THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Semiparametric survival models for routine register data

Frank Eriksson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2010

Semiparametric survival models for routine register data
Frank Eriksson
NO 2010:26
ISSN 1652-9715

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX.
Printed in Göteborg, Sweden 2010

Semiparametric survival models for routine register data

Frank Eriksson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg

Abstract

Routine registers offer researchers opportunities to carry out studies of covariate effects on lifetimes of rare diseases otherwise infeasible because of the large cohorts required. Familial relationships necessary for analysis of environmental or genetic factors can be identified by record linking. The vast amount of data and clustering of related individuals pose statistical challenges. As most statistical information is associated with the cases, an estimator based on a sample where cases are overrepresented can drastically reduce the sample size with only a minor loss of efficiency.

This thesis concerns regression of clustered cohort sampled survival data within the broad class of semiparametric transformation models. This class includes the proportional hazards and proportional odds models as special cases. Correlations within clusters are modeled by random effects.

We derive consistency and asymptotic normality of a weighted maximum likelihood estimator and provide a consistent estimator of its asymptotic variance. A likelihood ratio test for regression coefficients is also proposed. The method is shown to perform well on simulated data and is illustrated by application to a study on cardiovascular diseases among Swedish men.

Key words: Survival, transformation, clustered, likelihood, random effects, case-cohort, routine register, semiparametric.

Preface

This thesis consists of two parts. The purpose of the introductory part is to informally introduce the topics of the second part where an article constituting the scientific contribution of the thesis can be found. The introduction ends with a short summary of the article intended to connect part one and two. Part two aims at being self-contained in the sense that readers familiar with nonparametric maximum likelihood in survival analysis and routine registers can skip the introduction.

Acknowledgments

I would like to thank my supervisors Dragi Anevski and Marita Olsson.

Contents

I Introduction **vii**

1 Background **1**

1.1 Survival analysis 1

1.2 Register data 13

2 Summary of the paper **17**

II The paper **21**

1 Introduction 23

2 Data structure and model assumptions 25

3 Weighted maximum likelihood estimation 28

4 Simulation studies 33

5 An example 36

6 Discussion 39

A Proofs 40

Part I

Introduction

Chapter 1

Background

1.1 Survival analysis

In survival analysis the response variable is the time T to a specific event. Typically some event times are incompletely observed due to censoring. The most commonly encountered censoring is right censoring when we only observe an individual up to a possibly random censoring time C, i.e. we observe $Y = T \wedge C$ and an indicator $\Delta = I(Y = T)$ of whether or not censoring has occurred before the event time of interest. This may be because the subject has still to experience the event when the study is closed or because the subject is lost for follow-up due to other reasons. If T is the time to death from a given cause, then death from another cause may be regarded as a censored observation. Other types of censoring include interval censoring, left-censoring, truncation or filtering, see Andersen et al. (1993). Although not always necessary, we assume that the study ends at a time $\tau < \infty$ and that all individuals still alive at this time are censored.

It is not obvious at first glance how to incorporate censored observations into inference for the distribution of T. Estimation based only on the complete data may give biased results, so the censored observations need to be taken into account. Modeling of the hazard rate λ, the event rate at time t conditional on survival until time t, has proven to be highly successful for this purpose.

The hazard rate may be interpreted as the instantaneous failure rate among

those at risk and is given by

$$\begin{aligned} \lambda(t)dt &= P(t \leq T < t + dt | T \geq t) \\ &= -\frac{S'(t)dt}{S(t)} \\ &= -\frac{\partial}{\partial t} \log S(t), \end{aligned} \tag{1}$$

where $S(t) = P(T > t)$ is the survival function of T, the probability that the event of interest has not happened at time t. From (1), by integration and using $S(0) = 1$, we see that the survival function may be calculated from the hazard rate as

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(s)ds\right) \\ &= \exp\left(-\Lambda(t)\right), \end{aligned} \tag{2}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is called the cumulative hazard rate up to time t. Note that by (1) and (2), the survival function and hazard rate are alternative but equivalent representations and both completely specify the distribution of T.

1.1.1 Regression models for failure time data

A typical goal of a survival study is to relate the effect of explanatory variables on survival. It is convenient to build a regression model using the conditional hazard rate as target function. The model can then be used to examine various hypotheses about the impact of risk factors or estimate regression parameters that relate to the lifetimes, taking into account that some of the lifetimes are censored.

Semiparametric transformation models

The most popular survival model is Cox' proportional hazards model under which the hazard rate for an individual with covariate vector $X(\cdot)$ takes the form

$$\lambda(t|X(t)) = \lambda_0(t)e^{\beta^T X(t)},$$

where β is a vector of unknown regression parameters and λ_0 is a baseline hazard rate describing the shape of the hazard as a function of time and that is left unspecified. The model is thus *semiparametric* in that the baseline hazard rate is treated nonparametrically, while a parametric form is assumed

for the covariate effect. Correspondingly, the parameter contains an infinite dimensional component in addition to the finite dimensional covariate effect vector of particular interest.

When the covariates are time independent, the interpretation of the β vector is particularly easy. Assume that we observe two individuals with covariate vectors X and $\tilde{X}$, respectively. Then the ratio of their hazard rates is

$$\frac{\lambda(t|X)}{\lambda(t|\tilde{X})} = \frac{\lambda_0(t)\exp(\beta^T X)}{\lambda_0(t)\exp(\beta^T \tilde{X})} = \exp(\beta^T(X - \tilde{X})), \tag{3}$$

which is constant. Hence the name proportional hazards. The proportion (3) is called the relative risk or hazard ratio of the two individuals. For example, if the covariate vectors of two individuals differ only by a binary covariate X_1, then the risk of experiencing the event for the individual with $X_1 = 1$ relative to the individual with $X_1 = 0$ is $\exp(\beta_1)$.

The Cox model has had a monumental success in applied work. In some applications, however, the proportional hazards assumption on the effects of covariates may not be reasonable and there is therefore a need for alternative models. A popular alternative is the proportional odds model that constrains the ratio of the odds of survival associated with two sets of covariate values to be constant over time. Consequently, the ratio of the hazards converge to one with time. This is different from the proportional hazards model that constrains the hazard ratio to be constant while the odds ratio tends to zero or infinity. Both models are examples of semiparametric transformation models. In this broad class the cumulative hazard of T is related to $X(\cdot)$ by

$$\Lambda(t|\overline{X}(t)) = H\left(\int_0^t e^{X(s)^T\beta} d\Lambda_0(s)\right), \tag{4}$$

where the transformation H is a strictly increasing function such that $H(0) = 0$ and $H(\infty) = \infty$. The choice $H(x) = x$ and $H(x) = \log(1 + x)$ corresponds to the proportional hazards and proportional odds models, respectively.

One class of transformations is the Box-Cox transformations,

$$H(x) = \begin{cases} \{(1+x)^\rho - 1\}/\rho, & \rho > 0. \\ \log(1+x), & \rho = 0. \end{cases}$$

For $\rho > 1$ the covariate effects increase over time, for $\rho < 1$ the covariate effects decrease over time. Another useful set of transformations is the logarithmic transforms given by

$$H(x) = \begin{cases} \log(1+rx)/r, & r > 0. \\ x, & r = 0. \end{cases}$$

For $r > 0$, the covariate effects always decrease over time, with a higher rate of decrease for larger r. The choice $\rho = 1$ or $r = 0$ yields the proportional hazards model while the choice $\rho = 0$ and $r = 1$ yields the proportional odds model.

Expression (4) can generate very general models, but this generality often comes with a problem of a lack of transparency of the role of covariates. Except for in special cases, it is typically difficult to look at the expression for the cumulative hazard and gain any intuitive insight into how covariates influence the hazard.

Additive models

Although not considered further in this thesis, an overview of failure time regression is not complete without a mentioning of additive models. An alternative to transformation models is Aalen's model, assuming that the hazard takes the form

$$\lambda(t|X) = \beta(t)^T X(t),$$

where $\beta(t)$ is a vector of regression functions describing how the covariates affect the hazard rate at time t. Note that this model allows the covariate effects to change with time.

The additive model measures excess risk due to covariates on an *absolute* scale whereas the proportional hazards model measures additional risk in *relative* terms. It is possible that the additive risk varies with time while the relative risk is constant over time. Note that changes in absolute risk with time give no information on changes in relative risk with time.

Martinussen and Scheike (2006, Chapter 7) describe how additive and multiplicative approaches can be combined to achieve flexible models. One example is the Cox-Aalen model where the hazard takes the form

$$\lambda(t|X = (X^1, X^2)) = \left\{\beta(t)^T X^1(t)\right\} \times \exp(\beta^T X^2(t)),$$

where the covariate vector X is partitioned into the vectors X^1 of additive effects, and X^2 that acts multiplicatively.

1.1.2 Clustered failure time data

Clustered failure time data arise when subjects are sampled in clusters so that the failure times within the same cluster tend to be correlated. Medical examples include the onset of a genetic disease among family members with families serving as clusters. Sometimes one would assume common distributions for all

individuals in a cluster, while in other situations the cluster structure may be rather complex. For instance, when considering the lifetimes of parents and children in a family, individuals within the same cluster are not exchangeable and we have to distinguish between levels.

There are two main approaches to modeling cluster effects, marginal and conditional. The choice depends mainly on the purpose of the study. In marginal models the covariate effects are specified unconditionally and we assume that the regression model holds marginally for each individual, but that individuals within groups are associated. For the conditional approach we assume instead that the model holds for each individual conditional on some unobserved effect, which is modeled as random.

Marginal models

The marginal approach is well suited for the situation where one aims at estimating regression effects on the population level, and only have to deal with correlation to get valid standard errors to ensure correct inference. Then the cluster structure is ignored when estimating the covariance effects and is only used to derive correct standard errors. This approach is closely linked to the generalized estimating equations methodology (Liang and Zeger, 1986).

Most marginal models do not make any assumptions regarding the dependence structure. It can be seen as an advantage that we do not have to rely on a specific structure, but on the other hand such models cannot be used for assessment of dependence. It is however possible to extend the marginal models to provide estimates of the within cluster correlation. This can be achieved either by building a model that contains correlation as well as marginal regression parameters or by modeling the dependence of the marginals with a copula structure estimated in a two-step procedure. A copula model assumes that the joint survival function of failure times $T_1, \ldots, T_m$ within a cluster is given by

$$P\left(T_1 > t_1, \ldots, T_m > t_m\right) = C_\gamma \left\{S_1(t_1), \ldots, S_m(t_m)\right\},$$

where S_j denotes the marginal survivor functions and the copula C_γ is a m-dimensional survival function with uniform margins parameterized by γ. Different copulas give different joint distributions but the marginals are unaltered. A summary of marginal modeling with specified correlation structure can be found in Martinussen and Scheike (2006, Chapter 9).

Conditional models

In conditional models dependence is modeled by introducing unobserved random effects, often called frailties in the survival context, and the regression model is assumed valid conditional on the random effects. This is analogous to linear mixed models (Laird and Ware, 1982).

The simplest model is the shared frailty where all survival times that are related have the same level of frailty attached to them, corresponding to one realization of the frailty variable. The frailty measures the specific risk level for a cluster and conditional on the frailty the survival times are independent.

More complicated dependencies can be modeled through multivariate random effects with associated covariates. For example, by introducing the random variable $\mathbf{b}$ and additional covariates $Z(\cdot)$ in the model (4), we get the model

$$\Lambda(t|\overline{X}_{ij}(t), \overline{Z}_{ij}(t), \mathbf{b}_i) = H\left(\int_0^t e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}_i} d\Lambda_0(s)\right), \tag{5}$$

for the conditional cumulative hazard of individual i in cluster j.

Until recently, frailties have often been assumed to follow a Gamma distribution with mean one and unknown variance. The choice of the gamma distribution has been made mostly for mathematical convenience. This rather restrictive assumption does no longer appear necessary when all details have been worked out for a large class of distributions, including the Gaussian (Zeng and Lin, 2007, 2010).

1.1.3 Semiparametric maximum likelihood

Counting processes and martingale methods have traditionally been the main tools when studying asymptotics in survival analysis. Let $N(t)$ denote a generic process counting the number of events that have occurred for some unit of interest up to time t. We can construct a counting process per individual or cluster and another process counting the events for all subjects under study. The counting processes can be decomposed into a deterministic model part, the compensator A, and a random noise part M such that

$$M(t) = N(t) - A(t)$$

is a martingale. Many interesting quantities in survival analysis, including likelihoods and associated score functions, can be written as stochastic integrals

of the form

$$\int_0^t K(s)dM(s), \tag{6}$$

where K is a *predictable* stochastic process. Informally, a the process K is predictable if the value $K(t)$ is known given the history just prior to time t. Integrals of the form (6) are, under some conditions, themselves martingales and asymptotic theory can often be established by Robelledo's martingale central limit theorem (Andersen et al., 1993, p. 83). Large sample properties of maximal partial likelihood estimators of both the covariate effects and the cumulative baseline hazard rate in Cox's proportional hazards model were derived along these lines by Andersen and Gill (1982).

Although appealing due to their conceptual foundation, martingale methods are not always applicable. In particular, if the integrand in (6) is not predictable, then the integral is not a martingale. This is the case for example if K contains weights that depend on events that might not have occurred at time t.

Often (6) can alternatively be viewed as an empirical process and large sample properties then follow by modern empirical process techniques (van der Vaart and Wellner, 1996). There is a price to be paid for this however, as empirical processes in this setting pose the strong restriction of independence of sampling units (e.g. individuals or clusters) whereas martingales allow more complex dependencies on the past. For example, martingale techniques can be applied when the censoring mechanism depends on what happened previously to any individuals or clusters, even though this set up is clearly non-i.i.d. We conclude that none of the methods can fully replace the other. We will not pursue the martingale track any further and refer readers to Andersen et al. (1993); Martinussen and Scheike (2006).

The most common approach to efficient estimation in semiparametric models is based on modifications of maximum likelihood estimation. Likelihood functions for the transformation models discussed above can be written in the generic form as

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\lambda(Y_{ij})^{\Delta_{ij}}\Omega(\theta,\Lambda)[O_i]$$

where θ is a finite dimensional parameter and Ω is a functional of random processes O_i containing information on events and covariates of the n_i subjects in cluster i. If Λ_0 is known to be absolutely continuous, then, as in the case of density estimation, very high peaks at the observations would yield an arbitrarily large likelihood and there is no maximizer of the likelihood. Instead we maximize over all right continuous functions and replace $\lambda(t)$ with the jump

size at t, denoted $\Lambda\{t\}$. That is, we maximize

$$L_n(\theta, \Lambda) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \Lambda\{Y_{ij}\}^{\Delta_{ij}} \Omega(\theta, \Lambda)[O_i].$$

The best choice among the discrete distributions are Λ that jump at the points Y_{ij} with $\Delta_{ij} = 1$ only. This reduces the maximization problem to finding the jump sizes $\Lambda\{Y_{ij}\}$.

The maximizer

$$(\hat{\theta}_n, \hat{\Lambda}_n) = \operatorname{argmax} L_n(\theta, \Lambda) \tag{7}$$

is referred to as the nonparametric maximum likelihood estimate (NPLME). Due to the complications resulting from the presence of the infinite dimensional parameter, what we treat as a likelihood here is not really a likelihood in the sense of products of densities. Thus, we need to verify that the NPLME indeed behaves like a maximum likelihood estimate, i.e. we wish to establish consistency, asymptotically normality and efficiency.

Murphy (1994, 1995) used empirical process theory to prove consistency, asymptotic normality and efficiency for the NPMLE in the shared gamma-frailty model without covariates. Her work was generalized to the correlated gamma-frailty model allowing for covariates by Parner (1998). Many authors have since used similar techniques for various models. We will briefly outline the method of proof. Please bear in mind that despite the common general scheme, the technical details can be very different from model to model. See Zeng and Lin (2007, 2010) for a thorough exposition of NPLME based on inference in semiparametric transformation models.

Consistency

Given that the model is identifiable, a Wald type argument based on comparing the value of the likelihood at the maximum likelihood estimator and at the true value of the parameter is a classical type of consistency proof. In the present case this causes problems. First, $\hat{\Lambda}_n$ is not assumed bounded, that is the parameter space for Λ is not a priori known to be compact as required by a Wald type of proof. However, one can use ideas developed in Murphy (1994) and Parner (1998) to show that the form of the likelihood forces $\hat{\Lambda}_n$ to be bounded. Then $\hat{\Lambda}_n$ is relatively compact and Helly's selection theorem implies that, for any subsequence, we can choose a further subsequence (still denoted n) such that $\hat{\Lambda}_n$ converges to some Λ^*. We assume that $\theta \in \Theta$, where Θ is compact, so that $\hat{\theta}_n$ also converges to some θ^*.

We thus need to verify that $(\theta^*, \Lambda^*) = (\theta_0, \Lambda_0)$. In a Wald type of proof this is achieved by showing that the Kullback-Leibler divergence is zero and this is only possible if $(\theta^*, \Lambda^*) = (\theta_0, \Lambda_0)$. Again, this is not straightforward because the likelihood at the maximum likelihood estimator, a random discrete distribution, and at the true parameter are different in character and can not be compared directly. Instead we compare the NPLME with an intermediate random sequence converging to the true value, $(\theta_0, \tilde{\Lambda}_n)$ where $\tilde{\Lambda}_n$ is discrete and converges to Λ_0. The function $\tilde{\Lambda}_n$ is chosen similar in structure to $\hat{\Lambda}_n$, but also similar to Λ_0. If we can show that this log-likelihood difference converges to minus the Kullback-Leibler divergence we can finish of the proof by the same arguments as in the fully parametric scenario.

Asymptotic normality

To prove asymptotic normality of parametric maximum likelihood estimators we usually consider a system of estimating equations of the same dimension as the parameter. The solutions are asymptotically normal if the system is appropriately differentiable. A semiparametric model would require infinitely many estimating equations. As shown in van der Vaart (1998, section 25.12) and van der Vaart (1999, Lecture 10) it turns out that we can proceed much in the same way as a finite dimensional system, provided that we substitute functional analysis for multivariate calculus. The system is linearized in the estimators by a Taylor expansion around the true parameter, and the limit distribution involves the inverse of the derivative.

In order to present van der Vaart's master theorem we first introduce the following fundamental concepts from empirical process theory. Let $X_1, \ldots, X_n$ be a random sample from a probability distribution P. Given a measurable real valued function f we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure and Pf for the expectation under P,

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \text{ and } Pf = \int f dP.$$

The empirical process evaluated at f is defined as $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$. A class $\mathcal{F}$ of measurable real valued functions is called P-Donsker if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly to a tight limit process in $\ell^\infty(\mathcal{F})$, the space of bounded real valued functions on $\mathcal{F}$. The limit process $\{\mathbb{G}f : f \in \mathcal{F}\}$ is a zero mean Gaussian process with covariance $E[f(X)g(X)] - E[f(X)]E[g(X)]$ for $f, g \in \mathcal{F}$. $\mathbb{G}$ is known as the P-Brownian bridge.

Verifying that a class of functions is P-Donsker can be achieved by entropy

calculations. Fortunately, we do not need to calculate entropy for each new problem as there are a number of methods to determine if a class is P-Donsker based on whether the class is built up of classes that are known to be P-Donsker. For example, if $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, then $\{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$, $\{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$ and $\{f+g : f \in \mathcal{F}, g \in \mathcal{G}\}$ are also P-Donsker. Moreover, if $\mathcal{F}$ and $\mathcal{G}$ are bounded P-Donsker, then $\{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is P-Donsker (van der Vaart and Wellner, 1996, Section 2.10).

To set up the system of estimating equations, consider the set

$$\mathcal{H} = \{\mathbf{h} = (\mathbf{h}_\theta, h_\Lambda) : \mathbf{h}_\theta \in \mathbb{R}^d, h_\Lambda \in BV[0,\tau], \|\mathbf{h}_\theta\| + \|h_\Lambda\|_V \leq 1\},$$

where $BV[0,\tau]$ is the class of real valued functions of bounded variation in $[0,\tau]$ and $\|h_\Lambda\|_V$ denotes the total variation of h_Λ in $[0,\tau]$. Define

$$\psi(\theta,\Lambda)[\mathbf{h}_\theta, h_\Lambda] = \mathbf{h}_\theta^T \ell_\theta(\theta,\Lambda) + \ell_\Lambda(\theta,\Lambda)[h_\Lambda], \tag{8}$$

where l_θ is the score function for θ and l_Λ is a score operator for Λ. The finite dimensional parameter can be perturbed in the usual way and $\mathbf{h}_\theta^T l_\theta$ is the ordinary score function for $\mathbf{h}_\theta^T \theta$ treating Λ as fixed. The operator l_Λ is a little more involved. For each fixed (θ,Λ) and $h_\Lambda \in BV[0,\tau]$, $l_\Lambda(\theta,\Lambda)[h_\Lambda]$ corresponds to the score function for the one-dimensional submodel given by $\varepsilon \mapsto (\theta, \int(1+\varepsilon h_\Lambda)d\Lambda)$ and can be found as the directional derivative of the log likelihood in the direction h_Λ. Each choice of $(\mathbf{h}_\theta, h_\Lambda)$ in (8) corresponds to an estimating equation for (θ,Λ).

We identify $(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)$ as a random element in $\ell^\infty(\mathcal{H})$ by defining its value at $(\mathbf{h}_\theta, h_\Lambda)$ as $\mathbf{h}_\theta^T(\hat{\theta}_n - \theta_0) + \int h_\Lambda d(\hat{\Lambda}_n - \Lambda_0)$. Let P_0 denote the distribution of the random processes holding information on events and covariates in a cluster.

From van der Vaart and Wellner (1996, Theorem 3.3.1, Lemma 3.3.2) if $(\hat{\theta}_n, \hat{\Lambda}_n)$, is consistent for (θ_0,Λ_0) and if

(i) $\mathbb{P}_n\psi(\hat{\theta}_n, \hat{\Lambda}_n) = o_P(n^{-1/2})$ and $P_0\psi(\theta_0,\Lambda_0) = 0$,

(ii) $\{\psi(\theta_0,\Lambda_0)\mathbf{h} : \mathbf{h} \in \mathcal{H}\}$ is P_0-Donsker,

(iii) $(\theta,\Lambda) \mapsto \psi(\theta,\Lambda)$ is continuous in $L_2(P_0)$ at (θ_0,Λ_0) uniformly in $\mathcal{H}$,

(iv) The map $(\theta,\Lambda) \mapsto P_0\psi(\theta,\Lambda)$ is Fréchet differentiable (van der Vaart and Wellner, 1996, Example 3.9.2) at (θ_0,Λ_0) with a continuously invertible derivative $P_0\dot{\psi}_0$,

then

$$n^{1/2}\left((\hat{\theta}_n, \hat{\Lambda}_n) - (\theta_0,\Lambda_0)\right) \xrightarrow{\mathcal{L}} P_0\dot{\psi}_0^{-1}\mathbb{G}\psi(\theta_0,\Lambda_0)$$

in $\ell^\infty(\mathcal{H})$.

When all parameters can be estimated at $n^{1/2}$ rate we may treat the NPMLE as a parametric log-likelihood with θ_0 and the jump sizes of Λ_0 at the observed failure times as the parameters. The asymptotic covariance matrix of the NPMLEs for these parameters can be estimated by inverting the observed information matrix.

Profile likelihood

Now consider inference for the finite dimensional parameter, θ_0. Estimation of θ_0 in the semiparametric model is more taxing, meaning that the information is worse, than under any parametric submodel. If the information for a *regular* estimator is equal to the minimum of the information over all efficient estimators for all parametric submodels, then then estimator is called semiparametric efficient. A parametric model which achieves this minimum, if such a model exists, is called a least favorable submodel. For a definition of a *regular* estimator we refer to van der Vaart (1999, Lecture 2) and settle for claiming that most commonly encountered estimators are regular. Nonparametric maximum likelihood generally yields semiparametric efficient estimators.

The semiparametric log profile likelihood is defined as the semiparametric likelihood but where the infinite dimensional component is profiled out,

$$pl_n(\theta) = \sup_{\Lambda} \log L_n(\theta, \Lambda). \tag{9}$$

By taking the supremum in (9) in two steps, we note that the maximizer of (7) is the first component of $\operatorname{argmax}_{\theta,\Lambda} L_n(\theta, \Lambda)$, i.e. the NPMLE of θ_0.

Murphy and van der Vaart (2000) showed that under some conditions, the profile likelihood admits an expansion around the maximum likelihood estimator $\hat{\theta}_n$ in the form

$$\begin{aligned} \log pl_n(\tilde{\theta}_n) &= \log pl_n(\hat{\theta}_n) - \frac{1}{2} n(\tilde{\theta}_n - \hat{\theta}_n)^T \tilde{I}(\tilde{\theta}_n - \hat{\theta}_n) \\ &\quad + o_P(n^{1/2}\|\tilde{\theta}_n - \tilde{\theta}_n\| + 1)^2, \end{aligned} \tag{10}$$

where $\tilde{I}$ is the efficient information for estimating θ_0, for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

The asymptotic expansion suggests that the semiparametric profile likelihood asymptotically can be treated much like an ordinary likelihood. In particular, under some conditions the maximum profile likelihood estimator is consistent, asymptotically normal and efficient. Differentiation of the profile likelihood

yields consistent estimators of the efficient information matrix. A profile likelihood ratio statistic can be compared to percentiles of the χ^2 distribution to produce asymptotic hypothesis tests.

Weighted nonparametric maximum likelihood

The following development of Breslow and Wellner (2007) extend the ideas of the previous sections to data sets sampled in two phases. Typically the first phase sample contains incomplete information for a very large cohort. When using all subjects from the phase one sample is infeasible we can choose a subsample, the phase two sample, for further analysis. Based on the information from the first phase we might want to overrepresent subjects believed to hold more statistical information or otherwise ascertain enough subjects of specific characteristics. In the setting with routine registers, the first phase typically corresponds to the collection of the full register and the second phase to carefully selecting a subset from the register for further analysis.

Specifically, assume that the first phase consist of n observations. and that the cohort is partitioned into K strata depending on information available in the phase one sample. Let $\xi_i = 1$ indicate whether observation i was included in the subsample of the second phase and let $\pi_i = P(\xi_i = 1)$. The probabilities π_i depend on stratum membership of observation i. Then

$$\mathbb{P}_n^\pi f = \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{\pi_i} f(X_i)$$

is the expectation of the measurable function f under the inverse probability weighted (IPW) empirical measure. Define the IPW empirical process

$$\begin{aligned} \mathbb{G}_n^\pi &= \sqrt{n}\,(\mathbb{P}_n^\pi - P) \\ &= \sqrt{n}(\mathbb{P}_n - P) + \sqrt{n}(\mathbb{P}_n^\pi - \mathbb{P}_n). \end{aligned}$$

From Breslow and Wellner (2007, section 4) we have that if the population proportion of stratum k members, v_k, is positive for $k = 1, \ldots, K$, then

$$\mathbb{G}_n^\pi \xrightarrow{\mathcal{L}} \mathbb{G} + \sum_{k=1}^{K} \sqrt{v_k \frac{1-p_k}{p_k}}\,\mathbb{G}_k \tag{11}$$

in $\ell^\infty(\mathcal{F})$, where $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_K)$ is a vector of independent Brownian bridge processes, all indexed by a P-Donsker class $\mathcal{F}$. Specifically, $\mathbb{G}_k$ is a P_k-Brownian bridge process indexed by $\mathcal{F}$, where P_k denotes P conditional on membership of stratum k. Breslow and Wellner (2007, Proposition B.1) further states that if $\mathcal{F}$ is P-Donsker then $\mathcal{F}$ is P_k-Donsker on stratum k, $k = 1, \ldots, K$.

1.2 Register data

Cohort studies are usually based on detailed information gathered on a limited number of individuals. When the disease of interest is rare, a large cohort is required in order to accumulate sufficiently many cases for a meaningful statistical analysis. This will usually require a long period of time and tend to be very expensive.

The use of existing routine administrative registers in epidemiological studies may mean considerable cuts in total research costs. In the Nordic countries there are several registers of high quality that can be linked by the unique personal identification number assigned to each permanent resident used across all registers.

The central population registers collect and update information received from several different sources, for example hospitals, religious communities (marriages), law courts (civil marriages and divorces) and individual citizens (announcement of migration). Further, each individual can be linked to parents and children via the national multi-generation registers.

Causes of disease can be identified in the registers of disease. Examples include the national cancer registers and causes of death registers. The hospital discharge registers are collected from all public and private hospitals, and are based on inpatient care periods. The registers includes information on the length of stay in the hospital, diagnoses and procedures during hospitalization. The medical birth registers include information on mother's background, maternal health during pregnancy and delivery, medical interventions and newborn's outcome up to the age of seven days. Since the medical birth registers are routinely combined with the central population registers and the cause-of-death registers, they are complete in terms of births and deaths.

As routine registers are becoming increasingly common worldwide, the possibilities to use administrative data in epidemiological research is expanding. So is the need for statistical methods analyzing such data.

1.2.1 Cohort sampling designs

Standard use of regression models requires inclusion of covariate information on all individuals in a cohort even when only a small fraction of these actually experience the event of interest. As noted in the previous section, when studying rare diseases the cohorts must necessarily be large and an analysis based on the full cohort may demand unreasonable computer power resources or time.

Thus, when working with routine register data, a study design allowing for estimation of covariate effects without having to collect all data on all members of the cohort is desirable.

When the disease of interest is rare, the contribution of non failures (controls), in terms of statistical power may be close to negligible compared to that of failures. Standard case-control study designs exploit this principle efficiently. As most of the statistical information is contained in the cases, sampling designs that include only a portion of the controls and overrepresent cases may drastically reduce sample sizes but still be sufficient to give reliable answers of the questions of interest. There are two important classes of case-control sampling designs: nested case-control sampling and case-cohort sampling.

Nested case-control sampling

In a nested case-control design, one selects, whenever an event occurs, a typically small number of controls among those at risk. The set consisting of these controls together with the case is called the sampled risk set. Covariate information is collected on the individuals in the sampled risk sets but are not needed for the other individuals in the cohort.

The selection of controls is done independently at the different event times, so that subjects may serve as controls for multiple cases, and cases may serve as controls for other cases that experienced an event when the case was at risk. A crucial assumption is that at any time we do not make use of any information on events in the future. Nested case-control sampling is a prime example of a situation where martingale methods can not easily be replaced by empirical process techniques. This is because the sampling probabilities at each observed event depend on all individuals at risk and are therefore not independent.

If a surrogate measure of the exposure of main interest is available for everyone, then this information can be incorporated into the sampling process so that we obtain a more informative sample of controls. This stratified nested-case control design is called counter-matching and is described in Langholz and Borgan (1995).

Case cohort sampling

In a case-cohort design, covariates are obtained for individuals who experience the event and for a subcohort of controls selected from the full cohort at the outset of the study. In contrast to the nested-case control design we already

from the start have information about which individuals that will become cases and the same individuals are used as controls at all event times when they are at risk. Thus, subjects are sampled with unequal probabilities depending on whether or not they experience the event.

It is well known that one can improve the efficiency of the parameter estimates by stratifying according to the covariates of the members in the cohort. Borgan et al. (2000) present large sample results for stratified case-cohort estimators in Cox proportional hazards model. The asymptotic covariance matrix can be split into two components; the cohort covariance matrix and a covariance matrix due to sampling the subcohort from the full cohort.

Chapter 2

Summary of the paper

In the second part of this thesis we propose inferential procedures that can considerably reduce the resources needed to analyze clustered survival data from routine registers. We sample from registers with unequal inclusion probabilities in order to achieve an informative subsample of a modest size, so that it can be analyzed with reasonable resources. The sampling is performed in two stages and is similar to the stratified case-cohort design described in the previous chapter. When considering large registers, even if the cases are small in proportion they may be big in numbers and we might want to sample cases as well. This is readily achieved by our design.

The weights we use depend on stratum membership and are typically not determined until an individual experiences an event or is censored. Such weights are certainly not predictable and martingales are of no help. It turns out that the inverse probability weighted empirical process techniques of Breslow and Wellner (2007) are exactly what we need.

We consider the general class of semiparametric transformation models with clustering induced by random effects as given by (5). The regression parameters should thus be interpreted conditional on the random effects. Consistency and asymptotic normality of the nonparametric maximum likelihood estimator in this model were derived along the lines of Section 1.1.3 by Zeng et al. (2008). We combine the work of Zeng et al. (2008) and Breslow and Wellner (2007) and derive similar results for estimation based on two-phase sampled data. An asymptotic likelihood ratio test for testing hypothesized values of one or more regression parameters is also given.

We suggest consistent estimators of the asymptotic variance of the IPW maximum likelihood estimator. The variance is the sum of two components. The first component is the usual variability of an estimator based on random sampling from an infinite population whereas the second component represents the additional variability from selecting only a subsample in the second phase.

We present an extensive simulation study to illustrate the performance of the methods. We also apply the procedure to a data set of sibling pairs collected from Swedish routine registers in order to study components that might effect the risk of death in cardiovascular diseases.

Bibliography

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

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Borgan, Ø., Langholz, B., Samuelsen, S., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58.

Breslow, N. and Wellner, J. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34:86–102.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika*, 82(1):69–79.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Martinussen, T. and Scheike, T. (2006). *Dynamic Regression Models for Survival Data*. Springer.

Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22:712–731.

Murphy, S. (1995). Asymptotic theory for the frailty model. *Annals of Statistics*, 23:182–198.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26(1):183–214.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. (1999). Semiparametric statistics. In *Ecole d'Ete de Probailites de St. Flour XXIX*, volume 1781 of *Lectures on probability theory and statistics*, pages 331–457. Springer.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer.

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

Zeng, D. and Lin, D. (2010). A generalized asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*, 20:871–910.

Zeng, D., Lin, D., and Lin, X. (2008). Semiparametric transformation models with random. effects for clustered failure time data. *Statistica Sinica*, 18:355–377.

Part II

The paper

Semiparametric transformation models for clustered survival data from routine registers

Frank Eriksson[1]

[1]Chalmers University of Technology and University of Gothenburg

May 17, 2010

Abstract

Routine registers offer possibilities to study covariate effects on survival times for rare diseases otherwise unavailable because of the large cohorts required. The vast amount of data and clustering of related individuals pose statistical challenges. We adapt previous work on semiparametric regression with random effects to a flexible cohort sampling design that can drastically reduce the sample size needed with only a minor loss of efficiency. We develop weighted likelihood based inference procedures and illustrate their applicability on both simulated and real data sets.
Key words: survival, transformation, family, random effects, case-cohort, routine register, semiparametric, likelihood.

1 Introduction

The Nordic countries have a long tradition of collecting data on deaths and diseases. There are several epidemiological registers of high quality which routinely store information on diagnoses and possible covariates, e.g. cause of death, cancer, medical birth and hospital discharge registers. By linking registers we can obtain survival time data, including covariates, for millions of individuals. Multi-generation registers can be used to identify familial relationships needed for studying disease incidence and clustering due to environmental factors and/or genetics. A cohort generated in this way can potentially be very large while the proportion of cases may be relatively small for rare diseases. The massive amount of data and the correlation of failure times among related

individuals pose serious statistical challenges in assessment of covariate effects on failure time.

Analysis based on the full cohort may be intractable due to various reasons. The computer power and time it demands may be unreasonable, or we might want to include additional covariates not in the register that would be impractical to collect for all individuals. As most information is assumed to be associated with the cases, Prentice (1986) proposed the case-cohort design where all cases are included but only a sample of the controls. He showed that his estimator, although based on only a fraction of the data, is almost as efficient as one estimated on the full cohort. Since the introduction of the case-cohort design many authors have developed methods for analysis or proposed modifications, e.g. Prentice and Self (1988), Lin and Ying (1993), Borgan et al. (2000) among others.

There are two main approaches when modeling correlated data, namely marginal and random effects models. Marginal models are used when interest is in estimating population average effects. Standard errors are typically corrected for correlations without explicitly modeling dependencies. Marginal proportional hazards models for clustered case-cohort failure time data have been studied by e.g. Lu and Shih (2006), Lu and Wang (2002). In random effects models on the other hand, both covariate effects and characterization of dependency are concerned. The dependence structure is specified by incorporating unobserved random variables accommodating the dependence within clusters. Copula models combine the marginal approach with a model for the dependency structure. The joint survival function is modeled through the marginal survival functions and an association parameter. Andersen (2005, 2004) considered both parametric and semiparametric estimation in copula models for family register data. Moger et al. (2008) dealt with random effects models for family register data, but considered only parametric models.

In this paper we investigate regression of sampled clustered cohort data for a broad class of semiparametric models. In a typical scenario we sample families and want to oversample families with at least one uncensored failure time. Because of the large cohort sizes in register data, even when the cases are few in proportion they can be large in numbers and it can be desirable to include only a subsample in the analysis. Also, we might want to divide the cohort into strata that are sampled with unequal probabilities to increase efficiency (cf. Borgan et al. (2000)). Our design is flexible enough to allow this.

The structure of this paper is as follows. In Section 2 we introduce the model in detail and present the underlying assumptions. In Section 3 we develop the estimation theory. The estimators are shown to be consistent and asymptot-

ically normal. Consistent variance estimators are obtained and we present a likelihood ratio test. This is followed by a numerical study in Section 4 that reveal that the proposed estimators perform well for realistic sample sizes and that the efficiency loss is small compared to the computational cost of analyzing the full cohort. In section 5 we analyze a data set on death in cardiovascular diseases collected from Swedish routine registers. All proofs are found in the Appendix.

2 Data structure and model assumptions

Suppose the cohort is made up of n independent clusters, each consisting of n_i, $i = 1, \ldots, n$, study subjects, that are sampled at random from an infinite population. Let $X_{ij}(\cdot)$ be a d_1-vector of covariates and $Z_{ij}(\cdot)$ another set of covariates and let $\overline{X}_{ij}(t)$ and $\overline{Z}_{ij}(t)$ denote the corresponding histories over $[0, t]$. Assume the cumulative hazard of T_{ij}, the failure time of the jth subject in the ith cluster, is related to $X_{ij}(\cdot)$ and $Z_{ij}(\cdot)$ by

$$\Lambda(t|\overline{X}_{ij}(t), \overline{Z}_{ij}(t), \mathbf{b}_i) = H\left(\int_0^t e^{X_{ij}(s)^T\beta_0 + Z_{ij}(s)^T\mathbf{b}_i} d\Lambda_0(s)\right), \qquad (1)$$

where the transformation H is a strictly increasing function such that $H(0) = 0$ and $H(\infty) = \infty$, β_0 is an unknown vector valued regression parameter and $\mathbf{b}_i$ is a set of unobserved mean-zero random effects for the ith cluster with density $\eta(\cdot, \gamma_0)$ indexed by a d_2-dimensional parameter γ_0. The choice $H(x) = x$ and $H(x) = \log(1 + x)$ correspond to the proportional hazards and proportional odds models with random effects, respectively.

One class of transformations is the Box-Cox transformations,

$$H(x) = \begin{cases} \{(1+x)^\rho - 1\}/\rho, & \rho > 0 \\ \log(1+x), & \rho = 0 \end{cases} \qquad (2)$$

For $\rho > 1$ the covariate effects increase over time, for $\rho < 1$ the covariate effects decrease over time. Another useful set is the logarithmic transformations given by

$$H(x) = \begin{cases} \log(1+rx)/r, & r > 0 \\ x, & r = 0. \end{cases}$$

For $r > 0$, the covariate effects always decrease over time, with a higher rate of decrease for larger r. The choice $\rho = 1$ or $r = 0$ yields the proportional hazard model while the choice $\rho = 0$ and $r = 1$ yields the proportional odds model.

Assume T_{ij} is right censored by the censoring time C_{ij} so that we only observe the censored failure time $Y_{ij} = T_{ij} \wedge C_{ij}$ and its corresponding indicator $\Delta_{ij} = I(T_{ij} \leq C_{ij})$. Let O_i denote the observations in the ith cluster,

$$O_i = \left\{(Y_{ij}, \Delta_{ij}, \overline{X}_{ij}(Y_{ij}), \overline{Z}_{ij}(Y_{ij}))\right\}_{j=1}^{n_i}.$$

The model described above was studied in detail by Zeng et al. (2008). We will adapt their findings to two-phase sampling, a subsampling design where O_i is not fully observed for all n clusters and whose most basic semblance is the classical case-cohort design.

At phase one, we observe only a coarsening of O plus auxiliary variables that serve to determine the sampling strata. Let $V \in \mathcal{V}$ denote the variables actually observed for everyone. Suppose $\mathcal{V}$ is partitioned into $\mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_K$, and that the cohort is divided correspondingly into K strata, with the ith cluster in stratum k if $V_i \in \mathcal{V}_k$. Let M_k denote the number of clusters in the kth stratum. At phase two samples of size $m_k \leq M_k$ are drawn at random without replacement from the kth stratum, with sampling for different strata conducted independently. The full covariate histories and event times are observed for these clusters. A sequence of binary indicators $\{\xi_i\}_{i=1}^n$ indicates if cluster i was selected ($\xi_i = 1$) at phase two. Let π_i denote the selection probability of cluster i. If $k(i)$ denotes the stratum of cluster i, $\pi_i = m_{k(i)}/M_{k(i)}$. Let P_0 denote the distribution of the variables potentially available for everyone, but in fact fully observed only for those clusters in the phase two sample.

Note that the classical case-cohort design is a special case of two-phase sampling, with phase one corresponding to observing only censoring indicators.

We impose the following regularity conditions (A2-A11 are taken from Zeng et al. (2008); Zeng and Lin (2010)).

A1 The sampling fractions for stratum $k = 1, \ldots, K$, converge with $m_k/M_k \to p_k > 0$ as $n \to \infty$ and the population proportion of stratum k, v_k, is positive.

A2 Conditional on $\overline{X}_{ij}$ and $\overline{Z}_{ij}$, the censoring time C_{ij} is independent of the failure time T_{ij} and the random effect $\mathbf{b}_i$. Subjects still alive at time τ are censored. There exists some positive constant δ_0 such that

$$\begin{aligned} & P_0(\{\xi_i = 1\} \cap \{C_{ij} \geq \tau\} | \overline{X}_{ij}(\tau), \overline{Z}_{ij}(\tau)) \\ = \;& P_0(\{\xi_i = 1\} \cap \{C_{ij} = \tau\} | \overline{X}_{ij}(\tau), \overline{Z}_{ij}(\tau)) \\ \geq \;& \delta_0 \end{aligned}$$

almost surely.

A3 With probability one, $\mathbf{X}_{ij}(\cdot)$ and $\mathbf{Z}_{ij}(\cdot)$ are right continuous in $[0,\tau]$ with uniformly bounded right derivatives.

A4 $\Lambda_0(t)$ is a strictly increasing function in $[0,\tau]$ and is continuously differentiable and such that $\Lambda_0(0)=0$ and $\Lambda_0'(0)>0$.

A5 $\theta_0=(\beta_0,\gamma_0)$ belongs to the interior of a known compact set

$$\begin{aligned}\Theta &= \{\theta=(\beta,\gamma): \|\beta\|\le B \text{ for some constant } B \text{ and}\\ &\quad \gamma \text{ is in a known compact set } \Gamma_0\}.\end{aligned}$$

A6 The cluster size is independent of the survival and censoring variables and $\max_{1\le i\le n} n_i \le n_0$ for a constant n_0 almost surely.

A7 The function $G(x)=1-\exp(-H(x)):[0,\infty)\mapsto[0,1]$ is four times continuously differentiable in $[0,\infty)$ with $G(0)=0$, $G'(x)>0$ and $\sup_{x\ge 0}\left|G_0^{(k)}(x)\right|<\infty$ for $k=1,\ldots,4$.

A8 There exists a positive constant ρ_0 such that

$$\limsup_{x\to\infty}(1+x)^{\rho_0}(1-G(x))<\infty$$

and

$$\limsup_{x\to\infty}(1+x)^{(\rho_0+1)}G'(x)<\infty.$$

A9 The function $\eta(\mathbf{b};\gamma)$ is thrice-differentiable with respect to γ, and for $l=1,2,3$, $\int_{\mathbf{b}}\left|\eta^{(l)}(\mathbf{b};\gamma)\right|d\mathbf{b}$ is uniformly bounded for $\gamma\in\Gamma_0$.

A10 For any fixed constant κ,

$$\sup_{\gamma\in\Gamma_0}\int_{\mathbf{b}} e^{\kappa\|\mathbf{b}\|}\sum_{l=0}^{3}\left|\eta^{(l)}(\mathbf{b};\gamma)\right|d\mathbf{b}<\infty$$

and

$$\sup_{\gamma\in\Gamma_0}\int_{\mathbf{b}} e^{\kappa\|\mathbf{b}\|}\sum_{l=1}^{3}\left|\frac{\eta^{(l)}}{\eta}(\mathbf{b};\gamma)\right|d\mathbf{b}<\infty.$$

A11 For any pair of parameters (θ_1,Λ_1) and (θ_2,Λ_2), if with probability one,

$$\begin{aligned}&\int_{\mathbf{b}}\prod_{j=1}^{k}\left\{G\left(\int_0^{t_j} e^{X_{ij}(s)^T\beta_1+Z_{ij}^T\mathbf{b}}d\Lambda_1\right)\right\}\eta(\mathbf{b};\gamma_1)d\mathbf{b}\\ &= \int_{\mathbf{b}}\prod_{j=1}^{k}\left\{G\left(\int_0^{t_j} e^{X_{ij}(s)^T\beta_2+Z_{ij}^T\mathbf{b}}d\Lambda_2\right)\right\}\eta(\mathbf{b};\gamma_2)d\mathbf{b}\end{aligned}$$

for any $k \in \{1, \ldots, n_i\}$ and any $t_1, \ldots, t_k \in [0, \tau]$, then $\theta_1 = \theta_2$ and $\Lambda_1(t) = \Lambda_2(t)$ for $t \in [0, \tau]$.

A12 If $X_{ij}(t)^T \mathbf{h}_1 + h(t) = 0$ with probability one for some vector $\mathbf{h}_1$ and a function $h(t)$, then $\mathbf{h}_1 = 0$ and $h(t) = 0$. In addition, if there exist a vector $\mathbf{h}_2$ and functions $A_j(t, \mathbf{b})$, $j = 1, \ldots, n_i$ such that with probability one,

$$\int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ G\left(\int_0^{t_j} e^{X_{ij}(s)^T \beta_0 + Z_{ij}^T \mathbf{b}} d\Lambda_0 \right) \right\} \times \left\{ \sum_{j=1}^{k} A_j(t_j, \mathbf{b}) + \frac{\eta'(\mathbf{b}; \gamma_0)^T \mathbf{h}_2}{\eta(\mathbf{b}; \gamma_0)} \right\} d\mathbf{b} = 0$$

for any $k \in \{1, \ldots, n_i\}$ and any $t_1, \ldots, t_k \in [0, \tau]$, then $\mathbf{h}_2 = \mathbf{0}$ and $A_j(t, \mathbf{b}) = 0$, $j = 1, \ldots, n_i$.

Assumptions A2-A6 are standard conditions for this kind of problem. Assumptions A7 and A8 that concerns the transformation are quite mild and are satisfied for all commonly used transformations. Zeng and Lin (2010) verified A7 and A8 for the Box-Cox and logarithmic transformations as well as for the linear transformation model. Assumption A10 appertains to the distribution of the random effects and is easily seen to be satisfied for the Gaussian distribution. A11 and A12 ensure parameter identifiability and non-singularity of the Fisher information matrix. Zeng et al. (2008); Zeng and Lin (2010) discussed these assumptions and showed that they are fulfilled when the covariates are time independent, $\mathbf{Z}_{ij}$ are the same within clusters and the random effects are Gaussian, provided that the covariates are linearly independent.

3 Weighted maximum likelihood estimation

If the covariate histories and event times were available for all n clusters, we would estimate the parameters (θ_0, Λ_0) by maximizing the semiparametric likelihood

$$L_n(\theta, \Lambda) = \prod_{i=1}^{n} l(\theta, \Lambda)[O_i],$$

where

$$l(\theta,\Lambda)[O_i] = \int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{ G'\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta + \mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t) \right) e^{\mathbf{X}_{ij}(Y_{ij})^T\beta + \mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}} \Lambda\{Y_{ij}\} \right\}^{\Delta_{ij}} \times \left\{ 1 - G\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta + \mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t) \right) \right\}^{1-\Delta_{ij}} \eta(\mathbf{b};\gamma) d\mathbf{b},$$

with $\Lambda\{t\}$ denoting the jump size of Λ at time t, is the empirical likelihood for cluster i. As the unrestricted maximum likelihood estimator does not exist (cf. Zeng et al. (2008)), we maximize over the space of non-decreasing right continuous functions,

$$\begin{aligned} \mathbf{\Lambda} &= \{\Lambda : \Lambda \text{ a right continuous non-decreasing function in } [0,\tau], \\ &\quad \text{with } \Lambda(0) = 0\}. \end{aligned}$$

For two-phase sampled data we propose to instead maximize the the Inverse Probability Weighted (IPW) semiparametric likelihood, defined by

$$L_n^{\pi}(\theta,\Lambda) = \prod_{i=1}^{n} l(\theta,\Lambda)^{\xi_i/\pi_i}[O_i], \tag{3}$$

and denote the estimator

$$(\hat{\theta}_n, \hat{\Lambda}_n) = \underset{\theta\in\Theta,\Lambda\in\mathbf{\Lambda}}{\operatorname{argmax}} L_n^{\pi}(\theta,\Lambda).$$

In the Appendix we prove the following two theorems, the main findings of our paper.

Theorem 1 (Existence and consistency)**.** *Under A1-A12, the IPW Maximum Likelihood Estimate (IPWMLE)* $(\hat{\theta}_n, \hat{\Lambda}_n)$ *of (3) exists almost surely and* $\|\hat{\Lambda}_n - \Lambda_0\|_{\ell^\infty([0,\tau])} \xrightarrow{a.s.} 0$, $\|\hat{\theta}_n - \theta_0\| \xrightarrow{a.s.} 0$.

Theorem 2 (Weak convergence)**.** *Under A1-A12,* $n^{1/2}(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)$ *weakly converges to a zero-mean Gaussian process in* $\mathbb{R}^{d_1+d_2} \times \ell^\infty[0,\tau]$.

It is essential to estimate the variance of the limit in Theorem 2. The variation in estimating Λ_0 arises from the variation in estimating the jump sizes of Λ_0 at the uncensored event times. We regard the observed weighted likelihood function as a likelihood function indexed by the parameter θ and the jump sizes

of Λ at the Y_{ij} for which $\Delta_{ij} = 1$. Let $\mathbb{J}_n$ denote the negative Hessian matrix of $\log L_n^\pi$ and $\mathbb{K}_n$ the matrix of derivatives of $\{\log l(\theta, \Lambda)[O_i] : \xi_i = 1, i = 1, \ldots, n\}^T$, both with respect to $(\theta^T, \vec{\Lambda}^T)$ where the vector $\vec{\Lambda}$ denotes the jump sizes $\Lambda\{\cdot\}$ at the uncensored failure times in the phase two sample, evaluated at $(\hat{\theta}_n, \hat{\Lambda}_n)$.

For any constant vector $\mathbf{h}_\theta \in \mathbb{R}^{d_1+d_2}$, and any bounded function h_Λ, the ith row of the vector $n\mathbb{K}_n\mathbb{J}_n^{-1}\mathbf{h}_n$, where $\mathbf{h}_n = (\mathbf{h}_\theta^T, \overrightarrow{h_\Lambda \Lambda}^T)$ and $\overrightarrow{h_\Lambda \Lambda}$ is $h_\Lambda(\cdot)\hat{\Lambda}_n\{\cdot\}$ evaluated at the observed failure times, approximates the contribution from the ith cluster to the score for estimating $\mathbf{h}_\theta^T\theta_0 + \int h_\Lambda d\Lambda_0$. The following theorem gives an expression for consistently estimating the limiting variance.

Theorem 3 (Asymptotic variance)**.** *The asymptotic variance of*

$$n^{1/2}\mathbf{h}_\theta^T(\hat{\theta}_n - \theta_0) + n^{1/2}\int_0^\tau h_\Lambda d(\hat{\Lambda}_n - \Lambda_0)$$

can be consistently estimated by

$$n\mathbf{h}_n^T\mathbb{J}_n^{-1}\mathbf{h}_n + \sum_{k=1}^{K}\frac{M_k}{n}\frac{M_k - m_k}{m_k^2}\sum_{i:V_i\in\mathcal{V}_k,\xi_i=1}\left(\ell_i - m_k^{-1}\sum_{j:V_j\in\mathcal{V}_k,\xi_j=1}\ell_j\right)^2, \quad (4)$$

where ℓ_i represents the row of $n\mathbb{K}_n\mathbb{J}_n^{-1}\mathbf{h}_n$ corresponding to cluster i.

The first term in (4) is an estimate of the variance of the estimator based on the full cohort. The second component corresponds to the additional uncertainty due to subsampling.

3.1 Inference for finite dimensional parameters

Often focus is on inference for the finite dimensional parameter θ_0, while the infinite dimensional Λ_0 is treated as nuisance. Inference can then be facilitated conveniently by profile likelihood theory (Murphy and van der Vaart, 2000).

The IPW Profile Likelihood is defined as

$$pL_n^\pi(\theta) = \sup_{\Lambda\in\mathbf{\Lambda}}\prod_{i=1}^{n} l(\theta, \Lambda)^{\xi_i/\pi_i}[O_i].$$

By taking the supremum in two steps we observe that $\hat{\theta}_n$, the first component of the maximizer of (3), also maximizes $\theta \mapsto pL_n^\pi(\theta)$. Choosing h_Λ equal to zero in Theorem 2 we see that $\hat{\theta}_n$ is asymptotically normal. We have the

following corollary to Theorem 2 that concerns the efficient score function and the efficient information. For definitions and discussion of these concepts in the semiparametric setting we refer to Murphy and van der Vaart (2000) or van der Vaart (1998, 1999).

Corollary 1 (IPW Profile likelihood). *The IPWMLE of θ_0 is asymptotically normal and has the asymptotic expansion*

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2}\tilde{I}^{-1}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\tilde{\ell}[O_i] + o_P(1), \tag{5}$$

where $\tilde{\ell}$ is the efficient score and $\tilde{I}$ is the efficient information for estimating θ_0. The asymptotic variance of $n^{1/2}(\hat{\theta}_n - \theta_0)$ is

$$\tilde{V} = \tilde{I}^{-1} + \sum_{k=1}^{K} v_k \frac{1-p_k}{p_k} Var_k(\tilde{I}^{-1}\tilde{\ell}), \tag{6}$$

where v_k is the fraction of the population belonging to stratum k, p_k the limiting sampling probability for subjects in stratum k and Var_k is the variance conditional on membership in stratum k.

Equation (5) presents an alternative to Theorem 3 for estimating the asymptotic variance of the IPWMLE. If $\hat{\theta}_n^{-i}$ denotes the IPWMLE of θ_0 leaving out observation i, then $\hat{\theta}_n - \hat{\theta}_n^{-i} = n^{-1}\frac{\xi_i}{\pi_i}\tilde{I}^{-1}\tilde{\ell}[O_i] + o_P(1)$. The sample variance of these Jackknife estimates can be used to estimate both the unconditional and conditional variances in (6). Unfortunately, our numerical studies reveal that this variance estimator is computationally too slow for the sample sizes we have in mind.

Testing hypothesized values of the finite dimensional parameters is essential e.g. for model selection.

Theorem 4 (IPW Profile Likelihood Ratio). *Let $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{d_1+d_2}$, where $\theta_2 \in \mathbb{R}^r$ and partition $\tilde{V}$ as*

$$\tilde{V} = \begin{pmatrix} \tilde{V}_{11} & \tilde{V}_{12} \\ \tilde{V}_{21} & \tilde{V}_{22} \end{pmatrix},$$

where the subscripts correspond to the elements related to θ_1 and θ_2 in the obvious way. Under the null hypothesis $H_0 : \theta_2 = \theta_{20}$, i.e. $\theta \in \Theta_0 = \{(\theta_1, \theta_2) \in \Theta : \theta_2 = \theta_{20}\}$, the sequence

$$2\log\frac{\sup_{\theta\in\Theta} pL_n^{\pi}(\theta)}{\sup_{\theta\in\Theta_0} pL_n^{\pi}(\theta)}$$

is asymptotically distributed as $\lambda_1\chi_1^2 + \ldots + \lambda_r\chi_1^2$*, where the* λ*s are the eigenvalues of the matrix*

$$\tilde{V}_{22}^{1/2}\begin{pmatrix} \tilde{V}_{11}^{-1}\tilde{V}_{21} \\ \mathbf{1}_{r\times r} \end{pmatrix}^T \tilde{I}\begin{pmatrix} \tilde{V}_{11}^{-1}\tilde{V}_{21} \\ \mathbf{1}_{r\times r} \end{pmatrix}\tilde{V}_{22}^{1/2}.$$

Of course, we do not know the values of $\tilde{I}$ and $\tilde{V}$ in Theorem 4 and the result seems at first to be of little practical use. From Theorem 3, we do however have consistent estimators of both matrices. We have performed extensive simulations to investigate the validity of a test where the test statistic is compared to a distribution where $\tilde{I}$ and $\tilde{V}$ are replaced by estimates. We found this procedure to yield acceptable accuracy in all of the various models we tried, see Section 4. We have not been able to theoretically justify this.

We typically want to test $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 \neq 0$. The alternative states that at least one of the additional covariates have an effect on the failure times. Observe that according to assumption A5, θ_0 must be in the interior of Θ. Hence, we can not use the theorem for testing if the variance of one or more random effects is zero. For example, consider a model with d-dimensional normally distributed random effects. When the full cohort is used in analysis, the distribution of the likelihood ratio statistic for testing if one of the random effects is zero, i.e. testing if the corresponding column and row of the covariance matrix are zero against the alternative that the matrix is positive definite, follows a $50 : 50$ mixture of a χ^2_{d-1} and a χ^2_d distribution. We have yet not found the distribution of the statistic under two-phase sampling.

The asymptotic distribution in Theorem 4 is the distribution of the quadratic form $A^T\tilde{I}A$, where A has a multivariate normal distribution with zero mean, and is sometimes called a generalized χ^2 distribution. Note that if $\mathrm{Var}(A) = \tilde{I}^{-1}$, which it is when estimates are based on the full cohort data, the second matrix in Theorem 4 reduce to the identity matrix so that all eigenvalues have value one and the asymptotic distribution is the usual χ^2_r distribution. Under two-phase sampling however, the λs are greater than one, a consequence of the additional uncertainty from discarding data.

The proofs of Corollary 1 and Theorem 4 are based on an asymptotic expansion of the profile log likelihood of Murphy and van der Vaart (2000) and can be found in Appendix A.4.

3.2 Numerical methods

We have explored maximization of the IPW likelihood (3) with both an EM algorithm similar to that suggested by Zeng and Lin (2007a) as well as with the algorithm *fminunc* in the MATLAB optimization toolbox. The latter is a subspace trust-region method based on the interior-reflective Newton method described in Coleman and Li (1996, 1995). Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients. To avoid negative estimates of the jump sizes for Λ, we used the logarithms of the jump sizes as the parameters. All programming was done in MATLAB with parts of the code written in C. We numerically approximated the integrals for the normal random effects by Gaussian quadratures. After convergence, variances were estimated by the formula suggested in Theorem 3. We also tried estimating the variances with the method suggested in the discussion following Corollary 1. Simulations indicated that the estimates are consistent, but the algorithm is too slow to be of practical use.

At least for larger sample sizes, the EM algorithm was observed to be much faster than the interior-reflective Newton method when maximizing unweighted likelihoods. For weighted likelihoods, however, the EM algorithm appears to come reasonably close to the maximum value in just a few iterations, but requires many more rounds to actually converge. Stopping the algorithm too early had a severe impact on the likelihood ratios, less so on parameter estimates. The direct optimization on the other hand is extremely slow when the sample size is large. In the end, we settled for an optimization procedure where we first run the EM algorithm to find an approximate solution subsequently fine tuned by the subspace trust-region method.

Davies (1980) presents an algorithm for calculating the distribution of linear combinations of χ^2 random variables based on numerical inversion of the characteristic function. We used the C implementation of this algorithm that can be found on that article's author's web page to calculate the p-values of the IPW likelihood ratio test of Theorem 4.

4 Simulation studies

To assess the performance of the proposed inferential procedures we conducted numerical studies. We first considered a proportional odds model with Gaussian random effects shared within pairs,

$$\Lambda(t|X_{ij}, Z_{ij}, \mathbf{b}_i) = \log\left\{1 + \Lambda_0(t)\exp(\beta^T X_{ij} + \mathbf{b}_i)\right\}, \tag{7}$$

with $\Lambda_0(t) = t^2/20000$ corresponding to a Weibull distribution with scale parameter 1/20000 and shape parameter 2; $X_{ij} = (X_{ij1}, \ldots, X_{ij5})^T$, where $X_{ij1} \sim N(0,2)$, $X_{i12} = X_{i22} \sim N(0,1)$, $X_{i13} = X_{i23} = I(N(0,1) > 0.2)$, $X_{i14} = X_{i24} = I(N(0,1) > 0.5)$, $X_{ij5} = X_{ij4} + \zeta_{ij} I(\zeta_{ij} < 1.5) + 1.5 I(\zeta_{ij} > 1.5)$, where $\zeta_{ij} \sim N(0,1)$; $\beta = (0,0,0,1,-1)^T$ and $\mathbf{b}_i \sim N(0, \sigma^2 = 4)$.

Censoring times were generated from a normal distribution with mean 75 and variance 100. All remaining subjects were censored at time $\tau = 100$. This corresponds to a censoring rate of approximately 94%. The starting value for the finite dimensional parameter was set to $(0,0,0,0,0,1)$ and the starting value of the estimate of the jump size $\Lambda\{Y_{ij}\}$ was set to one divided by the number of subjects still at risk at time Y_{ij}.

				Bias	ESE	SEE	CP	Eff
$m_1 = M1$	β_1	=	0	0.015	0.047	0.051	0.946	0.889
$m_2 = 2M1$	β_2	=	0	-0.024	0.079	0.071	0.965	0.815
	β_3	=	0	-0.012	0.159	0.156	0.932	0.815
	β_4	=	1	-0.030	0.192	0.174	0.972	0.839
	β_5	=	-1	0.012	0.090	0.079	0.975	0.918
	σ^2	=	4	0.086	0.626	0.579	0.930	0.995
$m_1 = 0.5M1$	β_1	=	0	0.006	0.065	0.068	0.912	0.601
$m_2 = 1.5M1$	β_2	=	0	-0.020	0.105	0.102	0.935	0.541
	β_3	=	0	-0.005	0.213	0.232	0.938	0.541
	β_4	=	1	-0.061	0.259	0.238	0.938	0.561
	β_5	=	-1	0.005	0.123	0.126	0.978	0.629
	σ^2	=	4	-0.047	0.892	0.973	0.900	0.707
$m_1 = M1$	β_1	=	0	0.015	0.053	0.056	0.935	0.803
$m_2 = M1$	β_2	=	0	-0.021	0.092	0.099	0.935	0.698
	β_3	=	0	0.008	0.186	0.196	0.938	0.699
	β_4	=	1	-0.041	0.222	0.221	0.953	0.730
	β_5	=	-1	0.013	0.097	0.097	0.970	0.848
	σ^2	=	4	0.079	0.631	0.588	0.930	0.989

Table 1: Bias and ESE correspond to the bias and empirical standard error of the IPWMLE; SEE to the mean of the standard error estimates; CP to the coverage proportion of 95% confidence intervals; and Eff to efficiency compared to the cohort estimator.

We simulated 1000 cohorts, each consisting of 10000 pairs. Pairs were divided into two strata depending on whether any member experienced the event (cases) or not (controls). On average there were 430 case pairs per cohort. We selected

m_1 of the M_1 case pairs in the cohort and m_2 of the control pairs into a subcohort for analysis. Results are shown in Table 1. The bias is small in all designs and the standard error estimates and coverage proportions are fairly accurate, except possibly for the random effect component.

Efficiency was calculated from estimates of the standard deviation of the full cohort estimators, based on the subcohorts. In the first design, where we included all case pairs and twice as many control pairs into the subcohort, the efficiency was greater than 0.8 for all parameters even though only about 15% of the observations were used. It is interesting to compare the second and third designs where the size of the subcohort is the same, but the proportion of cases is different. Bias and standard error estimates agreed well, but the efficiency was seriously deflated when subsampling case pairs. This is expected as the case pairs carry most of the relevant information. The design including more cases is clearly preferable.

As a second model we generated survival time pairs from a proportional hazards model with bivariate normal random effects,

$$\Lambda(t|X_{ij}, Z_{ij}, \mathbf{b}_i) = \Lambda_0(t) \exp(\beta_1 X_{ij1} + \beta_2 X_{ij2} + Z_{ij}\mathbf{b}_{1i} + \mathbf{b}_{2i}),$$

with $\Lambda_0(t) = t/10000$ corresponding to an exponential distribution with parameter $1/10000$; $\beta_1 = 1$, $\beta_2 = -1$;

$$\begin{pmatrix} \mathbf{b}_{1i} \\ \mathbf{b}_{2i} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right);$$

$X_{ij1} \sim N(0,1)$, $X_{i12} = 1$, $X_{i22} = 0$ and $Z_{i1} = Z_{i2} \sim N(0,1)$.

We used the same censoring as for the proportional hazards model, now corresponding to a censoring rate of approximately 95%. The starting values for the β parameters were set zero, those pertaining to the covariance matrix of the random effect to the identity matrix and the inital jump sizes were set as for the previous model. Results are shown in Table 2. Again, the regression coefficient estimators were virtually unbiased and the corresponding estimated standard errors agreed well with the emprical standard errors. The coverage proportions also agreed with the theory. The estimates of the elements of the covariance matrix of the random effects were less impressive. We note that Zeng et al. (2005) experienced similar problems with covariance matrix estimation when considering a proportional odds model with a bivariate normal random effect but without subsampling.

To investigate the distribution of the likelihood ratio we considered testing $\beta_1 = \beta_2 = \beta_3 = 0$ in a proportional odds model identical to (7), exceot that now $\Lambda_0(t) = t^{3/2}/8000$ corresponding to a censoring of 93%. All case pairs and

				Bias	ESE	SEE	CP	Eff
$m_1 = M_1$	β_1	=	1	-0.001	0.076	0.079	0.947	0.748
$m_2 = M_1$	β_2	=	-1	-0.005	0.111	0.113	0.944	0.896
	σ_{11}	=	1	0.037	0.196	0.241	0.856	0.851
	σ_{12}	=	0.5	0.013	0.113	0.119	0.929	0.758
	σ_{22}	=	1	0.069	0.512	0.483	0.987	0.774

Table 2: Bias and ESE correspond to the bias and empirical standard error of the IPWMLE of the proportional hazards model; SEE to the mean of the standard error estimates; CP to the coverage proportion of 95% confidence intervals; and Eff to efficiency compared to the cohort estimator.

twice as many controls were inculuded in the phase two sample. We calculated p-values of the test for each of 1000 cohorts, consisting of 3500 pairs each, based on the distribution of Theorem 4 but with $\tilde{I}$ and $\tilde{V}$ estimated from data. 4.8% of the p-values were bellow 0.05 and 10.51% bellow 0.1. Figure 1 shows density estimates and a QQ-plot of the observed values of the likelihood ratio statisics against a simulated data set of the same size from the generalized χ^2-distribution of Theorem 4 with estimated parameters.

We have investigated the distribution of the likelihood ratio statistic in numerous models with various choices of Box-Cox transformations, different numbers and distributions of covariates and varying number of regression parameters set to zero and both univariate and bivariate Gaussian random effects (not shown). The conclusions are in all set ups the same as for the model described above. We conclude that the approximation is sufficiently accurate for practical use.

5 An example

As a hands on illustration, we considered a study on cardiovascular diseases (CVD) among Swedish men. In Sweden 30% of the premature deaths (45-64 years) among men are caused by CVD. It is well known that a family history of CVD elevates the risk of CVD. A negative impact of social class on CVD and death has been documented for both social class of origin and obtained social class. A detailed description of the data set as well as references to the litterature on social class as risk factor of CVD can be found in Tiikkaja et al. (2010).

The phase one sample consisted of 230942 pairs, corresponding to the two oldest brothers from families that were registered in the 1960 Swedish population

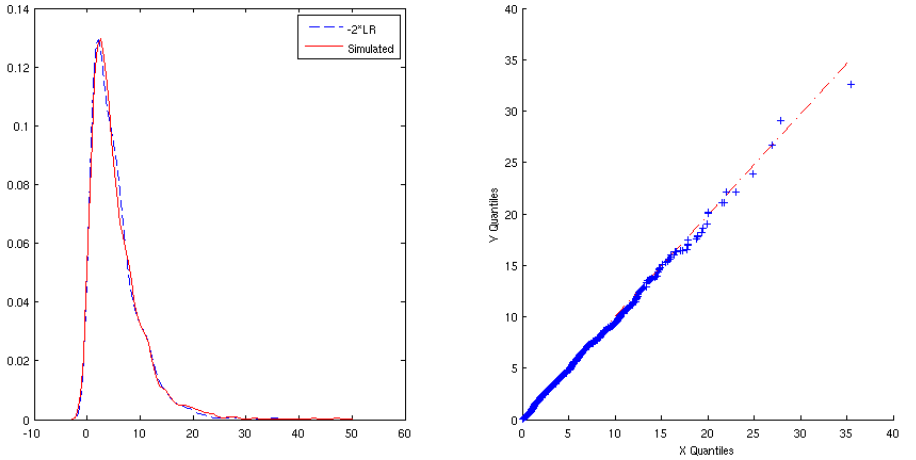

Figure 1: The plot to the left shows the kernel smoothed density (as given by the MATLAB command *ksdensity*) of the likelihood ratio statistic from 1000 simulations of for testing if $\beta_1 = \beta_2 = \beta_3 = 0$ in the proportional odds model against 1000 simulated realizations of the distribution given in Theorem 4 but with estimated parameters. On the right hand side is a QQ-plot of the same data set.

and housing census and where both siblings were residenting in Sweden 1990. Sibships were identified from the Swedish multi-generation register. There are 6313 deaths caused by CVD. In 106 sib-pairs both brothers died from CVD.

The covariates we considered were parental social class (1960) and own mid-life occupational class (1990), educational level and family history of CVD. All covariates were categorical and represented by dummy variables in our model. Social class of origin and obtained social class were categorized as unclassifiable (missing information or unclassifiable occupation), manual and non-manual (including also self-employed). The highest household class was used for parental class whereas the individual's own occupation was used for adult social class. Educational level was grouped into unknown, primary, secondary and university. A family history covariate indicated if any of the parents died from CVD.

We randomly split the data set and used one third of the lifetimes for model selection and the rest for parameter estimation in the final model. By using disjunct sets we avoid problems associated with model selection when interpreting the confidence intervals.

Education	Primary	β_1	0.556 (0.031)
	Secondary	β_2	0.186 (0.032)
	Unknown	β_3	0.080 (0.037)
Family history	CVD death	β_4	0.150 (0.008)
Parental class	Manual	β_5	-0.316 (0.047)
	Non-manual	β_6	-0.362 (0.048)
Adult class	Manual	β_7	-0.974 (0.016)
	Non-manual	β_8	-1.259 (0.015)
Clustering		σ^2	0.848 (0.024)

Table 3: Estimates in the final proportional hazards model. Standard errors shown in parentheses.

From the portion of the cohort intended for model selection we selected a subcohort consisting of all sib-pairs with at least one death due to CVD (2101 pairs) and a random sample of 1000 control pairs (1.3% of all available controls). As a first model we included the covariates family history of CVD, parental and obtained social class, and consider the class of Box-Cox transformations see (2) with a shared univariate Gaussian random effect. We let ρ vary from 0 to 1 in 0.1 increments and maximized the corresponding weighted observed likelihood. Since all models had the same number of parameters, model comparisons were based directly on the likelihood values. It turned out that $\rho = 1$, corresponding to the proportional hazards model, gave the maximal likelihood value.

Next, we investigated including additional covariates to the chosen model. Theorem 4 was applied (again with matrices $\tilde{I}$ and $\tilde{V}$ estimated from data) to evaluate the effect of adding the three dummies representing educational class. The test indicated a significant effect (p-value=0.017) and therefore we decided to include educational class in our final model.

The selected model was then fitted to a subcohort of the remaining sibling pairs. All case pairs (4145 pairs) were included in the subcohort together with 1000 randomly chosen control pairs (0.007% of the available control pairs). Results, summarised in Table 3, show that a family history of CVD increases the risk of death in CVD, higher education level decreases the risk, and a higher childhood social class, as well as adult, has a negative impact on the risk of death in CVD. The adult class seem to have a greater effect than childhood class, which is in line with findings reported elsewhere (see Tiikkaja et al. (2010)

and the references given there). Furthermore, there is a strong clustering effect among brothers.

6 Discussion

We have extended previous results on maximum likelihood estimation in semiparametric transformation models for clustered failure times to two-phase sampled data and rigorously derived the asymptotic properties of the estimator and a related likelihood ratio test. A consistent estimator of the asymptotic variances was proposed. The method was verified to perform well on simulated data and illustrated on a real world data set.

The variance estimator we propose requires inversion of the full observed information matrix, even if interest is restricted to the finite dimensional parameter. This is disappointing both because it requires specifying the Hessian of the log likelihood during implementation and also, when the number of observed failure times is large, because of problems associated with inversion of big matrices. When inference for the Euclidean parameter is based on the whole cohort, treatment of the full information matrix can elegantly be avoided by differentiation of the profile likelihood (Zeng et al., 2008; Murphy and van der Vaart, 2000). We have not yet been able to find equivalent results for two-phase sampled data.

In some applications we may be interested in testing whether the variance of a random effect is zero. The null hypothesis is then on the boundary of the parameter space, a violation of assumption A5. We believe that the distribution of the likelihood ratio statistic in this case is a mixture of weighted sums of χ^2_1 distributions, but further work is needed to find the values of the weights.

Our extension can easily be adapted to the broader class of semiparametric regression models for right censored failure time data described by Zeng and Lin (2007a, 2010). One obvious extension would be to let the transformation H depend on some unknown parameter estimated from data. This was considered by Zeng and Lin (2007b), and is readily achieved by modifications of the assumptions similar to theirs. An extension to allow for recurrent events within the same individual and also allowing for events of different types should be straightforward given the work of Zeng and Lin (2010).

In writing this paper we specifically had sampling from routine registers in mind. The main reason for subsampling in this set up is the computational cost associated with the large sample size, but the methods apply more generally. In epidemiological studies the motive for subsampling may instead be the cost,

or ethical issues, associated with measuring the covariate of actual interest. If auxiliary variables believed to be correlated with the covariates are more readily available, these can be used to increase statistical efficiency, either by calibrating the weights to cohort totals or by using these variables to estimate the weights, see e.g. Borgan et al. (2000), Breslow and Wellner (2007, 2008) and Breslow et al. (2009).

Acknowledgment

We thank Professor Pär Sparén, and PhD students Sanna Tiikkaja and Ninoa Malki at Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, for permission to use the sib-pair data and preparation of the data set.

A Proofs

We review some results from Breslow and Wellner (2007) that are indispensable in the following. Define the IPW empirical measure of the (potentially partial) observations from clusters $O_1, \ldots, O_n$, by

$$\mathbb{P}_n^\pi = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \delta_{O_i}$$

where δ_{O_i} denotes the Dirac measure placing unit mass on O_i. We will use operator notation in the following. That is, for any measurable function $f[O]$,

$$\mathbb{P}_n^\pi f = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} f[O_i]$$

and

$$P_0 f = Ef[O_1].$$

Define the IPW empirical process

$$\begin{aligned} \mathbb{G}_n^\pi &= \sqrt{n}\,(\mathbb{P}_n^\pi - P_0) \\ &= \sqrt{n}(\mathbb{P}_n - P_0) + \sqrt{n}(\mathbb{P}_n^\pi - \mathbb{P}_n). \end{aligned}$$

From Breslow and Wellner (2007, Section 4) we have that if $v_k = P1_{\mathcal{V}_k} > 0$, for $k = 1, \ldots, K$,

$$\mathbb{G}_n^\pi \xrightarrow{\mathcal{L}} \mathbb{G} + \sum_{k=1}^{K} \sqrt{v_k \frac{1-p_k}{p_k}} \mathbb{G}_k = \mathbb{G}^\pi \tag{8}$$

on $\ell^\infty(\mathcal{F})$, where $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_K)$ is a vector of independent Brownian bridge processes, all indexed by a P_0-Donsker class $\mathcal{F}$. Specifically, $\mathbb{G}_k$ is a $P_{0|k}$-Brownian bridge process, where $P_{0|k}(\cdot) = E(\cdot|V \in \mathcal{V}_k)$ indexed by $\mathcal{F}$. Breslow and Wellner (2007, Proposition B.1) further states that if $\mathcal{F}$ is P_0-Donsker and $v_k > 0$, then $\mathcal{F}$ is $P_{0|k}$-Donsker on stratum k, $k = 1, \ldots, K$.

We will repeatedly make use of the following Lemma with is an IPW analog to van der Vaart and Wellner (1996, Lemma 3.3.5) with a similar proof.

Lemma 1. *For each ϑ in subset of a normed space and every h in an arbitrary indexing set $\mathcal{F}$, let $x \mapsto \phi(\vartheta)\mathbf{h}[x]$ be a measurable function such that*

$$\{\phi(\vartheta)h - \phi(\vartheta_0)h : \|\vartheta - \vartheta_0\| < \delta, h \in \mathcal{F}\},$$

is P-Donsker for some $\delta > 0$ and

$$\sup_{h\in\mathcal{F}} P\{\phi(\vartheta)h - \phi(\vartheta_0)h\}^2 \to 0, \text{ as } \vartheta \to \vartheta_0.$$

Then if $\vartheta \xrightarrow{P} \vartheta_0$,

$$\|\mathbb{G}_n^\pi\{\phi(\vartheta) - \phi(\vartheta_0)\}\|_{\ell^\infty(\mathcal{F})} = o_P(1).$$

A.1 Proof of Theorem 1

The proof follows closely that of Zeng et al. (2008, Lemma 3, Theorem 1), where we will replace the empirical measure by the IPW empirical measure. The result will follow if we can verify steps $(i) - (iii)$ below.

(i) The maximum likelihood estimate $(\hat{\theta}_n, \hat{\Lambda}_n)$ exists.

(ii) With probability one, $\lim_{n\to\infty} \hat{\Lambda}_n(\tau) < \infty$.

(iii) If (ii) is true, by compactness of the parameter space for θ and the Helly selection theorem, there exists a subsequence of $\{n\}$, still denoted $\{n\}$, for which $\hat{\theta}_n$ converges to θ^* and $\hat{\Lambda}_n(t) \xrightarrow{a.s.} \Lambda^*(t)$ for $t \in [0, \tau]$ along that subsequence, for some θ^* and Λ^*. We show that any convergent subsequence of $(\hat{\theta}_n, \hat{\Lambda}_n)$ must converge to (θ_0, Λ_0).

Proof of (i). It is sufficient to show that the jump size of $\hat{\Lambda}_n$ at Y_{ij} such that $\Delta_{ij} = 1$ is finite. From assumption A8, $\{G'(x)x\}^{\Delta_{ij}}\{1-G(x)\}^{1-\Delta_{ij}}$ is bounded.

Choosing $x = \int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)$ in this expression we see that

$$
\begin{aligned}
& L_n^{\pi}(\theta,\Lambda) \\
&= \prod_{i=1}^{n}\left[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\left\{G'\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right.\right. \\
&\quad \left.\times\, e^{\mathbf{X}_{ij}(Y_{ij})^T\beta+\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}}\Lambda\{Y_{ij}\}\right\}^{\Delta_{ij}} \\
&\quad \left.\times\left\{1-G\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right\}^{1-\Delta_{ij}}\eta(\mathbf{b};\gamma)d\mathbf{b}\right]^{\xi_i/\pi_i} \\
&\leq \prod_{i=1}^{n}\left[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\left\{G'\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right.\right. \\
&\quad \left.\times\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right\}^{\Delta_{ij}} \\
&\quad \left.\times\left\{1-G\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right\}^{1-\Delta_{ij}}\eta(\mathbf{b};\gamma)d\mathbf{b}\right]^{\xi_i/\pi_i} \\
&= O(1).
\end{aligned}
$$

By assumption A2, there will eventually exist some (i,j) such that $Y_{ij} = \tau$ and $\xi_i = 1$ with probability one. That is, at least one integral in the expression

$$
\prod_{i=1}^{n}\left[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\left\{1-G\left(\int_0^{\tau} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right\}^{I(Y_{ij}=\tau)}\eta(\mathbf{b};\gamma)d\mathbf{b}\right]^{\xi_i/\pi_i} \tag{9}
$$

is present when n is large enough. Since $G(\infty) = 1$ by A7 such an integral has value zero if Λ has an infinite jump size for some failure time. We conclude that the jump sizes of Λ must be finite. Observing that (9) is a factor in $L_n^{\pi}(\theta,\Lambda)$, at least asymptotically, we obtain that $L_n^{\pi}(\theta,\Lambda)$ is bounded by a finite constant times

$$
\prod_{i=1}^{n}\left[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\left\{1-G\left(\int_0^{\tau} e^{\mathbf{X}_{ij}(t)^T\beta+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\right)\right\}^{I(Y_{ij}=\tau)}\eta(\mathbf{b};\gamma)d\mathbf{b}\right]^{\xi_i/\pi_i}.
$$

Together with assumption A5 that θ belongs to a compact set, this shows that the maximum likelihood estimate $(\hat{\theta}_n, \hat{\Lambda}_n)$ exists. □

Proof of (ii). In order to show that $\hat{\Lambda}_n(\tau)$ is bounded uniformly for all large n we first construct a step function $\overline{\Lambda}_n$ with jumps only at the Y_{ij} for which $\Delta_{ij} = 1$ such that $\overline{\Lambda}_n$ is close to the true function Λ_0.

Consider one-dimensional submodels for Λ defined by the map

$$\varepsilon \mapsto \Lambda^\varepsilon = \int (1 + \varepsilon h_\Lambda) d\Lambda$$

where h_Λ is an arbitrary nonnegative function of bounded variation. The directional derivative in the direction h_Λ of the weighted log likelihood with respect to ε evaluated at $\varepsilon = 0$ yields a score function for Λ. It takes the form

$$\begin{aligned} \ell_\Lambda(\theta, \Lambda)[h_\Lambda][O_i] &= \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log l\left(\theta, \int (1 + \varepsilon h_\Lambda) d\Lambda\right)[O_i] \qquad (10) \\ &= \sum_{j=1}^{n_i} \Delta_{ij} h_\Lambda(Y_{ij}) \\ &\quad + \int_{\mathbf{b}} \sum_{j=1}^{n_i} Q_{ij}(\beta, \Lambda, \mathbf{b}) \int_0^{Y_{ij}} e^{X_{ij}(t)^T\beta + Z_{ij}(t)^T\mathbf{b}} h_\Lambda(t) d\Lambda(t) d\mu_i(\theta, \Lambda, \mathbf{b}), \end{aligned}$$

where

$$\begin{aligned} Q_{ij}(\beta, \Lambda, \mathbf{b}) &= \frac{\Delta_{ij} G''(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}} d\Lambda(s))}{G'(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}} d\Lambda(s))} \\ &\quad - \frac{(1 - \Delta_{ij}) G'(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}} d\Lambda(s))}{1 - G(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}} d\Lambda(s))}, \end{aligned}$$

$$d\mu_i(\theta, \Lambda, \mathbf{b}) = \frac{R_1(\beta, \Lambda, \mathbf{b})[O_i]\eta(\mathbf{b}, \gamma) d\mathbf{b}}{\int_{\mathbf{b}} R_1(\beta, \Lambda, \mathbf{b})[O_i]\eta(\mathbf{b}, \gamma) d\mathbf{b}}.$$

and

$$\begin{aligned} &R_1(\beta, \Lambda, \mathbf{b})[O_i] = \\ &\quad \prod_{j=1}^{n_i} \left\{ G'\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta + \mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t)\right) e^{\mathbf{X}_{ij}(Y_{ij})^T\beta + \mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}} \right\}^{\Delta_{ij}} \\ &\quad \times \left\{ 1 - G\left(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\beta + \mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t)\right) \right\}^{1-\Delta_{ij}}. \end{aligned}$$

Since the weighted log likelihood is maximized at $(\hat{\theta}_n, \hat{\Lambda}_n)$ over the whole model, it is maximized at $\varepsilon = 0$ when evaluated on the submodel given by $(\hat{\theta}_n, \hat{\Lambda}_n^\varepsilon)$.

Therefore $\mathbb{P}_n^\pi \ell_\Lambda(\hat{\theta}_n, \hat{\Lambda}_n)[h_\Lambda] = 0$ for all h_Λ. With the choice $h_\Lambda(u) = I\{u \leq t\}$ and using (10) this becomes

$$\begin{aligned} & \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} 1(Y_{ij} \leq t) \qquad (11) \\ = & \int_0^t \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \int_{\mathbf{b}} \sum_{j=1}^{n_i} I(Y_{ij} > s) Q_{ij}(\beta, \Lambda, \mathbf{b}) \\ & \times e^{X_{ij}(s)^T \beta + Z_{ij}(s)^T \mathbf{b}} d\mu_i(\hat{\theta}_n, \hat{\Lambda}_n, \mathbf{b}) d\hat{\Lambda}_n(s). \end{aligned}$$

Changing the order of integration and solving for $\hat{\Lambda}_n$ in (11) results in the following expression for the IPWMLE of Λ_0

$$\hat{\Lambda}_n(t) = \int_0^t \frac{d\mathbb{P}_n^\pi N(s)}{\mathbb{P}_n^\pi Q(s; \hat{\theta}_n, \hat{\Lambda}_n)}, \qquad (12)$$

where

$$Q(t; \theta, \Lambda)[O_i] = \int_{\mathbf{b}} \sum_{j=1}^{n_i} Q_{ij}(\beta, \Lambda, \mathbf{b}) e^{X_{ij}(t)^T \beta + Z_{ij}(t)^T \mathbf{b}} d\mu_i(\theta, \Lambda, \mathbf{b}) I(Y_{ij} > t)$$

and $N(t)[O_i] = \sum_{j=1}^{n_i} \Delta_{ij} 1(Y_{ij} \leq t)$ denotes the number of events in cluster i up to time t.

We introduce the following step function, in which we have replaced estimates of θ and Λ with the true values θ_0 and Λ_0 in the right hand side of expression (10).

$$\overline{\Lambda}_n(t) = \int_0^t \frac{d\mathbb{P}_n^\pi N(s)}{\mathbb{P}_n^\pi Q(s; \theta_0, \Lambda_0)}. \qquad (13)$$

The class $\{Q(t; \theta_0, \Lambda_0) : t \in [0, \tau]\}$ is P_0-Donsker and uniformly bounded away from zero (Zeng et al., 2008, Technical report, Appendix). Thus,

$$\sup_{t \in [0,\tau]} |(\mathbb{P}_n^\pi - P_0) Q(t; \theta_0, \Lambda_0)| \overset{a.s.}{\to} 0$$

and

$$\sup_{t \in [0,\tau]} \left| \overline{\Lambda}_n(t) - E\left[\frac{\sum_{j=1}^{n_i} I(Y_{ij} < t) \Delta_{ij}}{P_0 Q(t; \theta_0, \Lambda_0)} \right] \right| \overset{a.s.}{\to} 0.$$

Since from Zeng et al. (2008, Technical report, pp. 8-9)

$$E \sum_{j=1}^{n_i} I(Y_{ij} < t) \Delta_{ij} / P_0 Q(t; \theta_0, \Lambda_0) = \Lambda_0(t),$$

$\overline{\Lambda}_n(t)$ converges to $\Lambda_0(t)$ uniformly in $t \in [0, \tau]$ almost surely.

By the compactness of the parameter space for θ, there exists a subsequence of $\{n\}$ for which $\hat{\theta}_n$ converges to some θ^* along that subsequence. Now choose a further subsequence, still denoted $\{n\}$, for which $\hat{\theta}_n \to \theta^*$ and such that $\hat{\Lambda}_n \to \infty$. We will work towards a contradiction by showing that the right hand side of

$$0 \leq n^{-1} \log L_n^{\pi}(\hat{\theta}_n, \hat{\Lambda}_n) - n^{-1} \log L_n^{\pi}(\theta_0, \overline{\Lambda}_n) \tag{14}$$

will become negative as $n \to \infty$.

First,

$$\begin{aligned}
& n^{-1} \log L_n^{\pi}(\theta_0, \overline{\Lambda}_n) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \log \prod_{j=1}^{n_i} \overline{\Lambda}_n\{Y_{ij}\}^{\Delta_{ij}} \int_{\mathbf{b}} R_1(\beta_0, \overline{\Lambda}_n, \mathbf{b})[O_i]\eta(\mathbf{b}; \gamma_0)\mu(\mathbf{b}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} \log \overline{\Lambda}_n\{Y_{ij}\} \\
&\quad + \mathbb{P}_n^{\pi} \log \int_{\mathbf{b}} R_1(\beta_0, \overline{\Lambda}_n, \mathbf{b})\eta(\mathbf{b}; \gamma_0)\mu(\mathbf{b}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} \log \frac{n^{-1}}{\mathbb{P}_n^{\pi} Q(Y_{ij}; \beta_0, \gamma_0, \Lambda_0)} + O(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} \log n^{-1} \\
&\quad + \int_0^{\tau} \log \frac{d\mathbb{P}_n^{\pi} N(t)}{\mathbb{P}_n^{\pi} Q(t; \theta_0, \Lambda_0)} + O(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} \log n^{-1} + O(1).
\end{aligned} \tag{15}$$

The boundedness (almost surely) follows from the smoothness of R_1 in Λ and the Donsker property of the classes $\{Q(t; \theta_0, \Lambda_0) : t \in [0, \tau]\}$ and $\{\log Q(t; \theta_0, \Lambda_0) : t \in [0, \tau]\}$ (Zeng et al., 2008, Technical report, Appendix).

Furthermore, Zeng et al. (2008, Lemma 1) states that

$$\int_{\mathbf{b}} R_1(\beta, \Lambda, \mathbf{b})[O_i]\eta(\mathbf{b}; \gamma) d\mathbf{b} \leq O(1) \prod_{j=1}^{n_i} \{1 + \Lambda(Y_{ij})\}^{-(\rho_0 + \Delta_{ij})} \tag{16}$$

for a finite constant $O(1)$ independent of (θ, Λ) with probability one. Using

(16) we obtain

$$n^{-1}\log L_n^{\pi}(\hat{\theta}_n,\hat{\Lambda}_n) \leq O(1)+\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}\log\hat{\Lambda}_n\{Y_{ij}\}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})\log\{1+\hat{\Lambda}_n(Y_{ij})\} \qquad (17)$$

Thus, (14), (15) and (17) imply

$$\begin{aligned}
0 \quad &\leq \quad O(1)+\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}\log\hat{\Lambda}_n\{Y_{ij}\}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})\log\{1+\hat{\Lambda}_n(Y_{ij})\}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}\log n^{-1}\\
&= \quad O(1)+\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}\log n\hat{\Lambda}_n\{Y_{ij}\}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})\log\{1+\hat{\Lambda}_n(Y_{ij})\}. \qquad (18)
\end{aligned}$$

We will work towards a contradiction by showing that the bound in (18) will be negative if $\hat{\Lambda}_n(\tau)$ diverges to ∞. The proof is based on the now classical partitioning idea of Murphy (1994) as used by Zeng et al. (2008). Following Zeng et al. (2008, p. 10-11) we can choose a finite sequence $\tau = s_0 > s_1 > s_2 > \ldots s_Q \geq s_{Q+1} = 0$ such that

$$\frac{1}{2}E\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I\{Y_{ij}=s_0\} \geq E\sum_{j=1}^{n_i}\Delta_{ij}I\{Y_{ij}\in[s_1,s_0)\} \qquad (19)$$

and

$$(1-\epsilon)E\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I\{Y_{ij}\in[s_q,s_{q-1})\} \geq E\sum_{j=1}^{n_i}\Delta_{ij}I\{Y_{ij}\in[s_{q+1},s_q)\}, \qquad (20)$$

for $q=1,\ldots,Q$, where $\epsilon\in(0,1)$ is a constant such that

$$\frac{\epsilon}{1-\epsilon} < \frac{E\sum_{j=1}^{n_i}I\{Y_{ij}\in[s_1,s_0)\}}{E\sum_{j=1}^{n_i}\Delta_{ij}I\{Y_{ij}\in[0,\tau)\}}.$$

Because all subjects are censored at time τ so that $\Delta_{ij} = 0$ whenever $Y_{ij} = \tau$ (see A2), the last two terms on the right hand side of (18) can be written

$$\begin{aligned}
&\sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \log n\hat{\Lambda}_n\{Y_{ij}\} \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} I\{Y_{ij} = \tau\} \rho_0 \log\{1 + \hat{\Lambda}_n(\tau)\} \qquad (21) \\
&\quad - \sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} (\rho_0 + \Delta_{ij}) I\{Y_{ij} \in [s_{q+1}, s_q)\} \log\{1 + \hat{\Lambda}_n(Y_{ij})\}
\end{aligned}$$

Since $\hat{\Lambda}_n$ is a non-decreasing function, $\hat{\Lambda}_n(Y_{ij}) \geq \hat{\Lambda}_n(s_{q+1})$ when $Y_{ij} \in [s_{q+1}, s_q)$, and (21) is bounded by

$$\begin{aligned}
&\sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \log n\hat{\Lambda}_n\{Y_{ij}\} \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} I\{Y_{ij} = \tau\} \rho_0 \log\{1 + \hat{\Lambda}_n(\tau)\} \\
&\quad - \sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} (\rho_0 + \Delta_{ij}) I\{Y_{ij} \in [s_{q+1}, s_q)\} \log\{1 + \hat{\Lambda}_n(s_{q+1})\}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \log n\hat{\Lambda}_n\{Y_{ij}\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \\
&\quad \times \frac{\sum_{k=1}^{n} \frac{\xi_k}{\pi_k} \sum_{l=1}^{n_k} \Delta_{kl} I\{Y_{kl} \in [s_{q+1}, s_q)\} \log n\hat{\Lambda}_n\{Y_{kl}\}}{\sum_{k=1}^{n} \frac{\xi_k}{\pi_k} \sum_{l=1}^{n_k} \Delta_{kl} I\{Y_{kl} \in [s_{q+1}, s_q)\}} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \\
&\quad \times \log \left\{ n \frac{\sum_{k=1}^{n} \frac{\xi_k}{\pi_k} \sum_{l=1}^{n_k} \Delta_{kl} I\{Y_{kl} \in [s_{q+1}, s_q)\} \hat{\Lambda}_n\{Y_{kl}\}}{\sum_{k=1}^{n} \frac{\xi_k}{\pi_k} \sum_{l=1}^{n_k} \Delta_{kl} I\{Y_{kl} \in [s_{q+1}, s_q)\}} \right\} \qquad (22)
\end{aligned}$$

where the inequality follows from Jensen's inequality. Recall that the jump

sizes of $\hat{\Lambda}_n$ are positive so that

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \hat{\Lambda}_n\{Y_{ij}\} \\
&\quad \leq \quad \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \leq s_q\} \hat{\Lambda}_n\{Y_{ij}\} \\
&\quad = \quad \hat{\Lambda}_n(s_q).
\end{aligned}$$

With this in mind we note that the right hand side of (22) is bounded by

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \\
&\qquad \times \left\{ \log \hat{\Lambda}_n(s_q) - \log \frac{1}{n} \sum_{k=1}^{n} \frac{\xi_k}{\pi_k} \sum_{l=1}^{n_k} \Delta_{kl} I\{Y_{kl} \in [s_{q+1}, s_q)\} \right\} \\
&= \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \log \hat{\Lambda}_n(s_q) + O(1)
\end{aligned}$$

Thus, right hand side of (18) is bounded from above by

$$\begin{aligned}
&O(1) + \sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_{q+1}, s_q)\} \log\{1 + \hat{\Lambda}_n(s_q)\} \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} I\{Y_{ij} = \tau\} \rho_0 \log\{1 + \hat{\Lambda}_n(\tau)\} \\
&\quad - \sum_{q=0}^{Q} \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} (\rho_0 + \Delta_{ij}) I\{Y_{ij} \in [s_{q+1}, s_q)\} \log\{1 + \hat{\Lambda}_n(s_{q+1})\} \\
&= \quad O(1) - \frac{1}{2n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} (\rho_0 + \Delta_{ij}) I\{Y_{ij} = \tau\} \log\{1 + \hat{\Lambda}_n(\tau)\} \\
&\quad - \left\{ \frac{1}{2n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} (\rho_0 + \Delta_{ij}) I\{Y_{ij} = \tau\} \right. \\
&\quad \left. - \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \sum_{j=1}^{n_i} \Delta_{ij} I\{Y_{ij} \in [s_1, s_0)\} \right\} \log\left\{1 + \hat{\Lambda}_n(\tau)\right\}
\end{aligned}$$

$$-\sum_{q=1}^{Q}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I\{Y_{ij}\in[s_q,s_{q-1})\}\right.$$
$$\left.-\frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}I\{Y_{ij}\in[s_{q+1},s_q)\}\right\}\log\left\{1+\hat{\Lambda}_n(s_q)\right\}$$

The second term diverges to $-\infty$ as $\hat{\Lambda}_n(\tau)\to\infty$. The third term is negative for large n as s_1 was chosen to satisfy (19). By the selection of s_q, $q=1,\ldots,Q$, such that (20) is fulfilled, the last term cannot diverge to ∞. Hence, the expression diverges to $-\infty$. We conclude that for all n large enough, $\hat{\Lambda}_n(\tau)<\infty$. □

Proof of (iii). We have that

$$\begin{aligned}0 &\leq n^{-1}\log L_n(\hat{\theta}_n,\hat{\Lambda}_n)-n^{-1}\log L_n(\theta_0,\overline{\Lambda}_n)\\ &= \mathbb{P}_n^{\pi}\log\left\{\int_{\mathbf{b}}R_1(\hat{\beta}_n,\hat{\Lambda}_n,\mathbf{b})\eta(\mathbf{b};\hat{\gamma}_n)d\mathbf{b}\right\}\\ &\quad -\mathbb{P}_n^{\pi}\log\left\{\int_{\mathbf{b}}R_1(\beta_0,\overline{\Lambda}_n,\mathbf{b})\eta(\mathbf{b};\gamma_0)d\mathbf{b}\right\}\\ &\quad +n^{-1}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}\sum_{j=1}^{n_i}\Delta_{ij}\log\left[\frac{\hat{\Lambda}_n\{Y_{ij}\}}{\overline{\Lambda}_n\{Y_{ij}\}}\right].\end{aligned}\tag{23}$$

We will show that the right hand side of (23) converges to

$$P_0\log\frac{\int_{\mathbf{b}}R_1(\beta^*,\Lambda^*,\mathbf{b})\eta(\mathbf{b},\gamma^*)d\mathbf{b}\prod_{j=1}^{n_i}\Lambda^{*\prime}(Y_{ij})^{\Delta_{ij}}}{\int_{\mathbf{b}}R_1(\beta_0,\Lambda_0,\mathbf{b})\eta(\mathbf{b},\gamma_0)d\mathbf{b}\prod_{j=1}^{n_i}\Lambda_0{}'(Y_{ij})^{\Delta_{ij}}}$$

which is the negative Kullback-Leibler information between the density for the model (1) with parameters (θ^*,Λ^*) and at the true parameter values. This cannot be positive and is hence zero. As a consequence

$$\int_{\mathbf{b}}R_1(\beta^*,\Lambda^*,\mathbf{b})\eta(\mathbf{b},\gamma^*)d\mathbf{b}\prod_{j=1}^{n_i}\Lambda^{*\prime}(Y_{ij})^{\Delta_{ij}}=$$
$$\int_{\mathbf{b}}R_1(\beta_0,\Lambda_0,\mathbf{b})\eta(\mathbf{b},\gamma_0)d\mathbf{b}\prod_{j=1}^{n_i}\Lambda_0{}'(Y_{ij})^{\Delta_{ij}}.$$

almost surely. The identifiability result of Zeng et al. (2008, Lemma 2) now implies that $\theta^*=\theta_0$ and $\Lambda^*=\Lambda_0$. This proves statement *(iii)*.

To prove convergence of the right hand side of (23) to the negative Kullback-Leibler information, note that as $\hat{\Lambda}_n(t)$ and $\overline{\Lambda}_n(t)$ both have positive jumps at and only at the Y_{ij} for which Δ_{ij}=1, $\hat{\Lambda}_n(t)$ is absolutely continuous with respect to $\overline{\Lambda}_n(t)$ and the construction of $\overline{\Lambda}_n(t)$ implies that

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\mathbb{P}_n^\pi Q(s;\theta_0,\Lambda_0)}{|\mathbb{P}_n^\pi Q(s;\hat{\theta}_n,\hat{\Lambda}_n)|} d\overline{\Lambda}_n(s). \tag{24}$$

We can take the absolute value because the jump sizes of $\hat{\Lambda}_n(t)$ are positive, cf. (12). From Zeng et al. (2008, Technical report, pp. 12) we have

$$\begin{aligned} &\left\| Q(t;\hat{\theta}_n,\hat{\Lambda}_n)[O_i] - Q(t;\theta^*,\Lambda^*)[O_i] \right\| \qquad (25)\\ &\leq \quad O(1)\left\{ \|\hat{\theta}_n - \theta^*\| + \| + \sum_{j=1}^{n_i} |\hat{\Lambda}_n(Y_{ij}) - \Lambda^*(Y_{ij})| + \int_0^\tau |\hat{\Lambda}_n(t) - \Lambda^*(t)| \right\}. \end{aligned}$$

From the point wise convergence of $\hat{\Lambda}_n$ to Λ^* and the dominated convergence theorem, $|\hat{\Lambda}_n(Y_{ij}) - \Lambda^*(Y_{ij})| \overset{a.s.}{\to} 0$ and $\int_0^\tau |\hat{\Lambda}_n(t) - \Lambda^*(t)| dt \overset{a.s.}{\to} 0$. In combination with (25) this shows

$$\sup_{t\in[0,\tau]} \left| \mathbb{P}_n^\pi \left[Q(t;\hat{\theta}_n,\hat{\Lambda}_n) - Q(t;\theta^*,\Lambda^*) \right] \right| \overset{a.s.}{\to} 0. \tag{26}$$

Since the class $\{Q(t;\theta_0,\Lambda_0) : t \in [0,\tau]\}$ is Glivenko-Cantelli (Zeng et al., 2008, Technical report, Appendix) we obtain

$$\sup_{t\in[0,\tau]} |(\mathbb{P}_n^\pi - P_0) Q(t;\theta_0,\Lambda_0)| \overset{a.s.}{\to} 0. \tag{27}$$

Furthermore, since the class $\{Q(t;\theta^*,\Lambda^*) : t \in [0,\tau]\}$ is Glivenko-Cantelli (Zeng et al., 2008, Technical report, Appendix),

$$\sup_{t\in[0,\tau]} \left| \mathbb{P}_n^\pi Q(t;\hat{\theta}_n,\hat{\Lambda}_n) - P_0 Q(\theta^*,\Lambda^*,t) \right| \overset{a.s.}{\to} 0.$$

We wish to take limits on both sides in (24). From Zeng et al. (2008, p. 13-14) we know that

$$\min_{t\in[0,\tau]} |P_0 Q(s;\theta^*,\Lambda^*)| > 0. \tag{28}$$

Thus, (28) implies that we can take limits in (24), using (26) and (27) to obtain

$$\Lambda^*(t) = \int_0^\tau \frac{P_0 Q(t;\theta_0,\Lambda_0)}{|P_0 Q(t;\theta^*,\Lambda^*)|} d\Lambda_0(t).$$

We conclude that $\Lambda^*(t)$ is absolutely continuous with respect to $\Lambda_0(t)$ so that $\Lambda^*(t)$ is differentiable with respect to t and $d\hat{\Lambda}_n(t)/d\overline{\Lambda}_n(t)$ converges to $d\Lambda^*(t)/d\overline{\Lambda}_n(t)$ uniformly in t.

From the Donsker, and thus also Glivenko-Cantelli, property of the class

$$\left\{ \int_{\mathbf{b}} R_1(\beta, \Lambda, \mathbf{b}) \eta(\mathbf{b}, \gamma) d\mathbf{b} : \theta \in \Theta, \Lambda \text{ increasing}, \Lambda(0), \Lambda(\tau) \text{ bounded} \right\}$$

and

$$\begin{aligned} &\left\| \int_{\mathbf{b}} R_1(\hat{\beta}_n, \hat{\Lambda}_n, \mathbf{b})[O_i] \eta(\mathbf{b}, \hat{\gamma}_n) d\mathbf{b} - \int_{\mathbf{b}} R_1(\beta^* \Lambda^*, \mathbf{b})[O_i] \eta(\mathbf{b}, \gamma^*) d\mathbf{b} \right\| \qquad (29) \\ &\leq \quad O(1) \left\{ \|\hat{\theta}_n - \theta^*\| + \sum_{j=1}^{n_i} |\hat{\Lambda}_n(Y_{ij}) - \Lambda^*(Y_{ij})| + \int_0^\tau |\hat{\Lambda}_n(t) - \Lambda^*(t)| \right\}. \end{aligned}$$

both shown in Zeng et al. (2008, Technical report, Appendix),

$$\mathbb{P}_n^\pi \log \int_{\mathbf{b}} R_1(\hat{\beta}_n, \hat{\Lambda}_n, \mathbf{b}) \eta(\mathbf{b}, \hat{\gamma}_n) d\mathbf{b} \overset{a.s.}{\to} P_0 \log \int_{\mathbf{b}} R_1(\beta^*, \Lambda^*, \mathbf{b}) \eta(\mathbf{b}, \gamma^*) d\mathbf{b}.$$

Thus,

$$\begin{aligned} & n^{-1} \log L_n(\hat{\theta}_n, \hat{\Lambda}_n) - n^{-1} \log L_n(\beta_0, \gamma_0, \overline{\Lambda}_n) \\ \to \quad & P_0 \log \frac{\int_{\mathbf{b}} R_1(\beta^*, \Lambda^*, \mathbf{b}) \eta(\mathbf{b}, \gamma^*) d\mathbf{b} \prod_{j=1}^{n_i} \Lambda^{*\prime}(Y_{ij})^{\Delta_{ij}}}{\int_{\mathbf{b}} R_1(\beta_0, \Lambda_0, \mathbf{b}) \eta(\mathbf{b}, \gamma_0) d\mathbf{b} \prod_{j=1}^{n_i} \Lambda_0{}'(Y_{ij})^{\Delta_{ij}}} \end{aligned}$$

and the proof is complete. □

We have proved steps $(i) - (iii)$ and conclude that $\|\hat{\theta}_n - \theta_0\| \overset{a.s.}{\to}$ and $\hat{\Lambda}_n(t) \overset{a.s.}{\to} \Lambda_0(t)$ for $t \in [0, \tau]$. Since Λ_0 is continuous in $[0, \tau]$, the latter can be strengthened to uniform convergence.

A.2 Proof of Theorem 2

Define an δ-neighborhood $\mathcal{U}_\delta$ of the true parameter (θ_0, Λ_0) as

$$\mathcal{U}_\delta = \left\{ (\theta, \Lambda) : \|\theta - \theta_0\| + \|\Lambda - \Lambda_0\|_{\ell^\infty[0,\tau]} < \delta \right\},$$

where $\|f\|_{\ell^\infty[0,\tau]} = \sup_{t \in [0,\tau]} |f(t)|$, for a small constant $\delta > 0$. From the consistency of $(\hat{\theta}_n, \hat{\Lambda}_n)$ we conclude that for every $\delta > 0$, $(\hat{\theta}_n, \hat{\Lambda}_n) \in \mathcal{U}_\delta$ with probability close to one when the sample size n is large.

Consider the set

$$\mathcal{H} = \{\mathbf{h} = (\mathbf{h}_\theta, h_\Lambda) : \mathbf{h}_\theta \in \mathbb{R}^{d_1+d_2}, h_\Lambda \in BV[0,\tau], \|\mathbf{h}\|_{\mathcal{H}} = \|\mathbf{h}_\theta\| + \|h_\Lambda\|_V \leq 1\},$$

where $\|h_\Lambda\|_V$ denotes the total variation of h_Λ in $[0,\tau]$. Define the map $\psi : \mathcal{U}_\delta \mapsto \ell^\infty(\mathcal{H})$ by

$$\psi(\theta,\Lambda)[\mathbf{h}_\theta, h_\Lambda] = \mathbf{h}_\theta^T \ell_\theta(\theta,\Lambda) + \ell_\Lambda(\theta,\Lambda)[h_\Lambda],$$

where l_θ and $l_\Lambda[h_\Lambda]$ is the score function for θ and the score operator for Λ in the direction h_Λ for a single cluster, respectively, i.e.

$$\mathbb{P}_n^\pi \psi(\theta,\Lambda)[\mathbf{h}_\theta, h_\Lambda] = \\ n^{-1} \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} \log \prod_{i=1}^n l\left(\theta + \varepsilon \mathbf{h}_\theta, \int (1+\varepsilon h_\Lambda) d\Lambda\right)^{\xi_i/\pi_i} [O_i].$$

Specifically, for the ith cluster,

$$\ell_\theta(\theta,\Lambda)[O_i] \quad = \quad \begin{pmatrix} \ell_\beta(\theta,\Lambda)[O_i] \\ \ell_\gamma(\theta,\Lambda)[O_i] \end{pmatrix},$$

where

$$\begin{aligned} &\ell_\beta(\theta,\Lambda)[O_i] = \textstyle\sum_{j=1}^{n_i} \{\Delta_{ij} X_{ij}(Y_{ij}) \\ &\quad + \textstyle\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b}) \int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(Y_{ij})^T\beta + \mathbf{Z}_{ij}(Y_{ij})^T \mathbf{b}} X_{ij}(t) d\Lambda(t) d\mu_i(\theta,\Lambda,\mathbf{b})\Big\}, \\ &\ell_\gamma(\theta,\Lambda)[O_i] = \textstyle\int_{\mathbf{b}} \eta'(\mathbf{b};\gamma)/\eta(\mathbf{b};\gamma) d\mu_i(\theta,\Lambda,\mathbf{b}) \end{aligned}$$

and $\ell_\Lambda(\theta,\Lambda)[h_\Lambda]$ was defined in equation (10).

Zeng et al. (2008, Theorem 3.2) argue that by the smoothness of $\eta(\mathbf{b};\gamma)$ and G, see A7 and A9, the Fréchet differentiability of $P_0\psi(\theta,\Lambda)$ at the true parameter values can be verified directly. Let $P_0\dot{\psi}_0$ denote the Fréchet derivative at (θ_0,Λ_0).

We treat $(\theta-\theta_0, \Lambda-\Lambda_0)$ as an element in $\ell^\infty(\mathcal{H})$ by defining its value at $\mathbf{h} = (\mathbf{h}_\theta, h_\Lambda)$ as $\mathbf{h}_\theta^T(\theta-\theta_0) + \int h_\Lambda d(\Lambda-\Lambda_0)$. Note that in this setting $P_0\dot{\psi}_0$ is a map from $\{(\theta-\theta_0,\Lambda-\Lambda_0) : (\theta,\Lambda) \in \mathcal{U}\} \in \ell^\infty(\mathcal{H})$ to $\ell^\infty(\mathcal{H})$. Straightforward calculations yield that

$$\begin{aligned} &P_0\dot{\psi}_0(\theta-\theta_0,\Lambda-\Lambda_0)[\mathbf{h}] = \\ &\quad = \quad -(\theta-\theta_0,\Lambda-\Lambda_0)[\sigma_0(\mathbf{h})] \\ &\quad = \quad -\sigma_{0\theta}(\mathbf{h})^T(\theta-\theta_0) - \int_0^\tau \sigma_{0\Lambda}(\mathbf{h}) d(\Lambda-\Lambda_0), \end{aligned} \tag{30}$$

where the operator

$$\sigma_0 = \begin{pmatrix} \sigma_{0\theta} \\ \sigma_{0\Lambda} \end{pmatrix} = P_0 \begin{pmatrix} \tilde{\sigma}_\theta(\theta_0, \Lambda_0) \\ \tilde{\sigma}_\Lambda(\theta_0, \Lambda_0) \end{pmatrix} : \mathcal{H} \mapsto \mathcal{H}$$

is called the information operator. See Appendix A.5 for an explicit expression of this operator.

Zeng et al. (2008) proved that σ_0 is continuously invertible by showing that σ_0 is one-to-one and can be written as a sum of a continuously invertible linear operator and a compact linear operator. The invertibility then follows from van der Vaart (1998, Theorem 25.93). From the invertibility of σ_0 we have that $\varepsilon\mathcal{H} \subset \sigma(\mathcal{H})$ (Kosorok, 2008, 6.16(i)) for some positive constant ε and thus

$$\begin{aligned} &\|P_0\dot{\psi}_0(\theta - \theta_0, \Lambda - \Lambda_0)\|_{\ell^\infty(\mathcal{H})} \\ &\quad = \sup_{\mathbf{h}\in\mathcal{H}} |(\theta - \theta_0, \Lambda - \Lambda_0)\sigma_0(\mathbf{h})| \\ &\quad = \|(\theta - \theta_0, \Lambda - \Lambda_0)\|_{\ell^\infty(\sigma_0(\mathcal{H}))} \\ &\quad \geq \varepsilon\|(\theta - \theta_0, \Lambda - \Lambda_0)\|_{\ell^\infty(\mathcal{H})}. \end{aligned} \tag{31}$$

for $(\theta - \theta_0, \Lambda - \Lambda_0) \in \ell^\infty(\mathcal{H})$. This shows that $P_0\dot{\psi}_0$ is continuously invertible on its range (Kosorok, 2008, 6.16(i)).

By the Fréchet differentiability of $P_0\psi(\theta_0, \Lambda_0)$ and (31),

$$\begin{aligned} &\left\|P_0\psi(\hat{\theta}_n, \hat{\Lambda}_n) - P_0\psi(\theta_0, \Lambda_0)\right\|_{\ell^\infty(\mathcal{H})} \\ &\quad = \left\|P_0\dot{\psi}_0(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\right\|_{\ell^\infty(\mathcal{H})} + o_{P_0}(\|(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\|_{\ell^\infty(\mathcal{H})}) \\ &\quad \geq \varepsilon\|(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\|_{\ell^\infty(\mathcal{H})} + o_P\left(\|(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\|_{\ell^\infty(\mathcal{H})}\right) \\ &\quad = \|(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\|_{\ell^\infty(\mathcal{H})}\left(\varepsilon + o_{P_0}(1)\right) \end{aligned} \tag{32}$$

On the other hand, since by construction, $\mathbb{P}_n^\pi\psi(\hat{\theta}_n, \hat{\Lambda}_n) = 0$ and $P_0\psi(\theta_0, \Lambda_0) = 0$ in $\ell^\infty(\mathcal{H})$,

$$\begin{aligned} &n^{1/2}\left\|P_0\psi(\hat{\theta}_n, \hat{\Lambda}_n) - P_0\psi(\theta_0, \Lambda_0)\right\|_{\ell^\infty(\mathcal{H})} \\ &\quad = n^{1/2}\left\|\mathbb{P}_n^\pi\psi(\hat{\theta}_n, \hat{\Lambda}_n) - P_0\psi(\hat{\theta}_n, \hat{\Lambda}_n)\right\|_{\ell^\infty(\mathcal{H})} \\ &\quad \leq \|\mathbb{G}_n^\pi\psi(\theta_0, \Lambda_0)\|_{\ell^\infty(\mathcal{H})} + \left\|\mathbb{G}_n^\pi\left\{\psi(\hat{\theta}_n, \hat{\Lambda}_n) - \psi(\theta_0, \Lambda_0)\right\}\right\|_{\ell^\infty(\mathcal{H})}. \end{aligned} \tag{33}$$

We will show that the right hand side of (33) is bounded in probability. Since $\mathcal{H}$ is a Donsker class and the functional ψ is a bounded Lipschitz functional

with respect to $\mathcal{H}$, the class $\mathcal{G} = \{\psi(\theta_0, \Lambda_0)\mathbf{h} : \mathbf{h} \in \mathcal{H}\}$ is P_0-Donsker, which implies boundedness of the first term.

Furthermore, from Zeng et al. (2008, p. 16), we know that

$$\{\psi(\theta, \Lambda)\mathbf{h} - \psi(\theta_0, \Lambda_0)\mathbf{h} : (\theta, \Lambda) \in \mathcal{U}_\delta, \mathbf{h} \in \mathcal{H}\}$$

is P_0-Donsker and that

$$\sup_{\mathbf{h}\in\mathcal{H}} P_0 \{\psi(\theta, \Lambda)\mathbf{h} - \psi(\theta_0, \Lambda_0)\mathbf{h}\}^2 \to 0,$$

when $\|\theta - \theta_0\| + \|\Lambda - \Lambda_0\|_{\ell^\infty[0,\tau]} \to 0$. This implies, by Lemma 1,

$$\left\|\mathbb{G}_n^\pi \left\{\psi(\hat{\theta}_n, \hat{\Lambda}_n) - \psi(\theta_0, \Lambda_0)\right\}\right\|_{\ell^\infty(\mathcal{H})} = o_{P_0}(1), \tag{34}$$

i.e. asymptotic negligibility of the second term.

Then, from (32) and the bounds for (33) we obtain

$$\begin{aligned} & n^{1/2}\|(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)\|_{\ell^\infty(\mathcal{H})} \\ &= n^{1/2} O_{P_0}\left(\left\|P_0\psi(\hat{\theta}_n, \hat{\Lambda}_n) - P_0\psi(\theta_0, \Lambda_0)\right\|_{\ell^\infty(\mathcal{H})}\right) \\ &= O_{P_0}\left(O_{P_0}(1) + o_{P_0}(1)\right) = O_{P_0}(1). \end{aligned} \tag{35}$$

For $\mathbf{h} = (\mathbf{h}_\theta, h_\Lambda) \in \mathcal{H}$, fixed but arbitrary,

$$\begin{aligned} & n^{1/2}\mathbf{h}_\theta^T(\hat{\theta}_n - \theta_0) + n^{1/2}\int_0^\tau h_\Lambda d(\hat{\Lambda}_n - \Lambda_0) \\ &= n^{1/2}(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)[\mathbf{h}] \\ &= -n^{1/2}P_0\dot{\psi}_0(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)[\sigma^{-1}(\mathbf{h})] \\ &= -n^{1/2}\left\{P_0\psi(\hat{\theta}_n, \hat{\Lambda}_n) - P_0\psi(\theta_0, \Lambda_0)\right\}[\sigma^{-1}(\mathbf{h})] \\ &\quad + o_P\left(n^{1/2}(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)[\sigma^{-1}(\mathbf{h})]\right) \\ &= \mathbb{G}_n^\pi\psi(\theta_0, \Lambda_0)[\sigma^{-1}(\mathbf{h})] + \mathbb{G}_n^\pi\left\{\psi(\hat{\theta}_n, \hat{\Lambda}_n - \psi(\theta_0, \Lambda_0))\right\}[\sigma^{-1}(\mathbf{h})] + o_{P_0}(1) \\ &= \mathbb{G}_n^\pi\psi(\theta_0, \Lambda_0)[\sigma^{-1}(\mathbf{h})] + o_{P_0}(1), \end{aligned} \tag{36}$$

where the second equality follows from (30), the third from the Fréchet differentiability of $P_0\psi$, the fourth from the second line in (33) and (35), and the fifth from (34).

The P_0-Donsker property of $\mathcal{G}$ and the limit distribution result (8) (cf. the discussion after (8)), implies

$$n^{1/2}(\hat{\theta}_n - \theta_0) + n^{1/2}(\hat{\Lambda}_n - \Lambda_0) \xrightarrow{\mathcal{L}} \mathbb{G}_n^{\pi}\psi(\theta_0, \Lambda_0)\sigma^{-1}$$

in $\mathbb{R}^{d_1+d_2} \times \ell^{\infty}[0,\tau]$.

A.3 Proof of Theorem 3

From (36) and (8) we know that

$$n^{1/2}\mathbf{h}_{\theta}^T(\hat{\theta}_n - \theta_0) + n^{1/2}\int h_{\Lambda} d(\hat{\Lambda}_n - \Lambda_0) \tag{37}$$

converges weakly to a zero mean normally distributed variable with variance

$$\mathrm{Var}\left(\psi(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})][O]\right) + \sum_{k=1}^{K} v_k \frac{1-p_k}{p_k}\mathrm{Var}_k\left(\psi(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})][O]\right)$$

where

$$\begin{aligned}\mathrm{Var}_k(f) &= P_{0|k}(f^{\otimes 2}) - P_{0|k}^{\otimes 2}(f) \\ &= P_0(1_{\mathcal{V}_k}(V)f^{\otimes 2})/v_k - P_0^{\otimes 2}(1_{\mathcal{V}_k}(V)f)/v_k^2.\end{aligned}$$

The first term in (37) corresponds to the usual variability in estimating the parameters, assuming observations were available for all clusters. By replacing $\mathbb{P}_n$ by its IPW analog $\mathbb{P}_n^{\pi}$ in the proof of Zeng et al. (2008, Technical report, Theorem 3.4) we see that this term can be consistently estimated by $n\mathbf{h}_n^T\mathbb{J}_n^{-1}\mathbf{h}_n$.

The second component of (37) represents additional variability due to subsampling. The expression given in matrix form in the statement of the theorem is equivalent to estimating σ_0 by its empirical version, i.e. $\mathbb{J}_n$, then inverting it and plugging it into the score operator and estimating the variance of the scores stratum wise. In order to justify the suggested procedure we need to verify that

$$\mathbb{P}_n^{\pi} 1_{\mathcal{V}_j}\psi(\hat{\theta}_n, \hat{\Lambda}_n)[\hat{\sigma}_n^{-1}(\mathbf{h})] \xrightarrow{P} P_0 1_{\mathcal{V}_j}\psi(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})] \tag{38}$$

and

$$\mathbb{P}_n^{\pi} 1_{\mathcal{V}_j}\psi^2(\hat{\theta}_n, \hat{\Lambda}_n)[\hat{\sigma}_n^{-1}(\mathbf{h})] \xrightarrow{P} P_0 1_{\mathcal{V}_j}\psi^2(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})] \tag{39}$$

where $\hat{\sigma}_n = \mathbb{P}_n^{\pi}\tilde{\sigma}(\hat{\theta}_n, \hat{\Lambda}_n)$.

In Appendix A.5 we prove that the class

$$\{\tilde{\sigma}(\theta,\Lambda)(\mathbf{h}) : (\theta,\Lambda) \in \mathcal{U}_\delta, \mathbf{h} \in \mathcal{H}\}$$

is P_0-Donsker and that $\tilde{\sigma}$ is continuous in the indexing sets by using arguments similar to those in Zeng et al. (2008, Technical report, Appendix). Thus $\|(\mathbb{P}_n^\pi - P_0)\tilde{\sigma}(\theta,\Lambda)\|_{\ell^\infty(\mathcal{H})} = o_{P_0}(1)$ for any $(\theta,\Lambda) \in \mathcal{U}_\delta$. Because of the boundedness and continuity of $\tilde{\sigma}$ in (θ,Λ) (cf. Appendix A.5) we have, by dominated convergence, that $\sup_{\mathbf{h}\in\mathcal{H}} \|\hat{\sigma}_n(\mathbf{h}) - \sigma_0(\mathbf{h})\|_{\mathcal{H}} = o_{P_0}(1)$.

Since $\hat{\sigma}_n$ converges to σ_0 and σ_0 is continuously invertible on $\mathcal{H}$ we have that $\hat{\sigma}_n$ is continuously invertible with probability converging to one as $n \to \infty$ (cf. Parner (1998, Proof of Theorem 3)). For any $\mathbf{g} \in \mathcal{H}$ we choose $\mathbf{h}_n = \hat{\sigma}_n^{-1}(\mathbf{g})$. Then,

$$\begin{aligned} &\left\|\hat{\sigma}_n^{-1}(\mathbf{g}) - \sigma_0^{-1}(\mathbf{g})\right\|_{\mathcal{H}} \\ &= \left\|\sigma_0^{-1}(\sigma_0(\mathbf{h}_n)) - \sigma_0^{-1}(\hat{\sigma}_n(\mathbf{h}_n))\right\|_{\mathcal{H}} \\ &\leq O(1) \sup_{\mathbf{h}\in\mathcal{H}} \|\sigma_0(\mathbf{h}) - \hat{\sigma}_n(\mathbf{h})\|_{\mathcal{H}} \end{aligned}$$

where the inequality follows from the continuity of σ_0^{-1}. We conclude that

$$\sup_{\mathbf{h}\in\mathcal{H}} \left\|\hat{\sigma}_n^{-1}(\mathbf{h}) - \sigma_0^{-1}(\mathbf{h})\right\|_{\mathcal{H}} = o_{P_0}(1). \tag{40}$$

Because $\mathcal{H}$ is a bounded Donsker class and the functional ψ is a bounded Lipschitz functional with respect to $\mathcal{H}$, the class $\{\psi(\theta,\Lambda)\mathbf{h} : \mathbf{h} \in \mathcal{H}\}$, $(\theta,\Lambda) \in \mathcal{U}_\delta$, is a bounded P_0-Donsker class. By van der Vaart and Wellner (1996, Example 2.10.8) products of bounded Donsker classes are Donsker. Thus,

$$\left\{1_{\mathcal{V}_j}\psi(\hat{\theta}_n,\hat{\Lambda}_n)[\mathbf{h}] : \mathbf{h} \in \mathcal{H}\right\} \text{ and } \left\{1_{\mathcal{V}_j}\psi^2(\hat{\theta}_n,\hat{\Lambda}_n)[\mathbf{h}] : \mathbf{h} \in \mathcal{H}\right\}$$

are P_0-Donsker classes and

$$(\mathbb{P}_n^\pi - P_0)1_{\mathcal{V}_j}\psi(\hat{\theta}_n,\hat{\Lambda}_n)[\hat{\sigma}^{-1}(\mathbf{h})] \xrightarrow{a.s.} 0 \tag{41}$$

and

$$(\mathbb{P}_n^\pi - P_0)1_{\mathcal{V}_j}\psi^2(\hat{\theta}_n,\hat{\Lambda}_n)[\hat{\sigma}^{-1}(\mathbf{h})] \xrightarrow{a.s.} 0. \tag{42}$$

The consistency result of Theorem 1, (40) and

$$\begin{aligned}
&\left|\psi(\tilde{\theta},\tilde{\Lambda})[\tilde{\mathbf{h}}][O_i] - \psi(\theta,\Lambda)[\mathbf{h}][O_i]\right| \\
&\leq O(1)\left\{\|\tilde{\mathbf{h}}_\theta - \mathbf{h}_\theta\| + \|\tilde{\theta}_n - \theta\| + \sum_{j=1}^{n_i}|\tilde{h}_\Lambda(Y_{ij}) - h_\Lambda(Y_{ij})|\right. \\
&\quad \left. + \int_0^\tau |\tilde{h}_\Lambda(t) - h_\Lambda(t)|dt + \sum_{j=1}^{n_i}|\tilde{\Lambda}(Y_{ij}) - \Lambda(Y_{ij})| + \int_0^\tau |\tilde{\Lambda}(t) - \Lambda(t)|dt\right\}
\end{aligned}$$

(Zeng et al., 2008, Technical report, p. 16) shows, by dominated convergence, that

$$P_0 1_{\mathcal{V}_j}\psi(\hat{\theta}_n,\hat{\Lambda}_n)[\hat{\sigma}_n^{-1}(\mathbf{h})] \to P_0 1_{\mathcal{V}_j}\psi(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})]$$

and

$$P_0 1_{\mathcal{V}_j}\psi^2(\hat{\theta}_n,\hat{\Lambda}_n)[\hat{\sigma}_n^{-1}(\mathbf{h})] \to P_0 1_{\mathcal{V}_j}\psi^2(\theta_0,\Lambda_0)[\sigma_0^{-1}(\mathbf{h})].$$

Combining the two displays above and (41), (42), we get the desired consistency (38) and (39).

A.4 Proof of Corollary 1 and Theorem 4

We will use minor modification of Murphy and van der Vaart (2000, Theorem 1) to conclude that the weighted profile likelihood has a similar expansion as the ordinary profile likelihood. In order to verify the conditions of Murphy and van der Vaart (2000, Theorem 1) we construct a least favorable submodel (van der Vaart, 1999, Definition 9.7) for estimating θ_0.

Choosing $\mathbf{h} = (\mathbf{e}_i, 0)$, $i \in \{1,\ldots,d_1+d_2\}$, where $\mathbf{e}_i$ is the ith unit vector, in (36) yields

$$\begin{aligned}
&n^{1/2}\mathbf{e}_i^T(\hat{\theta}_n - \theta_0) \\
&\quad = \mathbb{G}_n^\pi\psi(\theta_0,\Lambda_0)[\sigma^{-1}(\mathbf{e}_i,0)] + o_{P_0}(1) \\
&\quad = \mathbb{P}_n^\pi\left\{\sigma_\theta^{-1}(\mathbf{e}_i,0)^T\ell_\theta(\theta_0,\Lambda_0) + \ell_\Lambda(\theta_0,\Lambda_0)[\sigma_\Lambda^{-1}(\mathbf{e}_i,0)]\right\} + o_{P_0}(1) \\
&\quad = \sigma_\theta^{-1}(\mathbf{e}_i,0)^T n^{1/2}\mathbb{P}_n^\pi\left\{\ell_\theta(\theta_0,\Lambda_0) + \ell_\Lambda(\theta_0,\Lambda_0)[\sigma_\theta(\mathbf{e}_i,0)\sigma_\Lambda^{-1}(\mathbf{e}_i,0)]\right\} + o_{P_0}(1).
\end{aligned}$$

We conclude that $\mathbf{e}_i^T\hat{\theta}_n$ is an asymptotically linear estimator of the ith component of θ_0. Repeating this procedure d_1+d_2 times to extract all components of θ_0 we get

$$\begin{aligned}
&n^{1/2}(\hat{\theta}_n - \theta_0) \\
&\quad = \tilde{I}^{-1}n^{1/2}\mathbb{P}_n^\pi\left\{\ell_\theta(\theta_0,\Lambda_0) + \ell_\Lambda(\theta_0,\Lambda_0)[h_0]\right\} + o_{P_0}(1),
\end{aligned}$$

where

$$\tilde{I}^{-1} = \left(\sigma_\theta^{-1}(\mathbf{e}_1, 0) \ldots \sigma_\theta^{-1}(\mathbf{e}_{d_1+d_2}, 0)\right)$$

and

$$h_0 \quad = \quad \tilde{I} \begin{pmatrix} \sigma_\Lambda^{-1}(\mathbf{e}_1, 0) \\ \vdots \\ \sigma_\Lambda^{-1}(\mathbf{e}_{d_1+d_2}, 0) \end{pmatrix}$$

Thus, $\hat{\theta}_n$ is an asymptotically linear estimator for θ_0 and the influence function $\tilde{\ell} = \ell_\theta(\theta_0, \Lambda_0) + \ell_\Lambda(\theta_0, \Lambda_0)[h_0]$ belongs to the space spanned by the score functions. Consequently, h_0 is the least favorable direction for estimating θ_0 (van der Vaart, 1999, Section 2.2).

Our candidate for the approximately least favorable submodel is

$$\mathbf{\Lambda}_\xi(\theta, \Lambda) = \Lambda + (\xi - \theta) \circ \int h_0 d\Lambda, \ \xi \in \mathbb{R}^{d_1+d_2},$$

where $\circ$ represents component wise multiplication.

Murphy and van der Vaart (2000, Theorem 1) impose the following conditions.

(i) The map $\xi \mapsto \ell(\xi, \theta, \Lambda)[O] = \log l(\xi, \mathbf{\Lambda}_\xi(\theta, \Lambda))[O]$ is twice continuously differentiable with derivatives $\dot{\ell}$, $\ddot{\ell}$, both continuous at $(\theta_0, \theta_0, \Lambda)$.

(ii) There exists some neighborhood $\mathcal{U}$ of $(\theta_0, \theta_0, \Lambda_0)$ such that

$$\left\{\dot{\ell}(\xi, \theta, \Lambda) : (\xi, \theta, \Lambda) \in \mathcal{U}\right\}$$

is P_0-Donsker with square integrable envelope function and

$$\left\{\ddot{\ell}(\xi, \theta, \Lambda) : (\xi, \theta, \Lambda) \in \mathcal{U}\right\}$$

is P_0-Glivenko-Cantelli and is bounded in $L_1(P_0)$.

(iii) $\dot{\ell}(\theta_0, \theta_0, \Lambda_0)[O] = \tilde{\ell}[O]$

(iv) $\mathbf{\Lambda}_\theta(\theta, \Lambda) = \Lambda$ for every (θ, Λ)

(v) $\|\hat{\Lambda}_{\tilde{\theta}_n} - \Lambda_0\|_{\ell^\infty[0,\tau]} = o_{P_0}(1)$ for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$, where $\hat{\Lambda}_\theta = \text{argmax}_\Lambda \, L_n^\pi(\theta, \Lambda)$ is the maximizer of the IPW likelihood for θ fixed.

(vi) $P_0\dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\Lambda}_{\tilde{\theta}_n}) = o_{P_0}(n^{-1/2} + \|\tilde{\theta}_n - \theta_0\|)$

Zeng et al. (2008, Technical report, Proof of Theorem 3.3) outlines how to verify these conditions. Our development is similar but slightly more developed.

$$
\begin{aligned}
&\dot{\ell}(\tilde{\theta}, \theta, \Lambda)[O_i] \\
&= \frac{\partial}{\partial \tilde{\theta}} \log l\left(\tilde{\theta}, \mathbf{\Lambda}_{\tilde{\theta}}(\theta, \Lambda)\right)[O_i] \\
&= \ell_\theta(\tilde{\theta}, \mathbf{\Lambda}_{\tilde{\theta}}(\theta, \Lambda))[O_i] + \sum_{j=1}^{n_i} \frac{\Delta_{ij} h_0(Y_{ij})}{1 + (\tilde{\theta} - \theta) \circ h_0(Y_{ij})} \\
&\quad + \int_{\mathbf{b}} \sum_{j=1}^{n_i} Q_{ij}(\beta, \Lambda, \mathbf{b}) \int_0^{Y_{ij}} e^{X_{ij}(s)^T \beta + Z_{ij}(s)^T \mathbf{b}} h_0(s) d\Lambda(s) \mu_i(\tilde{\theta}, \mathbf{\Lambda}_{\tilde{\theta}}(\theta, \Lambda), \mathbf{h}),
\end{aligned}
$$

where the division in the second term on the right hand side is to be understood component wise. Using the same arguments as in Appendix A.5 we can show that $\dot{\ell}$ is continuous in the first argument. The same is true also for the second derivative, although the calculations are tedious. Again using arguments from Appendix A.5 we can show that $\left\{\dot{\ell}(\xi, \theta, \Lambda) : (\xi, \theta, \Lambda) \in \mathcal{U}\right\}$ and $\left\{\ddot{\ell}(\xi, \theta, \Lambda) : (\xi, \theta, \Lambda) \in \mathcal{U}\right\}$ are uniformly bounded P_0-Donsker classes. Thus conditions $(i) - (ii)$ are satisfied.

Evaluating $\dot{\ell}(\tilde{\theta}, \theta, \Lambda)$ at the true parameter values yields

$$
\begin{aligned}
&\dot{\ell}(\theta_0, \theta_0, \Lambda_0)[O_i] \\
&= \ell_\theta(\theta_0, \Lambda_0)[O_i] + \sum_{j=1}^{n_i} \Delta_{ij} h_0(Y_{ij}) \\
&\quad + \int_{\mathbf{b}} \sum_{j=1}^{n_i} Q_{ij}(\beta, \Lambda, \mathbf{b}) \int_0^{Y_{ij}} e^{X_{ij}(s)^T \beta + Z_{ij}(s)^T \mathbf{b}} h_0(s) d\Lambda(s) \mu_i(\tilde{\theta}, \mathbf{\Lambda}_{\tilde{\theta}}(\theta, \Lambda), \mathbf{h}) \\
&= \ell_\theta(\theta_0, \Lambda_0)[O_i] + \ell_\Lambda(\theta_0, \Lambda_0)[h_0][O_i]
\end{aligned}
$$

which is the efficient score function for estimating θ_0 and thus condition (iii) is fulfilled.

Consider any sequence $\|\tilde{\theta}_n - \theta_0\| \xrightarrow{P} 0$. Replacing equation (14) with $0 \leq n^{-1} \log L_n^\pi(\tilde{\theta}_n, \hat{\mathbf{\Lambda}}_{\tilde{\theta}_n}) - n^{-1} \log L_n^\pi(\tilde{\theta}_n, \overline{\Lambda}_n)$ in the proof of step (ii) of Theorem 1 we see that $\hat{\mathbf{\Lambda}}_{\tilde{\theta}_n}$ is bounded in probability by the same arguments as in that proof. Since

$$
n^{-1} \log L_n(\hat{\theta}_n, \hat{\Lambda}_n) \leq n^{-1} \log L_n(\tilde{\theta}_n, \hat{\mathbf{\Lambda}}_{\tilde{\theta}_n}) \leq n^{-1} \log L_n(\tilde{\theta}_n, \overline{\Lambda}_n),
$$

we can use arguments from the proof of step (iii) of Theorem 1 to show that $\hat{\mathbf{\Lambda}}_{\tilde{\theta}_n}$ is uniformly consistent for Λ_0, which is condition (v).

In order to prove the last condition, note that

$$\begin{aligned} & n^{1/2} P_0 \left\{ \ell_\Lambda(\theta_0, \hat{\Lambda}_{\tilde{\theta}_n}) - \ell_\Lambda(\theta_0, \Lambda_0) \right\} \qquad (43) \\ = \; & -n^{1/2} \left\{ \mathbb{P}_n^\pi \ell_\Lambda(\tilde{\theta}_n, \hat{\Lambda}_{\tilde{\theta}_n}) - P_0 \ell_\Lambda(\theta_0, \hat{\Lambda}_{\tilde{\theta}_n}) \right\} \\ = \; & -n^{1/2} \left\{ \mathbb{P}_n^\pi \ell_\Lambda(\tilde{\theta}_n, \hat{\Lambda}_{\tilde{\theta}_n}) - P_0 \ell_\Lambda(\tilde{\theta}_n, \hat{\Lambda}_{\tilde{\theta}_n}) \right\} + O_{P_0}(n^{1/2} \|\tilde{\theta}_n - \theta_0\|) \\ = \; & -n^{1/2} \left\{ \mathbb{P}_n^\pi \ell_\Lambda(\theta_0, \Lambda_0) - P_0 \ell_\Lambda(\theta_0, \Lambda_0) \right\} + O_{P_0}(1 + n^{1/2} \|\tilde{\theta}_n - \theta_0\|) \\ = \; & O_{P_0}(1 + n^{1/2} \|\tilde{\theta}_n - \theta_0\|). \end{aligned}$$

For the first equality we used $P_0 \ell_\Lambda(\theta_0, \Lambda_0) = 0$, $\mathbb{P}_n^\pi \ell_\Lambda(\tilde{\theta}_n, \hat{\Lambda}_{\tilde{\theta}_n}) = 0$. The second equality follows from the Fréchet differentiability of the score operator and the invertibility of the derivative, cf. (32). The third equality is an application of Lemma 1 of the Appendix and the last equality is due to the Donsker property of the class $\mathcal{G}$ discussed in the proof of Theorem 2.

By a similar argument as for the second equality in (43), again cf. (32),

$$\left\| P_0 \left\{ \ell_\Lambda(\theta_0, \hat{\Lambda}_{\tilde{\theta}_n}) - \ell_\Lambda(\theta_0, \Lambda_0) \right\} \right\|_{\ell^\infty(\mathcal{H})} \geq \varepsilon \sup_{\|h\|_V \leq 1} \left| \int h d \left(\hat{\Lambda}_{\tilde{\theta}_n} - \Lambda_0 \right) \right|$$

for some $\varepsilon > 0$. By the discussion leading to Murphy and van der Vaart (2000, equation (16)) this implies condition (vi).

We have now verified conditions $(i) - (vi)$. Therefore we can replace $\mathbb{P}_n$ by $\mathbb{P}_n^\pi$ in the proof of Murphy and van der Vaart (2000, Theorem 1) and conclude that

$$\begin{aligned} \log pL_n^\pi(\tilde{\theta}_n) = \log pL_n^\pi(\theta_0) + (\tilde{\theta}_n - \theta_0)^T n \mathbb{P}_n^\pi \tilde{\ell} \\ -\frac{1}{2} n (\tilde{\theta}_n - \theta_0)^T \tilde{I} (\tilde{\theta}_n - \theta_0) + o_P(n^{1/2} \|\tilde{\theta}_n - \theta_0\| + 1)^2 \qquad (44) \end{aligned}$$

for any random sequence $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

Having established the asymptotic expansion (44) we are prepared for the proof of Theorem 4. Corollary 1 follow in an intermediate step. We first look into the IPWMLEs under the two hypotheses. By replacing Δ_n in the proof of Murphy and van der Vaart (2000, Corollary 1) with the IPW version $n^{1/2} \mathbb{P}_n^\pi \tilde{\ell}$ we get

$$n^{1/2} (\hat{\theta}_n - \theta_0) = n^{-1/2} \tilde{I}^{-1} \mathbb{P}_n^\pi \tilde{\ell} + o_{P_0}(1). \qquad (45)$$

Asymptotic normality of $n^{1/2}(\hat{\theta}_n - \theta_0)$ follows from (36) and (8), cf. the proof of Theorem 2. This proves Corollary 1. From standard normal theory we know that the maximum likelihood estimate $\hat{\theta}^0_{n1}$ of θ_1 conditional on $\theta_2 = \theta_{20}$ is

$$\hat{\theta}^0_{n1} = \hat{\theta}_{n1} + \tilde{V}_{11}^{-1}\tilde{V}_{21}(\hat{\theta}_{n2} - \theta_{20}), \tag{46}$$

which is also normally distributed. Let $\hat{\theta}^0_2$ denote the maximum likelihood estimate under the null hypothesis, that is $\hat{\theta}^0_n = \theta_{20}$. From (46) we can write $\hat{\theta}_n - \hat{\theta}^0_n$ in terms of $\hat{\theta}_{n2}$ as

$$\begin{aligned} \hat{\theta}_n - \hat{\theta}^0_n &= \begin{pmatrix} \hat{\theta}_{n1} - \hat{\theta}^0_{n1} \\ \hat{\theta}_{n2} - \hat{\theta}^0_{n2} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{V}_{11}^{-1}\tilde{V}_{21}(\hat{\theta}_{n2} - \theta_{20}) \\ \hat{\theta}_{n2} - \theta_{20} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{V}_{11}^{-1}\tilde{V}_{21} \\ \mathbf{1}_{r\times r} \end{pmatrix} (\hat{\theta}_{n2} - \theta_{20}). \end{aligned} \tag{47}$$

Inserting $\mathbb{P}^\pi_n\tilde{\ell} = \tilde{I}(\hat{\theta}_n - \theta_0) + o_{P_0}(1)$, known from (45), in (44) we get

$$\begin{aligned} \log pL^\pi_n(\tilde{\theta}_n) = \log pL^\pi_n(\theta_0) + n(\tilde{\theta}_n - \theta_0)^T\tilde{I}(\hat{\theta}_n - \theta_0) \\ -\frac{1}{2}n(\tilde{\theta}_n - \theta_0)^T\tilde{I}(\tilde{\theta}_n - \theta_0) + o_P(n^{1/2}\|\tilde{\theta}_n - \theta_0\| + 1)^2 \end{aligned} \tag{48}$$

for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$. Writing out the difference of expression (48) in the two points $\hat{\theta}_n$ and $\hat{\theta}^0_n$, we get, after canceling identical terms,

$$\begin{aligned} &\log pL^\pi_n(\hat{\theta}_n) - \log pL^\pi_n(\hat{\theta}^0_n) \\ &= -\frac{1}{2}n(\hat{\theta}_n - \hat{\theta}^0_n)^T\tilde{I}(\hat{\theta}_n - \hat{\theta}^0_n) \\ &\quad + o_{P_0}\left((n^{1/2}\|\hat{\theta}_n - \theta_0\| + 1)^2 + (n^{1/2}\|\hat{\theta}^0_n - \theta_0\| + 1)^2\right) \\ &= -\frac{1}{2}n(\hat{\theta}_n - \hat{\theta}^0_n)^T\tilde{I}(\hat{\theta}_n - \hat{\theta}^0_n) + o_{P_0}(1), \end{aligned} \tag{49}$$

where the last equality in (49) follows since from (44), $n^{1/2}\|\hat{\theta}_n - \theta_0\| = O_{P_0}(1)$ and similarly, $n^{1/2}\|\hat{\theta}^0_n - \theta_0\| = O_{P_0}(1)$.

Thus,

$$
\begin{aligned}
& 2\log\frac{\sup_{\theta\in\Theta} pL_n^\pi(\theta)}{\sup_{\theta\in\Theta_0} pL_n^\pi(\theta)} \qquad (50)\\
&= 2\left(\log pL_n^\pi(\hat\theta_n) - \log pL_n^\pi(\hat\theta_n^0)\right)\\
&= n(\hat\theta_n^0-\hat\theta_n)^T\tilde I(\hat\theta_n^0-\hat\theta_n)+o_{P_0}(1)\\
&= n^{1/2}(\hat\theta_{n2}-\theta_{20})^T\begin{pmatrix}\tilde V_{11}^{-1}\tilde V_{21}\\ \mathbf{1}_{r\times r}\end{pmatrix}^T I\begin{pmatrix}\tilde V_{11}^{-1}\tilde V_{21}\\ \mathbf{1}_{r\times r}\end{pmatrix} n^{1/2}(\hat\theta_{n2}-\theta_{20})+o_{P_0}(1).
\end{aligned}
$$

The second equality in (50) is due to (49) and the last due to (47).

Now, there exists an orthonormal matrix O such that

$$
\tilde V_{22}^T\begin{pmatrix}\tilde V_{11}^{-1}\tilde V_{21}\\ \mathbf{1}_{r\times r}\end{pmatrix}^T I\begin{pmatrix}\tilde V_{11}^{-1}\tilde V_{21}\\ \mathbf{1}_{r\times r}\end{pmatrix}\tilde V_{22} = OAO^T,
$$

where $A=\mathrm{diag}(\lambda_1,\ldots,\lambda_r)$. Since $n^{1/2}(\hat\theta_{n2}-\theta_{20})$ is asymptotically zero mean normal with covariance $\tilde V_{21}$, the right hand side of (50) is asymptotically distributed as

$$
\mathbf{Z}^TO^TAO\mathbf{Z} = \|A^{1/2}O\mathbf{Z}\|^2 \stackrel{\mathcal{L}}{=} \sum_{i=1}^r(\sqrt{\lambda_i}Z_i)^2 = \sum_{i=1}^r\lambda_iZ_i^2,
$$

where $\mathbf{Z}=(Z_1,\ldots,Z_r)$ is a vector of independent standard normal random variables. This proves the statement of Theorem 4.

A.5 The Information operator

In this appendix we give the expression for the information operator and also argue that the the operator forms a P_0-Donsker class when indexed by $\mathcal{U}_\delta\times\mathcal{H}$. We further show that the operator is continuous.

$$
\tilde\sigma(\theta,\Lambda)[\mathbf{h}] = \begin{pmatrix}\ell_{\beta\beta}(\theta,\Lambda) & \ell_{\beta\gamma}(\theta,\Lambda) & \ell_{\Lambda\beta}(\theta,\Lambda)\\ \ell_{\gamma\beta}(\theta,\Lambda) & \ell_{\gamma\gamma}(\theta,\Lambda) & \ell_{\Lambda\gamma}(\theta,\Lambda)\\ \ell_{\beta\Lambda}(\theta,\Lambda) & \ell_{\gamma\Lambda}(\theta,\Lambda) & \ell_{\Lambda\Lambda}(\theta,\Lambda)\end{pmatrix}\begin{bmatrix}\mathbf{h}_\theta\\ h_\Lambda\end{bmatrix}
$$

In order to simplify the presentation we introduce

$$
\begin{aligned}
&P_{ij}(\beta,\Lambda,\mathbf{b}) = \\
&\quad \Delta_{ij}\frac{G'''(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}{G'(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))} \\
&\quad -\Delta_{ij}\left(\frac{G''(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}{G'(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}\right)^2 \\
&\quad -\frac{(1-\Delta_{ij})G''(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}{1-G(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))} \\
&\quad -(1-\Delta_{ij})\left(\frac{G'(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}{1-G(\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s))}\right)^2
\end{aligned}
$$

The elements of the operator evaluated at the ith cluster are

$$
\begin{aligned}
&\ell_{\beta\beta}(\theta,\Lambda)[O_i] = \\
&\quad \int_{\mathbf{b}}\left(\sum_{j=1}^{n_i} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}(s)d\Lambda(s)\right)^{\otimes 2} d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad -\left(\int_{\mathbf{b}}\sum_{j=1}^{n_i} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}(s)d\Lambda(s)d\mu_i(\theta,\Lambda,\mathbf{b})\right)^{\otimes 2} \\
&\quad +\int_{\mathbf{b}}\sum_{j=1}^{n_i} P_{ij}(\beta,\Lambda,\mathbf{b})\left(\int_0^{Y_{ij}} X_{ij}(s)e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s)\right)^{\otimes 2} d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad +\int_{\mathbf{b}}\sum_{j=1}^{n_i} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}^{\otimes 2}(s)d\Lambda(s)d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\ell_{\beta\gamma}(\theta,\Lambda)[O_i] = \\
&\quad \int_{\mathbf{b}}\frac{\eta'(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}\sum_{j=1}^{n_i} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}^T(s)d\Lambda(s)d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad -\int_{\mathbf{b}}\frac{\eta'(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad \times\int_{\mathbf{b}}\sum_{j=1}^{n_i} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}(s)^T d\Lambda(s)d\mu_i(\theta,\Lambda,\mathbf{b})
\end{aligned}
$$

$$
\begin{aligned}
&\ell_{\beta\Lambda}(\theta,\Lambda)(t)[O_i] = \\
&\quad \sum_{j=1}^{n_i}\Bigg\{\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}(s)d\Lambda(s) \\
&\qquad \times\sum_{k=1}^{n_i} Q_{ik}(\beta,\Lambda,\mathbf{b})e^{X_{ik}(t)^T\beta+Z_{ik}(t)^T\mathbf{b}}d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\qquad +\int_{\mathbf{b}} P_{ij}(\beta,\Lambda,\mathbf{b}) \\
&\qquad \times\int_0^{Y_{ij}} X_{ij}^T(s)e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}d\Lambda(s)e^{X_{ij}(t)\beta+Z_{ij}(t)^T\mathbf{b}}d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\qquad +\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b})e^{X_{ij}(t)^T\beta+Z_{ij}(t)^T\mathbf{b}}d\mu_i(\theta,\Lambda,\mathbf{b})X_{ij}(t)^T \\
&\qquad -\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b})\int_0^{Y_{ij}} e^{X_{ij}(s)^T\beta+Z_{ij}(s)^T\mathbf{b}}X_{ij}^T(s)d\Lambda(s)d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\qquad \times\int_{\mathbf{b}}\sum_{k=1}^{n_i} Q_{ik}(\beta,\Lambda,\mathbf{b})e^{X_{ik}(t)^T\beta+Z_{ik}(t)^T\mathbf{b}}d\mu_i(\theta,\Lambda,\mathbf{b})\Bigg\} I(Y_{ij}>t) \\
&\ell_{\gamma\beta}(\theta,\Lambda)[O_i] = (\ell_{\beta\gamma}(\theta,\Lambda)[O_i])^T \\
&\ell_{\gamma\gamma}(\theta,\Lambda)[O_i] = \int_{\mathbf{b}}\frac{\eta''(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}d\mu_i(\theta,\Lambda,\mathbf{b}) - \left(\int_{\mathbf{b}}\frac{\eta'(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}d\mu_i(\theta,\Lambda,\mathbf{b})\right)^{\otimes 2} \\
&\ell_{\gamma\Lambda}(\theta,\Lambda)(t)[O_i] = \\
&\quad \sum_{j=1}^{n_i}\Bigg\{\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b})e^{X_{ij}(t)^T\beta+Z_{ij}(t)^T\mathbf{b}}\frac{\eta'^T(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\qquad -\int_{\mathbf{b}}\frac{\eta'^T(\mathbf{b},\gamma)}{\eta(\mathbf{b},\gamma)}d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\qquad \times\int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b})e^{X_{ij}(t)^T\beta+Z_{ij}(t)^T\mathbf{b}}d\mu_i(\theta,\Lambda,\mathbf{b})\Bigg\} I(Y_{ij}>t) \\
&\ell_{\Lambda\beta}(\theta,\Lambda)[h_\Lambda](t)[O_i] = \int (\ell_{\beta\Lambda}(\theta,\Lambda)(t)[O_i])^T h_\Lambda(t)d\Lambda(t) \\
&\ell_{\Lambda\gamma}(\theta,\Lambda)[h_\Lambda](t)[O_i] = \int (\ell_{\gamma\Lambda}(\theta,\Lambda)(t)[O_i])^T h_\Lambda(t)d\Lambda(t)
\end{aligned}
$$

$$
\begin{aligned}
&\ell_{\Lambda\Lambda}(\theta,\Lambda)[h_\Lambda](t)[O_i] = \\
&\quad \sum_{j=1}^{n_i} \int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b}) e^{X_{ij}(t)^T\beta + Z_{ij}(t)^T\mathbf{b}} d\mu_i(\theta,\Lambda,\mathbf{b}) I(Y_{ij} > t) h_\Lambda(t) \\
&\quad + \int_0^\tau \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \Bigg\{ \int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b}) e^{X_{ij}(t)^T\beta + Z_{ij}(t)^T\mathbf{b}} \\
&\quad \times Q_{ik}(\beta,\Lambda,\mathbf{b}) e^{X_{ik}(s)^T\beta + Z_{ik}(s)^T\mathbf{b}} d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad - \int_{\mathbf{b}} Q_{ij}(\beta,\Lambda,\mathbf{b}) e^{X_{ij}(t)^T\beta + Z_{ij}(t)^T\mathbf{b}} d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad \times \int_{\mathbf{b}} Q_{ik}(\beta,\Lambda,\mathbf{b}) e^{X_{ik}(s)^T\beta + Z_{ik}(s)^T\mathbf{b}} d\mu_i(\theta,\Lambda,\mathbf{b}) \\
&\quad + \int_{\mathbf{b}} P_{ij}(\beta,\Lambda,\mathbf{b}) e^{X_{ij}(s)^T\beta + Z_{ij}(s)^T\mathbf{b}} e^{X_{ij}(t)^T\beta + Z_{ij}(t)^T\mathbf{b}} d\mu_i(\theta,\Lambda,\mathbf{b}) \Bigg\} \\
&\quad \times I(Y_{ik} > s) h_\Lambda(s) d\Lambda(s) I(Y_{ij} > t)
\end{aligned}
$$

From Zeng et al. (2008, Technical report, Appendix) classes of the form

$$
\left\{ \int_{\mathbf{b}} f(\theta,\Lambda)\eta^l(\gamma,\mathbf{b}) d\mathbf{b} : l = 1,2,\ \theta \in \Theta, \Lambda \text{ increasing}, \Lambda(0) = 0, \Lambda(\tau) \text{ bounded} \right\},
$$

where

$$
\begin{aligned}
&|f(\theta_1,\Lambda_1)[O_i] - f(\theta_2,\Lambda_2)[O_i]| \\
&\quad \leq \quad O(1) e^{M\|\mathbf{b}\|} \Bigg\{ \|\beta_1 - \beta_2\| + \int_0^\tau |\Lambda_1(t) - \Lambda_2(t)| dt \\
&\qquad\qquad + \sum_{j=1}^{n_i} |\Lambda_1(Y_{ij}) - \Lambda_2(Y_{ij})| \Bigg\},
\end{aligned}
$$

are P_0-Donsker. Note that by the arguments in Zeng et al. (2008, Technical report, Appendix) all elements of $\tilde{\sigma}$ are built up from elements of either classes of this type or of the type $I(Y_{ij} > \cdot)$. Because of the assumptions A5, A10 and A3 the classes of interest are bounded and bounded away from zero. Further, $\{I(Y_{ij} > t) : t \in [0,\tau]\}$ is of bounded variation and hence a P_0-Donsker class. Donsker properties of Donsker classes are preserved under addition and multiplication as well as division whenever the classes are bounded and bounded away from zero (van der Vaart and Wellner, 1996, Examples 2.10.7, 2.10.10). Thus, the class

$$
\{\tilde{\sigma}(\mathbf{h}_\theta, h_\Lambda)(\theta,\Lambda) : \mathbf{h} \in \mathcal{H}, \theta \in \Theta, \Lambda \text{ increasing}, \Lambda(0) = 0, \Lambda(\tau) \text{ bounded}\}
$$

is P_0-Donsker.

Also,

$$
\begin{aligned}
&|\tilde{\sigma}(\mathbf{h}_1)(\theta_1,\Lambda_1) - \tilde{\sigma}(\mathbf{h}_2)(\theta_2,\Lambda_2)| \\
&\quad \leq \quad O(1)\left\{\|\mathbf{h}_{1\theta} - \mathbf{h}_{2\theta}\| + \|\theta_1 - \theta_2\|\right. \\
&\qquad + \sum_{j=1}^{n_i} |h_{1\Lambda}(Y_{ij}) - h_{2\Lambda}(Y_{ij})| + \int_0^\tau |h_{1\Lambda}(t) - h_{2\Lambda}(t)|dt \\
&\qquad \left. + \sum_{j=1}^{n_i} |\Lambda_1(Y_{ij}) - \Lambda_2(Y_{ij})| + \int_0^\tau |\Lambda_1(t) - \Lambda_2(t)|dt \right\}
\end{aligned}
$$

Bibliography

Andersen, E. (2004). Composite likelihood and two-stage estimation in family studies. *Biostatistics*, 5(1):15–30.

Andersen, E. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, 11(3):333–350.

Borgan, Ø., Langholz, B., Samuelsen, S., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009). Improved horvitzŰthompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1:32–49.

Breslow, N. and Wellner, J. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34:86–102.

Breslow, N. and Wellner, J. (2008). A z-theorem with estimated nuisance parameters and correction note for Śweighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 35:186Ű192.

Coleman, T. and Li, Y. (1995). On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 67(2):189–224.

Coleman, T. and Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445.

Davies, R. (1980). Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics*, 87(1):323Ű333.

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

Lin, D. and Ying, D. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):1341–1349.

Lu, S. and Shih, J. (2006). Case-cohort designs and analysis for clustered failure time data. *Biometrics*, 62:1138–1148(11).

Lu, S. and Wang, M. (2002). Cohort case-control design and analysis for clustered failure-time data. *Biometrics*, 58(4):764–772.

Moger, T., Pawitan, Y., and Borgan, Ø. (2008). Case-cohort methods for survival data on families from routine registers. *Statistics in Medicine*, 27(7):1062–1074.

Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22:712–731.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26(1):183–214.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Prentice, R. and Self, S. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16(1):64–81.

Tiikkaja, S., Olsson, M., Malki, N., Modin, B., and Sparén, P. (2010). Familial risk of premature cardiovascular mortality and the impact of intergenerational class mobility. *Submitted*.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. (1999). Semiparametric statistics. In *Ecole d'Ete de Probailites de St. Flour XXIX*, volume 1781 of *Lectures on probability theory and statistics*, pages 331–457. Springer.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer.

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

Zeng, D. and Lin, D. (2007b). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102:167–180.

Zeng, D. and Lin, D. (2010). A generalized asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*, 20:871–910.

Zeng, D., Lin, D., and Lin, X. (2008). Semiparametric transformation models with random. effects for clustered failure time data. *Statistica Sinica*, 18:355–377.

Zeng, D., Lin, D., and Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association*, 100:470–483.