $\hat{\gamma}_n$ denote the maximizer of $pl_n^s(\gamma)$. We now study the asymptotic property of $\hat{\gamma}_n$ by first obtaining the asymptotic limit $pl^s(\gamma)$ of $pl_n^s(\gamma)$ given in Theorem 4.1.1. The following regularity conditions are imposed:

(C1) $\gamma_0 \in \Gamma$, where $\Gamma \in \mathbb{R}^q$ is compact.

(C2) $J_n/\sqrt{n} \to 0, J_n \to \infty$.

(C3) $\lambda_0(t,u)$ has non-zero partial derivative with respect to u for any $t \in [0,\tau]$; Moreover, $\int_0^\tau \lambda_0(t,u)dt < \infty$, for any $u \in \mathbb{R}$.

(C4) Given covariates W, T and C are independent.

(C5) $P(T > \tau) < 1$, where τ denotes the end of the study.

Theorem 4.1.1. *If conditions (C1) and (C2) hold, then* $\sup_\gamma |pl_n^s(\gamma) - pl^s(\gamma)| \to_P 0$, *where*

$$pl^s(\gamma) = -E\left[\Delta \log P\left(Y \geq y | F_{\gamma^T W}(\gamma^T W)\right)|_{y=Y}\right] = -E\left[\Delta \log P\left(Y \geq y | \gamma^T W\right)|_{y=Y}\right].$$

Here $F_{\gamma^T W}(\cdot)$ *is the distribution function of* $\gamma^T W$.

Note that the difference between $pl^{loc}(\gamma)$ given in Theorem 2.3.1 and $pl^s(\gamma)$ is that the latter does not involve the density $f_{\gamma^T W}(\gamma^T W)$ in the log function. The reason is that the stratified method is based on $F_{\gamma^T W}(\gamma^T W)$ which follows the uniform distribution on $[0, 1]$.

From Theorem 4.1.1, we conclude similarly that $\hat{\gamma}_n$ converges to the maximizer of $pl^s(\gamma)$ in probability. Next, we show in Theorem 4.1.2 that $\hat{\gamma}_n$ is consistent for γ_0 under certain restrictive conditions.

Theorem 4.1.2. *Assume conditions (C1) and (C2)-(C5) hold and suppose* C *is independent of* W, *then* $\hat{\gamma}_n \to_P \gamma_0$.

Remark 4.1.1. $\partial/\partial u \lambda_0(t, u) = 0$ implies $\lambda_0(t, u)$ is constant in u and thus W has no effect on the hazard function. Therefore, assuming the first part of condition (C3) is not unreasonable. The second part of condition (C3) ensures a positive probability of censoring. Also, we have assumed the independence between C and W. Without this assumption, we conjecture that the stratified likelihood approach would lead to biased estimation. In fact, the simulation studies suggest that this approach can fail if C and W are dependent.

We reconsider the 4 simulation settings in Section 2.4, using the aforementioned profile stratified likelihood. A grid search with step size 0.01 is used since the objective function

Table 4.1: Simulation results of stratified likelihood in single-index hazards model

			Stratified likelihood		Cox model	
Simulation settings	Parameter	Sample size	Bias	SE	Bias	SE
(i) $C \perp W$	γ_1	2000	.003	.076	-.001	.026
$\lambda_0(t,u) = 0.5e^u t$		5000	-.000	.059	.000	.017
$cov(W) = 0$		10000	-.001	.039	.000	.012
(ii) $C \perp W$	γ_1	2000	.001	.066	-.001	.030
$\lambda_0(t,u) = 0.5e^u t$		5000	.004	.049	.000	.019
$cov(W) = 0.5$		10000	.002	.034	.001	.014
(iii) $C \perp W$	γ_1	2000	.015	.121	.500	.032
$\lambda_0(t,u) = 0.25(t+u^2)$		5000	.008	.078	.501	.020
$cov(W) = 0.5$		10000	.003	.057	.500	.014
(iv) $C \not\perp W$	γ_1	2000	-.312	.030	.446	.033
$\lambda_0(t,u) = 0.25(t+u^2)$		5000	-.310	.018	.448	.022
$cov(W) = 0.5$		10000	-.310	.014	.447	.015

NOTE: Each entry is based on 500 replicates.

is not continuous in γ. The number of strata is chosen from 4, 8 or 12 for $\gamma^T W$. We report the results from the number of strata yielding the smallest bias.

Table 4.1 summarizes the simulation results in setting (i)-(iv) with sample sizes 2000, 5000 and 10000, where γ_1 is the first component of the γ vector. As expected by Theorem 4.1.2, the stratified approach works in simulation settings (i)-(iii). However, it fails in setting (iv) probably due to the dependence between C and W. Figure 4.1 shows the profile stratified likelihood function in each simulation setting based on a simulated data set of size 10000. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The number of strata is 12 for $\gamma^T Z$. Again, the stratified profile likelihood curves are maximized around the true value -0.5 of γ_1 in the first three simulation settings, suggesting that the stratified approach yields estimators with little bias in these settings, but it gives biased estimation in setting (iv) where C depends on W.

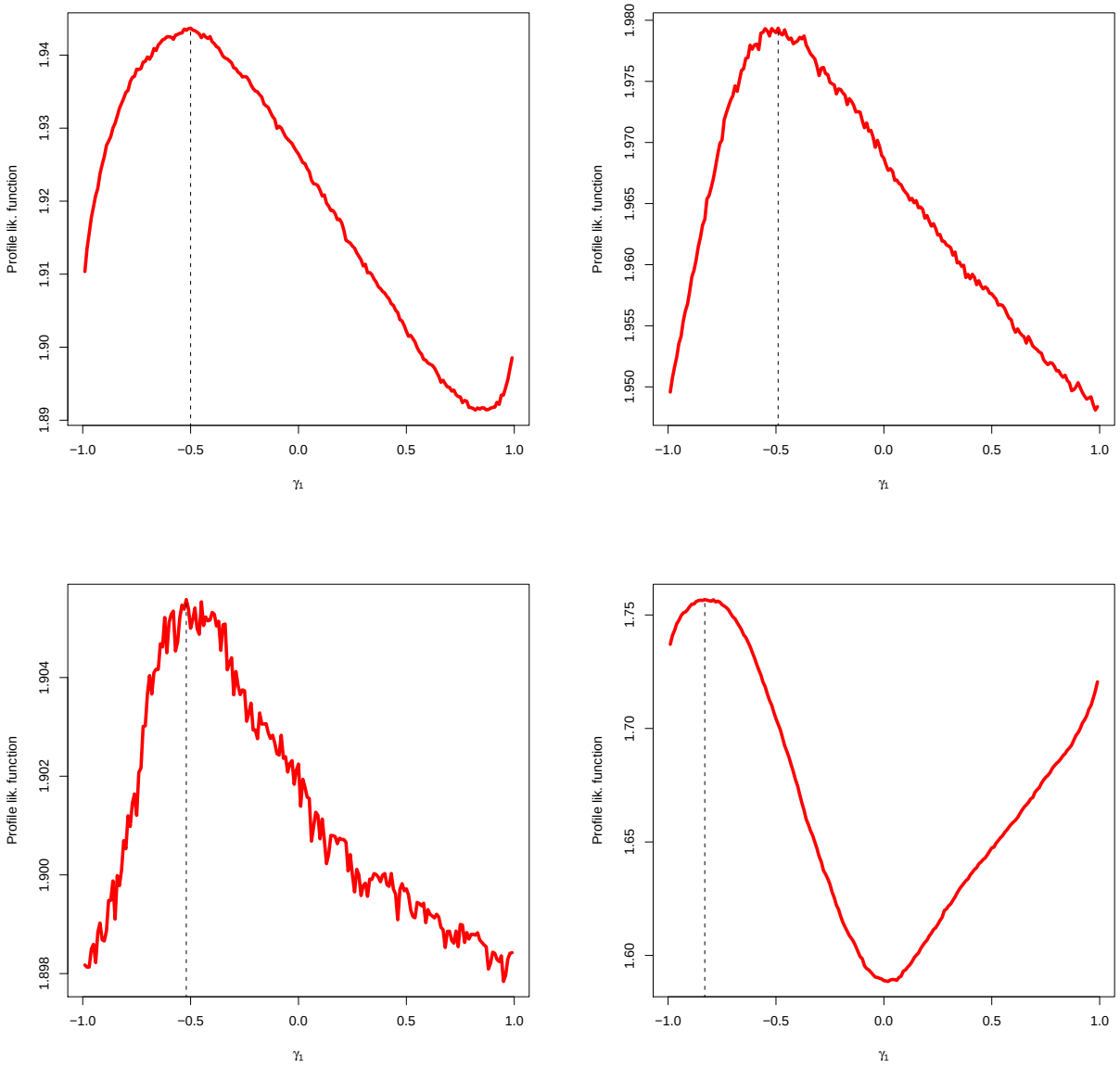

Figure 4.1: Profile stratified likelihood curve of γ_1 in single-index hazards model

4.2 Partly Proportional Single-Index Hazards Model

4.2.1 Method

Similarly, for any fixed γ, model (3.1) can be viewed as a stratified Cox model with strata defined by values of $\gamma^T Z$. Given the data, we consider stratifying the range R of $\gamma^T Z_i$, $i = 1, \cdots, n$ based on the sample quantiles. Specifically, for any fixed γ, let $\min_{1\le i\le n} \gamma^T Z_i = t_0 < \cdots < t_{J_n} = \max_{1\le i\le n} \gamma^T Z_i$ be a partition of R into J_n subintervals $[t_{k-1}, t_k), k = 1, \cdots, J_n$ such that $t_k = \mathbb{F}_n^{-1}(k/J_n)$, where $\mathbb{F}_n(\cdot)$ is the empirical distribution of $\gamma^T Z_i, i = 1, \cdots, n$. Note that t_i is a random variable depending on γ. Also, it is clear by the above construction that $I(\gamma^T Z_i \in [t_{k-1}, t_k)) = I(\mathbb{F}_n(\gamma^T Z_i) \in [(k-1)/J_n, k/J_n))$. Let $S_k \equiv [(k-1)/J_n, k/J_n)$. We assume for any t, u,

$$\lambda(t, u) = \sum_{k=1}^{J_n} I(u \in [t_{k-1}, t_k))\lambda_k(t) = \sum_{k=1}^{J_n} I(\mathbb{F}_n(u) \in S_k)\lambda_k(t),$$

thus we assume that the baseline hazard function takes different forms on different strata defined by the sample quantiles of $\gamma^T Z$. Plug this into (3.2) and then in the setting of NPMLE, we maximize

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left(\beta^T X_i + \sum_{k=1}^{J_n}\log\Lambda_k\{Y_i\}I(\mathbb{F}_n(\gamma^T Z_i)\in S_k)\right)\right.$$
$$\left. - e^{\beta^T X_i}\sum_{k=1}^{J_n} I(\mathbb{F}_n(\gamma^T Z_i)\in S_k)\sum_{Y_j\le Y_i}\Lambda_k\{Y_j\}\right]. \quad (4.3)$$

The maximizer for $\Lambda_k\{Y_i\}$ is

$$\hat{\Lambda}_k\{Y_i\} = \frac{\Delta_i I(\mathbb{F}_n(\gamma^T Z_i)\in S_k)}{\sum_{Y_j\ge Y_i} e^{\beta^T X_j} I(\mathbb{F}_n(\gamma^T Z_j)\in S_k)}. \quad (4.4)$$

After plugging (4.4) into (4.3), we obtain, up to a constant, the profile stratified likelihood function as

$$\begin{aligned} pl_n^s(\beta,\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i\beta^T X_i \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\Delta_i\sum_{k=1}^{J_n}\left[I\left(\mathbb{F}_n(\gamma^T Z_i)\in S_k\right)\log\left(\frac{J_n}{n}\sum_{Y_j\geq Y_i}e^{\beta^T X_j}I\left(\mathbb{F}_n(\gamma^T Z_j)\in S_k\right)\right)\right]. \end{aligned}$$

Note that $pl_n^s(\beta,\gamma)$ is not smooth in γ so that we have to use grid search to find the maximum point. However, grid search becomes infeasible when the dimension of Z increases.

4.2.2 Bias Analysis

We impose the following regularity conditions:

(C1) $\beta_0\in\mathcal{B},\gamma_0\in\Gamma$, where $\mathcal{B}\in\mathbb{R}^p,\Gamma\in\mathbb{R}^q$ are compact.

Theorem 4.2.1. *If condition (C1) holds, $J_n/\sqrt{n}\longrightarrow 0$ and $J_n\longrightarrow\infty$, then $\sup_{\beta,\gamma}|pl_n^s(\beta,\gamma)-pl^s(\beta,\gamma)|\longrightarrow_P 0$, where*

$$\begin{aligned} pl^s(\beta,\gamma) &= E\left[\Delta\left(\beta^T X+\log\frac{1}{E\left(I(Y\geq y)e^{\beta^T X}|F_{\gamma^T Z}(\gamma^T Z)\right)|_{y=Y}}\right)\right] \\ &= E\left[\Delta\left(\beta^T X+\log\frac{1}{E\left(I(Y\geq y)e^{\beta^T X}|\gamma^T Z\right)|_{y=Y}}\right)\right]. \end{aligned}$$

Here $F_{\gamma^T Z}(\cdot)$ is the distribution function of $\gamma^T Z$.

Notice that the difference between $pl^{loc}(\beta,\gamma)$ given in Theorem 3.2.1 and $pl^s(\beta,\gamma)$ is that the latter does not have the density $f_{\gamma^T Z}(\gamma^T Z)$ in the denominator of the log function. The reason is that the stratified method is based on $F_{\gamma^T Z}(\gamma^T Z)$ which follows the uniform distribution on $[0,1]$.

From Theorem 4.2.1, we conclude similarly that the profile stratified likelihood estimator converges to the maximizer of $pl^s(\beta, \gamma)$ in probability. Next, we show in Theorem 4.2.2 that in some special cases, $pl^s(\beta, \gamma)$ indeed has zero derivatives at the true parameter (β_0, γ_0).

Theorem 4.2.2. *Suppose C is independent of (X, Z), and Z is independent of X, then (i) $\frac{\partial}{\partial \beta}|_{\beta=\beta_0,\gamma=\gamma_0} pl^s(\beta, \gamma) = 0$; (ii) $\frac{\partial}{\partial \gamma}|_{\beta=\beta_0,\gamma=\gamma_0} pl^s(\beta, \gamma) = 0$.*

Remark 4.2.1. This theorem is based on the restrictive independent censoring assumption and the critical assumption that X and Z are independent. When X and Z are dependent, we conjecture that part (ii) in Theorem 4.2.2 no longer holds. In fact, our simulation results will show that the profile stratified likelihood method fails when X and Z are correlated.

We now reconsider the 4 simulation settings studied in Section 3.2.2, using the aforementioned profile stratified likelihood. A grid search with step size 0.01 was used since the objective function is not continuous in γ. The number of strata was chosen from 4, 8 or 12 for $\gamma^T Z$. We reported the results from the number of strata yielding the smallest bias.

Table 4.2 summarizes the simulation results in setting (i)-(iv) with sample sizes 100, 200 and 400, where γ_1 is the first coordinate of the γ vector. As expected from Theorem 4.2.2, the profile stratified method works in setting (i) and (ii) since the censoring time C is independent of covariates (X, Z) and X is independent of Z. However, this methods fails in setting (iii) and (iv) due to the dependence between X and Z. Figure 4.2 shows the profile stratified likelihood curve (of γ_1) in each setting based on a simulated dataset with $n = 5000$. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The number of strata is 12 for $\gamma^T Z$. It is observed again that the profile stratified likelihood approach works in setting (i) and (ii), but fails in setting (iii) and (iv).

Table 4.2: Simulation results of stratified likelihood in PPSIH model

			Stratified likelihood		Cox model	
Simulation settings	Sample size	Parameters	Bias	SE	Bias	SE
(i) $X \perp Z$	100	β	.033	.178	.030	.163
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.022	.236	.015	.134
$cov(Z) = 0$	200	β	-.014	.110	.018	.105
		γ_1	-.013	.169	.009	.090
	400	β	.005	.077	.006	.072
		γ_1	-.006	.118	.005	.063
(ii) $X \perp Z$	100	β	.014	.178	.028	.163
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.034	.304	.014	.154
$cov(Z) = 0.5$	200	β	-.033	.110	.018	.104
		γ_1	-.015	.226	.010	.103
	400	β	-.003	.079	.006	.072
		γ_1	-.003	.153	.006	.072
(iii) $X \not\perp Z$	100	β	.330	.211	.027	.193
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.964	.447	.014	.164
$cov(Z) = 0.5$	200	β	.254	.137	.019	.123
		γ_1	-.964	.397	.008	.109
	400	β	.210	.097	.006	.083
		γ_1	-1.005	.333	.005	.077
(iv) $X \not\perp Z$	100	β	.233	.189	-.125	.163
$\lambda_0(t,u) = 0.25e^{e^u t}$		γ_1	-.949	.464	-.048	.147
$cov(Z) = 0.5$	200	β	.163	.120	-.139	.103
		γ_1	-.952	.419	-.044	.104
	400	β	.126	.084	-.151	.073
		γ_1	-1.021	.364	-.041	.070

NOTE: Each entry is based on 1000 replicates.

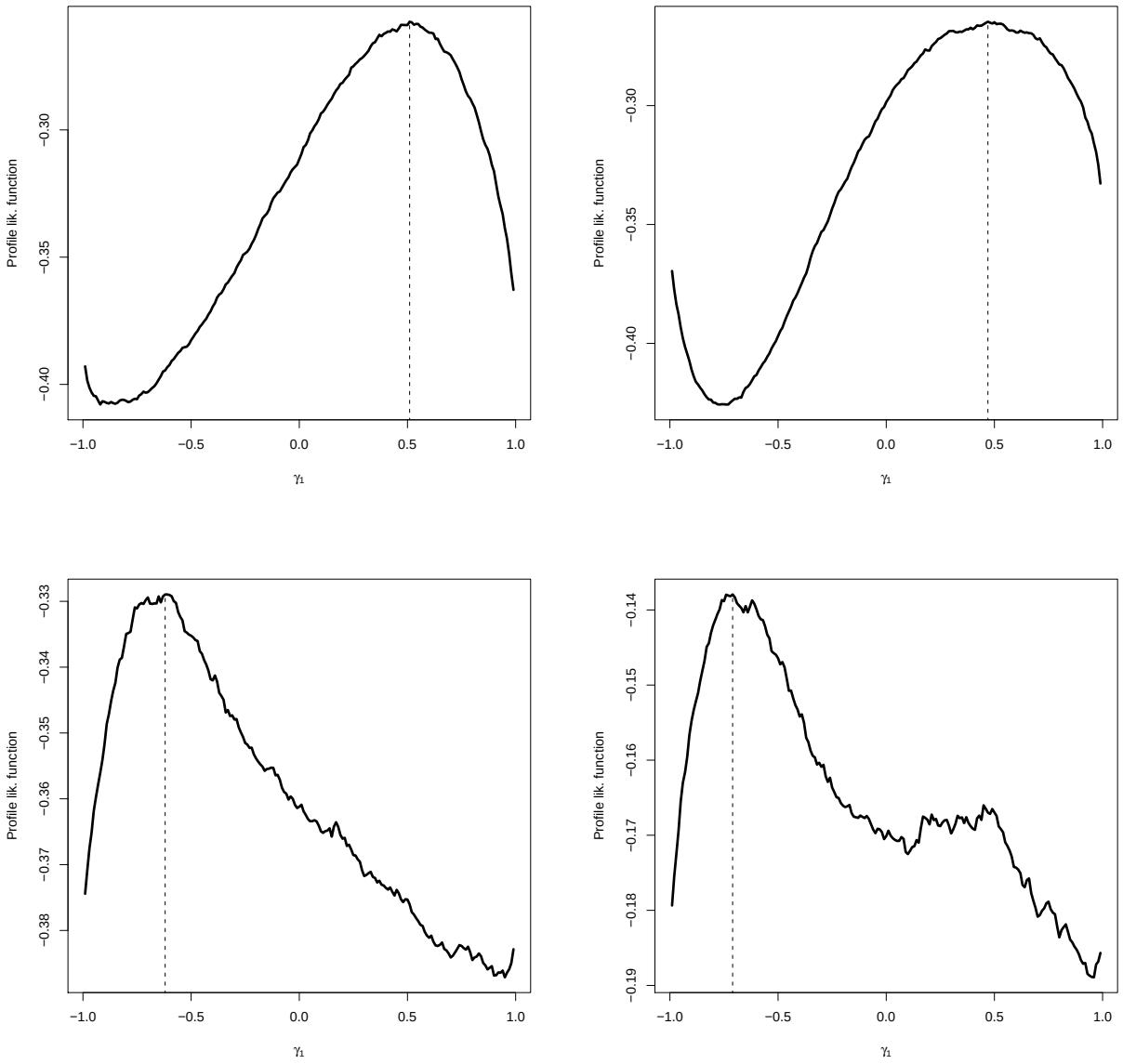

Figure 4.2: Profile stratified likelihood curve of γ_1 in PPSIH model

4.2.3 Bias Correction

The bias can be corrected by using essentially the same argument as those in Section 3.3.1. Recall that the corrected asymptotic limit is

$$cpl^s(\beta,\gamma) \equiv E\left[\Delta\left(\beta^T X + \log\frac{\frac{d}{dy}E\left(I(\Delta=1,Y\le y)|\gamma^T Z\right)}{E\left(I(Y\ge y)e^{\beta^T X}|\gamma^T Z\right)}\Big|_{y=Y}\right)\right].$$

Note that the difference between $cpl^s(\beta,\gamma)$ and $pl^s(\beta,\gamma)$ is

$$E\left[\Delta\log\left(\frac{d}{dy}\Big|_{y=Y}E\left(I(\Delta=1,Y\le y)|F_{\gamma^T Z}(\gamma^T Z)\right)\right)\right],$$

which can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i\sum_{k=1}^{J_n}\sum_{m=1}^{I_n}\Bigg[I\left(\mathbb{F}_n(\gamma^T Z_i)\in S_k, Y_i\in T_m\right) \times\log\left(\frac{I_n J_n}{nM}\sum_{j=1}^{n}\Delta_j I\left(\mathbb{F}_n(\gamma^T Z_j)\in S_k, Y_j\in T_m\right)\right)\Bigg]$$

uniformly in β and γ, where $S_k=[(k-1)/J_n, k/J_n)$, $T_m=[(m-1)M/I_n, mM/I_n)$, I_n and J_n are the number of strata for Y and $\gamma^T Z$ respectively and $[0,M]$ contains all values of Y. Note that, unlike for $\gamma^T Z$, we have equally partitioned the range of Y in the stratified approach. Hence, the corrected profile stratified likelihood function is

$$cpl_n^s(\beta,\gamma)=pl_n^s(\beta,\gamma)+\frac{1}{n}\sum_{k=1}^{J_n}\sum_{m=1}^{I_n}\Bigg[\left(\sum_{i=1}^{n}\Delta_i I\left(\mathbb{F}_n(\gamma^T Z_i)\in S_k, Y_i\in T_m\right)\right) \times\log\left(\frac{I_n J_n}{nM}\sum_{i=1}^{n}\Delta_i I\left(\mathbb{F}_n(\gamma^T Z_i)\in S_k, Y_i\in T_m\right)\right)\Bigg].$$

We denote its point of maximum as $(\tilde{\beta}_n,\tilde{\gamma}_n)$. We will show $(\tilde{\beta}_n,\tilde{\gamma}_n)$ is consistent with the following two additional conditions:

(C2) $\lambda_0(t,u)$ has non-zero partial derivative with respect to u.

(C3) The column vectors of the matrix $[1, X]$ and the column vectors of the matrix $[1, Z]$ are linearly independent. Furthermore, the support of Z given X contains the zero-vector.

Theorem 4.2.3. *Under conditions (C1)-(C3), suppose* $I_n/\sqrt{n} \to 0, J_n/\sqrt{n} \to 0, I_n \to \infty, J_n \to \infty$, *then* $\tilde{\beta}_n \xrightarrow{P} \beta_0$ *and* $\tilde{\gamma}_n \xrightarrow{P} \gamma_0$.

We again consider the same settings studied in Section 3.2.2, but now applying the corrected profile stratified likelihood method. A grid search with step size 0.01 was used to obtain the point of maximum. The number of strata for $\gamma^T Z$ and the number of strata for Y were chosen from $\{4, 8, 12\}$. In each setting for each sample size, we reported the best result. Simulation results reported in Table 4.3 suggest that the corrected profile stratified likelihood method works very well under every simulation setting. The profile stratified likelihood curves (corrected and uncorrected) in each setting based on a dataset of size 5000 were also plotted in Figure 4.3. The upper two panels pertain to case (i) and (ii), respectively; The bottom two panels pertain to case (iii) and (iv), respectively. The number of strata is 12 for $\gamma^T Z$. In each case, the corrected curve is maximized around the true value 0.5 of γ_1, suggesting that the corrected profile likelihood methods give estimators with little bias.

4.2.4 Data Application

In Section 3.4, by treating patient's ethnicity and baseline age as covariates of main interest while treating the remaining 4 biomarkers as "nuisance" covariates, the partly proportional single-index hazards model (PPSIH Model (2)) fits the MACS data set well. The results reported in Table 3.3 suggest that the "nuisance" covariates log(CD4) and neopterin are not significant at the .05 significance level. Thus it appears reasonable to consider a model (PPSIH model (3)) by still treating patient's ethnicity and baseline age as covariates of main interest, but only controlling for the significant "nuisance" covariates

Table 4.3: Simulation results of corrected stratified likelihood in PPSIH model

			Corrected stratified likelihood	
Simulation settings	Sample size	Parameters	Bias	SE
(i) $X \perp Z$	100	β	-.006	.185
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.016	.228
$cov(Z) = 0$	200	β	.006	.120
		γ_1	-.013	.171
	400	β	-.004	.079
		γ_1	-.004	.108
(ii) $X \perp Z$	100	β	-.014	.184
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.030	.278
$cov(Z) = 0.5$	200	β	-.002	.121
		γ_1	-.001	.223
	400	β	-.000	.082
		γ_1	-.001	.158
(iii) $X \not\perp Z$	100	β	.127	.227
$\lambda_0(t,u) = 0.5e^u t$		γ_1	-.101	.369
$cov(Z) = 0.5$	200	β	.079	.138
		γ_1	-.037	.233
	400	β	.035	.093
		γ_1	-.017	.175
(iv) $X \not\perp Z$	100	β	.075	.207
$\lambda_0(t,u) = 0.25e^{e^u t}$		γ_1	-.092	.357
$cov(Z) = 0.5$	200	β	.035	.135
		γ_1	-.027	.212
	400	β	.010	.086
		γ_1	-.017	.131

NOTE: Each entry is based on 1000 replicates.

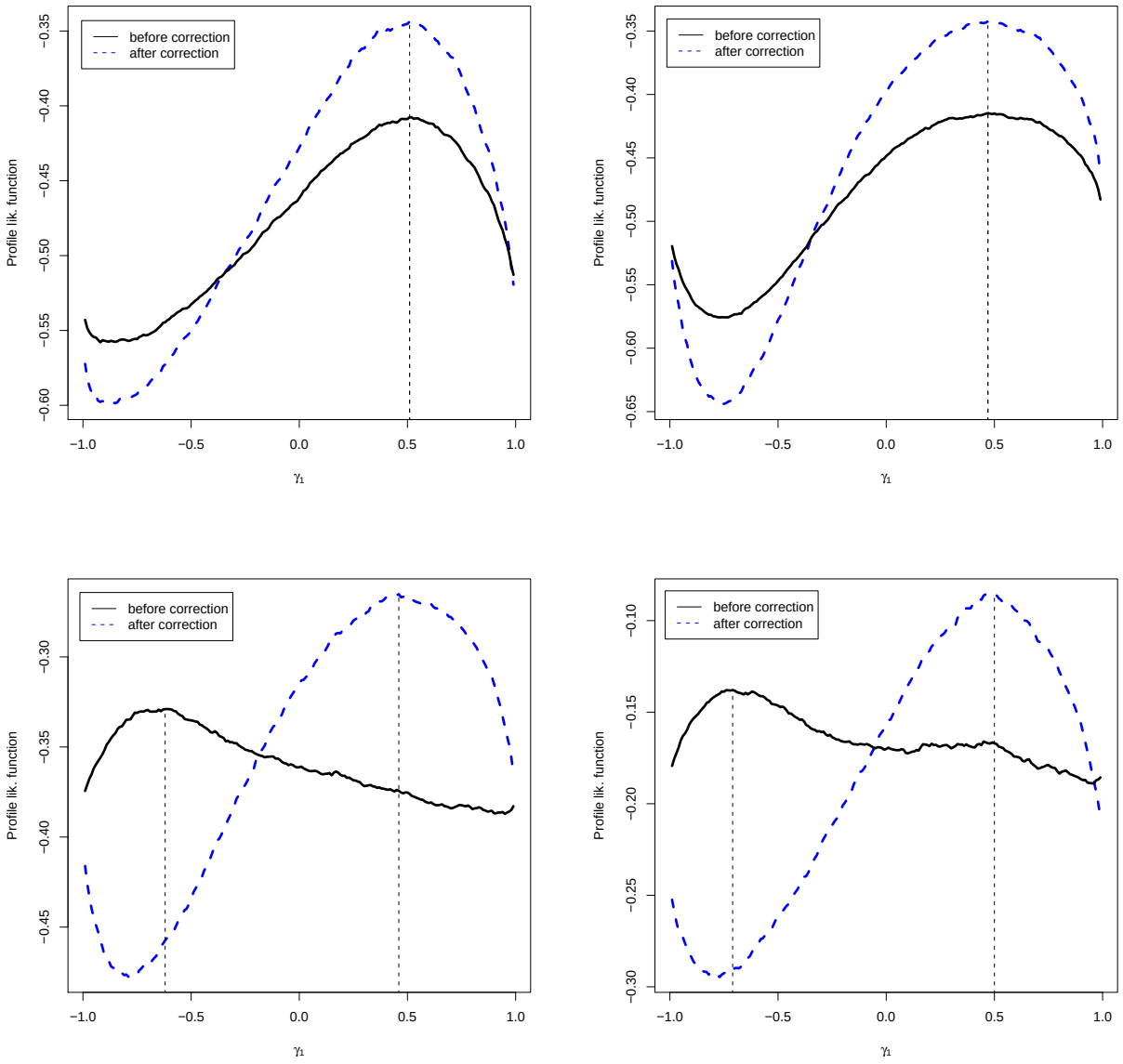

Figure 4.3: Profile stratified likelihood curves (corrected and uncorrected) of γ_1 in PPSIH model

Table 4.4: Analysis of MACS Data under PPSIH Model (3)

	Corrected Profile Stratified			Corrected Profile Local		
Parameter	Est.	SE	p-value	Est.	SE	p-value
age	.022	.053	.678	.016	.065	.805
white	.198	.258	.443	.296	.264	.264
log(viral)	.903	.116	<.001	.990	.007	<.001
microgloburin	.285	.299	.341	.139	.053	.009

NOTE: "white" is an indicator for whites. Est. and SE denote the parameter estimate and (estimated) standard error, respectively.

log(viral) and microgloburin. We fit this model using both the corrected profile stratified likelihood and the corrected profile local likelihood methods.

For the corrected profile stratified likelihood method, we choose the number of strata for the single-index and the survival time Y to be 8 and 12, respectively. The variances are estimated using 500 bootstrap samples. For the corrected profile local likelihood method, the bandwidths are $c_i \times IQR_i \times n^{-1/3}$ for point estimation and $d_i \times IQR_i \times n^{-1/4}$ for variance estimation, where c_i is selected using cross-validation and we set $d_i = c_i$ for simplicity, $i = 1, 2$. We choose $c_1 = 1$ and $c_2 = 3.5$.

Table 4.4 summarize the results. It is observed that the variance estimators for "nuisance" covariates (log(viral) and microgloburin) are much larger under the stratification method than those under the local likelihood method. This again suggests that, compared to the corrected profile local likelihood method, the corrected profile stratified method may not produce efficient parameter estimators, especially the estimators for the single-index coefficient γ. For the MACS data, one would conclude that the covariate microgloburin is not significant using the stratification approach while it is highly significant under the local likelihood approach.

4.3 Proofs of Theorems

Proof of Theorem 4.1.1

$$-pl_n^s(\gamma) = \sum_{k=1}^{J_n} \mathbb{P}_n \left[\Delta I\left(\mathbb{F}_n(\gamma^T W) \in S_k\right) \log\left(\frac{J_n}{n} \sum_j I(Y_j \geq Y) I\left(\mathbb{F}_n(\gamma^T W_j) \in S_k\right)\right)\right].$$

Since $J_n/\sqrt{n} \to 0$ and by Donsker arguments, we have

$$\left| \frac{J_n}{n} \sum_j I(Y_j \geq y) I\left(\mathbb{F}_n(\gamma^T W_j) \in S_k\right) - J_n E\left[I(Y \geq y) I(F_{\gamma^T W}(\gamma^T W) \in S_k)\right] \right| \xrightarrow{P} 0,$$

uniformly in y and γ, where $F_{\gamma^T W}(\cdot)$ is the distribution function of $\gamma^T W$. Note that

$$J_n E\big(I(Y \geq y) I(F_{\gamma^T W}(\gamma^T W) \in S_k)\big)$$
$$= E\left[I(Y \geq y) | F_{\gamma^T W}(\gamma^T W) = (k-1)/J_n \right] + o(1). \quad \text{Hence,}$$

$$\left| \frac{J_n}{n} \sum_j I(Y_j \geq y) I\left(\mathbb{F}_n(\gamma^T W_j) \in S_k\right) - E\left[I(Y \geq y) | F_{\gamma^T W}(\gamma^T W) = (k-1)/J_n \right] \right| \xrightarrow{P} 0,$$

uniformly in y and γ. Next, by either Glivenko-Cantelli or Donsker arguments,

$$\left| -pl_n^s(\gamma) - \sum_{k=1}^{J_n} E\left[\Delta I(F_{\gamma^T W}(\gamma^T W) \in S_k) \log E\left[I(Y \geq y) | F_{\gamma^T W}(\gamma^T W) = \frac{k-1}{J_n} \right] \Big|_{y=Y} \right] \right|$$
$$\xrightarrow{P} 0,$$

uniformly in γ. The second term inside the absolute value of the above display equals

$$\sum_{k=1}^{J_n} \frac{1}{J_n} E\left[\Delta \log E\left[I(Y \geq y) | F_{\gamma^T W}(\gamma^T W) \right] \Big|_{y=Y} \Big| F_{\gamma^T W}(\gamma^T W) = (k-1)/J_n \right],$$

which converges to

$$\begin{aligned}\int_0^1 E\left[\Delta \log E\left[I(Y \geq y)|F_{\gamma^T W}(\gamma^T W)\right]\Big|_{y=Y}\Big| F_{\gamma^T W}(\gamma^T W) = u\right] f_{F_{\gamma^T W}(\gamma^T W)}(u)du \\ = E\left\{E\left[\Delta \log E\left[I(Y \geq y)|F_{\gamma^T W}(\gamma^T W)\right]\Big|_{y=Y}\Big| F_{\gamma^T W}(\gamma^T W)\right]\right\} \\ = E\left[\Delta \log E\left[I(Y \geq y)|F_{\gamma^T W}(\gamma^T W)\right]\Big|_{y=Y}\right].\end{aligned}$$

Hence, we have shown that $pl_n^s(\gamma)$ converges uniformly in γ to $pl^s(\gamma)$.

Proof of Theorem 4.1.2

By Theorem 2.12 of Kosorok (2008) and Theorem 4.1.1, it suffices to show that γ_0 is the unique maximizer of $pl^s(\gamma)$. Since

$$\begin{aligned}pl^s(\gamma) &= -E\left[\int \log P\left(Y \geq t|\gamma^T W\right) f_{T|W}(t)G_C(t)dt\right] \\ &= -E\left[\int \log\left(G_C(t)E(S_{T|W}(t)|\gamma^T W)\right) f_{T|W}(t)G_C(t)dt\right] \\ &= -E\left[\int \log E(S_{T|W}(t)|\gamma^T W) f_{T|W}(t)G_C(t)dt\right] - E\left[\int \log G_C(t) f_{T|W}(t)G_C(t)dt\right] \\ &= -E\left[\int \left(E(f_{T|W}(t)|\gamma^T W)\log E(S_{T|W}(t)|\gamma^T W)\right) G_C(t)dt\right] \\ &\quad - E\left[\int \log G_C(t) f_{T|W}(t)G_C(t)dt\right]\end{aligned}$$

and

$$\begin{aligned}E(f_{T|W}(t)|\gamma^T W)\log E(S_{T|W}(t)|\gamma^T W) = &-\frac{d}{dt}\left[E(S_{T|W}(t)|\gamma^T W)\log E(S_{T|W}(t)|\gamma^T W)\right] \\ &- E\left(f_{T|W}(t)|\gamma^T W\right),\end{aligned}$$

$$pl^s(\gamma) = \int \frac{d}{dt} \left\{ E\left[E(S_{T|W}(t)|\gamma^T W) \log E(S_{T|W}(t)|\gamma^T W) \right] \right\} G_C(t)dt + \int E\left[f_{T|W}(t) \right] G_C(t)dt$$
$$- E\left[\int \log G_C(t) f_{T|W}(t) G_C(t)dt \right].$$

The first term of the previous display equals

$$\begin{aligned}
&\int G_C(t)dE\left[E(S_{T|W}(t)|\gamma^T W) \log E(S_{T|W}(t)|\gamma^T W) \right] \\
&\qquad = \int E\left[E(S_{T|W}(t)|\gamma^T W) \log E(S_{T|W}(t)|\gamma^T W) \right] f_C(t)dt \\
&\qquad \leq \int E\left[E\left(S_{T|W}(t) \log S_{T|W}(t)|\gamma^T W \right) \right] f_C(t)dt \\
&\qquad = \int E\left(S_{T|W}(t) \log S_{T|W}(t) \right) f_C(t)dt,
\end{aligned}$$

where the inequality follows from Jensen's inequality since $g(x) \equiv x \log x$ is a convex function. Therefore, for any γ, $pl^s(\gamma)$ is less than or equal to

$$\begin{aligned}
&\int E\left(S_{T|W}(t) \log S_{T|W}(t) \right) f_C(t)dt + \int E\left[f_{T|W}(t) \right] G_C(t)dt \\
&\qquad - E\left[\int \log G_C(t) f_{T|W}(t) G_C(t)dt \right] \\
&= - \int E\left(S_{T|W}(t) \log S_{T|W}(t) \right) dG_C(t) + \int E\left[f_{T|W}(t) \right] G_C(t)dt \\
&\qquad - E\left[\int \log G_C(t) f_{T|W}(t) G_C(t)dt \right] \\
&= \int G_C(t)dE\left(S_{T|W}(t) \log S_{T|W}(t) \right) + \int E\left[f_{T|W}(t) \right] G_C(t)dt \\
&\qquad - E\left[\int \log G_C(t) f_{T|W}(t) G_C(t)dt \right] \\
&= - \int G_C(t) E\left[f_{T|W}(t) \log P(Y \geq t|W) \right] dt \\
&= - E\left[\Delta \log P(Y \geq t|\gamma_0^T W)|_{t=Y} \right] \\
&= pl^s(\gamma_0).
\end{aligned}$$

Suppose $pl^s(\gamma^*) = pl^s(\gamma_0)$, then conditional on $\gamma^{*T} W$, $S_{T|W}(t)$ is a constant almost surely

since the function $g(\cdot)$ is strictly convex. That is, $S_{T|W}(t) = h(t, \gamma^{*T}W)$ almost surely for some function $h(\cdot)$. After taking the derivative with respect to W on both sides, we obtain $\gamma^* \propto \gamma_0$. The proof is complete in view of the requirements that γ^* has a unit norm with one positive component.

Proof of Theorem 4.2.1

Denote $pl_n^s(\beta, \gamma)$ by (1) − (2). (1) converges uniformly on a compact set of β to $E[\Delta\beta^T X]$ as in Theorem 2. Since $J_n/\sqrt{n} \to 0$ and by Donsker arguments, we have

$$\left| \frac{J_n}{n} \sum_j I(Y_j \geq y) e^{\beta^T X_j} I\left(\mathbb{F}_n(\gamma^T Z_j) \in S_k\right) - E\left[I(Y \geq y) e^{\beta^T X} | F_{\gamma^T Z}(\gamma^T Z) = \frac{k-1}{J_n} \right] \right| \xrightarrow{P} 0,$$

uniformly in y, β and γ. Next, by Donsker arguments again,

$$\left| (2) - \sum_{k=1}^{J_n} E\left(\Delta I(F_{\gamma^T Z}(\gamma^T Z) \in S_k) \log E\left[I(Y \geq y) e^{\beta^T X} | F_{\gamma^T Z}(\gamma^T Z) = \frac{k-1}{J_n} \right] \Big|_{y=Y} \right) \right| \xrightarrow{P} 0,$$

uniformly in β and γ. The second term inside the absolute value of the above display equals

$$\sum_{k=1}^{J_n} \frac{1}{J_n} E\left[\Delta \log E\left[I(Y \geq y) e^{\beta^T X} | F_{\gamma^T Z}(\gamma^T Z) \right] \Big|_{y=Y} \Big| F_{\gamma^T Z}(\gamma^T Z) = (k-1)/J_n \right],$$

which converges to $E\left[\Delta \log E\left[I(Y \geq y) e^{\beta^T X} | F_{\gamma^T Z}(\gamma^T Z) \right] \Big|_{y=Y} \right]$. Hence, we have shown $pl_n^s(\beta, \gamma)$ converges uniformly in β and γ to $pl^s(\beta, \gamma)$.

Proof of Theorem 4.2.2

Part (i) follows by using the same arguments as used in the proof of Theorem 3.2.2.

The independence between C and (X, Z) together with the independence between X and Z imply that

$$-\frac{\partial}{\partial\gamma}\Big|_{\beta_0,\gamma_0} pl^s(\beta,\gamma) \;=\; \iint \lambda_0(y,\gamma_0^T w) f_Z(w) \nabla_\gamma \left(E(I(Y \geq y) e^{\beta_0^T X} | \gamma^T Z = \gamma^T w) \right) dy dw.$$

Note the gradient equals

$$\begin{aligned}
&\nabla_\gamma \left[\frac{E(I(Y \geq y) e^{\beta_0^T X} | \gamma^T Z = \gamma^T w) f_{\gamma^T Z}(\gamma^T w)}{f_{\gamma^T Z}(\gamma^T w)} \right] \\
&\qquad = \nabla_\gamma \left[\frac{\lim_{h\to 0} \frac{1}{h} E_Z \left[K\left(\frac{\gamma^T Z - \gamma^T w}{h} \right) g(y, \gamma_0^T Z) \right]}{f_{\gamma^T Z}(\gamma^T w)} \right] \\
&\qquad = g_2'(y, \gamma_0^T w)(w - r(\gamma_0^T w)),
\end{aligned}$$

where $g(y, \gamma_0^T Z)$ and $r(\cdot)$ are defined in the proof of Theorem 3.2.2 and we also use the kernel representation of $f_{\gamma^T Z}(\gamma^T w)$ by $\lim_{h\to 0} E[K(\gamma^T(Z - w)/h)/h]$. Therefore,

$$-\frac{\partial}{\partial\gamma}\Big|_{\beta=\beta_0,\gamma=\gamma_0} pl^s(\beta,\gamma) = \int E_Z \left[\lambda_0(y, \gamma_0^T Z) g_2'(y, \gamma_0^T Z) \left(Z - r(\gamma_0^T Z) \right) \right] dy = 0.$$

Proof of Theorem 4.2.3

By Theorem 4.2.1, $\sup_{\beta,\gamma} |pl_n^s(\beta,\gamma) - pl^s(\beta,\gamma)| \to_P 0$, as $n \to \infty$. Similar arguments to those used in the proof of Theorem 4.2.1 can be used to show

$$\frac{I_n J_n}{nM} \sum_{j=1}^{n} \Delta_j I\left(\mathbb{F}_n(\gamma^T Z_j) \in S_k, Y_j \in T_m \right)$$

converges uniformly in β and γ to

$$E\left(\Delta | Y, F_{\gamma^T Z}(\gamma^T Z) \right) f_{Y, F_{\gamma^T Z}(\gamma^T Z)}\left(Y, F_{\gamma^T Z}(\gamma^T Z) \right) \Big|_{Y=(m-1)/I_n, F_{\gamma^T Z}(\gamma^T Z)=(k-1)/J_n}.$$

Denote the above quantity by $g((m-1)/I_n,(k-1)/J_n)$. Also,

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i\sum_{k=1}^{J_n}\sum_{m=1}^{I_n}\Bigg[I\left(\mathbb{F}_n(\gamma^T Z_i)\in S_k, Y_i\in T_m\right) \times\log\left(\frac{I_nJ_n}{nM}\sum_{j=1}^{n}\Delta_j I\left(\mathbb{F}_n(\gamma^T Z_j)\in S_k, Y_j\in T_m\right)\right)\Bigg]$$

converges uniformly in β and γ to

$$\begin{aligned}&\sum_{k=1}^{J_n}\sum_{m=1}^{I_n}E\left(\Delta I\left(F_{\gamma^TZ}(\gamma^TZ)\in S_k, Y\in T_m\right)\log g\left((m-1)/I_n,(k-1)/J_n\right)\right)\\ &\qquad\longrightarrow\iint g(y,u)\log g(y,u)dydu\\ &\qquad= E\left[\Delta\log\left(\frac{d}{dy}\Big|_{y=Y}E\left(I(\Delta=1,Y\le y)|F_{\gamma^TZ}(\gamma^TZ)\right)\right)\right].\end{aligned}$$

The last equality follows since $F_{\gamma^TZ}(\gamma^TZ)$ follows a uniform distribution. Therefore, the corrected profile stratified likelihood $cpl^s(\beta,\gamma)$ converges uniformly to

$$E\left[\Delta\left(\beta^TX+\log\frac{\frac{d}{dy}E\left(I(\Delta=1,Y\le y)|\gamma^TZ\right)}{E\left(I(Y\ge y)e^{\beta^TX}|\gamma^TZ\right)}\Big|_{y=Y}\right)\right].$$

This coincides with the limit of the corrected profile local likelihood $cpl^{loc}(\beta,\gamma)$. The rest of proof is identical to that given in Theorem 3.3.1.

Chapter 5

Discussion

In this dissertation, we have proposed the single-index hazards model for right-censored survival data. The commonly used profile local likelihood approach was considered. Even under the restrictive condition that the censoring time C and the covariate vector W are independent, the profile local likelihood method gives inconsistent estimation in general. Therefore, this method should not be used for this single-index hazards model. In contrast, under this independent censoring assumption and some other regularity conditions, the profile stratified likelihood method always yields consistent estimation. We note that the stratification needs to be based on sample quantiles of the single-index since an equally spaced stratification on the original scale of the single-index would lead to the same limiting profile likelihood function as under the local likelihood approach and thus the same estimation bias would occur.

In addition to the independent censoring assumption, another requirement in order for the stratified likelihood approach to be a consistent procedure in the single-index hazards model is a positive probability of censoring in the data, as guaranteed by the second part of condition (C3) in Section 4.1.2. If there is no censoring present, it can be shown that $pl_n^s(\gamma)$ is free of γ and converges to the constant 1 (the limit function $pl^s(\gamma) \equiv 1$ as well.), and thus cannot be used for parameter estimation. That the presence of censoring is required to achieve consistency is quite surprising.

We note that the independent censoring assumption is crucial for the stratified likelihood approach to work. Without this assumption, we conjecture that the stratified likelihood method fails, which is demonstrated numerically. Therefore, one should not use the stratified likelihood approach either unless the independent censoring can be reasonably assumed. One possible way to relax this restrictive assumption is to modify the limiting profile stratified likelihood function so that the modified function has a unique maximizer at the true parameter value without assuming the independence between C and W. We then can make a corresponding modification in the original profile stratified likelihood function and use it for estimation.

Besides the aforementioned methods, the spline method can also possibly be used for parameter estimation in our single-index hazards model. For example, Yu and Ruppert (2002) considered the penalized spline method in a partially linear single-index model and they showed that their method outperforms the local likelihood method adopted by Carroll et al. (1997). It would then be worthwhile to examine a spline method for our model and to investigate whether or not the estimation bias issue still exists.

The existence and nature of the failures of the two commonly used estimation approaches considered is somewhat surprising and suggest that nonstandard approaches may be needed. In addition to the aforementioned spline approach, there are yet other approaches which may need to be considered in order to find an estimator that is consistent under realistically general conditions.

One drawback of the single-index hazards model is that the interpretation of covariate effects is in general difficult. In practice, it is of great interest to have a model which can address the effect of covariates of primary interest, while allows for flexible modeling of effects of "nuisance" covariates. In this spirit, we have proposed the partly proportional single-index model. The conventional profile-kernel method was studied under this model and the profile likelihood formed by plugging the baseline hazard estimator (3.6) into the nonparametric maximum likelihood leads to biased estimation in the regression

parameters. However, it is interesting, as shown in Theorem 3.3.3, that the baseline hazard estimator (3.6) is in fact consistent provided that consistent estimators for β and γ can be obtained. Thus the bias is not due to "bad" estimation of the infinite dimensional parameter, but due to an "unbalanced" structure resulting from combining a local baseline hazard estimator and the nonparametric maximum likelihood function. Note that this conventional profile local likelihood method may work under some stringent conditions (e.g. those stated in Theorem 3.2.2). Similar phenomena was also observed in the longitudinal data setting (Lin and Carroll 2001).

Since the partly proportional single-index hazards model can be viewed as a stratified Cox model, we can construct another profile likelihood based on stratification of the single-index. Simulation results reveal that this approach may work under the stringent conditions of independent censoring and the independence between covariates of primary interest and "nuisance" covariates, but in general it can lead to biased estimation in the regression parameters as well.

An ad hoc approach has been proposed to correct the bias in both the profile local likelihood method and the profile stratified likelihood method. Simulation studies suggest that the standard errors using the corrected profile stratified likelihood approach are always bigger than those using the corrected profile local likelihood approach. Thus the former method may yield estimators not as efficient as those estimated from the latter method. However, it would still be worthwhile to investigate the asymptotic properties (semiparametric efficiency in particular) of this stratification-based method.

In the partly proportional single-index hazards model, covariates of primary interest need to satisfy the proportional hazards assumption. Although this assumption can be checked in view of the similarity between our model and the stratified (on the single-index) Cox model, development of direct model checking techniques would be of great interest.

Due to the so-called "curse of dimensionality", dimension reduction is an important

issue in model estimation. Recent work in survival analysis includes Sun et al. (2008) who studied a partially linear proportional hazards model in which a single-index was used for dimension reduction and Xia et al. (2010) who proposed a novel dimension reduction method to estimate the conditional hazard function via estimation of the central subspace in a general model which includes the transformation model (Zeng and Lin 2007b) and the accelerated failure time model (Cox and Oakes 1984, chap. 5) as its special cases. Similarly in our partly proportional single-index hazards model, the single-index is introduced for dimension reduction so that the nonparametric estimation of the baseline hazard function becomes feasible. In some sense, a single-index can be viewed as a principle component of the "nuisance" covariate vector. When the dimension of the "nuisance" vector is high, one may wish to include multiple principle components into the model. Thus it may be attractive to consider a partly proportional multiple-index hazards model. Models involving multiple single-indices have been recently studied by Ichimura and Lee (1991), Horowitz (1998) and Xia (2008), among others. The challenges in this partly proportional multiple-index model setting would then be the choice of the number of single-indices and the statistical inference.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics,* 10, 1100–1120.

Bennett, 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*, New York: Springer.

Buckley, J., and James, I. (1979), "Linear Regression with Censored Data," *Biometrika,* 66, 429–436.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association,* 92, 477–489.

Chen, X., Wang, L., Smith, J. D., and Zhang, B. (2008), "Supervised Principle Component Analysis for Gene Set Enrichment of Microarray Data with Continuous or Survival Outcomes," *Bioinformatics,* 24, 2474–2481.

Chiou, J. and Müller, H. (2004), "Quasi-Likelihood Regression with Multiple Indices and Smooth Link and Variance Functions," *Scandinavian Journal of Statistics,* 31, 367–386.

Chow, T., Kereiakes, D. J., Bartone, C., Booth, T., Schloss, E.J., Waller, T., Chung, E.S., Menon, S., Nallamothu, B. K., and Chan, P.S. (2006), "Prognostic Utility of Microvolt T-Wave Alternans in Risk Stratification of Patients With Ischemic Cardiomyopathy," *Journal of the American College of Cardiology,* 47, 1820–1827.

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics,* 30, 455–474.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of Royal Statistical Society, Ser. B,* 34, 187–220.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika,* 62, 269–276.

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman & Hall.

Dabrowska, D. M. (1997), "Smoothed Cox Regression," *The Annals of Statistics,* 25, 1510–1540.

Dabrowska, D. M., and Doksum, K. A. (1988). Partial Likelihood in Transformation Models with Censored Data. *Scandinavian Journal of Statistics,* 15, 1–23.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956), "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of Mathematical Statistics,* 27, 642–669.

Fan, J., Gijbels, I., and King, M. (1997), "Local Likelihood and Local Partial Likelihood in Hazard Regression," *The Annals of Statistics,* 25, 1661–1690.

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Greenland, S. (1989), "Modeling and Variable Selection in Epidemiologic Analysis," *American Journal of Public Health,* 79, 340–349.

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics,* 21, 157–178.

Heller, G. (2001), "The Cox Proportional Hazards Model with a Partly Linear Relative Risk Function," *Lifetime Data Analysis,* 7, 255–277.

Horowitz, J. (1998), *Semiparametric Methods in Econometrics*, New York: Springer.

Huang, J. (1999), "Efficient Estimation of the Partly Linear Additive Cox Model," *The Annals of Statistics,* 27, 1536–1563.

Huang, J.Z., and Liu, L. (2006), "Polynomial Spline Estimation and Inference of Proportional Hazards Regression Models with Flexible Relative Risk Form," *Biometrics,* 62, 269–276.

Ichimura, H., and Lee, L. F. (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics,* eds. W. A. Barnett, J. Powell, and G. Tauchen, Cambridge, U.K.: Cambridge University Press.

Kong, S. W., Pu, W. T., Park, P. J. (2006), "A Multivariate Approach for Integrating Genome-Wide Expression Data and Biological Knowledge," *Bioinformatics,* 22, 2373–2380.

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), "Robust Inference for Univariate Proportional Hazards Frailty Regression Models," *The Annals of Statistics,* 32, 1448–1491.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993), "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals," *Biometrika,* 80, 557–572.

Lin, X., and Carroll, R. J. (2001), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association,* 96, 1045–1056.

Lu, X., Chen, G., Song, X.-K., and Singh, R. S. (2006), "A Class of Partially Linear Single-Index Survival Models," *The Canadian Journal of Statistics,* 34, 97–112.

Lu, X, Singh, R. S., and Desmond, A. F. (2001), "A Kernel Smoothed Semiparametric Survival Model," *Journal of Statistical Planning and Inference,* 98, 119–135.

Ma, S., and Kosorok, M. R. (2009), "Identification of Differential Gene Pathways with Principle Component Analysis," *Bioinformatics,* 25, 882–889.

Mellors, J. W., Munoz, A., Giorgi, J. V., Margolick, J. B., Tassoni, C. J., Gupta, P., Kingsley, L. A., Todd, J. A., Saah, A. J., Detels, R., Phair, J. P., and Rinaldo Jr., C. R. (1997), "Plasma Viral Load and CD4+ Lymphocytes as Prognostic Markers of HIV-1 Infection," *The Annals of Internal Medicine,* 126, 946–954.

Motzer, R. J., Mazumdar, M., Bacik, J., Berg, W., Amsterdam, A., and Ferrara, J. (1999), "Survival and Prognostic Stratification of 670 Patients with Advanced Renal Cell Carcinoma," *Journal of Clinical Oncology,* 17, 2530–2540.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximal Likelihood Estimation in the Proportional Odds Model," *Journal of the American Statistical Association,* 92, 968–976.

Naik, P. A., and Tsai, C. (2001), "Single-Index Model Selection," *Biometrika,* 88, 821–832.

Nielsen, J. P., and Linton, O. B. (1995), "Kernel Estimation in a Nonparametric Marker Dependent Hazard Model," *The Annals of Statistics,* 23, 1735–1748.

Nielsen, J. P., Linton, O. B., and Bickel, P. J. (1998), "On a Semiparametric Survival Model with Flexible Covariate Effect," *The Annals of Statistics,* 26, 215–241.

Pettitt, A. N. (1984), "Proportional Odds Models for Survival Data and Estimates Using Ranks," *Applied Statistics,* 33, 169–175.

Prentice, R. L. (1978), "Linear Rank Tests with Right Censored Data," *Biometrika,* 65, 167–179.

Sasieni, P. (1992a), "Information Bounds for the Conditional Hazard Ratio in a Nested Family of Regression Models," *Journal of Royal Statistical Society, Ser. B,* 54, 617–635.

Sasieni, P. (1992b), "Non-orthogonal Projections and Their Application to Calculating the Information in a Partly Linear Cox Model," *Scandinavian Journal of Statistics,* 19, 215–233.

Scharfstein, D. O., Tsiatis, A. A., and Gilbert, P. B. (1998), "Semiparametric Efficient Estimation in the Generalized Odds-Rate Class of Regression Models for Right-Censored Time-to-Event Data," *Lifetime Data Analysis,* 4, 355–391.

Slud, E. V., and Vonta, F. (2004), "Consistency of the NPML Estimator in the Right-Censored Transformation Model," *Scandinavian Journal of Statistics,* 31, 21–41.

Sun, J., Kopciuk, K. A., and Lu, X. (2008), "Polynomial Spline Estimation of Partially Linear Single-Index Proportional Hazards Regression Models," *Computational Statistics and Data Analysis,* 53, 176–188.

Wang, W. (2004), "Proportional Hazards Regression Model with Unknown Link Function and Time-Dependent Covariates," *Statistica Sinica,* 14, 885–905.

Xia, Y. (2008), "A Multiple-Index Model and Dimension Reduction," *Journal of the American Statistical Association,* 103, 1631–1640.

Xia, Y., and Härdle, W. (2006), "Semi-parametric Estimation of Partially Linear Single-Index Models," *Journal of Multivariate Analysis,* 97, 1162–1184.

Xia, Y., Tong, H., and Li, W. K. (1999), "On Extended Partially Linear Single-Index Models," *Biometrika,* 86, 831–842.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of Royal Statistical Society, Ser. B,* 64, 363–410.

Xia, Y., Zhang, D., and Xu, J. (2010), "Dimension Reduction and Semiparametric Estimation of Survival Models," *Journal of the American Statistical Association,* 105, 278–290.

Yin, X., and Cook, R. D. (2002), "Dimension Reduction for the Conditional k-th Moment in Regression," *Journal of Royal Statistical Society, Ser. B,* 64, 159–175.

Yu, Y., and Ruppert, D. (2002), "Penalized Spline Estimation for Partially Linear Single-Index Models," *Journal of the American Statistical Association,* 97, 1042–1054.

Zeng, D., and Lin, D. Y. (2007a), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association,* 102, 1387–1396.

Zeng, D., and Lin, D. Y. (2007b), "Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data" (with discussion), *Journal of Royal Statistical Society, Ser. B,* 69, 507–564.