Semiparametric Methods for the Analysis of Progression-Related Endpoints

by

Audrey Boruvka

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

Use of progression-free survival in the evaluation of clinical interventions is hampered by a variety of issues, including censoring patterns not addressed in the usual methods for survival analysis. Progression can be right-censored before survival or interval-censored between inspection times. Current practice calls for imputing events to their time of detection. Such an approach is prone to bias, underestimates standard errors and makes inefficient use of the data at hand. Moreover a composite outcome prevents inference about the actual treatment effect on the risk of progression. This thesis develops semiparametric and sieve maximum likelihood estimators to more formally analyze progression-related endpoints. For the special case where death rarely precedes progression, a Cox-Aalen model is proposed for regression analysis of time-to-progression under intermittent inspection. The general setting considering both progression and survival is examined with a Markov Cox-type illness-death model under various censoring schemes. All of the resulting estimators globally converge to the truth slower than the parametric rate, but their finite-dimensional components are asymptotically efficient. Numerical studies suggest that the new methods perform better than their imputation-based alternatives under moderate to large samples having higher rates of censoring.

ACKNOWLEDGEMENTS

I am indebted to my supervisor Richard Cook for suggesting this research problem and providing sound advice along the way, to my thesis examiners Jerry Lawless, Mary Thompson, Charmaine Dean and Stephen McColl for their constructive comments and helpful suggestions, and to Ker-Ai Lee for her assistance with the application data. I also owe thanks to my friends and family for their patience, understanding and encouragement.

To Bradford

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

BACKGROUND

The exact time of disease progression in a clinical trial is often difficult to establish. Maintaining frequent assessments of progression status to study closure may be prohibitive, resulting in dual right-censoring times acting separately on progression and overall survival. More manageable assessment schedules can be achieved by spacing inspections further apart, but this leads to interval-censored progression times. To evaluate the effect of an intervention on the risk of this non-terminal event, standard practice calls for analysis of *progression-free survival* (PFS), a composite endpoint given by the time elapsed between treatment initiation and earliest detection of progression or death (FDA 2007). Use of PFS is fraught with issues, including the reliance on systematically-imputed progression times (Fleming et al. 2009). This thesis develops maximum likelihood methods for the analysis of alternatives to PFS that mitigate imputation and incorporate all of the data at hand.

We begin with a brief overview of the relevant statistical framework. Section 1.1 describes the general approach to specifying counting process models from event history data. Section 1.2 collects results later used to derive asymptotic properties and computational algorithms. Section 1.3 summarizes related work, with emphasis on non- and semiparametric methods. Some motivating problems and a plan for the remaining chapters are outlined in Section 1.4.

1.1 EVENT HISTORY ANALYSIS

An *event history* is a record of events arising from some underlying stochastic process. These events typically coincide with transitions an individual makes between a finite number of states (Figure 1.1). The process terminates once the individual reaches a state that is impossible to leave, an event known as *absorption*. *Event history analysis* is the study of a collection of individual event histories. When individuals are observed continuously over a time period $\mathcal{T} = [0, \tau]$, $0 < \tau < \infty$, events can be registered by jumps in a (possibly multivariate) counting process $N = \{N(t) : t \in \mathcal{T}\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$. N is assumed to be *adapted* to a *filtration* $\{\mathcal{F}_t : t \in \mathcal{T}\}$; for each $t \in \mathcal{T}$, $N(t)$ is $\mathcal{F}_t$-measurable with $\mathcal{F}_0$ containing all P-null sets in $\mathcal{F}$, $\mathcal{F}_t = \cap_{h>0} \mathcal{F}_{t+h}$ and $\mathcal{F} = \mathcal{F}_\tau$. Any stochastic process X is trivially adapted to its history $\{\mathcal{X}_t\}$, $\mathcal{X}_t = \{X(s) : s \leq t\}$.

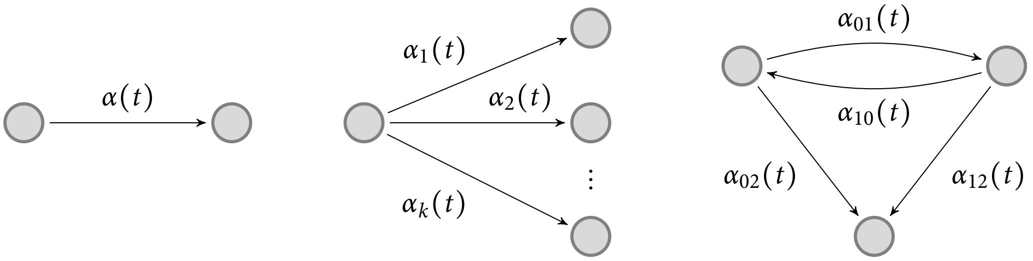


FIGURE 1.1 Survival, competing risks and illness-death models.

A *stopping time* T with respect to $\{\mathcal{F}_t\}$ is a nonnegative random variable such that $\{T \leq t\} \in \mathcal{F}_t$, for all $t \in \mathcal{T}$. Any property of an $\{\mathcal{F}_t\}$-adapted stochastic process X is said to hold *locally* if there exists a sequence of $\{\mathcal{F}_t\}$-stopping times $T_1 < T_2 < \ldots$ with $\lim_{k\to\infty} T_k = \infty$, almost surely, such that the *stopped process* $X^{T_k} = \{X(t \wedge T_k) : t \geq 0\}$ satisfies the property for each k.

Analysis is carried out on the basis of the identity

$$N = \Lambda + M, \tag{1.1}$$

where Λ is a unique nondecreasing right-continuous *compensator* whose value at t is $\mathcal{F}_{t-}$-measurable and M is such that $\mathrm{E}(M(t) \mid \mathcal{F}_s) = M(s)$ for all $s \leq t$. Since $\Lambda(t)$ is nonrandom given $\mathcal{F}_t$, Λ is said to be *predictable*. M satisfies a property that characterizes *martingale* processes. N itself is a *submartingale* since $\mathrm{E}(N(t) \mid \mathcal{F}_s) \geq N(s)$ for $s \leq t$. The identity (1.1) follows from the *Doob-Meyer decomposition*, a result that guarantees the unique decomposition of a nonnegative right-continuous local submartingale into a nondecreasing right-continuous predictable process and a right-continuous local martingale. This is analogous to the statistical decomposition: *data = model + noise.*

1.1.1 *The multiplicative intensity model*

Often we wish to specify a model on an *observed* filtration $\{\mathcal{G}_t\}$ that does not necessarily contain $\{\mathcal{F}_t\}$. The time N is under observation can be represented by the $\{\mathcal{G}_t\}$-predictable indicator process or *filter* Y. Let $A = \int \alpha$ denote the cumulative intensity function of N on $\{F_t\}$. If the *filtered* $\{\mathcal{G}_t\}$-adapted counting process $\int Y \, \mathrm{d}N$ has compensator $\int Y \, \mathrm{d}A$, the filtering of Y is said to be *independent* (Martinussen and Scheike 2006, Definition 3.1.1). This preservation of functional form enables us to specify $\mathrm{E}(Y(t)\,\mathrm{d}N(t) \mid \mathcal{G}_{t-})$ via the *multiplicative intensity model*

$$\lambda(t, \theta) = Y(t)\alpha(t, \theta), \tag{1.2}$$

where α is a nonnegative locally integrable *intensity function* determined by the parameter θ. Under the special case where N is a univariate one-jump counting process, the intensity process λ is referred to as the *hazard rate* and the jump in N coincides with a failure or survival time.

1.1 EXAMPLE (Independent right-censoring). Consider the survival time T with hazard rate $1(T \geq t)\alpha(t)$. Suppose C is a right-censoring time. Let $Y(t) = 1(T \wedge C \geq t)$. If the observed process $\int Y \,\mathrm{d}N$ has compensator $\int Y \,\mathrm{d}A$, then C follows an *independent* right-censoring mechanism. ◻

The stochastic integrals such $\int Y \,\mathrm{d}N$ or $\int Y \,\mathrm{d}A$ reduce to Lebesgue-Stieltjes integrals. With any two processes X and Y, $\int_s^t Y(u)\,\mathrm{d}X(u)$ is Lebesgue-Stieltjes provided that $\int_s^t |Y(u)||\,\mathrm{d}X(u)| < \infty$, for each $\omega \in \Omega$ and $s \leq t$, and Y is a cadlag finite variation process. If Y is predictable and X is a local martingale then the process $\int Y \,\mathrm{d}X$ is, under some weak conditions, also a local martingale.

1.1.2 *Likelihood construction*

A multiplicative analog of stochastic integration is the *product integral*, which can be defined using various identities. Product integration is central to likelihood construction, as demonstrated in Theorem 1.3 below.

1.2 DEFINITION (Product integration, Gill and Johansen 1990). Let A be a $m \times m$ matrix of cadlag finite variation processes. The *product integral* $\prod(I + \mathrm{d}A)$ on the interval $(s, t]$ is defined by the *product-limit*

$$\prod_{(s,t]}(I + \mathrm{d}A) = \lim_{|\mathcal{S}|\to 0} \prod_i (I + (A(t_i) - A(t_{i-1}))),$$

taken over refinements of the partition $\mathcal{S}$ of $(s, t]$. Equivalently $\prod(I + \mathrm{d}A)$ is the unique solution to either the *forward* or *backward equations* given by

$$X(s,t) - I = \int_s^t X(s, u-)\,\mathrm{d}A(u) = \int_s^t \mathrm{d}A(u) X(u,t),$$

respectively. In the special *commutative case* where the collection of matrices A on $(0, t]$ commute (this holds trivially when $m = 1$),

$$\prod_{(s,t]}(I + \mathrm{d}A) = \exp\Big\{\int_s^t \mathrm{d}(A - \Delta A)(t)\Big\} \prod_{u\in(s,t]}(I + \Delta A),$$

where $A - \Delta A$ is the continuous part of A. ◻

1.3 THEOREM (Jacod's likelihood ratio, Andersen et al. 1993, Theorem II.7.2). *Let $N = (N_h)$ be a multivariate counting process indexed by the time interval $[0, \tau]$ with*

$N(\tau) < \infty$. Put $\mathcal{F}_t = \mathcal{F}_0 \vee \sigma\{N(s) : s \le t\}$. On $\mathcal{F}_\tau$, consider two probability measures Q and P under which N has finite compensators Λ^Q and Λ^P, respectively. Suppose that $P \ll Q$, $\Lambda_h^P \ll \Lambda_h^Q$ Q-a.s. for each h and $\Delta\Lambda_\cdot^Q(t) = 1$ for any t. Then $\Delta\Lambda_\cdot^P(t) = 1$, Q-a.s. and

$$\frac{\mathrm{d}P}{\mathrm{d}Q} = \frac{\mathrm{d}P}{\mathrm{d}Q}\bigg|_{\mathcal{F}_0} \frac{\mathop{\Large\pi}_{(0,\tau]} \prod_h \mathrm{d}\Lambda_h^P(t)^{\Delta N_h(t)}(1 - \mathrm{d}\Lambda_\cdot^P(t))^{1-\Delta N_h(t)}}{\mathop{\Large\pi}_{(0,\tau]} \prod_h \mathrm{d}\Lambda_h^Q(t)^{\Delta N_h(t)}(1 - \mathrm{d}\Lambda_\cdot^Q(t))^{1-\Delta N_h(t)}}. \tag{1.3}$$

In (1.3), P is the measure of interest and Q serves as a convenient reference. Q is often chosen to make components of N i.i.d. standard Poisson processes. Jacobsen (2006, Theorem 5.2.1) essentially ensures existence of Q for any given P and Λ^P within the class of multiplicative intensity models (1.2). Under the parametric model $\{P_\theta : \theta \in \Theta\}$ with intensity function $\alpha_\theta = (\alpha_h^\theta)$, an independent filter $Y = (Y_h)$ and some weak condtions, we obtain from (1.3) the *likelihood process*

$$L(\theta, t) = L(\theta, 0) \mathop{\Large\pi}_{s \le t} \prod_h \mathrm{d}\Lambda_h(s, \theta)^{\Delta N_h(s)}(1 - \mathrm{d}\Lambda_\cdot(s, \theta))^{1-\Delta N_h(s)} \tag{1.4}$$

$$= L(\theta, 0) \prod_{h,s} \mathrm{d}\Lambda_h(s, \theta)^{\Delta N_h(s)} \mathop{\Large\pi}_{s \le t} (1 - \mathrm{d}\Lambda_\cdot(s, \theta)) \tag{1.5}$$

$$= L(\theta, 0) \prod_{h,s} Y_h(s)\alpha_h(s, \theta)^{\Delta N_h(s)} \exp(-\Lambda_\cdot(t, \theta)), \tag{1.6}$$

where (1.5) and (1.6) correspond respectively to the cases where $\Lambda_h^{\theta_0}$ is P_{θ_0}-a.s. continuous and P_{θ_0}-a.s. differentiable on $[0, \tau]$. Typically we assume differentiability and obtain a maximum likelihood estimator $\hat{\theta}_n$ from the *score process*

$$U(\theta, t) = \frac{\partial}{\partial\theta} \log L(\theta, t)$$

$$= \sum_h \int_0^t \frac{\partial}{\partial\theta} \log(\lambda_h(s, \theta))(\mathrm{d}N_h(s) - \lambda_h(s, \theta)\,\mathrm{d}s) \tag{1.7}$$

$$= \sum_h \int_0^t W_h(s, \theta) \frac{\partial}{\partial\theta} \lambda_h(s, \theta)(\mathrm{d}N_h(s) - \lambda_h(s, \theta)\,\mathrm{d}s), \tag{1.8}$$

where $W_h(t, \theta) = Y_h(t)/\lambda_h(t, \theta)$. The score process may lead directly to a valid score function $\dot{\ell}_\theta = U(\theta, \tau)$. Otherwise an alternative (but predictable) weight matrix $W_h(t, \theta)$ is selected and the estimator is derived using the fact that (1.8) is a $(P_\theta, \mathcal{F}_t)$-local martingale. In either case large sample properties are obtained by application of Rebolledo's (1980) central limit theorem for martingales, which implies that the sequence $\mathbb{P}_n \dot{\ell}_\theta$ converges weakly to a mean-zero Gaussian process with variance determined by α_θ.

In principle the only constraint for parameter estimation from (1.4) is the Doob-Meyer decomposition; whatever remains unexplained by the model we specify should be reasonably approximated by a martingale. So, in particular, $\int Y \,\mathrm{d}N$ can have correlated components and any covariates of the intensity function $\alpha(t, \theta)$ can depend arbitrarily on $\mathcal{G}_{t-}$. Estimates for the distribution function of the corresponding event times can be recovered from $\hat{\theta}_n$ provided that this dependence is confined to *external covariates* (Yashin and Arjas 1988). A covariate process Z is said to be external if

$$\mathrm{E}(\mathrm{d}N_j(u) \mid Z(u), Y_j(u) = 1) = \mathrm{E}(\mathrm{d}N_j(u) \mid Z(t), Y_j(u) = 1), \tag{1.9}$$

for all $0 < u \leq t$, otherwise Z is deemed *internal* (Kalbfleisch and Prentice 2002, Section 6.3.2). When this identity is satisfied for univariate N, the product integral

$$1 - F(t) = \prod_{s \leq t} (1 - \mathrm{d}A(s)) = \exp(-A(t)),$$

can be interpreted as the probability of no event up to time t. Extension to multivariate N requires some consideration of the multistate structure. In competing risks with external covariates, for example, $\exp(-A_{\cdot}(t))$ corresponds to the survivor function. In a general multistate model, product integrals yield transition probabilities provided that the process is *Markov*; only current information on the state occupied is required to specify the transition intensities.

1.4 THEOREM (Andersen et al. 1993, Theorem II.6.7). *Let $A = (A_{hj})$ be the $p \times p$ intensity measure of a multistate process X with state space $\{1, \dots, p\}$, $p < \infty$. Define the transition matrix*

$$P(s, t) = \prod_{(s,t]} (I + \mathrm{d}A(u)), \quad s \leq t.$$

If X is Markov *then*

$$\mathrm{P}(X(t) = j \mid X(s) = h, \mathcal{X}_{s-}) = \mathrm{P}(X(t) = j \mid X(s) = h) = P_{hj}(s, t), \quad s \leq t.$$

Moreover given that X is in state h at time s, the process remains in h for a duration with cumulative hazard function

$$-(A_{hh}(t) - A_{hh}(s)), \quad s \leq t \leq \inf\{u \leq s : \Delta A_{hh}(u) = 1\}.$$

If $X(t-) = h$ and $X(t) \neq h$, then the new state occupied at time t is $j \neq h$ with probability $\mathrm{d}A_{hj}(t)/ - \mathrm{d}A_{hh}(t)$.

The empirical transition intensity measure is given by the Nelson-Aalen estimator (Aalen 1975, 1978)

$$\hat{A}_{hj}(t) = \int_0^t \frac{1(Y_h(s) > 0)}{Y_h(s)} \, \mathrm{d}N_{hj}(s), \quad h \neq j,$$

where Y_h is the at-risk process $Y_h(t) = 1(X(t) = h)$. Its large sample properties follow easily from the martingale

$$\begin{aligned} \hat{A}_{hj}(t) - A^*_{hj}(t) &= \int_0^t \frac{1(Y_h(s) > 0)}{Y_h(s)} \, \mathrm{d}N_{hj}(s) - \int_0^t 1(Y_h(s) > 0) \, \mathrm{d}A_{hj}(s) \\ &= \int_0^t \frac{1(Y_h(s) > 0)}{Y_h(s)} \, \mathrm{d}M_{hj}(s). \end{aligned}$$

A martingale representation for the corresponding Aalen-Johansen estimator (Aalen and Johansen 1978) of the transition probability matrix

$$\hat{P}(s,t) = \prod_{(s,t]} (I + \mathrm{d}\hat{A}(u))$$

is obtained on the basis of the following result, which can be derived from Fubini's theorem and the forward and backward equations from Definition 1.2.

1.5 THEOREM (Duhamel's equation, Gill and Johansen 1990, Theorem 6). *Let A_1 and A_2 be intensity measures. Then*

$$\prod_{(s,t]} (I + \mathrm{d}A_1) - \prod_{(s,t]} (I + \mathrm{d}A_2) = \int_{(s,t]} \prod_{(s,u)} (I + \mathrm{d}A_1)(A_1 - A_2)(\mathrm{d}u) \prod_{(u,t]} (I + \mathrm{d}A_2).$$

1.1.3 *Semiparametric intensity-based models*

The multiplicative intensity model (1.2) can be partly specified by $(\theta, \Lambda) \in \Theta \times H$, where θ is Euclidean and Λ is some function of time. The score function for θ is defined in the usual manner, with $\dot{\ell}_{\theta,\Lambda} = \partial \log L((\theta,\Lambda),\tau)/\partial\theta$. A likelihood equation for Λ is obtained by considering a one-dimensional submodel $s \mapsto \Lambda_s$, where Λ_s depends on some bounded measurable function of time b running through an index set H. The score with respect to s at $s = 0$ defines a *score operator* $b \mapsto B_{\theta,\Lambda} b$ for Λ. The maximum likelihood estimator $(\hat{\theta}_n, \hat{\Lambda}_n)$ is the solution of $\mathbb{P}_n \, \dot{\ell}_{\theta,\Lambda} = 0$ and

$$\mathbb{P}_n \, B_{\theta,\Lambda} b - P \, B_{\theta,\Lambda} b = 0, \quad b \in H.$$

Typically the form of the intensity function is chosen so that this system of likelihood equations reduces to a finite set of estimating equations. The following examples consider estimation in a univariate counting process N. Extension to multivariate $N = (N_h)$ is straightforward.

1.6 EXAMPLE (Cox model). Cox (1972) proposed a model for the hazard rate λ having the form

$$\lambda(t \mid Z) = Y(t)\,\lambda(t)\exp(Z(t)^\top\theta),$$

where Z is a covariate process, θ is a vector of unknown regression coefficients and λ is a nonnegative integrable but otherwise unspecified baseline intensity function. When Z is time-invariant, the model reduces to *proportional* hazards and the hazard ratio θ is interpreted as a relative risk. Taking the model as the intensity of a counting process N, the score function for θ is given by

$$\sum_{i=1}^{n}\int_0^\tau Z_i(t)(\mathrm{d}N_i(t) - Y_i(t)\exp(Z_i(t)^\top\theta)\,\mathrm{d}\Lambda(t)). \tag{1.10}$$

The score of the submodel $s \mapsto \mathrm{d}\Lambda_s = (1+sb)\,\mathrm{d}\Lambda$ for some $b \in H$ is

$$\int_0^\tau \frac{\partial}{\partial s}\log(\lambda_s(t)\exp(Z_i(t)^\top\theta))(\mathrm{d}N_i(t) - Y_i(t)\exp(Z_i(t)^\top\theta)\,\mathrm{d}\Lambda_s(t)).$$

At $s = 0$ the likelihood equation is

$$\sum_{i=1}^{n}\int_0^\tau b(t)\,\mathrm{d}N(t) = \sum_{i=1}^{n}\int_0^\tau b(t)Y_i(t)\exp(Z_i(t)^\top\theta)\,\mathrm{d}\Lambda(t).$$

The system for $b \in H$ is satisfied for

$$\int \mathrm{d}\Lambda(t) = \int \frac{\mathrm{d}N.(t)}{\sum_{i=1}^{n} Y_i(t)\exp(Z_i(t)^\top\theta)}. \tag{1.11}$$

Profiling out Λ in (1.10) gives the likelihood equation for θ

$$\sum_{i=1}^{n}\int_0^\tau \left\{Z_i(t) - \frac{\sum_{j=1}^{n} Y_j(t)Z_j(t)\exp(Z_j^\top\theta)}{\sum_{j=1}^{n}\exp(Z_j(t)^\top\theta)}\right\}\mathrm{d}N_i(t) = 0. \tag{1.12}$$

The solution to (1.12) is the maximum profile likelihood estimator $\hat\theta_n$. With $\theta = \hat\theta_n$, (1.11) is the maximum profile likelihood estimator $\hat\Lambda_n$. Equivalence between $(\hat\theta_n, \hat\Lambda_n)$ and the partial likelihood estimators originally proposed by Breslow (1972) and Cox (1972) was first shown by Johansen (1983). Using martingale methods Andersen and Gill (1982) proved that $(\hat\theta_n, \hat\Lambda_n)$ is uniformly consistent and weakly converges to a Gaussian process. ▫

A similar approach may be used to derive estimating equations for *nonparametric* models, specified only by an infinite-dimensional parameter.

1.7 EXAMPLE (Aalen model). Aalen (1980, 1989) specified the intensity of an arbitrary counting process by the nonparametric additive model

$$\lambda(t \mid Z) = Y(t)\, Z(t)^\top \lambda(t),$$

where Z is a vector-valued covariate process, λ is a vector of integrable regression functions constrained by $YZ^\top\lambda \geq 0$, almost surely, but is otherwise unspecified. Typically the first element in Z is fixed at 1 and the remaining components of Z are appropriately scaled so that the first component of λ can be interpreted as a baseline intensity function. Additional regression functions account for departures from this baseline level of risk. Let $\Lambda(t) = \int_0^t \lambda(s)\, \mathrm{d}s$. Based on the submodel $s \mapsto \mathrm{d}\Lambda_s = sb + \mathrm{d}\Lambda$, the score for Λ is

$$\int_0^\tau W_i(t,\lambda) Z_i(t)^\top b(t)(\mathrm{d}N_i(t) - Z_i(t)^\top \mathrm{d}\Lambda(t)),$$

where $W_i(t,\lambda) = Y_i(t)/Z_i(t)^\top\lambda(t)$. The corresponding likelihood equation is satisfied for any $b \in H$ provided that

$$\sum_{i=1}^n W_i(t,\lambda) Z_i(t)(\mathrm{d}N_i(t) - Z_i(t)^\top \mathrm{d}\Lambda(t)) = 0. \tag{1.13}$$

An *approximate* score function can be obtained by replacing λ in W_i by a suitable estimate (Huffer and McKeague 1991; Sasieni 1992). If $W_i(t,\lambda)$ is predictable, then (1.13) is the increment of a local martingale process. Aalen (1980, 1989) proposed use of $W_i(t,\lambda) = Y_i(t)$ to obtain an estimating equation with closed form solution

$$\int \mathrm{d}\Lambda(t) = \int \sum_{i=1}^n \Big\{ \sum_{j=1}^n Y_j(t) Z_j(t) Z_j(t)^\top \Big\}^{-1} Z_i(t)\, \mathrm{d}N_i(t). \tag{1.14}$$

The resulting estimator is $\sqrt{n}$-consistent and converges in distribution to a Gaussian process (Aalen 1980). □

The Aalen model is suited for exploratory analysis, since the influence of each covariate can depend arbitrarily on time. Some of this flexibility can be traded for easier interpretation through a restricted Aalen model. Lin and Ying (1994) examined the case where only the first component of λ is time-dependent. McKeague and Sasieni (1994) considered the special case of a fixed covariate $Z(t) = Z$ with at least one component having a fixed effect. The following example attempts a balance between the Cox and Aalen models.

1.8 EXAMPLE (Cox-Aalen model). Scheike and M.-J. Zhang (2002) proposed the form

$$\lambda(t \mid Z) = Y(t)\, X(t)^\top \lambda(t) \exp(Z(t)^\top \theta),$$

which combines the features of the Aalen and Cox models. Based on the submodel $s \mapsto \mathrm{d}\Lambda_s = sb + \mathrm{d}\Lambda$, the score for Λ is

$$\int_0^\tau \frac{Y_i(t)}{X_i(t)^\top \lambda(t)} X_i(t)^\top b(t) (\mathrm{d}N_i(t) - e^{Z_i(t)^\top \theta} X_i(t)^\top \mathrm{d}\Lambda(t)).$$

The corresponding likelihood equation is met for any $b \in H$ if

$$\sum_{i=1}^n \frac{Y_i(t)}{X_i(t)^\top \lambda(t)} X_i(t) (\mathrm{d}N_i(t) - e^{Z_i(t)^\top \theta} X_i(t)^\top \mathrm{d}\Lambda(t)) = 0.$$

Replacing $1/X_i(t)^\top \lambda(t)$ with a predictable weight $W_i(t)$ yields an estimating equation for Λ given θ. Scheike and M.-J. Zhang's (2002) choice $W_i(t) = 1$ yields the solution

$$\int \mathrm{d}\Lambda(t) = \int \sum_{i=1}^n Y_i(t) X(\theta, t)^{-1} X_i(t)\, \mathrm{d}N_i(t), \tag{1.15}$$

for given θ, where $X(\theta, t)^{-1}$ is the inverse of the matrix

$$X(\theta, t) = \sum_{i=1}^n Y_i(t) e^{Z_i(t)^\top \theta} X_i(t) X_i(t)^\top.$$

The estimating equation for θ

$$\sum_{i=1}^n \int_0^\tau \Big\{ Z_i(t) - \sum_{j=1} Y_j(t) Z_j(t) e^{Z_j(t)^\top \theta} (X_i(t)^\top X(\theta, t)^{-1} X_j(t)) \Big\}\, \mathrm{d}N_i(t) = 0 \tag{1.16}$$

is obtained by profiling out Λ in the score for θ. □

Inverse weights corresponding to a uniformly consistent estimator for the baseline intensity function achieve the information bound under the Aalen model and its variants (Greenwood and Wefelmeyer 1991; Sasieni 1992), though in practice modest efficiency gains are typically seen only with large samples (Huffer and McKeague 1991; Martinussen and Scheike 2006, p. 115).

1.1.4 *Alternatives to intensity-based models*

Not all models proposed in the literature are intensity-based. The *linear transformation model*, for example, assumes that the survivor function $S(t \mid Z) = \mathrm{P}(T > t \mid Z)$ satisfies

$$S(t \mid Z) = g(h(t) + Z^\top \theta), \tag{1.17}$$

where g is a known continuous decreasing function and h is increasing but otherwise unknown. This model can be equivalently specified as

$$h(t) = Z(t)^\top \theta + \varepsilon,$$

where ε is an error term from the distribution function $1 - g^{-1}$. With ε from the extreme value distribution $g(t) = \exp(-\exp(t))$, (1.17) corresponds to the Cox model. If g is the standard logistic function $g(t) = \exp(t)/(1 + \exp(t))$, then (1.17) reduces to the *proportional odds model*

$$\frac{S(t \mid Z)}{1 - S(t \mid Z)} = \frac{S_0(t)}{1 - S_0(t)} \exp(-Z^\top \theta), \tag{1.18}$$

where S_0 is a baseline survivor function. Under the *accelerated failure time model* the survival time follows a log-linear function

$$\log(T) = Z^\top \theta + \varepsilon, \tag{1.19}$$

where the distribution of ε can be left unspecified.

1.2 SEMIPARAMETRIC MAXIMUM LIKELIHOOD

Most large sample properties for estimators from independently right-censored event history data can be derived using martingale methods, without formal consideration of semiparametric theory. This is generally not the case for other censoring patterns, such as those found in interval-censored data. The current section collects results used to derive and compute semiparametric maximum likelihood estimators. The presentation is largely taken from van der Vaart (1998, 2002) and van der Vaart and Wellner (1996).

1.2.1 *Empirical processes*

Uniform versions of the law of large numbers and the central limit theorem over an infinite-dimensional parameter space are defined by way of empirical process theory. We give some basic definitions and results.

For a given probability space $(\Omega, \mathcal{A}, P)$, consider a random element X taking values in a metric space $(\mathbb{D}, d)$ and a measurable function $f : \mathbb{D} \to \mathbb{R}$. X itself need not be Borel-measurable. We define the *outer expectation* of f by

$$\mathrm{E}^* f(X) = \inf\{\mathrm{E}\, Y : Y \geq f(X), Y : \Omega \to \mathbb{D}, \mathrm{E}\, Y \text{ exists}\}.$$

Similarly the *outer probability* of a set $B \subset \Omega$ is given by

$$\mathrm{P}^*(B) = \inf\{\mathrm{P}(A) : B \subset A, A \in \mathcal{A}\}.$$

The sequence X_n *converges in probability* to X, $X_n \overset{P}{\to} X$, if $\mathrm{P}^*(d(X_n, X) > \varepsilon) \to 0$ for every $\varepsilon > 0$. X_n *converges almost surely* to X, $X_n \overset{\mathrm{as}^*}{\to} X$, if there is a sequence of measurable random variables Δ_n with $d(X_n, X) \le \Delta_n$ and $\Delta_n \overset{\mathrm{as}}{\to} 0$.

Consider a random sample $X_1, \dots, X_n$ from P. The expectation of f under the *empirical measure* $\mathbb{P}_n$ is

$$f \mapsto \mathbb{P}_n f = \int f \, \mathrm{d}\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Similarly we write $Pf = \int f \, \mathrm{d}P$ for the expectation of f under P. The *empirical process* $\mathbb{G}_n$ at f is

$$f \mapsto \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf).$$

Provided that Pf exists, the law of large numbers is $\mathbb{P}_n f \overset{\mathrm{as}}{\to} Pf$. A class $\mathcal{F}$ of measurable functions $f : \Omega \to \mathbb{R}$ is *P-Glivenko-Cantelli* if it satisfies the uniform law of large numbers

$$\| \mathbb{P}_n f - Pf \|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{\mathrm{as}^*}{\to} 0.$$

If $Pf^2 < \infty$, then it follows from the central limit theorem that $\mathbb{G}_n f \rightsquigarrow N(0, P(f - Pf)^2)$. Consider $\ell^\infty(\mathcal{F})$, the set of all functions $g : \mathcal{F} \to \mathbb{R}$ with $\|g\|_{\mathcal{F}} < \infty$. Suppose that $\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty$ for every $x \in \mathbb{D}$, a condition that can be simply met if $\mathcal{F}$ has an integrable *envelope function* F with $|f(x)| \le F(x) < \infty$ for every x and $f \in \mathcal{F}$. Then $f \mapsto \mathbb{G}_n f$ is a map into $\ell^\infty(\mathcal{F})$. If the sequence $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges in distribution to a tight Gaussian process in $\ell^\infty(\mathcal{F})$, then the class $\mathcal{F}$ satisfies a uniform version of the central limit theorem and is said to be *P-Donsker*.

In general, the Glivenko-Cantelli and Donsker properties hold for classes that are sufficiently “small”. We quantify the size of a class using *entropy numbers*.

1.9 DEFINITION (Covering number). The *covering number* $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε needed to cover the set $\mathcal{F}$. The centers of the balls need not belong to $\mathcal{F}$. The *entropy* (*without bracketing*) is the logarithm of the covering number. The *entropy integral* is $J(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|)} \, \mathrm{d}\varepsilon$. □

1.10 DEFINITION (Bracketing number). The *bracket* $[l, u]$ for two given functions l and u is the set of all $f \in \mathcal{F}$ with $l \le f \le u$. An ε-bracket is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$. The *bracketing number* $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. The lower and upper bounds l and u must have finite norms, but need not belong to $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number. The *entropy integral with bracketing* or simply the *bracketing integral* is $J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\varepsilon$. □

Typically $\|\cdot\|$ is the $L_r(P)$ norm, given by $\|f\|_{P,r} = (\int |f|^r \,\mathrm{d}P)^{1/r}$. This and any other norm we consider satisfy the *Riesz property*: if $|f| \le |g|$ then $\|f\| \le \|g\|$. So a 2ε-bracket $[l, u]$ is contained in the ε-ball around $(l + u)/2$. Thus

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) \le N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|).$$

The following results give sufficient conditions for a class to be Glivenko-Cantelli or Donsker in terms of entropy numbers.

1.11 THEOREM (van der Vaart 1998, Theorem 19.4). *Every class $\mathcal{F}$ of measurable functions with $N_{[\,]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$ is P-Glivenko-Cantelli.*

1.12 THEOREM (van der Vaart 1998, Theorem 19.5). *Every class $\mathcal{F}$ of measurable functions with $J_{[\,]}(1, \mathcal{F}, L_2(P)) < \infty$ is P-Donsker.*

Entropy numbers for specific types of classes can be derived up to proportionality. The semiparametric parameter spaces we encounter later are not much bigger than the class of uniformly bounded, monotone functions.

1.13 THEOREM (van der Vaart and Wellner 1996, Theorem 2.7.5). *Let $\mathcal{F}$ be the class of monotone functions $f : \mathbb{R} \to [0,1]$. Then*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}, L_r(P)) \le \frac{K}{\varepsilon},$$

for every probability measure P, every $r \ge 1$ and some constant K depending only on r.

1.14 THEOREM (van der Vaart and Wellner 2000, Theorem 3). *If $\mathcal{F}_1, \dots, \mathcal{F}_k$ are Glivenko-Cantelli classes with and $\varphi : \mathbb{R}^k \to \mathbb{R}$ is continuous, then the class $\varphi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$ is Glivenko-Cantelli provided that it has an integrable envelope.*

1.15 THEOREM (van der Vaart and Wellner 1996, Theorem 2.10.6). *If $\mathcal{F}_1, \dots, \mathcal{F}_k$ are Donsker classes with integrable envelopes and $\varphi : \mathbb{R}^k \to \mathbb{R}$ is Lipschitz, then the class $\varphi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$ is Donsker provided that it has a square-integrable envelope.*

1.2.2 *Consistency and rate of convergence*

On $(\Omega, \mathcal{A}, P)$ suppose we specify the set of all possible values for P via the semiparametric model $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where Θ is Euclidean and H is some infinite-dimensional set. Suppose that $m_{\theta,\eta} : \Omega \to \mathbb{R}$ is measurable for every (θ, η) in the metric space $(\Theta \times H, d)$. Consistency of a semiparametric maximum likelihood estimator (SPMLE) can be established using the following generalization of Wald's (1949) proof.

1.16 THEOREM (van der Vaart 1998, Theorem 5.7). *Let $\{m_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ be a P-Glivenko-Cantelli class. Suppose there exists $(\theta_0, \eta_0) \in \Theta \times H$ such that*

$$\sup\{P\, m_{\theta,\eta} : d((\theta, \eta), (\theta_0, \eta_0)) > \varepsilon\} < P\, m_{\theta_0,\eta_0},$$

for every $\varepsilon > 0$. Then for any estimator $(\hat\theta_n, \hat\eta_n)$, $\mathbb{P}_n\, m_{\hat\theta_n,\hat\eta_n} \geq \mathbb{P}_n\, m_{\theta_0,\eta_0} - o_P(1)$ implies that $d((\hat\theta_n, \hat\eta_n), (\theta_0, \eta_0)) \overset{\text{as}}{\to} 0$.

The SPMLE maximizes the random criterion

$$\log \mathrm{lik}_n(\theta, \eta) \equiv n\, \mathbb{P}_n \log \mathrm{lik}(\theta, \eta),$$

where $\mathrm{lik}(\theta, \eta)(X)$ is some suitably-defined likelihood function for a single observation X from (θ_0, η_0). However note that Theorem 1.16 only needs $(\hat\theta_n, \hat\eta_n)$ to "nearly" maximize $\mathbb{P}_n\, m_{\theta,\eta}$ in a neighbourhood of the truth (θ_0, η_0). Often this weaker requirement is used to exploit a function $m_{\theta,\eta}$ that is technically more convenient than $\log \mathrm{lik}(\theta, \eta)$.

A consistent SPMLE from interval-censored data typically converges slower than the parametric rate $\sqrt{n}$, though it may be possible to show that the finite-dimensional component $\hat\theta_n$ is asymptotically efficient. Theory used to derive efficiency imposes an upper bound for the global rate of convergence, which can be obtained by application of the result below.

1.17 THEOREM (van der Vaart and Wellner 1996, Theorem 3.2.5). *Let $(\hat\theta_n, \hat\eta_n)$ be a consistent estimator satisfying $\mathbb{P}_n\, m_{\hat\theta_n,\hat\eta_n} \geq \mathbb{P}_n\, m_{\theta_0,\eta_0}$. Suppose that, for every (θ, η) close to (θ_0, η_0) and sufficiently small $\delta > 0$,*

$$P(m_{\theta,\eta} - m_{\theta_0,\eta_0}) \lesssim -d((\theta, \eta), (\theta_0, \eta_0)), \tag{1.20}$$

$$\mathrm{E}^* \sup_{d((\theta,\eta),(\theta_0,\eta_0))<\delta} |\mathbb{G}_n(m_{\theta,\eta} - m_{\theta_0,\eta_0})| \lesssim \varphi_n(\delta), \tag{1.21}$$

where $\delta \mapsto \varphi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ independent of n. Then $d((\hat\theta_n, \hat\eta_n), (\theta_0, \eta_0)) = O_P^*(1/r_n)$ with any sequence of positive numbers r_n such that and $r_n^2 \varphi_n(1/r_n) \to \infty$ no faster than $\sqrt{n}$ for every n.

The SPMLE presumes that we can *a priori* identify a finite set on which the corresponding empirical distribution potentially concentrates its mass. Moreover the parameter (θ, η) must be jointly estimable at each of these candidate support points. Under the multiple censoring schemes studied in Chapters 3 and 4, the task of support-finding proves intractable. One way out is to maximize the log-likelihood over a finite-dimensional *sieve* H_n whose size increases to H as $n \to \infty$, an approach generally known as the "method of sieves" (Geman and Hwang 1982; Grenander 1981). The resulting *sieve maximum likelihood estimator* (SMLE) is consistent, but its rate of convergence is slower than the parametric rate; for fixed n the SMLE converges to some finite-dimensional approximation $(\theta_0, \eta_{0,n}) \in \Theta \times H_n$ of the truth $(\theta_0, \eta_0) \in \Theta \times H$. A variant of Theorem 1.17 in this setting is provided by van der Vaart and Wellner (1996, Section 3.4).

1.18 THEOREM (van der Vaart and Wellner 1996, Theorem 3.4.1). *Let $(\hat\theta_n, \hat\eta_n)$ be a consistent estimator satisfying $\mathbb{P}_n m_{\hat\theta_n, \hat\eta_n} \geq \mathbb{P}_n m_{\theta_0, \eta_{0,n}}$. Suppose that, for every n and $d((\theta_0, \eta_{0,n}), (\theta_0, \eta_0)) \lesssim \delta < \infty$,*

$$\sup_{\substack{\delta/2 < d((\theta,\eta),(\theta_0,\eta_{0,n}))<\delta, \\ (\theta,\eta)\in\Theta\times H_n}} P(m_{\theta,\eta} - m_{\theta_0,\eta_{0,n}}) \leq -\delta^2, \tag{1.22}$$

$$\mathrm{E}^* \sup_{\substack{\delta/2 < d((\theta,\eta),(\theta_0,\eta_{0,n}))<\delta \\ (\theta,\eta)\in\Theta\times H_n}} |\mathbb{G}_n(m_{\theta,\eta} - m_{\theta_0,\eta_{0,n}})| \lesssim \varphi_n(\delta), \tag{1.23}$$

where $\delta \mapsto \varphi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ independent of n. Then $d((\hat\theta_n, \hat\eta_n), (\theta_0, \eta_0)) = O_P^*(1/r_n) + d((\theta_0, \eta_{0,n}), (\theta_0, \eta_0))$ with any sequence of positive numbers r_n such that $1/r_n \gtrsim d((\theta_0, \eta_{0,n}), (\theta_0, \eta_0))$ and $r_n^2 \varphi_n(1/r_n) \leq \sqrt{n}$ for every n.

The following results are useful for verifying the conditions of Theorems 1.16 to 1.18.

1.19 THEOREM (cf. van der Vaart and Wellner 1996, Corollary 2.3.12 and Problem 2.1.5). *Let $\mathcal{F}$ be a P-Donsker class with integrable envelope. Then*

$$\sup\{|\mathbb{G}_n(f-g)| : f, g \in \mathcal{F}, P(f-g)^2 < \delta_n\} \xrightarrow{P} 0,$$

for any sequence $\delta_n \to 0$ as $n \to \infty$.

1.20 DEFINITION (Hellinger distance). Let P and Q be probability measures possessing densities p and q with respect to a common σ-finite dominating measure ν. The *Hellinger distance* $d_{\mathrm{H}}(p, q)$ between P and Q is defined by

$$d_{\mathrm{H}}^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\nu.$$

Let $d_{\mathrm{TV}}(p,q) = \int |p-q|\,\mathrm{d}\nu$ denote the *total variation distance* between P and Q. Then

$$d_{\mathrm{H}}^2(p,q) \le d_{\mathrm{TV}}(p,q) \le \sqrt{2}d_{\mathrm{H}}(p,q), \tag{1.24}$$

where the first inequality follows from $|\sqrt{p}-\sqrt{q}|^2 \le |p-q|$ for any $p,q \ge 0$ and the second is a consequence of the Cauchy-Schwarz inequality. Since $\log x \le 2(\sqrt{x}-1)$ for every $x \ge 0$ we also have

$$P\log(q/p) \le 2\int \sqrt{pq}\,\mathrm{d}\nu - 2 = -\int (\sqrt{p}-\sqrt{q})^2\,\mathrm{d}\nu = -d_{\mathrm{H}}^2(p,q), \tag{1.25}$$

with equality only if $p = q$, ν-a.e. □

1.21 LEMMA (Murphy and van der Vaart 1997, Lemma A.6). *Suppose that h, g_1 and g_2 are measurable functions with $c_1 \le g_0 \le c_2$ and $(P\,g_1g_2)^2 \le c\,P\,g_1^2\,P\,g_2^2$ for constants $c < 1$ and $c_1 < 1 < c_2$ close to 1. Then*

$$P(g_0g_1 + g_2)^2 \ge K(P\,g_1^2 + P\,g_2^2),$$

where K is a constant that depends on (c, c_1, c_2) and approaches $(1-\sqrt{c})$ as $c_1 \nearrow 1$ and $c_2 \searrow 1$.

1.22 LEMMA (van der Vaart and Wellner 1996, Lemma 3.4.2). *Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $P\,f^2 < \delta^2$ and $\|f\|_\infty \le M$. Then*

$$\mathrm{E}_P^* \sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \lesssim J_{[\,]}(\delta,\mathcal{F},L_2(P))\Big(1 + \frac{J_{[\,]}(\delta,\mathcal{F},L_2(P))}{\delta^2\sqrt{n}}M\Big).$$

1.2.3 *Profile likelihood and semiparametric efficiency*

A *submodel* for $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is a parametric subset of $\mathcal{P}$ containing the truth $P_0 \equiv P_{\theta_0,\eta_0}$. Estimation of the Euclidean parameter θ in a submodel is certainly easier than estimation in $\mathcal{P}$. Whatever "information" is available to estimate θ in $\mathcal{P}$ should be no larger than the infimum of the Fisher information among all submodels. A submodel achieving this infimum is *least favourable*. The information bound for θ in $\mathcal{P}$ can be expressed in terms of the "efficient" score function, defined by the ordinary score function for θ minus its projection onto the score-space for η. A semiparametric maximum likelihood estimator $\hat{\theta}_n$ is deemed *asymptotically efficient* if it is asymptotically linear in the efficient score function.

The ordinary score function for θ in $\mathcal{P}$ is defined in the usual manner, as the partial derivative of the log-likelihood with respect to θ

$$\dot{\ell}_{\theta,\eta}(x) \equiv \frac{\partial}{\partial\theta}\log \mathrm{lik}(\theta,\eta)(x).$$

Consider a path $t \mapsto \eta_t(\theta, \eta)$ in H that induces a differentiable submodel $t \mapsto P_{\theta+at,\eta_t}$ with

$$\frac{\partial}{\partial t}_{|t=0} \log \operatorname{lik}(\theta + at, \eta_t)(x) = a^\top \dot{\ell}_{\theta,\eta}(x) + B_{\theta,\eta} h(x),$$

where $h \in L_2^0(\eta)$ and $B_{\theta,\eta} : L_2(\eta) \to L_2(P_{\theta,\eta})$ is a continuous, linear *score operator* for η. We take $B_{\theta,\eta}h$ as a score function for η with θ fixed and $L_2^0(\eta)$ an index of "directions" in which η_t approaches η. The *efficient score function* for θ is

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta},$$

where $\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$ is the projection of the ordinary score function for θ onto the closed linear span of the score functions for η. The *efficient information matrix* for θ is $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top$. Let $B_{\theta,\eta}^* : L_2(P_{\theta,\eta}) \to L_2(\eta)$ be the adjoint of the score operator for η, $B_{\theta,\eta}$, characterized by

$$P_{\theta,\eta} h(B_{\theta,\eta} g) = \eta(B_{\theta,\eta}^* h) g, \tag{1.26}$$

for every $h \in L_2^0(\eta)$ and $g \in L_2^0(P_{\theta,\eta})$. If the *information operator* $B_{\theta,\eta}^* B_{\theta,\eta}$ is continuously invertible, then the orthogonal projection $\Pi_{\theta,\eta}$ is $(B_{\theta,\eta}^* B_{\theta,\eta})^{-1} B_{\theta,\eta}^*$, the *least favourable direction* is $-(B_{\theta,\eta}^* B_{\theta,\eta})^{-1} B_{\theta,\eta}^* \dot{\ell}_{\theta,\eta}$ and the efficient score function for θ is given by

$$\tilde{\ell}_{\theta,\eta} = (I - B_{\theta,\eta}(B_{\theta,\eta}^* B_{\theta,\eta})^{-1} B_{\theta,\eta}^*)\dot{\ell}_{\theta,\eta}.$$

More often than not this condition fails to hold. For missing data problems, such as interval censoring, it is convenient to consider the observation X arising from *information loss model* (van der Vaart 1998, Section 25.5.2). In particular, we take $X = m(Y)$, where m is a known measurable transformation and Y is an unobservable variable from (θ, η). Then a score function $B_{\theta,\eta}h$ for the induced model $t \mapsto P_{\theta+at,\eta_t}$ at $t = 0$ is the conditional expectation of the score function for $t \mapsto \eta_t(\theta, \eta)$ at $t = 0$

$$B_{\theta,\eta} h(x) = \mathrm{E}_{\theta,\eta}(h(Y) \mid X = x).$$

The adjoint score operator $B_{\theta,\eta}^*$ reverses the roles of X and Y in the conditional expectation

$$B_{\theta,\eta}^* B_{\theta,\eta} h(y) = \mathrm{E}_{\theta,\eta}(B_{\theta,\eta} h(x) \mid Y = y).$$

The efficient score for θ is then

$$\tilde{\ell}_{\theta,\eta}(x) = \dot{\ell}_{\theta,\eta}(x) - B_{\theta,\eta} h_{\theta,\Lambda}(x), \tag{1.27}$$

where $h_{\theta,\Lambda}$ is the least favourable direction satisfying

$$\mathrm{E}_{\theta,\eta}(B_{\theta,\eta} h_{\theta,\Lambda}(X) \mid Y = y) = \mathrm{E}_{\theta,\eta}(\dot{\ell}_{\theta,\eta}(X) \mid Y = y). \tag{1.28}$$

for almost every y.

Now consider a random sequence $\tilde{\theta}_n \xrightarrow{P} \theta_0$. Asymptotic normality and semiparametric efficiency can be established by way of the expansion

$$\begin{aligned}\log \operatorname{plik}_n(\tilde{\theta}_n) = \log \operatorname{plik}_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^\top \sum_{i=1}^n \tilde{\ell}_0(X_i) \\ - \tfrac{1}{2}(\tilde{\theta}_n - \theta_0)^\top \tilde{I}_0 (\tilde{\theta}_n - \theta_0) + o_{P_0}(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1)^2, \end{aligned} \tag{1.29}$$

where $\log \operatorname{plik}_n(\theta) = n \sup_{\eta \in H} \mathbb{P}_n \log \operatorname{lik}(\theta, \eta)$ is the *profile* log-likelihood. Conditions under which (1.29) will hold can be made with respect to an *approximately least favourable* submodel, defined as follows.

For each (θ, η), consider a map $t \mapsto \eta_t(\theta, \eta)$ from a fixed neighbourhood of θ to H that passes through (θ, η) at $t = \theta$

$$\eta_\theta(\theta, \eta) = \eta. \tag{1.30}$$

Assume that this submodel ensures $t \mapsto \ell(t, \theta, \eta)(x) \equiv \log \operatorname{lik}(t, \eta_t(\theta, \eta))(x)$ is twice continuously differentiable for every x and the first derivative at the truth corresponds to the efficient score function for θ

$$\dot{\ell}(\theta_0, \theta_0, \eta_0) = \tilde{\ell}_{\theta_0, \eta_0}. \tag{1.31}$$

Moreover, for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$,

$$\hat{\eta}_{\tilde{\theta}_n} \equiv \operatorname*{arg\,max}_{\eta \in H} \log \operatorname{lik}_n(\tilde{\theta}_n, \eta) \xrightarrow{P} \eta_0 \tag{1.32}$$

and

$$P_0 \dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n}) = o_P(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}). \tag{1.33}$$

In other words, the maximizer of the profile log-likelihood at any consistent estimator $\tilde{\theta}_n$ is also consistent and the corresponding score at $t = \theta_0$ tends to zero with order $o_P(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2})$.

1.23 THEOREM (Murphy and van der Vaart 2000, Theorem 1). *Suppose that there is an approximately least favourable submodel for which* (1.30)–(1.33) *hold. Further assume that there exists a neighbourhood V of $(\theta_0, \theta_0, \eta)$ where $\{\dot{\ell}(t, \theta, \eta) : (t, \theta, \eta) \in V\}$ is P_0-Donsker with square-integrable envelope and $\{\ddot{\ell}(t, \theta, \eta) : (t, \theta, \eta) \in V\}$ is P_0-Glivenko-Cantelli and bounded in $L_1(P_0)$. Then* (1.29) *is satisfied.*

1.24 COROLLARY (cf. Murphy and van der Vaart 2000, Corollary 1). *If* (1.29) *holds, $\tilde{I}_0$ is invertible and $\hat{\theta}_n$ is consistent, then $\hat{\theta}_n$ is asymptotically normal with*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) + o_{P_0}(1).$$

1.25 COROLLARY (Murphy and van der Vaart 2000, Corollary 2). *If* (1.29) *holds, $\tilde{I}_0$ is invertible and $\hat{\theta}_n$ is consistent, then under the null hypothesis $H_0 : \theta = \theta_0$, the asymptotic distribution of $2\log(\operatorname{plik}(\hat{\theta}_n)/\operatorname{plik}(\theta_0))$ is chi-square with d degrees of freedom.*

1.26 COROLLARY (Murphy and van der Vaart 2000, Corollary 3). *If* (1.29) *holds and $\hat{\theta}_n$ is consistent, then for any sequence $v_n \xrightarrow{P} v$ in $\mathbb{R}^d$ and $\rho_n \xrightarrow{P} 0$ such that $(\sqrt{n}\rho_n)^{-1} = O_P(1)$,*

$$-2\frac{\log \operatorname{plik}_n(\hat{\theta}_n + \rho_n v_n) - \log \operatorname{plik}_n(\hat{\theta}_n)}{n\rho_n^2} \xrightarrow{P} v^\top \tilde{I}_0 v. \tag{1.34}$$

If $P_{\theta,\eta}\dot{\ell}(\theta,\theta,\eta) = 0$ for every (θ,η), then the "no-bias" condition (1.33) reduces to

$$P_0\dot{\ell}(\theta_0,\theta_0,\hat{\eta}_{\tilde{\theta}_n}) = o_P(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}). \tag{1.35}$$

Moreover if $\eta - \eta_0$ is a valid direction, then with (1.30)

$$\begin{aligned} P_0\dot{\ell}(\theta_0,\theta_0,\eta) = P_0\Big[\frac{p_0 - p_{\theta_0,\eta}}{p_0}\big(\dot{\ell}(\theta_0,\theta_0,\eta) - \dot{\ell}(\theta_0,\theta_0,\eta_0)\big)\Big] \\ - P_0\dot{\ell}(\theta_0,\theta_0,\eta_0)\Big[\frac{p_{\theta_0,\eta} - p_0}{p_0} - B_0(\eta - \eta_0)\Big]. \end{aligned} \tag{1.36}$$

The above expression is $O_P(\|\eta - \eta_0\|^2)$ when $\eta \mapsto p_{\theta_0,\eta}$ is twice differentiable and $\eta \mapsto \dot{\ell}(\theta_0,\theta_0,\eta)$ is differentiable at η_0. In this case a sufficient condition for (1.35) is

$$\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\| = O_P(\|\tilde{\theta}_n - \theta_0\|) + o_P(n^{-1/4}).$$

So, under some additional regularity conditions, verifying (1.33) often reduces to showing that $\hat{\eta}_n$ is consistent and converges to the truth η_0 at a rate no slower than $n^{1/4}$.

1.2.4 *Convex models*

Under right-censored data, the semiparametric intensity-based models summarized in Section 1.1.3 have closed-form estimating equations for the cumulative regression functions. This is rarely the case with interval-censored data. However the distribution function defined by any non- or semiparametric maximum likelihood estimator has no more than n support points and, in examining the likelihood function, we can often reduce the dimension further. The resulting finite-dimensional optimization problem may offer some simplifying features such as convexity. A common example is the *convex model*, where the objective function is convex and the parameters lie in a convex set.

This section summarizes some useful results for establishing convergence of algorithms to compute nonparametric maximum likelihood estimators under convex models. Throughout, suppose that the aim is to find a minimizer $\hat{\phi}$ of the objective function $L : \mathcal{K} \to (-\infty, \infty]$ over some metric space $(\mathcal{K}, d)$.

A sequential computational algorithm for $\hat{\phi}$ is specified by an initial value $\phi^{(0)} \in \mathcal{K}$, an *algorithmic map* $A : \phi^{(r)} \mapsto \phi^{(r+1)} \in \mathcal{K}$ and a solution set $\hat{\mathcal{K}}^{(r)} = \{\phi \in \mathcal{K} : d(\phi, \phi^{(r)}) < \varepsilon\}$ for some fixed $\varepsilon > 0$. To avoid overshooting the minimum, the map A is typically the composition of a minimizer B and *line search* C. Typically B is given by

$$B(\phi^{(r)}) = \operatorname*{arg\,min}_{\phi \in \mathcal{K}} Q(\phi, \phi^{(r)}), \tag{1.37}$$

where $Q(\phi, y)$ is a quadratic approximation of the difference $L(\phi) - L(y)$

$$Q(\phi, y) = (\phi - y)^\top \nabla L(y) + \tfrac{1}{2}(\phi - y)^\top D(y)(\phi - y), \tag{1.38}$$

for some positive definite matrix $D(y)$. The line search $C(B(\phi^{(r)}))$ is simply the identity map whenever $B(\phi^{(r)})$ satisfies

$$L(B(\phi^{(r)})) < L(\phi^{(r)}) + (1 - \varepsilon)(B(\phi^{(r)}) - \phi^{(r)})^\top \nabla L(\phi^{(r)}). \tag{1.39}$$

Otherwise $C(B(\phi^{(r)}))$ is an element from the segment $\{\phi : \phi = \phi^{(r)} + \rho(B(\phi^{(r)}) - \phi^{(r)}), 0 \le \rho \le 1\}$ such that

$$(1 - \varepsilon)(\phi - \phi^{(r)})^\top \nabla L(\phi^{(r)}) \le L(\phi) - L(\phi^{(r)}) \le \varepsilon(\phi - \phi^{(r)})^\top \nabla L(\phi^{(r)}),$$

for fixed $0 \le \varepsilon \le 1/2$. The following result is a generalization of convergence conditions based on the Fenchel dual (Groeneboom 1996, Lemma 2.1).

1.27 PROPOSITION (cf. Dümbgen et al. 2006, Section 3). *Let $L : \mathbb{R}^m \to (-\infty, \infty]$ be a continuous convex function on the extended real line and $(\mathcal{K}, d)$ be a metric space. Put $\mathcal{K}_0 = \{\phi \in \mathcal{K} : L(\phi) < \infty\}$. Suppose that $\mathcal{K}_0$ is non-empty, L is continuously differentiable on $\mathcal{K}_0$ and the set $\{\phi \in \mathcal{K} : L(\phi) \le c\}$ is compact for each $c \in \mathbb{R}$. Then the set $\hat{\mathcal{K}} = \operatorname{arg\,min}_{\phi \in \mathcal{K}} L(\phi)$ is nonempty and compact. Consider an algorithmic map $A = B \circ C$ satisfying $A(\phi) \in \hat{\mathcal{K}}$ for any $\phi \in \hat{\mathcal{K}}$ and $B(\phi) \in \mathcal{K}$ for each $\phi \in \mathcal{K}_0 \setminus \hat{\mathcal{K}}$. If, for any $y \to \phi$ in $\mathcal{K}$, $\limsup_{y \to \phi} \|B(y) - y\| < \infty$ and*

$$\liminf_{y \to \phi} (B(y) - y)^\top \nabla L(\phi) < 0, \tag{1.40}$$

then the sequence given by $\phi^{(0)} \in \mathcal{K}_0$ and $\phi^{(n)} := A(\phi^{(n-1)})$ for $n = 1, 2, \ldots$, is such that

$$\lim_{n \to \infty} \min_{\hat{\phi} \in \hat{K}} d(\phi^{(n)}, \hat{\phi}) = 0.$$

1.3 INTERVAL-CENSORED DATA

The martingale methods described in Section 1.1 have limited application in constructing estimators from interval-censored data. The current section examines existing methods to address interval censoring. This review is by no means exhaustive; the aim is to provide a broad summary of general strategies with emphasis placed on non- and semiparametric methods.

1.3.1 *Censoring patterns and mechanisms*

Inference from interval-censored data invariably requires some assumptions about the type of interval censoring and the mechanism generating the observation times. "Case 1" or *current status* is the simplest and most extreme form of interval censoring in which each subject is inspected for the occurrence of events at only one point in time. If we determine a subject's survival status at the inspection time U, then all we observe about the survival time T is $(U, 1_{(0,U]}(T))$. Inclusion of an additional inspection time $V > U$ yields "case 2" interval censoring, where we observe $(U, V, 1_{(0,V]}(T), 1_{(U,V]}(T))$. Exact event times can sometimes be observed at random, yielding *partly* or "mixed" interval-censored data. *Doubly-censored* data is related to this observation scheme, but differs in that T is observed exactly whenever $U < T \leq V$. If a censored observation cannot be expressed as an interval of time, the data are generally said to be *coarsened*. We later study a form of partly coarsened data that arises due to different censoring mechanisms acting on each event type.

1.28 EXAMPLE (Progression subject to earlier right-censoring, cf. Bebchuk and Betensky 2001). A progressive illness-death process (Figure 1.2) can be represented by the trivariate counting process $N = (N_{01}, N_{02}, N_{12})$ tracking the number of state transitions over time. Let $T_{hj} = \inf\{t : N_{hj}(t) = 1\}$ be the time of the transition from state h to state j, $S = T_{01} \wedge T_{02}$ denote the time of progression and $T = T_{02} \wedge T_{12}$ be the time of death. Suppose that observation of S is right-censored at C and T is right-censored at D with $C \leq D$. If we observe C and $T \wedge D$ such that $C < T \wedge D$ and $C < S$, then the exact progression time T_{01} is unknown; either $T_{01} \equiv \infty$ (that is, $S = T$) or $T_{01} \in (C, T \wedge D)$. Given a fixed covariate process Z, we assume that the censoring times (C, D) are conditionally independent of (S, T). It is easy to show that this condition is stronger than a natural extension of independent right-censoring (Example 1.1). □

A generalization of case 2 interval-censored data arises when an individual's status is observed at a random number $K \geq 1$ times over some observation period $[\sigma, \tau]$.

Schick and Yu (2000) call this "mixed case" interval censoring. A more common term, particularly under multistate processes, is *panel observation*. The "potential" inspection times can be represented by the triangular array $\mathbf{Y} = \{Y_{k,j} : k = 1, 2, \ldots, j = 1, \ldots, k\}$ of random variables.

1.29 EXAMPLE (Failure under panel observation, Schick and Yu 2000). Let T be a failure time under the panel observation scheme $(K, \mathbf{Y}_K)$. Put $Y_{K,0} \equiv 0$ and $Y_{K,K+1} \equiv \infty$. For $j = 1, \ldots, K+1$, we observe $1_{(Y_{K,j-1}, Y_{K,j}]}(T)$. □

1.30 EXAMPLE (Progression under panel observation, Joly et al. 2002). Consider the illness-death model where transition into the intermediate state corresponds to an irreversible progression in disease status. Let S denote the exit time from the initial state and T be the time of death. Suppose that T is known exactly whenever it occurs before τ, but progression is detected only through inspection times over time period $[\sigma, \tau]$. We then observe $Y = T \wedge \tau$, $1_{(0,\tau]}(T)$ and $1_{(Y_{K,j-1}, Y_{K,j}]}(S)$ for $Y_{K,j} < Y$. □

The potential observation scheme $(K, \mathbf{Y}_K)$ is *ignorable* provided that the joint likelihood with the event $\{K = k, \mathbf{Y}_K = \mathbf{y}_k\}$ is proportional to the likelihood obtained by treating (K, Y_K) as though it were fixed in advance. This is a standard assumption of methods for interval-censored data. Note that the relevant inspection times in panel-observed survival data essentially reduce to case 2 interval censoring. However the relevant observation times $\{Y_{K,j} : 1_{(Y_{K,l-1}, Y_{K,l}]}(T), l = j, j+1\}$ are clearly not ignorable. One obvious way ignorability can be met is to assume that, conditional on any information known at time zero, $(K, \mathbf{Y}_K)$ is independent of the event times.

1.31 DEFINITION (Conditionally independent panel observation). Let T be an event time and Z a fixed covariate process. Suppose that T is under a panel observation scheme $(K, \mathbf{Y}_K)$ satisfying

$$T \perp\!\!\!\perp (K, \mathbf{Y}_K) \mid Z,$$

then $(K, \mathbf{Y}_K)$ is said to be *conditionally independent* of T. □

Conditional independence is a typical assumption made by standard non- and semiparametric methods. Parametric alternatives often invoke a weaker form of ignorable panel observation described in Grüger et al. (1991). It is analogous to independent right-censoring, where we typically allow the censoring time to depend on the observed event history.

1.32 DEFINITION (Noninformative panel observation, Grüger et al. 1991). Let T be an event time and Z a covariate process. Suppose that T is under the panel observation scheme $(K, \mathbf{Y}_K)$. Put $\Delta_{K,j} = 1_{(Y_{K,j-1}, Y_{K,j}]}(T)$, $\Delta_j = (\Delta_{K,1}, \dots, \Delta_{K,j})$ and $Y_j = (Y_{K,1}, \dots, Y_{K,j})$. If

$$\begin{aligned} \mathrm{P}(T \le y_j \mid K = k, Y_{K,1} = y_1, \dots, Y_{K,j} = y_j, \{Z(u) = z(u) : u \le y_j\}, T > y_{k,j-1}) \\ = \mathrm{P}(T \le y_j \mid \{Z(u) = z(u) : u \le y_j\}, T > y_{j-1}), \end{aligned}$$

holds and

$$\mathrm{P}(Y_{K,j} = y_j \mid K = k, Y_{K,1} = y_1, \dots, Y_{K,j} = y_j, \{Z(u) = z(u) : u \le y_j\}, T > y_{k,j-1})$$

is functionally independent of the parameters specifying the distribution of T, then $(K, \mathbf{Y}_K)$ is said to be *noninformative* about T. If Z is time-invariant and the second condition is replaced by

$$\begin{aligned} \mathrm{P}(Y_{K,j} = y_j \mid K = k, Y_{K,1} = y_1, \dots, Y_{K,j} = y_j, Z = z, T > y_{k,j-1}) \\ = \mathrm{P}(Y_{K,j} = y_j \mid K = k, Y_{K,1} = y_1, \dots, Y_{K,j} = y_j, Z = z), \end{aligned}$$

then it is straightforward to show that $(K, \mathbf{Y}_K)$ is conditionally independent of T. □

1.3.2 *Nonparametric estimation*

Computation of the empirical survivor function from current status data dates back to the pooled-adjacent violators algorithm of Ayer et al. (1955). Turnbull (1976) later considered estimation from arbitrarily truncated and interval-censored data. This entailed characterizing the support of the nonparametric maximum likelihood estimator (NPMLE) and iteratively solving a set of "self-consistency" equations. Turnbull attributed the approach to Efron (1965), but it can be more formally motivated by the EM algorithm (Dempster et al. 1977). From the results of Wu (1983), it follows that the self-consistent estimator under case 1 and 2 interval censoring maximizes the observed data likelihood provided that the starting values put positive mass on the candidate support set of the NPMLE (Groeneboom and Wellner 1992).

Groeneboom (1991), Groeneboom and Wellner (1992) and Jongbloed (1998) develop the *iterative convex minorant* (ICM) algorithm for direct computation of the NPMLE via isotonic regression. It can be considered an instance of the general convex optimization routine described in Section 1.2.4. Groeneboom and Wellner (1992) show that the (modified) ICM algorithm computes the NPMLE from current status and case 2 interval-censored data. From Schick and Yu (2000) and van der Vaart and Wellner (2000) this result can be extended to panel-observed survival times, though Dümbgen et al. (2006) construct a general framework encompassing ICM to compute the NPMLE from panel data.

Turnbull's (1976) self-consistency algorithm has been extended to doubly-censored data (De Gruttola and Lagakos 1989), partly interval-censored data (Huang 1999), competing risks (Frydman and Liu 2013; Hudgens et al. 2001), the progressive Markov illness-death model (Frydman 1995; Frydman and Szarek 2009), progressive semi-Markov processes (Griffin and Lagakos 2010; Sternberg and Satten 1999) and the conditional survivor function on a continuous covariate (Dehghan and Duchesne 2011). Likewise extensions of isotonic regression with current status data have been obtained for estimating the subdistribution functions of the competing risks model (Jewell et al. 2003; Maathuis 2006), the sojourn time distribution in a Markov multistate model (Datta et al. 2009) and the state occupancy probabilities of a possibly non-Markov progressive process (Datta and Sundaram 2006).

The NPMLE from interval-censored data converges to the truth at $O_P(n^{1/3})$ (van de Geer 1993; Groeneboom 1991). Groeneboom (1991) proves that the limiting distribution of the NPMLE under case 1 and 2 interval censoring can be derived from the slope of the convex minorant of a two-sided standard Brownian motion process with parabolic drift. Groeneboom et al. (2008) show that Jewell et al.'s (2003) competing risks estimator and the corresponding NPMLE converge to a distribution characterized by a convex minorant, but only the former is based on a drifted Brownian motion process. With the availability of exact observation times, the NPMLE can generally achieve the parametric rate of convergence $\sqrt{n}$. Yu et al. (1998) show that the NPMLE of the survivor function is pointwise asymptotically normal provided that the underlying distribution is continuous and the observation times yielding independent case 2 interval-censoring are discrete. Similar results hold without discrete observation times under doubly-censored (Gu and C.-H. Zhang 1993) and partly case 2 interval-censored (Huang 1999) data. In the absence of exact data, the rate of convergence can be increased via a restricted NPMLE. Dümbgen and Rufibach (2009) show that $O_P(n^{2/5})$ is achieved under a unimodal distribution. Alternatively, if a parametric or semiparametric model of the independent censoring mechanism can be correctly specified, it is possible to obtain a locally $\sqrt{n}$-efficient estimator of the marginal survivor distribution via an extension of Robins and Rotnitzky's (1992) generalized estimating function approach (van der Laan and Hubbard 1997; van der Laan and Robins 1998).

The limiting distribution is useful for constructing bootstrap or analytic confidence intervals and bands. Various nonparametric tests for comparing survivor curves from right-censored data have been extended to independent case 1 and 2 interval censoring. These include the log-rank (J. Sun et al. 2005; Q. Zhao and J. Sun 2004; X. Zhao et al. 2008), survival-based (Fang et al. 2002; Petroni and Wolfe 1994; Yuen et al. 2006) and likelihood ratio (Groeneboom 2012) tests. Another approach is to exploit the structure of the observed data. Andersen and Ronn (1995), J. Sun (1999) and J. Sun and Kalbfleisch (1993) construct specialized tests for current status data. Dümbgen et al. (2006) propose a permutation test from panel observations.

1.3.3 *Semiparametric maximum likelihood*

Imposing some structure through a finite-dimensional parameter θ offers a parsimonious way to quantify sample comparisons and examine covariate effects. Provided that the class of functions containing the infinite-dimensional parameter is sufficiently "small", we can use the results of Section 1.2 to verify asymptotic normality of the maximum likelihood estimator $\hat{\theta}_n$.

Maximum likelihood estimation of the proportional hazards model (Example 1.6) from case 2 interval-censored data was first proposed by Finkelstein (1986). Study of asymptotic properties of the Cox model from case 1 and 2 interval-censored data largely originates from Huang's (1994) PhD dissertation. This work made use of developments in empirical processes (van der Vaart and Wellner 1996) and semiparametric theory (Bickel et al. 1993; van der Vaart 1988). Huang (1996) derived efficient estimators with current status data. Huang and Wellner (1995) and Kim (2003) obtained similar results under case 2 and partly interval-censored data, respectively. Wellner and Y. Zhang (2007) contributed additional large sample theory to obtain a Cox-type model for panel-observed count data. Huang and Wellner's (1995) estimator has been further extended to competing risks (J. Sun and J. S. Shen 2009) and to covariates with measurement error (Wen 2012).

Alternatives to the Cox model have also been examined. Rabinowitz et al. (1995) and Rossini and Tsiatis (1996) consider the accelerated failure time (1.19) and proportional odds models (1.18), respectively. Estimating functions based on current status and case 2 interval-censored data for the more general linear transformation model have been constructed by J. Sun and L. Sun (2005) and Z. Zhang et al. (2005). Lin et al. (1998) and Martinussen and Scheike (2002) estimate Lin and Ying's (1994) restricted Aalen model (Example 1.7) from current status data in which the monitoring process follows a Cox model. This censoring mechanism enables the use of martingale methods in deriving large sample results. Such simplifying assumptions are unfortunately difficult to find elsewhere. Zeng et al. (2006) consider estimation of

Lin and Ying's (1994) model from case 2 interval-censored survival data. They derive asymptotic results using the basic techniques found in Huang and Wellner (1995), with some application of updated profile likelihood theory by Murphy and van der Vaart (2000).

1.3.4 *Alternatives*

Non- and semiparametric maximum likelihood is computationally intensive when the sample size large. The dimension of the parameter space can be tempered with the use of a sieve (Section 1.2.2). X. Shen (2000) derived sieve estimates for the accelerated failure time model from current status data. Huang and Rossini (1997) and Y. Zhang et al. (2010) consider sieves for estimating the proportional odds and hazards models, respectively, from case 2 interval-censored survival data. They show that the sieve estimator can have faster asymptotic convergence, though still slower than the parametric rate. Moreover their empirical results suggest a substantial reduction in computing time relative to SPMLE. Y. Zhang et al. (2010) additionally achieved smaller finite-sample bias by smoothing via a spline-based sieve.

Smooth alternatives to the NPMLE can also be derived from penalized and local likelihood. Cai and Betensky (2003) use penalized maximum likelihood to estimate the proportional hazards model under partly case 2 interval-censored data. This approach was later extended to estimate the intensity measure of Markov and semi-Markov processes (Joly et al. 2002, 1998). Groeneboom et al. (2010) and Murphy et al. (1999) construct penalized likelihood estimators from current status data. Murphy et al. (1999) consider the semiparametric accelerated failure time model (1.19). Groeneboom et al. (2010) study the empirical distribution of a survival time. They show that the smoothed estimators may be biased, but are pointwise asymptotically normal with mean and variance depending on the densities of the event and inspection times. In local likelihood the degree of smoothing is determined by a user-defined bandwidth parameter rather than a penalty term. Betensky et al. (1999) obtain local likelihood estimates of the proportional hazards model using a local EM algorithm. Tolusso and Cook (2009) extend this to Markov multistate processes.

Methods akin to EM based on Cox's (1972) partial likelihood (Example 1.6) can avoid estimation of the infinite-dimensional parameter entirely. Heller (2011) constructs estimating functions weighted by the inverse probability of censoring. Goggins et al. (1998), Satten (1996) and Satten and Sternberg (1999) sample survival time ranks compatible with the observed data. Bebchuk and Betensky (2000) simulate survival times and apply local likelihood methods for right-censored data to obtain smooth estimates of the hazard function. Data augmentation also plays a role in Bayesian methods as the posterior distribution often must be evaluated numerically, even with a parametric prior (Calle and Gómez 2001; Gómez et al. 2004). Han et al. (2013) propose estimation of generalized linear models on the basis of pseudo-observations (Andersen et al. 2003) generated from the empirical distribution function under interval-censored survival data.

The weakly parametric Cox model permits use of closed-form estimating equations for baseline intensities while maintaining much of the flexibility of a semiparametric estimator (Lawless 2003, Section 7.4; Sutradhar and Cook 2008). "Stronger" parametric forms are relatively difficult to motivate with interval-censored data. However if a parametric model can be reasonably justified, it may offer greater insight into the underlying event process. Fully-specified models can also be parameterized in such a way to make maximum likelihood estimation relatively straightforward. The multinomial interpretation of the observed data likelihood, for example, permits the use of generalized linear models in fitting a weakly parametric intensity function (Farrington 1996; Lindsey and Ryan 1998; J. Sun 1997). Kalbfleisch and Lawless (1985) proposed the use of transition intensity models having an analytic solution to the (differential) forward equation (Definition 1.2). Extensions of this approach have been recently studied by Chen and Zhou (2011) and Titman (2011).

1.4 MOTIVATION AND PLAN

Our motivation is drawn from clinical studies in which the endpoint of interest considers what we loosely call "progression": a transition into an irreversible, non-terminal disease state. Progression-related endpoints can offer a pragmatic alternative to overall survival in terms of sample size demands, requisite follow-up time and (perhaps most importantly) relevance to the intervention under study. However evaluation of such outcomes poses a variety of challenges not encountered with overall survival (Fleming et al. 2009). *Time to progression* (TTP), defined as the time from randomization to progression, is a convenient endpoint when death is rare or arises only as a result of progression. Otherwise TTP is right-censored at death before progression (FDA 2007, p. 5). The underlying TTP then measures the time to progression after the subject's death. As this value has no reasonable interpretation the preferred endpoint is *progression-free survival* (PFS), measured from randomization to the earliest of progression and death.

As a *composite endpoint*, standard time-to-event methods are applied to evaluate PFS. Such an approach does not directly measure the individual effects of treatment

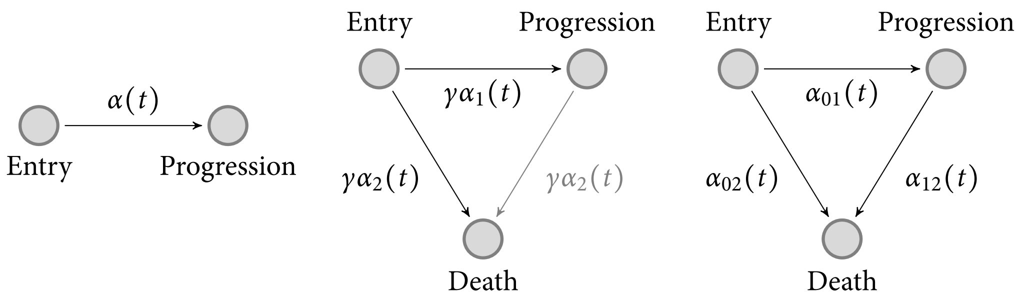


FIGURE 1.2 Time-to-event, semicompeting risks and illness-death models for progression-related endpoints.

on progression or survival. This limitation has been examined in the literature. *Semi-competing risks* partially specify the joint distribution of progression and death via a copula model (e.g. Day et al. 1997; Fine et al. 2001; Peng and Fine 2007; W. Wang 2003). This is largely a hypothetical construct in which the density of progression following death is deemed unobservable. One need not look very far to find alternatives that avoid latent distributions; Frydman (1995) and Xu et al. (2010) endorse the use of the traditional illness-death model. Xu et al. (2010) further show that semicompeting risks model is closely related to a rather restrictive illness-death process (Figure 1.2).

Evaluation of progression-related endpoints is complicated by censoring. When progression is censored by a mechanism separate from the one acting on survival, PFS is subject to multiple right-censoring schemes (Example 1.28). In both TTP and PFS progression is commonly observed through a discrete inspection process, leading to interval censoring (Examples 1.29 and 1.30). Standard practice is to systematically impute event times to the right-endpoint of the censoring interval (e.g. FDA 2007, Appendix 3). Variance estimates are likely too optimistic, unless some attempt is made to reflect error in the imputed values. Potential bias can probably be deemed small if sensitivity analysis gives similar results, but it is unclear how investigators should proceed when a variety of imputation rules offer meaningfully different estimates.

Section 1.3 summarized methods to more formally account for interval-censored data. These offer a variety of ways to deal with interval censored data in TTP, provided that progression is inspected at one or two points in time. Frydman and Szarek (2009) addresses panel-observed progression times in PFS through the construction of a self-consistency algorithm under a Markov illness-death process. The asymptotic properties of this nonparametric estimator are unknown. Dejardin et al. (2010) incorporate covariates in a three-state process related to the illness-death model by setting $\alpha_{02}(t) = 0$ and $\alpha_{01}(t) = \alpha_{12}(t) = \gamma\alpha(t)$.

Bebchuk and Betensky (2001, 2002) examine local likelihood and multiple imputation for handling different censoring schemes in the analysis of progression and survival times. Bebchuk and Betensky (2005) construct comparison tests in a Markov

Cox-type illness-death model where baseline intensities are assumed constant and departures from homogeneity are incorporated with a time-dependent covariate. Yuan et al. (2012) propose Bayesian estimation for covariate effects on the marginal distribution of progression and survival. A shortcoming of existing methods is that they either invoke parametric assumptions not often made in evaluating clinical endpoints or have limited provision for the censoring scheme encountered in practice.

This thesis extends variants of the Cox model (Example 1.6) to censoring patterns typically encountered with TTP and PFS via semiparametric or sieve maximum likelihood. Throughout we make the simplifying assumption that the censoring mechanism is (conditionally) independent and the covariates of interest do not vary with time. Chapter 2 considers estimation from interval-censored time-to-event data (Example 1.29). The resulting methods are appropriate for intent-to-treat analysis of TTP when death rarely ever precedes progression. Chapters 3 and 4 evaluate PFS via the illness-death model. Chapter 3 considers the case where progression and death are observed under different right-censoring schemes (Example 1.28). Nonparametric estimation of the illness-death model from interval-censored progression times has been previously studied by Frydman and Szarek (Example 1.30). In Chapter 4 we propose the extension of this work to account for covariates under a refinement of this observation scheme.

CHAPTER 2

INTERVAL-CENSORED TIME-TO-EVENT DATA

"Mixed case" interval-censoring arises when failure status is assessed a random number of times (Example 1.29). Up to two of these times specify the censoring interval, so case 2 methods constructed on the basis of Grüger et al.'s (1991) noninformative censoring mechanism (Definition 1.32) easily permit mixed case interval-censored data. These are primarily limited to weakly parametric models (Lawless 2003, Section 7.4). Case 2 non- and semiparametric methods derived under (conditionally) independent censoring (Definition 1.31) could likely be extended through some adjustment to the derivation of asymptotic properties. Schick and Yu (2000) and van der Vaart and Wellner (2000) have already proved consistency for the nonparametric maximum likelihood estimator. A consistent NPMLE allows for use of Han et al.'s (2013) jacknife pseudo-observations to fit a variety of semiparametric transformation models (Section 1.1.4) from mixed case interval-censored data, but their approach requires a completely independent censoring mechanism.

Semiparametric methods developed specifically to address mixed case interval-censoring are primarily limited to the Cox model. Wen (2012) derives an estimator for the proportional hazards model under both mixed case interval-censoring and covariate error. Cox-type models for panel count data have been examined under various censoring mechanisms and estimation schemes (e.g. Hu et al. 2003; J. Sun et al. 2007; J. Sun and Wei 2000; Wellner and Y. Zhang 2007). Although failure is essentially a one-jump counting process, these methods are constructed on the basis of a Poisson assumption and thus suited for recurrent events, rather than failure times.

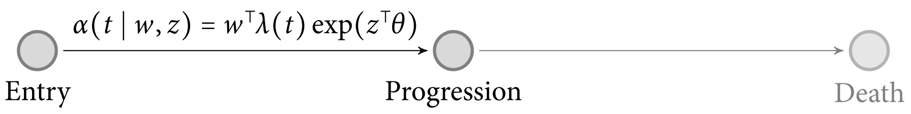

$\alpha(t \mid w, z) = w^\top \lambda(t) \exp(z^\top \theta)$
Entry
Progression
Death

FIGURE 2.1 Time to progression in a chain-of-events model.

The Aalen model and its variants (Examples 1.7 and 1.8) offer a flexible alternative to the Cox model. Their estimation schemes currently address only right-censored data (Aalen 1989; Huffer and McKeague 1991; Scheike and M.-J. Zhang 2002). This chapter considers maximum likelihood estimation for the Cox-Aalen model under mixed case interval-censored data and fixed covariates. The resulting estimator is relevant to the analysis of time to progression, particularly when death rarely ever precedes this event (Figure 2.1).

2.1 MODEL AND OBSERVATION SCHEME

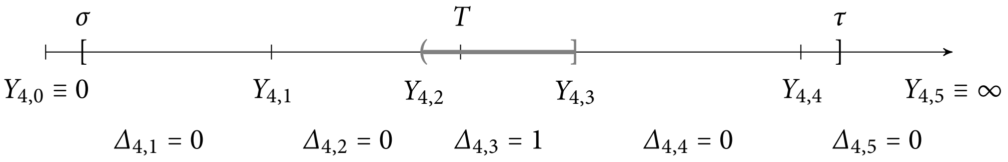
Consider a failure time T with hazard function

$$\alpha(t \mid W, Z) = w^\top \lambda(t) \exp(Z^\top \theta), \tag{2.1}$$

where $W = (W_1 \equiv 1, W_2, \ldots, W_{d_w})^\top$ and $Z = (Z_1, \ldots, Z_{d_z})^\top$ are vectors of fixed covariates, θ is a d_z-vector of regression coefficients and $\Lambda = \int \lambda$ is a d_w-variate cumulative regression function such that the hazard $W^\top \Lambda$ is almost surely nondecreasing. This corresponds to a restricted Cox-Aalen model; in its most general form both W and Z can depend arbitrarily on time (Example 1.8).

Instead of observing T exactly, suppose that we can only assess the event status up to K times, at $Y_K = (Y_{K,1}, \ldots, Y_{K,K})$ with $K \geq 1$ and $0 \equiv Y_{K,0} < Y_{K,1} < \cdots < Y_{K,K} < Y_{K,K+1} \equiv \infty$, but (K, Y_K) is otherwise random. Let $\mathbf{Y}$ denote the triangular array of "potential" inspection times $\{Y_{k,j} : j = 1, \ldots, k, k = 1, 2, \ldots\}$ and $\Delta_{K,j} = 1_{(Y_{K,j-1}, Y_{K,j}]}(T)$ denote the failure status at $Y_{K,j}$, $j = 1, 2, \ldots, K+1$ (Figure 2.2).

FIGURE 2.2 Intermittent inspection of an event time.



From Definition 1.2 and Theorem 1.3, the status vector $\Delta_K = (\Delta_{K,1}, \ldots, \Delta_{K,K+1})$ given (K, Y_K, W, Z) follows a multinomial distribution with one "trial" and "cell" probabilities

$$1 - \exp(-W^\top \Lambda(Y_{K,1}) e^{Z^\top \theta}), \exp(-W^\top \Lambda(Y_{K,1}) e^{Z^\top \theta}) - \exp(-W^\top \Lambda(Y_{K,2}) e^{Z^\top \theta}), \ldots,$$
$$\exp(-W^\top \Lambda(Y_{K,K-1}) e^{Z^\top \theta}) - \exp(-W^\top \Lambda(Y_{K,K}) e^{Z^\top \theta}), \exp(-W^\top \Lambda(Y_{K,K}) e^{Z^\top \theta}).$$

Let $x = (\delta_k, y_k, k, w, z)$ denote a realization of the observation $X = (\Delta_K, Y_K, K, W, Z)$. Assume that:

A1 The event time T is conditionally independent of $(K, \mathbf{Y})$ given (W, Z).

Then the density of $X = x$ with respect to a dominating measure ν determined by the distribution of $(K, \mathbf{Y}, W, Z)$ is

$$p_{\theta,\Lambda}(x) = \left\{1 - \exp(-w^\top \Lambda(y_{k,1}) e^{z^\top \theta})\right\}^{\delta_{k,1}}$$

$$
\begin{aligned}
&\quad\times \prod_{j=2}^{k}\left\{\exp(-w^{\top}\Lambda(y_{k,j-1})e^{z^{\top}\theta}) - \exp(-w^{\top}\Lambda(y_{k,j})e^{z^{\top}\theta})\right\}^{\delta_{k,j}} \\
&\quad\times \left\{\exp(-w^{\top}\Lambda(y_{k,k})e^{z^{\top}\theta})\right\}^{\delta_{k,k+1}} \qquad (2.2) \\
&= \delta_{k,1}\left\{1 - \exp(-w^{\top}\Lambda(y_{k,1})e^{z^{\top}\theta})\right\} \\
&\quad+ \sum_{j=2}^{k+1}\delta_{k,j}\left\{\exp(-w^{\top}\Lambda(y_{k,j-1})e^{z^{\top}\theta}) - \exp(-w^{\top}\Lambda(y_{k,j})e^{z^{\top}\theta})\right\}, \\
&\quad+ \delta_{k,k+1}\exp(-w^{\top}\Lambda(y_{k,k})e^{z^{\top}\theta}).
\end{aligned}
$$

Let $X_i = (\Delta^i_{K_i}, Y^i_{K_i}, K_i, W_i, Z_i)$, $i = 1, \dots, n$, be n iid observations of X from (θ_0, Λ_0), where $Y^i_{K_i} = (Y^i_{K_i,1}, \dots, Y^i_{K_i,K_i})$ and $\Delta^i_{K_i} = (\Delta^i_{K_i,1}, \dots, \Delta^i_{K_i,K_i+1})$. The corresponding log-likelihood function is

$$
\begin{aligned}
\log \operatorname{lik}_n(\theta, \Lambda) = \sum_{i=1}^{n}&\Delta^i_{K_i,1}\log\left\{1 - \exp(-W_i^{\top}\Lambda(Y^i_{K_i,1})e^{Z_i^{\top}\theta})\right\} \\
&+ \sum_{j=2}^{K_i}\Delta^i_{K_i,j}\log\left\{\exp(-W_i^{\top}\Lambda(Y^i_{K_i,j-1})e^{Z_i^{\top}\theta}) - \exp(-W_i^{\top}\Lambda(Y^i_{K_i,j})e^{Z_i^{\top}\theta})\right\} \\
&-\Delta^i_{K_i,K_i+1}W_i^{\top}\Lambda(Y^i_{K_i,K_i})e^{Z_i^{\top}\theta}. \qquad (2.3)
\end{aligned}
$$

2.1 REMARK. The expression in (2.3) reduces to same likelihood function obtained under the noninformative censoring mechanism from Definition 1.32 (Lawless 2003, p. 65). The stronger requirement in A1 simplifies the derivation of asymptotic properties. It may be motivated by the setting in which individuals are assessed according to a predetermined schedule, with the completion and exact timing of assessments determined by some random process related to T only via (W, Z). ▫

2.2 MAXIMUM LIKELIHOOD ESTIMATION

The regression model (2.1) is a valid intensity function provided that $W^{\top}\Lambda$ is almost surely nondecreasing. Estimating equations derived from the likelihood process (1.4) often dispense with the requirement entirely, but (1.4) applies only to filters. A discrete inspection process necessitates constrained maximization of the likelihood. To address this complication, consider two simplifying assumptions:

A2 The support of F_W, $\mathcal{W} \equiv \operatorname{supp}(F_W)$, is a bounded subset of $\mathbb{R}^{d_w}$. In particular, there exists some known $w_0, w_1 \in \mathcal{W}$ such that $\mathrm{P}(w_0 \le W \le w_1) = 1$.

A3 $F_{W_2} \times \cdots \times F_{W_{d_w}} \ll F_W$ and, for every $w \in \mathcal{W}$, we have $\mathrm{P}(T > \tau \mid W = w) > 0$.

2.2 REMARK. Condition A3 essentially implies that we need $w^{\top}\Lambda$ nondecreasing for every $w \in \mathcal{W}$. Condition A2 allows us to ensure monotonicity only in $\mathbf{w}^{\top}\Lambda$, where $\mathbf{w}$ is a matrix whose entries are determined from the values in w_0 and w_1. The choice of w_0 and w_1 is relatively straightforward by standardizing any continuous covariates. In general W is appropriately scaled so that the first entry in Λ, Λ_1, is a baseline cumulative hazard function. □

Not every inspection time contributes information to the likelihood function. As in Groeneboom and Wellner (1992, Part II, Definition 1.1) irrelevant inspections can be discarded to obtain a "thinned" set of observation times.

2.3 DEFINITION. Let $Y_{(1)}, \dots, Y_{(m)}$ be the order statistics of

$$\Upsilon = \{Y^i_{K_i,j}, j = 1, \dots, K_i, i = 1, \dots, n : \Delta^i_{K_i,j} + \Delta^i_{K_i,j+1} = 1\}$$

and $(W_{(i)}, \Delta_{(i)})$ denote the $(W, \Delta_{K,j})$ corresponding to the ith order statistic $Y_{(i)}$. □

If $\Delta_{(1)} = 0$, then the Λ maximizing (2.3) should satisfy $\Lambda(Y_{(1)}) = 0$. If $\Delta_{(m)} = 1$, then the maximizing Λ satisfies $w^{\top}\Lambda(Y_{(m)}) = \infty$ for every $w \in \mathcal{W}$ or, in other words, $\Lambda_1(Y_{(m)}) = \infty$. When combined with the remaining observations in Definition 2.3, these cases contribute nothing to the likelihood. So without loss of generality assume that $\Delta_{(1)} = 1$ and $\Delta_{(m)} = 0$.

Let Θ and H denote the set all possible θ and Λ, respectively. In particular H is the set of all zero-at-time-zero cadlag functions $\{\Lambda\}$ on $[0, \tau]$ with $\mathbf{w}^{\top}\Lambda$ nondecreasing. Since $\Delta_{(m)} = 0$ we can (again without loss of generality) assume that each $\Lambda \in H$ is uniformly bounded with $0 < \mathbf{w}^{\top}\Lambda(\tau) < \infty$. Under conditions A1 to A3 the *maximum likelihood estimator* $(\hat{\theta}_n, \hat{\Lambda}_n)$ is defined by

$$\log \mathrm{lik}_n(\hat{\theta}_n, \hat{\Lambda}_n) = \max_{\theta \in \Theta, \Lambda \in H} \log \mathrm{lik}_n(\theta, \Lambda).$$

Since the likelihood depends on Λ only through its value at the inspection times, we take $(\hat{\theta}_n, \hat{\Lambda}_n)$ as the *semiparametric maximum likelihood estimator* (SPMLE) that concentrates its distribution function on a subset of Υ. The *maximal* subset can be identified by adapting Turnbull (1976, Lemmas 1 and 2).

2.4 DEFINITION. Let $(L, R]$ denote the *censoring interval* $(Y_{K,j-1}, Y_{K,j}]$ satisfying $\Delta_{K,j} = 1$; that is, $T \in (Y_{K,j-1}, Y_{K,j}] = (L, R]$. From the random sample $X_1, \dots, X_n$ put $\mathcal{L} = \{L_1, \dots, L_n\}$ and $\mathcal{R} = \{R_1, \dots, R_n\}$. Let $\mathcal{I} = \{(s_1, t_1], \dots, (s_d, t_d]\}$ be the *maximal intersections* (Figure 2.3) given by the set of disjoint intervals whose left- and right-endpoints are selected respectively from $\mathcal{L}$ and $\mathcal{R}$ such that

$$(s_j, t_j] \cap (L_i, R_i] = \begin{cases} (s_j, t_j], & \text{or} \\ \varnothing, & \end{cases}$$

for every $j = 1, \dots, d$ and $i = 1, \dots, n$. □

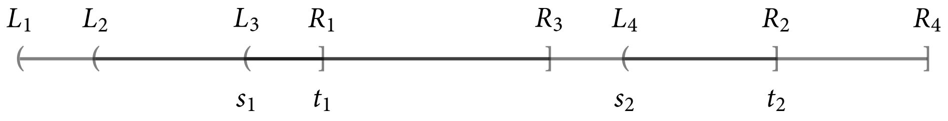


FIGURE 2.3 Maximal intersections.

2.5 PROPOSITION. *$W^\top \hat{\Lambda}_n$ is almost surely constant outside $\mathcal{I}$. Moreover for fixed $\hat{\Lambda}_n$ on the boundary of $\mathcal{I}$, the likelihood is invariant to the behaviour of $\hat{\Lambda}_n$ on the interior of $\mathcal{I}$.*

Proof (cf. Alioum and Commenges 1996, Lemmas 1 and 2). Fix some $(s_{j-1}, t_{j-1}]$ and $(s_j, t_j]$ in $\mathcal{I}$. Consider $\bar{\Lambda}, \tilde{\Lambda} \in H$ with $\bar{\Lambda}$ is constant outside $\mathcal{I}$ and $\bar{\Lambda} = \tilde{\Lambda}$ except on $(s_{j-1}, t_j]$. In particular suppose that $W^\top \tilde{\Lambda}$ almost surely increases on $(t_{j-1}, s_j]$. Then there is some $u_j \in (t_{j-1}, s_j]$ such that $u_j > r \in \mathcal{R}$, $u_j < l \in \mathcal{L}$ and one of the following hold almost surely

$$W^\top \tilde{\Lambda}(t_{j-1}) < W^\top \tilde{\Lambda}(u_j) = W^\top \bar{\Lambda}(t_{j-1})$$
$$W^\top \bar{\Lambda}(s_j) = W^\top \tilde{\Lambda}(u_j) < W^\top \tilde{\Lambda}(s_j).$$

This implies that $\text{lik}_n(\theta, \bar{\Lambda}) > \text{lik}_n(\theta, \tilde{\Lambda})$. The last statement follows from the fact that $\text{lik}_n(\theta, \Lambda)$ depends on Λ only through its value at the inspection times. ■

The maximal set to which $\hat{\Lambda}_n$ assigns mass is given by $\mathcal{T} = \Upsilon \cap \mathcal{I} = \{t_1, \dots, t_d\}$. Estimation reduces to a finite-dimensional optimization problem with objective function (2.3) continuous in the set of feasible solutions. The SPMLE $(\hat{\theta}_n, \hat{\Lambda}_n)$ therefore exists. Uniqueness is established in the following result.

2.6 PROPOSITION. *Let H_0 be the set of all possible Λ satisfying $\text{lik}_n(\theta, \Lambda) > 0$ for every θ. Then $\text{lik}_n(\theta, \Lambda)$ is log-concave in $\Lambda \in H_0$ for each θ. Moreover for fixed $\Lambda \in H_0$, $\text{lik}_n(\theta, \Lambda)$ is log-concave in θ.*

Proof. The function $g(\theta) = p_{\theta,\Lambda}(x)$ satisfies $g(\theta)g''(\theta) \le g'(\theta)^2$ for each x and any fixed $\Lambda \in H_0$. This inequality is strict unless $Z_i = 0$, $i = 1, \dots, n$. Let $e_1, \dots, e_{d_w}$ be the unit vectors in $\mathbb{R}^{d_w}$. For the jth component in $\Lambda \in H_0$, $j = 1, \ldots, d_w$, consider the path $\Lambda + s_j\varphi \in H$ with $s_j = se_j$, s sufficiently small and φ some arbitrary function. It is straightforward to show that the second partial derivative of $\log p_{\theta,\Lambda}(x)$ with respect to s_j is bounded above by zero if x corresponds to an interval- or right-censored observation. Moreover if x is left-censored, the second derivative is strictly negative. Since $\Delta_{(1)} = 0$ the log-likelihood is concave in each component of Λ, holding all remaining entries and θ fixed. ▪

2.3 ASYMPTOTIC PROPERTIES

This section shows that, under some additional regularity conditions, the maximum likelihood estimator $(\hat{\theta}_n, \hat{\Lambda}_n)$ is globally $n^{1/3}$-consistent and $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). This exercise largely amounts to adapting derivations from Huang and Wellner (1995), Murphy and van der Vaart (1997, Section A.3), van der Vaart and Wellner (2000, Section 5) and Wellner and Y. Zhang (2007). The limiting distribution of $n^{1/3}(\hat{\Lambda}_n - \Lambda_0)$ remains an open problem.

2.3.1 *Consistency*

Surely we cannot make any statements about the consistency of $\hat{\Lambda}_n$ outside the support of the inspection times, $\mathrm{supp}(F_Y)$. One method to address this limitation is to simply assume that $\mathrm{supp}(F_Y) = (0, \tau]$. An alternative is to consider convergence in measure. Following van der Vaart and Wellner (2000) and Wellner and Y. Zhang (2007), define for any $B \in \mathcal{B}(0, \tau]$ and $C \in \mathcal{B}(\mathbb{R}^{d_w + d_z})$

$$\begin{aligned}\mu(B \times C) = \int_C \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid W = w, Z = z)& \\ \times \sum_{j=1}^{k} \mathrm{P}(Y_{k,j} \in B \mid W = w, Z = z)\,\mathrm{d}F_{W,Z}(w,z)&, \\ \tilde{\mu}(B \times C) = \int_C \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid W = w, Z = z)& \\ \times \frac{1}{k}\sum_{j=1}^{k} \mathrm{P}(Y_{k,j} \in B \mid W = w, Z = z)\,\mathrm{d}F_{W,Z}(w,z)&,\end{aligned}$$

and $\mu_y(B) = \mu(B \times \mathbb{R}^{d_w + d_z})$.

A4 θ_0 lies in the interior of Θ and Θ is a compact subset of $\mathbb{R}^{d_z}$.

A5 There is $0 < \sigma < \tau$ and $0 < M < \infty$ such that $1/M < \mathbf{w}^{\top}\Lambda_0(\sigma-) < \mathbf{w}^{\top}\Lambda_0(\tau) < M$.

A6 $\mathrm{E}(K) < \infty$.

A7 The support of F_Z, $\mathcal{Z}$, is a bounded subset of $\mathbb{R}^{d_z}$.

A8 $\mu_y \times F_W \times F_Z \ll \mu$.

A9 $\mathrm{P}(Z^\top a \neq c) > 0$ for all $a \in \mathbb{R}^{d_z}$ with $a \neq 0$ and $c \in \mathbb{R}$. Similarly, if $d_w > 1$, then $\mathrm{P}(W^\top a \neq c) > 0$ for all $a \in \mathbb{R}^{d_w}$, $a \neq 0$, and $c \in \mathbb{R}$.

2.7 THEOREM. *Under the conditions listed previously, $\hat{\theta}_n \overset{\mathrm{as}}{\to} \theta_0$ and $\hat{\Lambda}_n \to \Lambda_0$, μ_y-a.e.*

Proof. We verify the requirements of Theorem 1.16 with criterion function

$$m_{\theta,\Lambda} = \log \frac{p_{\theta,\Lambda} + p_0}{2}.$$

Under conditions A2, A4, A5 and A7, p_0 is bounded away from zero and $p_{\theta,\Lambda}$ is bounded above by 1. So $m_{\theta,\Lambda}(x)$ is uniformly bounded in (θ, Λ) and x. Since the logarithm is concave and the SPMLE $(\hat{\theta}_n, \hat{\Lambda}_n)$ is the unique maximizer of the log-likelihood function $n\,\mathbb{P}_n \log p_{\theta,\Lambda}$,

$$\begin{aligned}
\mathbb{P}_n\, m_{\hat{\theta}_n,\hat{\Lambda}_n} - \mathbb{P}_n\, m_{\theta_0,\Lambda_0} &= \mathbb{P}_n \log \frac{p_{\theta,\Lambda} + p_0}{2p_0} \\
&\geq \mathbb{P}_n \tfrac{1}{2} \log \frac{p_{\hat{\theta}_n,\hat{\Lambda}_n}}{p_0} + \mathbb{P}_n \tfrac{1}{2} \log \frac{p_0}{p_0} = \tfrac{1}{2}(\mathbb{P}_n \log p_{\hat{\theta}_n,\hat{\Lambda}_n} - \mathbb{P}_n \log p_0) \\
&\geq 0.
\end{aligned}$$

Since $\log x \leq 2(\sqrt{x} - 1)$ for every $x \geq 0$,

$$\begin{aligned}
P\, m_{\theta,\Lambda} - P\, m_{\theta_0,\Lambda_0} &= P \log \frac{p_{\theta,\Lambda} + p_0}{2p_0} \\
&\leq 2 \int \sqrt{\frac{p_{\theta,\Lambda} + p_0}{2p_0}}\, \mathrm{d}P - 2 = 2 \int \sqrt{\tfrac{1}{2}(p_{\theta,\Lambda} + p_0)p_0}\, \mathrm{d}\nu - 2 \\
&\leq -\int \left[\sqrt{\tfrac{1}{2}(p_{\theta,\Lambda} + p_0)} - \sqrt{p_0}\right]^2 \mathrm{d}\nu = -d_{\mathrm{H}}^2(\tfrac{1}{2}(p_{\theta,\Lambda} + p_0), p_0) \\
&\leq -d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) \qquad (2.4) \\
&\leq 0,
\end{aligned}$$

where the last inequality holds with equality only if $p_{\theta,\Lambda}$ and p_0 define the same measure. By Lemma 2.8 below, this is equivalent to $(\theta, \Lambda) = (\theta_0, \Lambda_0)$. It remains to show that $\{m_{\theta,\Lambda} : \theta \in \Theta, \Lambda \in H\}$ is P-Glivenko-Cantelli. Each $\Lambda \in H$ can be written as $\Lambda = \Lambda^+ - \Lambda^-$ with both Λ^+ and Λ^- bounded and monotone on $[\sigma, \tau]$. By Theorem 1.13 we can cover $\{\Lambda^+, \Lambda^-\}$ with $\exp(K/\varepsilon)^{d_w}$ cubes of $L_r(P)$-size ε. Thus

$$N_{[\,]}(\varepsilon, \Theta \times H, L_r(P)) \lesssim (\operatorname{diam}\Theta/\varepsilon)^{d_z} \times \exp(2 d_w K/\varepsilon). \tag{2.5}$$

Let $\Lambda_L \leq \Lambda \leq \Lambda_R$ be a bracket for Λ in H. From conditions A2 and A5 and the Cauchy-Schwarz inequality

$$|m_{\theta,\Lambda_L} - m_{\theta,\Lambda_R}|^2 \lesssim \int_\sigma^\tau (\Lambda_L - \Lambda_R)^2(t)\,\mathrm{d}t. \tag{2.6}$$

Similarly if (θ_L, θ_R) is a bracket for θ in Θ then

$$|m_{\theta_L,\Lambda} - m_{\theta_R,\Lambda}|^2 \lesssim \|\theta_L - \theta_R\|, \tag{2.7}$$

from conditions A4 and A7 and the mean value theorem. So $\{m_{\theta,\Lambda} : \theta \in \Theta, \Lambda \in H\}$ is P-Glivenko-Cantelli with bracketing number proportional to (2.5) in $L_2(P)$. ▪

2.8 LEMMA. *For every $(\theta, \Lambda) \neq (\theta_0, \Lambda_0)$ on (σ, τ), $p_{\theta,\Lambda} \neq p_0$, almost surely.*

Proof. Let $S_{\theta,\Lambda}(y \mid w, z) = \exp(-w^\top \Lambda(y) e^{z^\top \theta})$ denote the survivor function with $S_{\theta,\Lambda}(0 \mid w, z) \equiv 1$, $S_{\theta,\Lambda}(\infty \mid w, z) \equiv 0$ and $F_{\theta,\Lambda} = 1 - S_{\theta,\Lambda}$. Then $p_{\theta,\Lambda} = p_0$ almost surely implies that

$$\begin{aligned} 0 &= \int |p_{\theta,\Lambda} - p_0|\,\mathrm{d}\nu \\ &= \int \sum_{k=1}^\infty \mathrm{P}(K = k \mid W = w, Z = z) \\ &\qquad \times \sum_{j=1}^{k+1} \int |(F_{\theta,\Lambda} - F_0)(y_{k,j-1}, y_{k,j}]|\,\mathrm{d}F_{Y_K|K}(y_k \mid k, w, z)\,\mathrm{d}F_{W,Z}(w, z). \end{aligned}$$

A lower bound for the inner summation is given by condition A6 and the following inequalities (cf. van der Vaart and Wellner 2000, Lemma 4).

$$\begin{aligned} &\sum_{j=1}^{k+1} \int |(F_{\theta,\Lambda} - F_0)(y_{k,j-1}, y_{k,j}]|\,\mathrm{d}F_{Y_K|K}(y_k \mid k) \\ &\qquad \geq \max_{1 \leq j \leq k} \int |F_{\theta,\Lambda}(y_{k,j}) - F_0(y_{k,j})|\,\mathrm{d}F_{Y_{K,j}|K}(y_{k,j} \mid k) \\ &\qquad\qquad \geq \frac{1}{k} \sum_{j=1}^{k} \int |S_{\theta,\Lambda}(y_{k,j}) - S_0(y_{k,j})|\,\mathrm{d}F_{Y_{K,j}|K}(y_{k,j} \mid k). \end{aligned}$$

Thus

$$0 = \int |p_{\theta,\Lambda} - p_0|\,\mathrm{d}\nu \geq \int |S_{\theta,\Lambda} - S_0|\,\mathrm{d}\tilde{\mu}. \tag{2.8}$$

Conditions A2, A6 and A7 ensure that both $\tilde{\mu}$ and μ are finite. Since $\mu \ll \tilde{\mu}$, the dominated convergence theorem gives

$$\int |S_{\theta,\Lambda} - S_0| \, \mathrm{d}\mu = 0.$$

From A5, $W^\top \Lambda_0$ is μ-almost everywhere bounded away from zero. So $e^{z^\top(\theta_0 - \theta)} = w^\top \Lambda(y)/w^\top \Lambda_0(y)$, μ-a.e., and $Z^\top(\theta_0 - \theta)$ is then degenerate given $Y \sim \mu_y$. Under A8 and A9 this implies that $\theta = \theta_0$ and hence $w^\top(\Lambda(y) - \Lambda_0(y)) = 0$, μ-a.e. Appealing to conditions A8 and A9 again yields $\Lambda = \Lambda_0$, μ_y-a.e. ▪

2.3.2 *Rate of convergence*

The global rate of convergence for $(\hat{\theta}_n, \hat{\Lambda}_n)$ follows from Theorem 1.17. The requirements for this result are verified largely by adaptation of Murphy and van der Vaart (1997, Section A.3) and Wellner and Y. Zhang (2007, Section 5). This requires one additional assumption.

A10 For $(Y, W, Z) \sim \mu_{w,z} = \mu/\mu(\mathcal{W} \times \mathcal{Z} \times (0, \tau])$, there exists some $0 < \rho < 1$ such that $a^\top \mathrm{Var}(Z \mid Y, W) a \leq \rho a^\top \mathrm{E}(ZZ^\top \mid Y, W) a$, almost surely, for all $a \in \mathbb{R}^{d_z}$.

2.9 REMARK (cf. Wellner and Y. Zhang 2007, Remark 3.4). The condition can be reasonably justified as follows. The matrix $\mathrm{E}(ZZ^\top)$ is positive definite by A9 and the Markov inequality. Assume that $\mathrm{Var}_{\mu_{w,z}}(Z \mid Y, W)$ is also positive definite. Let $\lambda_{\min}$ denote the smallest eigenvalue of $\mathrm{Var}_{\mu_{w,z}}(Z \mid Y)$ and $\lambda_{\max}$ the largest eigenvalue of $\mathrm{E}_{\mu_{w,z}}(ZZ^\top \mid Y, W)$. Then $0 < \lambda_{\min} \leq \lambda_{\max}$. Suppose that the ratio $\lambda_{\min}/\lambda_{\max}$ is bounded away from zero uniformly in (Y, W). Then A10 holds with ρ equal to this uniform lower bound. ▫

2.10 THEOREM. *Under the above conditions* $\|\hat{\theta}_n - \theta_0\| + \|\hat{\Lambda}_n - \Lambda_0\|_{\mu_y,2} = O_P(n^{-1/3})$, *where*

$$\|\hat{\Lambda}_n - \Lambda_0\|_{\mu_y,2} = \sum_{j=1}^{d_w} \left[\int |\hat{\Lambda}_{n,j} - \Lambda_{0,j}|^2 \, \mathrm{d}\mu_y \right]^{1/2}.$$

is the $L_2(\mu_y)$ *distance between* $\hat{\Lambda}_n$ *and* Λ_0.

Proof. We verify the requirements of Theorem 1.17 with the same criterion function $m_{\theta,\Lambda}$ from the proof of Theorem 2.7. From (2.4) and Lemma 2.11 below, $P(m_{\theta,\Lambda} - m_{\theta_0,\Lambda_0}) \lesssim -\|\theta - \theta_0\|^2 - \|\Lambda - \Lambda_0\|^2_{\mu_y,2}$. This gives (1.20). From (2.6) and (2.7)

$$\int (m_{\theta,\Lambda} - m_{\theta_0,\Lambda_0})^2 \, \mathrm{d}\nu \lesssim \|\theta - \theta_0\| + \|\Lambda - \Lambda_0\|_2,$$

and with (2.5),

$$J_{[\,]}(\delta, \{m_{\theta,\Lambda} : \theta \in \Theta, \Lambda \in H\}, L_2(P)) \lesssim \int_0^\delta \sqrt{1/\varepsilon}\,\mathrm{d}\varepsilon \lesssim \sqrt{\delta}.$$

Since $m_{\theta,\Lambda}$ is uniformly bounded, (1.21) is satisfied for

$$\varphi(\delta_n) = \sqrt{\delta_n}\left(1 + \frac{1}{\delta_n\sqrt{\delta_n n}}\right),$$

by Lemma 1.22. ■

2.11 LEMMA. *Let d_{H} denote the Hellinger distance (Definition 1.20). Under the previous conditions, $d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) \gtrsim \|\theta - \theta_0\|^2 + \|\hat{\Lambda}_n - \Lambda_0\|_{\mu_y,2}^2$.*

Proof. The squared Hellinger distance can be rewritten as

$$d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) = \int \frac{(p_{\theta,\Lambda} - p_0)^2}{(\sqrt{p_{\theta,\Lambda}} + \sqrt{p_0})^2}\,\mathrm{d}\nu.$$

Under A4, A5 and A7, p_0 is bounded away from zero and $p_{\theta,\Lambda}$ is bounded above by 1, so the denominator in the above integrand is uniformly bounded. Thus

$$d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) \gtrsim \int (p_{\theta,\Lambda} - p_0)^2\,\mathrm{d}\nu \gtrsim \int (S_{\theta,\Lambda} - S_0)^2\,\mathrm{d}\mu,$$

where the second inequality up to proportionality follows from (2.8), $\mu \ll \tilde{\mu}$ with $\mu, \tilde{\mu} < \infty$ and the inequality $|p-q|^2 \leq |p-q|$ for every $p, q \in [0,1]$. Let $\theta_t = t\theta + (1-t)\theta_0$ and $\Lambda_t = t\Lambda + (1-t)\Lambda_0$. From the mean value theorem, there is some $t \in (0,1)$ depending on (y, w, z) such that

$$\begin{aligned}(S_{\theta,\Lambda} - S_0)(y \mid w, z) &= \frac{\partial}{\partial t} S_{\theta_t,\Lambda_t}(y \mid w, z)\\ &= \exp(-w^\top \Lambda_t(y) e^{z^\top\theta_t}) t e^{z^\top\theta_t}\\ &\quad \times \{[1 + t(\theta - \theta_0)^\top z] w^\top(\Lambda - \Lambda_0)(y) + (\theta - \theta_0)^\top z w^\top \Lambda_0(y)\}.\end{aligned}$$

For $(Y, W, Z) \sim \mu_{w,z}$, define $g_0(Z) = 1 + t(\theta - \theta_0)^\top Z$, $g_1(Y, W) = W^\top(\Lambda - \Lambda_0)(Y)$ and $g_2(Y, W, Z) = (\theta - \theta_0)^\top Z W^\top \Lambda_0(Y)$. So $(S_{\theta,\Lambda} - S_0)(Y \mid W, Z)$ is equal to $g_0(Z) g_1(Y, W) + g_2(Y, W, Z)$ up to the factor $\exp(-W^\top \Lambda_t(y) e^{Z^\top\theta_t}) t e^{Z^\top\theta_t}$, which is bounded away from zero under A4, A5 and A7. By the Cauchy-Schwarz inequality,

$$\begin{aligned}&[\mathrm{E}_{\mu_{w,z}}(g_1 g_2)]^2\\ &= [\mathrm{E}_{\mu_{w,z}}(\mathrm{E}_{\mu_{w,z}}(g_1 g_2 \mid Y, W))]^2\\ &\leq \mathrm{E}_{\mu_{w,z}}(g_1^2)\, \mathrm{E}_{\mu_{w,z}}\big([W^\top \Lambda_0(Y)]^2 [\mathrm{E}_{\mu_{w,z}}((\theta - \theta_0)^\top Z \mid Y, W)]^2\big)\end{aligned}$$

$$\begin{aligned}
&= \mathrm{E}_{\mu_{w,z}}(g_1^2)\,\mathrm{E}_{\mu_{w,z}}\big([W^\top \Lambda_0(Y)]^2 \\
&\qquad\qquad \times \mathrm{E}_{\mu_{w,z}}((\theta-\theta_0)^\top\{Z-[Z-\mathrm{E}_{\mu_{w,z}}(Z \mid Y, W)]\}^{\otimes 2}(\theta-\theta_0) \mid Y, W)\big) \\
&\le (1-\rho)\,\mathrm{E}_{\mu_{w,z}}(g_1^2)\,\mathrm{E}_{\mu_{w,z}}\big([W^\top\Lambda_0(Y)]^2(\theta-\theta_0)^\top \mathrm{E}_{\mu_{w,z}}(ZZ^\top \mid Y, W)(\theta-\theta_0)\big) \\
&= (1-\rho)\,\mathrm{E}_{\mu_{w,z}}(g_1^2)\,\mathrm{E}_{\mu_{w,z}}(g_2^2).
\end{aligned}$$

where the last inequality appeals to A10 (cf. Wellner and Y. Zhang 2007, pp. 2126–27). Since ρ is bounded away from zero and $g_0(z)$ is uniformly close to 1 for θ close to θ_0, applying Lemma 1.21 gives

$$\int (S_{\theta,\Lambda} - S_0)^2 \,\mathrm{d}\mu \approx \mu(g_0 g_1 + g_2)^2 \gtrsim \mu g_2^2 + \mu g_1^2 \gtrsim \|\theta - \theta_0\| + \|\Lambda - \Lambda_0\|_{\mu_y}^2,$$

where the last inequality up to a constant holds under A2 and A5. ■

2.3.3 *Asymptotic normality*

This section derives the asymptotic distribution of $\hat\theta_n$ using the profile likelihood method described in Section 1.2.3. Having already established consistency of $(\hat\theta_n, \hat\Lambda_n)$ and the rate of convergence for $\hat\Lambda_n$, this task reduces to identifying an approximately least favourable submodel that meets the structural requirements of Theorem 1.23. We begin with the last of our regularity conditions.

A11 Λ_0 is continuously differentiable with bounded derivative λ_0 satisfying $\mathbf{w}^\top\lambda_0 > 0$ on $[\sigma, \tau]$.

A12 There is a constant $y_0 > 0$ such that $\mathrm{P}(T_{K,j} - T_{K,j-1} \ge y_0 : j = 1, \dots, K, Z) = 1$, almost surely.

A13 For $k = 1, 2, \dots$, $j = 2, \dots, k$ the conditional density functions $f_{Y_{k,1}|W,Z}$, $f_{Y_{k,j}|W,Z}$ and $f_{Y_{k,j-1},Y_{k,j}|W,Z}$ exist. Moreover the partial derivatives of the conditional expectations $\mathrm{E}_{K|W,Z}(\sum_{j=1}^K f_{Y_{K,j}|W,Z}(u \mid w, z))$ and $\mathrm{E}_{K|W,Z}(\sum_{j=2}^K f_{Y_{K,j-1},Y_{K,j}|W,Z}(u, v \mid w, z))$ with respect to u and v are uniformly bounded in (w, z).

2.12 REMARK. Conditions A11 to A13 greatly simplify the proof of Theorem 2.13 below, but their requirements have practical implications. A consequence of A12 is that the event times must be strictly interval-censored, prohibiting any exactly-observed times. Condition A13 precludes consideration of any discretely-distributed inspection process, though methods for grouped time-to-event data (e.g. Lawless 2003, Section 7.3) may be better suited in this setting. □

2.13 THEOREM. *Under the above conditions the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). In particular the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $\Sigma = \tilde{I}_0^{-1}$.*

Proof. The likelihood function offers no (convenient) closed-form expression for the efficient score. So we prove the existence of a least favourable submodel satisfying the requirements of Theorem 1.23 and Corollary 1.24 under the assumption that $\theta \in \mathbb{R}$; that is, $d_z = 1$. The case where $d_z > 1$ follows by application of the results here to each of the d_z entries in θ. The score function for θ takes the form

$$\dot{\ell}_{\theta,\Lambda}(x) = ze^{z\theta}\Bigg[\delta_{k,1}\frac{\exp(-w^\top\Lambda(y_{k,1})e^{z\theta})}{1-\exp(-w^\top\Lambda(y_{k,1})e^{z\theta})}w^\top\Lambda(y_{k,1}) - \delta_{k,k+1}w^\top\Lambda(y_{k,k}) \\ + \sum_{j=2}^{k}\delta_{k,j}\frac{w^\top\Lambda(y_{k,j-1})\exp(-w^\top\Lambda(y_{k,j-1})e^{z\theta}) - w^\top\Lambda(y_{k,j})\exp(-w^\top\Lambda(y_{k,j})e^{z\theta})}{\exp(-w^\top\Lambda(y_{k,j-1})e^{z\theta}) - \exp(-w^\top\Lambda(y_{k,j})e^{z\theta})}\Bigg].$$

Perturbing each entry in Λ generates a tangent set with respect to the product space $\{\Lambda_1 \times \cdots \times \Lambda_{d_w}\}$ of which the class of cumulative regression functions H is a subset. Consider a one-dimensional submodel $s \mapsto \Lambda_{s,1}\times\cdots\times\Lambda_{s,d_w}$ satisfying $h_j = \partial/\partial s_{|s=0}\Lambda_{s,j}$, $j = 1, \ldots, d_w$. For now assume that $h = (h_1, \ldots, h_{d_w})$ is chosen so that $\Lambda_s \in H$; that is, $w^\top\Lambda_s$ is a bounded cumulative hazard function on $[0, \tau]$ for each $w \in \mathcal{W}$. Then a score function for Λ is $B_{\theta,\Lambda}h = \sum_{j=1}^{d_w} B_{\theta,\Lambda}h_j$, where

$$B_{\theta,\Lambda}h_j(x) = e^{z\theta}\Big[\delta_{k,1}\frac{\exp(-w^\top\Lambda(y_{k,1})e^{z\theta})}{1-\exp(-w^\top\Lambda(y_{k,1})e^{z\theta})}w_jh_j(y_{k,1}) - \delta_{k,k+1}w_jh_j(y_{k,k}) \\ + \sum_{l=2}^{k}\delta_{k,l}\frac{w_jh_j(y_{k,l-1})\exp(-w^\top\Lambda(y_{k,l-1})e^{z\theta}) - w_jh_j(y_{k,l})\exp(-w^\top\Lambda(y_{k,l})e^{z\theta})}{\exp(-w^\top\Lambda(y_{k,l-1})e^{z\theta}) - \exp(-w^\top\Lambda(y_{k,l})e^{z\theta})}\Big].$$

Following Section 1.2.3 and Huang and Wellner (1995, Section 5), consider the underlying event time T as the unobserved variable in an information loss model so that the the adjoint $B^*_{\theta,\Lambda}$ of the score operator $B_{\theta,\Lambda}$ is given by $B^*_{\theta,\Lambda}g(t) = \mathrm{E}_{\theta,\Lambda}(g(X) \mid T = t)$. Then from (1.28) least favourable direction $h_{\theta,\Lambda}$ satisfies

$$B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}(t) = B^*_{\theta,\Lambda}B_{\theta,\Lambda}h_{\theta,\Lambda}(t). \tag{2.9}$$

By condition A1,

$$B^*_{\theta,\Lambda}B_{\theta,\Lambda}h(y) = \mathrm{E}_{\theta,\Lambda}(B_{\theta,\Lambda}h(X) \mid T = t) \\ = \mathrm{E}_{W,Z}\Big(\sum_{k=1}^{\infty}\mathrm{P}(K = k \mid W, Z)\,\mathrm{E}_{\theta,\Lambda}(B_{\theta,\Lambda}h(X) \mid T = t, W, Z, K = k)\Big).$$

For $u < v$ and $j = 1, \ldots, d_w$, put

$$\begin{aligned} C_{1,j}(u) &= \frac{\exp(-W^\top \Lambda(u) e^{Z\theta})}{1 - \exp(-W^\top \Lambda(u) e^{Z\theta})} W_j e^{Z\theta} \sum_{k=1}^{\infty} P(K = k \mid W, Z) f_{Y_{k,1}|W,Z}(u \mid W, Z), \\ C_{2,j}(u) &= W_j e^{Z\theta} \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid W, Z) f_{Y_{k,k}|W,Z}(u \mid W, Z), \\ C_{3,j}(u, v) &= \frac{\exp(-W^\top \Lambda(u) e^{Z\theta})}{\exp(-W^\top \Lambda(u) e^{Z\theta}) - \exp(-W^\top \Lambda(v) e^{Z\theta})} \\ &\qquad \times W_j e^{Z\theta} \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid W, Z) \sum_{l=2}^{k} f_{Y_{k,l-1}, Y_{k,l}|W,Z}(u, v \mid W, Z), \\ C_{4,j}(u, v) &= \frac{\exp(-W^\top \Lambda(v) e^{Z\theta})}{\exp(-W^\top \Lambda(u) e^{Z\theta}) - \exp(-W^\top \Lambda(v) e^{Z\theta})} \\ &\qquad \times W_j e^{Z\theta} \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid W, Z) \sum_{l=2}^{k} f_{Y_{k,l-1}, Y_{k,l}|W,Z}(u, v \mid W, Z). \end{aligned}$$

Then

$$\begin{aligned} &\mathrm{E}_{\theta,\Lambda}(B_{\theta,\Lambda} h(X) \mid T = t, W, Z, K = k) \\ &= \sum_{j=1}^{d_w} \int_t^\tau h_j(u) C_{1,j}(u)\, \mathrm{d}u - \int_\sigma^t h_j(u) C_{2,j}(u)\, \mathrm{d}u \\ &\qquad - \int_{u=\sigma}^{t} \int_{v=t}^{\tau} [h_j(u) C_{3,j}(u, v) - h_j(v) C_{4,j}(u, v)] 1(v - u \geq y_0)\, \mathrm{d}v\, \mathrm{d}u. \end{aligned}$$

With $D_{i,j} = Z C_{i,j}$,

$$\begin{aligned} &\mathrm{E}_{\theta,\Lambda}(\dot{\ell}_{\theta,\Lambda}(X) \mid T = t, W, Z, K = k) \\ &= \sum_{j=1}^{d_w} \int_t^\tau \Lambda_j(u) D_{1,j}(u)\, \mathrm{d}u - \int_\sigma^t \Lambda_j(u) D_{2,j}(u)\, \mathrm{d}u \\ &\qquad - \int_{u=\sigma}^{t} \int_{v=t}^{\tau} [\Lambda_j(u) D_{3,j}(u, v) - \Lambda_j(v) D_{4,j}(u, v)] 1(v - u \geq y_0)\, \mathrm{d}v\, \mathrm{d}u, \end{aligned}$$

by conditions A1 and A12. Let $c_{i,j}$ and $d_{i,j}$ denote the expectation of $C_{i,j}$ and $D_{i,j}$ with respect to the distribution of (W, Z). By Leibniz's rule,

$$\begin{aligned} q(t) &\equiv \frac{\partial}{\partial t} B^*_{\theta,\Lambda} \dot{\ell}_{\theta,\Lambda}(t) \\ &= \sum_{j=1}^{d_w} -\Lambda_j(t) d_{1,j}(t) - \Lambda_j(t) d_{2,j}(t) \\ &\qquad - \int_t^\tau [\Lambda_j(t) d_{3,j}(t, v) - \Lambda_j(v) d_{4,j}(t, v)] 1(t - u \geq y_0)\, \mathrm{d}v \\ &\qquad + \int_\sigma^t [\Lambda_j(u) d_{3,j}(u, t) - \Lambda_j(t) d_{4,j}(u, t)] 1(u - t \geq y_0)\, \mathrm{d}u, \end{aligned}$$

$$
\begin{aligned}
r(t) &\equiv \frac{\partial}{\partial t} B^*_{\theta,\Lambda} B_{\theta,\Lambda} h(t) \\
&= \sum_{j=1}^{d_w} -h_j(t)c_{1,j}(t) - h_j(t)c_{2,j}(t) \\
&\qquad - \int_t^\tau \left[h_j(t)c_{3,j}(t,v) - h_j(v)c_{4,j}(t,v)\right] 1(t-u \geq y_0)\,\mathrm{d}v \\
&\qquad + \int_\sigma^t \left[h_j(u)c_{3,j}(u,t) - h_j(t)c_{4,j}(u,t)\right] 1(u-t \geq y_0)\,\mathrm{d}u,
\end{aligned}
$$

Let $q_j(t)$ and $r_j(t)$ denote the jth contribution to $q(t)$ and $r(t)$, respectively. It then follows from conditions A2, A6 and A7 that (2.9) is satisfied if $q_j = r_j$ for each $j = 1, \ldots, d_w$. Thus the least favourable direction is $h_{\theta,\Lambda} = h^1_{\theta,\Lambda}, \ldots, h^{d_w}_{\theta,\Lambda}$, where each $h^j_{\theta,\Lambda}$, $j = 1, \ldots, d_w$, satisfies the Fredholm integral equation of the second kind

$$
h^j_{\theta,\Lambda}(t) = g_j(t) + \int K_j(s,t) h^j_{\theta,\Lambda}(s)\,\mathrm{d}s, \tag{2.10}
$$

on $[\sigma, \tau]$, where

$$
\begin{aligned}
g_j(t) &= -q_j(t)/a_j(t), \\
K_j(u,t) &= \left[c_{3,j}(u,t)\,1(t-u \geq y_0) - c_{4,j}(t,u)\,1(u-t \geq y_0)\right]/a_j(t),
\end{aligned}
$$

with

$$
a_j(t) = c_{1,j}(t) + c_{2,j}(t) + \int_t^\tau c_{3,j}(t,v)\,1(v-t \geq y_0)\,\mathrm{d}v + \int_\sigma^t c_{4,j}(u,t)\,1(t-u \geq y_0)\,\mathrm{d}u.
$$

At the truth (θ_0, Λ_0), $g_j = g_{0,j}$ and $K_j = K_{0,j}$ are bounded by A2, A5, A7 and A12. From Fredholm's first theorem (e.g. Kanwal 1997, p. 48), (2.10) at (θ_0, Λ_0) has the μ_y-a.e. unique solution

$$
h^j_0(t) = g_{0,j}(t) + \int \Gamma_{0,j}(u,t) g_{0,j}(u)\,\mathrm{d}u,
$$

where $\Gamma_{0,j}$ is completely determined by $K_{0,j}$ and is identically zero only if $g_{0,j} = 0$. From (1.27) the efficient score for θ at (θ_0, Λ_0) is $\tilde{\ell}_0 = \dot{\ell}_0 - B_{\theta_0,\Lambda_0} h_0$ and the efficient information matrix $\tilde{I}_0 = P_0\, \tilde{\ell}_0 \tilde{\ell}_0^\top$ is positive definite. We now identify a submodel that is indexed by h_0 and satisfies the structural requirements of Theorem 1.23. Extending the arguments of Huang (1996, pp. 563–64) and van der Vaart (1998, p. 411), consider

$$
\Lambda_s(\theta, \Lambda) = \Lambda + (\theta - s)\varphi(\Lambda)(h_0 \circ \Lambda_{0,1}^{-1} \circ \Lambda_1), \tag{2.11}
$$

where $\Lambda_{0,1}$ and Λ_1 are the first components of Λ_0 and Λ, respectively, and φ is a smooth approximation to $1_{(0,M)}(\mathbf{w}^\top y)$ ensuring that $0 \leq \mathbf{w}^\top \Lambda_s(\theta, \Lambda) \leq M$ on $[\sigma, \tau]$

and $\partial/\partial s_{|s=0}\Lambda_s(\theta,\Lambda_0) = h_0$. In particular $\varphi(\Lambda) = 1$ on $[\Lambda_0(\sigma),\Lambda_0(\tau)]$, $\Lambda \mapsto \varphi(\Lambda)$ is Lipschitz and, for every $\Lambda \in H$, $0 \le \mathbf{w}^\top\Lambda\varphi(\Lambda) \lesssim \mathbf{w}^\top\Lambda \wedge (M - \mathbf{w}^\top\Lambda)$ with the last inequality satisfied up to a constant depending only on (θ_0,Λ_0). From condition A5, φ exists. From condition A11, $\Lambda_{0,1}$ is strictly increasing and continuous, so its inverse is well-defined. Moreover with A13, $h_0 \circ \Lambda_{0,1}^{-1}$ is bounded and Lipschitz. Since the composition $h_0 \circ \Lambda_{0,1}^{-1} \circ \Lambda_1$ has the same jump discontinuities as Λ_1 we have, for every $u \le v$ and s sufficiently close to θ,

$$\mathbf{w}^\top(\Lambda_s(\theta,\Lambda)(u) - \Lambda_s(\theta,\Lambda)(v)) \le \mathbf{w}^\top(\Lambda(u) - \Lambda(v))(1 - |\theta - s|c_0),$$

where c_0 is the Lipschitz constant of $\Lambda \mapsto \varphi(\Lambda)h_0 \circ \Lambda_{0,1}^{-1}(\Lambda)$. Thus (2.11) defines an approximately least favourable submodel such that $\Lambda_\theta(\theta,\Lambda) = \Lambda$ and the map $s \mapsto \log\mathrm{lik}(s,\Lambda_s(\theta,\Lambda))(x) \equiv \ell(s,\theta,\Lambda)(x)$ is twice continuously differentiable with $\dot\ell(\theta_0,\theta_0,\Lambda_0) = \tilde\ell_0$. This gives (1.30) and (1.31). The limit in probability (1.32) follows from Theorem 2.7 and the fact that θ and Λ are variation independent. Under Lemma 2.8 the no-bias condition (1.33) can be established verifying (1.35). Fix some x with $\delta_{k,j} = 1$. Then each term on the right-hand side of (1.36) depends on Λ only through one or both of $\Lambda(y_{k,j-1})$ and $\Lambda(y_{k,j})$. Without loss of generality suppose that $1 < j < k$. Following Murphy and van der Vaart (2000, p. 460), ordinary Taylor expansions at the vector $(\Lambda(y_{k,j-1},\Lambda(y_{k,j})^\top$ yield the inequalities

$$\begin{aligned}|p_{\theta_0,\Lambda} - p_0|(x) &\lesssim |\Lambda - \Lambda_0|(y_{k,j-1}) + |\Lambda - \Lambda_0|(y_{k,j}),\\ |\dot\ell(\theta_0,\theta_0,\Lambda) - \dot\ell(\theta_0,\theta_0,\Lambda_0)|(x) &\lesssim |\Lambda - \Lambda_0|(y_{k,j-1}) + |\Lambda - \Lambda_0|(y_{k,j}),\\ |p_{\theta_0,\Lambda} - p_0 - B_0(\Lambda - \Lambda_0)p_0|(x) &\lesssim |\Lambda - \Lambda_0|^2(y_{k,j-1}) + |\Lambda - \Lambda_0|^2(y_{k,j}),\end{aligned}$$

since the first and second derivatives with respect to $(\Lambda(y_{k,j-1},\Lambda(y_{k,j})^\top$ are uniformly bounded under A2, A5 and A7. From (1.36)

$$P_0\dot\ell(\theta_0,\theta_0,\Lambda) \lesssim \|\Lambda - \Lambda_0\|^2_{\mu_y,2}.$$

Theorem 2.10 showed that the right-hand side is $O_P(n^{-2/3})$, which is more than enough to establish (1.35). For the same x, $\dot\ell(s,\theta,\Lambda)(x)$ and $\ddot\ell(s,\theta,\Lambda)(x)$ are Lipschitz in z, $e^{z\theta}$, $w^\top\Lambda(y_{k,j-1})$ and $w^\top\Lambda(y_{k,j})$. By Theorems 1.14 and 1.15, $\ddot\ell(s,\theta,\Lambda)(x)$ and $\dot\ell(s,\theta,\Lambda)(x)$ then form P_0-Glivenko-Cantelli and Donsker classes, respectively, for (θ,Λ) running through $\Theta \times H$. ▪

Theorem 2.13 and Corollary 1.26 give a consistent estimator for the profile information matrix of $\hat\theta_n$.

2.14 COROLLARY. *Let $e_1, \ldots, e_{d_z}$ be the unit vectors in $\mathbb{R}^{d_z}$ and ρ_n be a symmetric d_z-matrix whose entries ρ_{ij}, $i, j = 1, \ldots, d_z$, satisfy $(\sqrt{n}\rho_{ij})^{-1} = O_P(1)$. A consistent estimator for the (ij)th entry of $\tilde{I}_0$ is*

$$\begin{aligned}
&-\frac{1}{n\rho_{ij}^2}[\log \operatorname{plik}_n(\hat{\theta}_n + \rho_{ij}(e_i + e_j)) \log \operatorname{plik}_n(\hat{\theta}_n)] \\
&+ \frac{1}{n\rho_{ii}^2}[\log \operatorname{plik}_n(\hat{\theta}_n + \rho_{ii} e_i) - \log \operatorname{plik}_n(\hat{\theta}_n)] \\
&+ \frac{1}{n\rho_{jj}^2}[\log \operatorname{plik}_n(\hat{\theta}_n + \rho_{jj} e_j) - \log \operatorname{plik}_n(\hat{\theta}_n)], \qquad (2.12)
\end{aligned}$$

for $i, j = 1, \ldots, d_z$.

2.4 COMPUTATION

Use of standard methods to compute $(\hat{\theta}_n, \hat{\Lambda}_n)$ is complicated by the size of the parameter space and constraints on Λ. The latter cannot be eliminated through transformation, but can be expressed as a linear inequality. From Proposition 2.6, $\log \operatorname{lik}_n(\theta, \Lambda)$ is concave, so computation of $\hat{\Lambda}_n$ reduces to quadratic programming (QP). Cheng et al. (2011) recently applied QP to obtain Wellner and Y. Zhang's (2007) semiparametric estimators from panel count data. They proposed jointly updating estimates for θ and Λ using Pan's (1999) extension of the iterative convex minorant algorithm (Jongbloed 1998). The approach proposed here is similar, but the quadratic approximation is based on the relatively flexible Lagrangian framework of Dümbgen et al. (2006).

2.4.1 *Parameter estimates*

Let $\lambda_j = \Lambda(t_j)$, where t_j is the right-endpoint of the jth maximal intersection from Definition 2.4. By A2 the almost-sure constraints $W^\top \lambda_j \geq 0$ and $W^\top \Lambda(t_j) \leq W^\top \Lambda(t_k)$, $j < k$, amount to the inequality $\mathbf{A}\lambda \geq 0$, where $\lambda = (\lambda_1^\top, \ldots, \lambda_d^\top)^\top$ and $\mathbf{A}$ is the block diagonal matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{w} & 0 & 0 & 0 & \cdots & 0 \\ -\mathbf{w} & \mathbf{w} & 0 & 0 & \cdots & 0 \\ 0 & -\mathbf{w} & \mathbf{w} & 0 & \cdots & 0 \\ & & \cdots & & & \\ 0 & 0 & \cdots & 0 & -\mathbf{w} & \mathbf{w} \end{bmatrix},$$

with $\mathbf{w}$ as described in Remark 2.2. In practice the minimum w_0 and maximum w_1 from $\mathbf{w}$ can be drawn from the observed values in the sample.

For brevity put $\phi = (\theta^\top, \lambda^\top)^\top$ and let $\log\mathrm{lik}_n(\phi) \equiv \log\mathrm{lik}_n(\theta, \lambda)$. Following the results of Section 1.2.4, we specify a computational algorithm by an initial value $\phi^{(0)}$, a candidate step $\eta^{(r)} = (\eta_\theta^\top, \eta_\lambda^\top)^\top$, a line search finding $\phi^{(r+1)} \in \mathrm{seg}(\phi^{(r)}, \phi^{(r)} + \eta^{(r)})$ such that $\log\mathrm{lik}_n(\phi^{(r+1)}) \geq \log\mathrm{lik}_n(\phi^{(r)})$, and a stopping rule $d(\phi^{(r)}, \phi^{(r+1)}) < \varepsilon$.

Applying the framework of Dümbgen et al. (2006, Section 3), the candidate step for $\lambda^{(r)}$, $\eta_\lambda^{(r)}$, is based on a quadratic approximation. In particular

$$\begin{aligned}\eta_\lambda^{(r)} &= \operatorname*{arg\,max}_{\eta_\lambda : \mathbf{A}(\eta_\lambda + \lambda^{(r)}) \geq 0} \nabla_\lambda \log\mathrm{lik}_n(\phi^{(r)})^\top \eta_\lambda + \tfrac{1}{2}\eta_\lambda^\top \nabla_\lambda^2 \log\mathrm{lik}_n(\phi^{(r)})\eta_\lambda \\ &\approx \operatorname*{arg\,max}_{\lambda : \mathbf{A}\lambda \geq 0} \log\mathrm{lik}_n(\theta^{(r)}, \lambda) - \log\mathrm{lik}_n(\phi^{(r)}) - \lambda^{(r)}.\end{aligned} \tag{2.13}$$

$\theta^{(r)}$ is updated via the Newton-Raphson step

$$\eta_\theta^{(r)} = -\nabla_\theta^2 \log\mathrm{lik}_n(\phi^{(r)})^{-1} \nabla_\theta \log\mathrm{lik}_n(\phi^{(r)}). \tag{2.14}$$

Following Jongbloed (1998) overshoot is avoided using the step-halving line search, based on a variant of Armijo's (1966) rule. It is given by

$$\phi^{(r+1)} = \phi^{(r)} + \eta^{(r)}/2^j, \tag{2.15}$$

where j is the smallest nonnegative integer satisfying

$$\log\mathrm{lik}_n(\phi^{(r)}) - \log\mathrm{lik}_n(\phi^{(r)} + \eta^{(r)}/2^j) \leq \alpha \nabla_\phi \log\mathrm{lik}_n(\phi^{(r)})^\top \eta^{(r)}/2^j.$$

Here α is a fixed parameter set to some positive value less than the step factor: $0 < \alpha < 1/2$. Its value can affect the number of iterations needed to achieve the stopping rule, but is otherwise inconsequential (Fletcher 1987, p. 30).

2.15 ALGORITHM. Set $r := 0$, $\theta^{(0)} = 0$ and $\lambda_j^{(0)} = (t_j/\tau, 0_{d_w-1}^\top)^\top$. Let $\eta^{(r)}$ be the candidate step with components given by (2.14) and (2.13) and $\phi^{(r+1)}$ be the result of the line search (2.15). If

$$\|\phi^{(r+1)} - \phi^{(r)}\|_\infty \leq \varepsilon, \tag{2.16}$$

for small positive value ε, then stop. Otherwise, put $r := r + 1$. □

Convergence of Algorithm 2.15 to the maximum likelihood estimator follows from Propositions 1.27 and 2.6. Alternative convergence criteria to (2.16) can be based on the characterization of the SPMLE implied by Proposition 1.27:

$$|\nabla_\phi \log\mathrm{lik}_n(\phi^{(r)})^\top \phi^{(r)}| \leq \varepsilon. \tag{2.17}$$

Constrained Newton methods generally require many more iterations than the standard Newton-Raphson algorithm. Computing time is largely determined by processing power and the software used to carry out QP. The C routines available with IBM's (2012) CPLEX Optimization Studio offer a reasonably fast solution.

2.4.2 *Variance estimates*

The variance estimator for $\hat{\theta}_n$ given by (2.12) is based the curvature of the profile log-likelihood. This requires repeated evaluation of the profile log-likelihood

$$\log \operatorname{plik}_n(\theta) = \sup_{\lambda : \mathbf{A}\lambda \geq 0} \log \operatorname{lik}_n(\theta, \lambda),$$

by fixing $\theta^{(r)}$ at θ in Algorithm 2.15. Since we need to approximate the only value of the profile likelihood and not the profile maximizer, the stopping rule (2.16) is replaced by

$$\left| 1 - \frac{\log \operatorname{lik}_n(\theta, \lambda^{(r+1)})}{\log \operatorname{lik}_n(\theta, \lambda^{(r)})} \right| \leq \varepsilon.$$

This can reduce the computation time considerably since the log-likelihood often converges faster than $\lambda^{(r)}$.

The tuning parameter ρ_n in (2.12) determines the values around $\hat{\theta}_n$ used to assess the curvature of the profile log-likelihood. Standard practice calls for a scalar value $\rho_n \approx n^{-1/2}$ with proportionality constant chosen empirically. Some informal experimentation suggests that variance estimates are not highly sensitive to the choice of ρ_n, particularly with larger sample sizes and frequent inspections. This also seems apparent in numerical studies from Zeng et al. (2006). However for the sake of convenience, a data-driven selection method is desirable. Borrowing methods from numerical differentiation we adopt the matrix form of ρ_n and reduce the choice to specifying broad parameters describing the magnitude of θ.

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function. In the finite-difference approximation

$$f'(x) \approx \frac{f(x+\rho) - f(x)}{\rho},$$

it is standard practice to select $\rho \sim \sqrt{\epsilon}\,\operatorname{curv}(x)$, where ϵ is the error in evaluating f and $\operatorname{curv} = \sqrt{f/f''}$ is the "curvature scale" of f. This choice is a minimizer of the truncation error $\rho^3 f''$ in the above first-order approximation, plus the "round-off" error $\epsilon |f(x)/\rho|$ (Press et al. 2007, Section 5.7). When little is known about f'' one can simply set $\rho \sim \sqrt{\epsilon} x$ or, for x close to zero,

$$\rho \sim \sqrt{\epsilon}\,\operatorname{sign}(x) \max(|x|, \operatorname{typ} x),$$

where $\operatorname{typ} x$ is a typical absolute value for x (Dennis and Schnabel 1996, p. 98).

In (2.12) the curvature of the profile log-likelihood is evaluated with a second-order finite difference approximation. The corresponding curvature scale is based on the ratio of the profile log-likelihood and its third derivative, which can be evaluated easily. When the curvature scale too small or too large, the size of θ serves as fallback value. In particular, the (ij)th entry of ρ_n is

$$\rho_{ij} = n^{-1/2}\operatorname{sign}(\operatorname{curv}_{ij}(\hat{\theta}_n)) \times \max\{\min(|\operatorname{curv}_{ij}(\hat{\theta}_n)|, \sup\theta), |\hat{\theta}_{n,i}|, |\hat{\theta}_{n,j}|, \operatorname{typ}\theta\}, \quad (2.18)$$

where

$$\operatorname{curv}_{ij}(\hat{\theta}_n) = \left\{\frac{-2\log \operatorname{plik}_n(\hat{\theta}_n)}{(e_i \vee e_j)^{\top}\nabla_{\theta}^{3}\log\operatorname{plik}_n(\hat{\theta}_n)}\right\}^{1/3},$$

and $e_i \vee e_j$ is the element-wise maximum of the unit vectors e_i and e_j. Here the choice of ρ_n is reduced to setting the fixed scalar parameters typ θ and sup θ representing the typical and maximum magnitudes for the entries in θ, respectively.

2.5 SIMULATION STUDY

This section considers numerical properties of the SPMLE under variants of the "scheduled" observation scheme described in Remark 2.1. These consider the same cumulative intensity function

$$\Lambda(t \mid W, Z) = (t^{3/2} + Wt^{2/3})\exp(\theta_1 Z_1 + \theta_2 Z_2),$$

where $\theta_1 = \log(2)$, $\theta_2 = -\log(2)$, W is uniform on $(0,1)$, Z_1 is standard normal and Z_2 is uniform on $\{0,1\}$. Failure status is inspected on the basis of k "scheduled" visits, evenly spaced on $(0,\tau)$ with $\tau = 2$. "Actual" visit times follow k independent normal distributions centred at the scheduled times with common standard deviation

$$\sigma_k = \frac{\tau}{4(k+1)} = \frac{1}{2(k+1)}$$

and truncated at zero, $\tau = 2$ and the midpoints between consecutive scheduled times. This setup ensured that the actual inspection times were continuously distributed on $[\sigma,\tau]$ with most times occurring close to its scheduled target. Every inspection after the first was missed with probability $p(W,Z)$, where

$$\operatorname{logit}(p(W,Z)) = \beta_0 + \beta_1 Z_2.$$

One thousand Monte Carlo replicates of the sample sizes $n = 100$, 200 and 500 were generated under three scenarios:

- an independent inspection process with $k = 8$, $\beta_0 = \log(1/9)$ and $\beta_1 = 0$;

- an independent inspection process with $k = 4$, $\beta_0 = \log(1/9)$ and $\beta_1 = 0$; and
- a conditionally independent inspection process with $k = 8$, $\beta_0 = \log(1/4)$ and $\beta_1 = \log(4/9)$.

With $\exp(\beta) = (1/9, 1)$ the probability of a missing inspection is $p = 0.1$, irrespective of (W, Z). Under $\exp(\beta) = (1/4, 4/9)$ the probability remains the same for subjects with $Z_2 = 1$. Those having $Z_2 = 0$ are twice as likely to miss a scheduled inspection.

Estimates for each sample were obtained using a C implementation of Algorithm 2.15. This routine draws from Anderson et al.'s (1999) LAPACK library for matrix inversion and IBM's (2012) CPLEX Callable Library to carry out quadratic programming. For the tuning parameters, $\alpha = 1/3$ was used in the line search, $\varepsilon = 10^{-7}$, $\operatorname{typ}\theta = 1$ and $\sup\theta = 10$. This ensured convergence within a reasonable number of iterations over all scenarios and sample sizes.

In addition to the SPMLE we fit the same Cox-Aalen model to right-censored variants of the data (Example 1.8) via Martinussen and Scheike's (2006) `cox.aalen` routine from the `timereg` package. Estimates were based on four different right-censored data sets:

- underlying or "latent" event times right-censored only by τ,
- right-censoring times and midpoints of (finite) censoring intervals (MID),
- right-censoring times and right-endpoints of censoring intervals (END), and
- a variant of END obtained by discarding inspections after two missed visits (TTP).

Note that the last three approaches obtain right-censored data by systematic imputation. The scheme for TTP is achieved by right-censoring times from END by the last inspection before two or more visits missed in succession. This can be considered a loose adaptation of the FDA's (2007) guideline devised to mitigate bias in the analysis of progression-free survival times.

TABLE 2.1 Average censoring rates over 1000 replicates with $n = 500$.

Censoring parameters			Censoring rate		
k	β_0	β_1	Left	Interval	Right
8	$\log(1/9)$	0	0.167	0.622	0.211
4	$\log(1/9)$	0	0.265	0.462	0.273
8	$\log(1/4)$	$\log(4/9)$	0.167	0.619	0.214

The underlying rate of administrative censoring among the samples was roughly 15%, so right censoring in the "latent" data was low. The interval censoring rate in the observed data under $k = 8$ was approximately 60%. With half as many scheduled visits this decreased to just over 45% (Table 2.1).

The simulation results for the SPMLE $\hat{\theta}_n$ summarized by Table 2.3 are compatible with the asymptotic properties derived in Section 2.3. Bias becomes negligible with larger sample size. Monte Carlo sample standard deviations for the SPMLE decrease with increasing n and are reasonably approximated by the standard error estimates. Empirical coverage probabilities of the 95% confidence intervals are close to the nominal level. Estimators from rudimentary imputation performed relatively well in smaller samples, but under fewer inspection times their bias and coverage probabilities degrade with increasing sample size. Midpoint imputation generally outperformed the two alternative imputation schemes based on the right endpoint.

Pointwise means and percentiles for the SPMLE of the cumulative regression functions are depicted in Figure 2.4. Bias and variability in $\hat{\Lambda}_n$ decrease with increasing sample size and number of inspections. These also tend to be smaller closer to the scheduled visit times, where inspections are more frequent. Estimates of the cumulative coefficient Λ_2 vary considerably more than estimates of the baseline regression function Λ_1.

k	n	CPU time (minutes)	Number of iterations	$\nabla_\phi \log \operatorname{lik}_n(\hat{\phi})^\top \hat{\phi} \times 10^5$
4	100	0.13	210	0.79
	200	0.72	201	0.88
	500	6.60	177	1.05
8	100	0.22	356	6.20
	200	1.32	360	2.75
	500	14.02	372	4.61

TABLE 2.2 Average convergence results for Algorithm 2.15. CPU time covers variance estimation.

Although the SPMLE performs better under larger sample sizes having relatively frequent inspections, much more time is needed to converge. Table 2.2 gives the average CPU time needed to carry out Algorithm 2.15 and variance estimation on an AMD Opteron 6200 processor with each core rated at 3GHz. The CPLEX QP solver can be spread over multiple cores, but this feature offered little to no reduction in computing time. The reported averages are based on times achieved with multi-threading disabled. Estimation based on $n = 500$ was over fifty times slower than with $n = 100$. The rate at which computing time increases with n is sharper under more frequent inspection times. Since the number of iterations to convergence remains fairly stable with n, the poor scaling of Algorithm 2.15 is likely due to the computational demands of the QP step. The maximum norm typically reached $\varepsilon = 10^{-7}$ faster than the inner product $\nabla_\phi \log \operatorname{lik}_n(\hat{\phi})^\top \hat{\phi}$. Considering the computation time, the stopping rule based on the maximum norm (2.16) is preferable to the alternative (2.17).

TABLE 2.3
Bias, standard deviation (SD), average standard error estimate (ASE) and empirical coverage probability of 95% confidence intervals (CP) for the SPMLE of θ over 1000 replicates. Results from estimates based on midpoint imputed (MID), right-endpoint imputed (END, TTP) and latent event times are provided for comparison.

k,			θ_1				θ_2			
e^{β_1}	n	Method	Bias	SD	ASE	CP	Bias	SD	ASE	CP
8, 1	100	SPMLE	0.056	0.157	0.160	0.956	-0.035	0.257	0.257	0.952
		MID	0.004	0.139	0.135	0.943	0.012	0.241	0.241	0.950
		END	-0.007	0.138	0.135	0.941	0.034	0.240	0.243	0.952
		TTP	-0.003	0.139	0.136	0.949	0.031	0.242	0.245	0.953
	200	SPMLE	0.031	0.103	0.104	0.944	-0.030	0.175	0.174	0.950
		MID	-0.006	0.095	0.092	0.945	0.004	0.167	0.166	0.944
		END	-0.017	0.094	0.092	0.941	0.020	0.168	0.167	0.944
		TTP	-0.014	0.095	0.093	0.945	0.017	0.170	0.168	0.945
	500	SPMLE	0.013	0.058	0.062	0.959	-0.010	0.106	0.107	0.955
		MID	-0.012	0.054	0.057	0.950	0.012	0.102	0.104	0.950
		END	-0.023	0.054	0.057	0.935	0.024	0.103	0.104	0.950
		TTP	-0.020	0.054	0.057	0.945	0.021	0.104	0.105	0.949
4, 1	100	SPMLE	0.081	0.181	0.186	0.957	-0.056	0.288	0.287	0.957
		MID	-0.024	0.141	0.137	0.929	0.037	0.250	0.251	0.941
		END	-0.059	0.144	0.137	0.901	0.087	0.259	0.256	0.924
		TTP	-0.057	0.143	0.137	0.907	0.085	0.259	0.256	0.927
	200	SPMLE	0.045	0.114	0.119	0.958	-0.044	0.194	0.191	0.952
		MID	-0.035	0.097	0.093	0.916	0.029	0.172	0.173	0.952
		END	-0.073	0.100	0.092	0.832	0.076	0.176	0.174	0.921
		TTP	-0.070	0.100	0.093	0.842	0.073	0.176	0.175	0.921
	500	SPMLE	0.023	0.063	0.070	0.964	-0.021	0.117	0.117	0.942
		MID	-0.037	0.054	0.058	0.898	0.035	0.108	0.108	0.941
		END	-0.075	0.057	0.057	0.718	0.075	0.108	0.108	0.891
		TTP	-0.072	0.056	0.057	0.737	0.072	0.108	0.109	0.900
8, 4/9	100	SPMLE	0.057	0.159	0.161	0.955	-0.038	0.259	0.258	0.951
		MID	0.002	0.139	0.135	0.939	0.019	0.242	0.241	0.945
		END	-0.012	0.138	0.135	0.939	0.068	0.240	0.243	0.945
		TTP	-0.002	0.140	0.138	0.948	0.073	0.244	0.248	0.940
	200	SPMLE	0.032	0.104	0.105	0.947	-0.031	0.175	0.175	0.945
		MID	-0.008	0.096	0.092	0.942	0.012	0.166	0.167	0.939
		END	-0.024	0.096	0.092	0.932	0.055	0.169	0.168	0.927
		TTP	-0.014	0.097	0.094	0.941	0.060	0.172	0.170	0.924
	500	SPMLE	0.013	0.058	0.062	0.958	-0.010	0.107	0.108	0.951
		MID	-0.015	0.054	0.057	0.952	0.020	0.103	0.104	0.947
		END	-0.031	0.054	0.056	0.916	0.059	0.104	0.104	0.903
		TTP	-0.021	0.054	0.058	0.945	0.065	0.106	0.106	0.897
	100	Latent	0.018	0.135	0.133	0.940	0.001	0.236	0.232	0.948
	200		0.010	0.095	0.091	0.940	-0.008	0.163	0.160	0.939
	500		0.004	0.054	0.056	0.957	-0.002	0.101	0.100	0.951

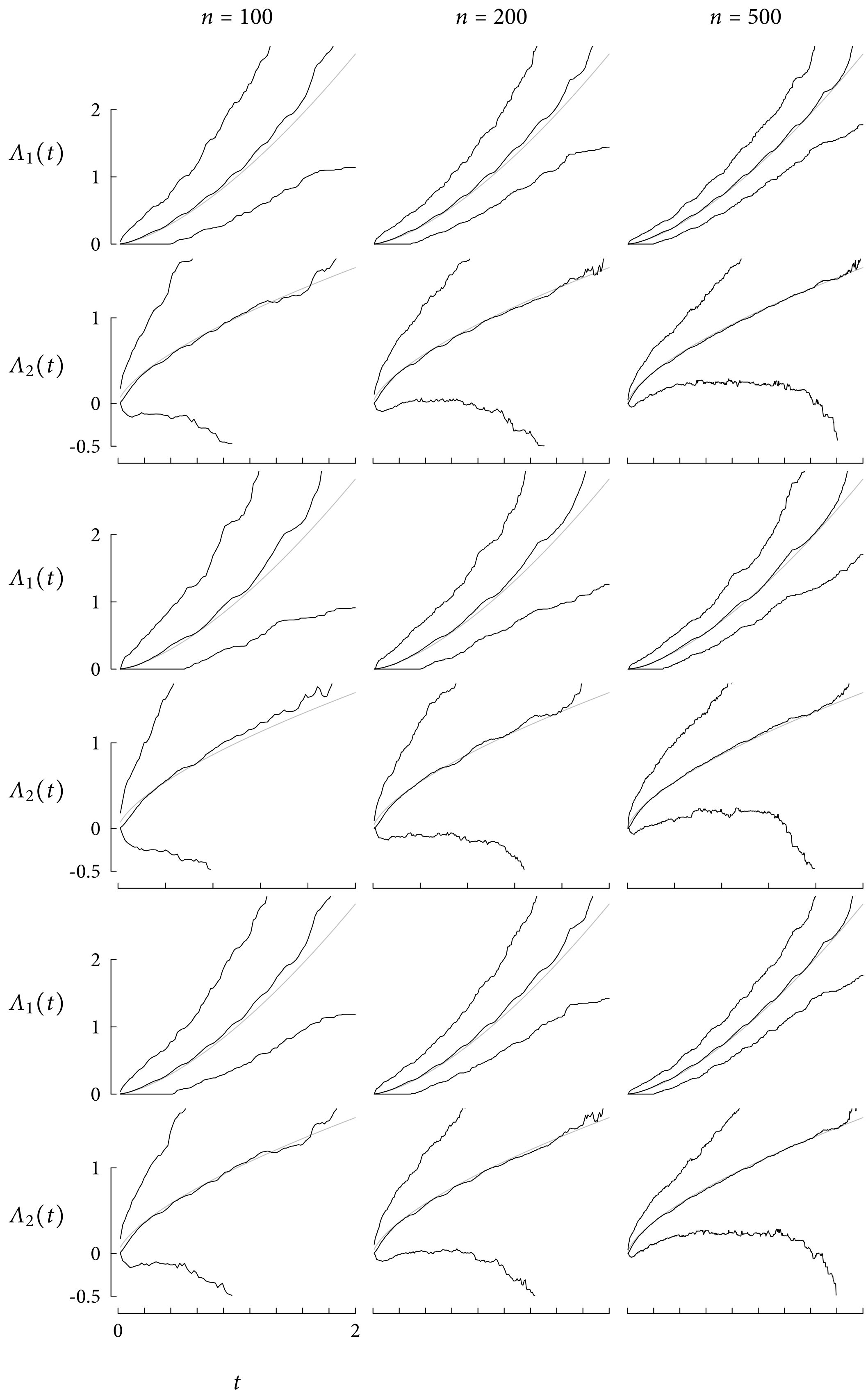


FIGURE 2.4
True values for Λ (–) depicted with pointwise lower and upper 2.5th percentiles and pointwise means for the SPMLE of Λ based on 1000 replicates with $k = 8$ and $e^{\beta} = (1/9, 1)$ (top), $k = 4$ and $e^{\beta} = (1/9, 1)$ (middle), and $k = 8$ and $e^{\beta} = (1/4, 4/9)$ (bottom).

2.6 APPLICATION

Hortobagyi et al. (1996) evaluated a placebo-controlled trial of pamidronate, a nitrogen-containing bisphosphonate, in reducing skeletal complications due to bone lesions. This efficacy analysis considered a sample of 380 women with breast cancer metastatic to bone, followed over a two-year period for the occurrence of skeletal-related events (SREs). These included pathologic fractures, radiation to bone or bone surgery. Radiographic surveys were scheduled at three- to six-month intervals after selected treatment cycles (Figure 2.5). Each survey provided information on the number, size and type of bone lesions. The analysis presented here examines the effect of pamidronate on the time to the first new bone lesion in a subgroup of 321 women assessed for the presence of new lesions at least once during the trial's SRE and survival follow-up period. Observations are right-censored at the last negative inspection preceding death or end of follow-up.

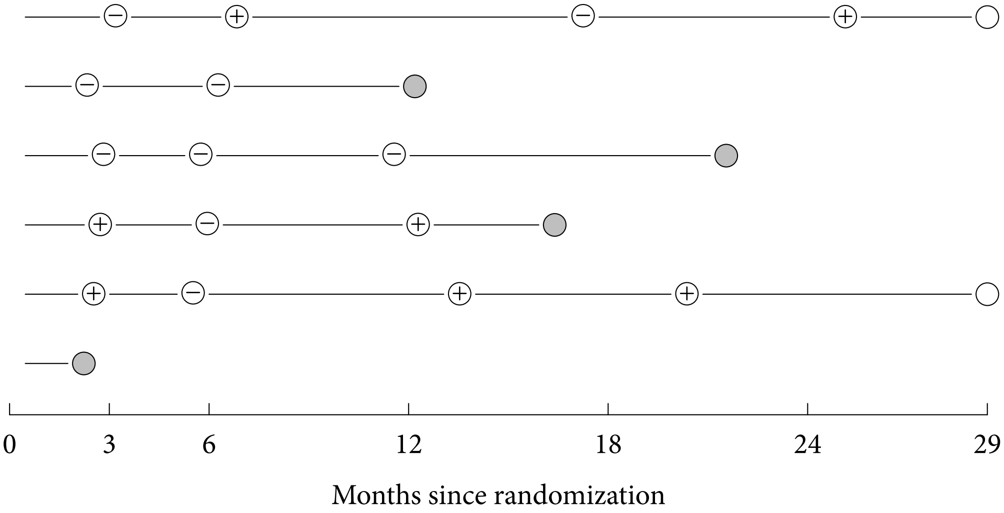


FIGURE 2.5 Periodic inspections for the formation of new bone lesions. Time to the first positive assessment ⊕ is right-censored at the last negative assessment ⊖ before death ● or end of follow-up ○.

An R function called `icsurv` was devised to give a user interface to the C estimation routine described in Section 2.5. Regression models are specified with syntax similar to Therneau's (2012) well-known `survival` package for R. The five relevant observations depicted in Figure 2.5 are represented using the R data frame

```
  left right mid end trt    age
1   90   180 135 180   1 -0.411
2  180    NA 180  NA   0  0.976
3  330    NA 330  NA   0 -0.394
4    0    60  30  60   0 -0.666
```

```
 5    0    60  30  60   0  0.050
```

where the endpoints of the censoring intervals are given by the variables `left` and `right`. For the right-censored observations, the time of right-censoring is stored in the variable `left` and `right` is set to the missing value `NA`. Additional variables measure imputed new lesion times and covariates. Note that the times are measured in days. The frame shown here does not give an excerpt of the actual data; values have been rounded and randomly generated for confidentiality.

The variables `left` and `right` are combined into a response variable using an `interval2`-type `Surv` object. The code fragment below specifies a model with two terms: an indicator of treatment with pamidronate (`trt`) and standardized age at study entry (`age`). Covariates having a multiplicative effect are identified using the identity function `prop`, defined by Martinussen and Scheike's (2006) `timereg` package for R. So the terms `prop(trt)` and `age` correspond respectively to Z and W_2 in our notation above. The "intercept" term here is W_1, which is always equal to 1.

```
 > fit <- icsurv(Surv(left, right, type = 'interval2') ~ prop(trt) + age,
          data = p19, eps = 1e-9, coef.typ = 1/2, coef.max = 2,
          rcsurv = list(Surv(mid, !is.na(right)) ~ .,
                        Surv(end, !is.na(right)) ~ .))
```

Additional models based on right-censored data can specified by the list argument `rcsurv`. These are fit using the `cox.aalen` function from the `timereg` package. Here the Cox-Aalen model is fit to the midpoint-imputed and right-endpoint–imputed first new lesion times stored in the variables `mid` and `end`, respectively. The model predictor (`~ .`) is a shorthand for the terms already specified in `icsurv`'s first argument, so the same model is fit throughout.

The remaining arguments set various tuning parameters. Here ε is given a value of 10^{-9} by `eps = 1e-9`. The typical typ θ and large values sup θ for θ needed by the variance estimation method described in Section 2.4.2 are set with `coef.typ = 1/2` and `coef.max = 2`, respectively. The `icsurv` function returns a class-`icsurv` object; a list of items representing the model fit. Its `print` method reproduces the `icsurv` function call, summarizes the multiplicative regression coefficients, reports the log-likelihood (2.3) at the initial and final parameter values, and gives the rates of left-, interval- and right-censoring in the provided data frame.

From the output below $\log \mathrm{lik}_n(\phi^{(0)}) = -353$ and, after $r = 104$ iterations, the log-likelihood at $\phi^{(r)}$ is -323. Time to the first new lesion is right-censored for the majority (58.6%) of the 291 subjects included in the analysis. Based on the output for the regression coefficient $\hat{\theta}_n = -0.378$, there is moderate evidence in the available data to suggest that a treated individual is less likely to develop a new lesion sooner than someone similar in age who did not receive treatment, with hazard ratio of 0.685 (95% confidence interval 0.486–0.966).

```
> fit
Call:
icsurv(formula = Surv(left, right, type = "interval2") ~ prop(trt) +
    age, data = p19, rcsurv = list(Surv(mid, !is.na(right)) ~ .,
    Surv(end, !is.na(right)) ~ .), ... = list(eps = 1e-09,
    coef.typ = 1/2, coef.max = 2))

            coef se(coef)     z     p exp(coef)  2.5% 97.5%
prop(trt) -0.378    0.176 -2.15 0.031     0.685 0.486 0.966

Based on n = 321
Initial log-likelihood: -352.546
Log-likelihood after 104 iterations: -322.921

               Left Interval Right
Censoring rate 0.19    0.224 0.586

Estimation from imputed data via timereg’s cox.aalen function

Formula:
Surv(mid, !is.na(right)) ~ prop(trt) + age

            coef se(coef)    z     p exp(coef) 2.5% 97.5%
prop(trt) -0.369    0.176 -2.1 0.036     0.691 0.49 0.975

Formula:
Surv(end, !is.na(right)) ~ prop(trt) + age

            coef se(coef)     z     p exp(coef)  2.5% 97.5%
prop(trt) -0.391    0.177 -2.21 0.027     0.676 0.478 0.957
```

The remaining `print` output gives the same coefficient summary for the models fit by `cox.aalen`. The estimates of θ based on midpoint- and right-endpoint–imputed data are -0.369 and -0.391, similar to $\hat{\theta}_n$.

The SPMLE for $\hat{\Lambda}_n$ is stored in a data frame given by the `icsurv` object's list argument `bhaz`. The following code fragment prints the first three rows from the data frames representing estimates for Λ based on the interval-censored and midpoint-imputed data. Note that the time variable is measured in days.

```
> fit$bhaz[1:3, ]
  time intercept         age
1  0.0 0.0000000  0.00000000
2 49.0 0.2296681 -0.07655604
3 54.9 0.2296681 -0.07655604
```

```
> fit$rcfit[[1]]$bhaz[1:3, ]
  time   intercept          age
1  0.0 0.000000000  0.000000000
2 24.5 0.003690588 -0.005289915
3 27.5 0.007390612 -0.004374364
```

Values from these data frames are fully displayed in Figure 2.6. The SPMLE for Λ shows an excess in risk with younger age. Since the limiting distribution of $\hat{\Lambda}_n$ is unknown, we cannot formally test properties of Λ_2 using the observed data. With right-censored new lesion times, the `cox.aalen` routine from the `timereg` package carries out a resampling-based test of significance $\sup_{t \le \tau} |\Lambda_j(t)| = 0$, and a Kolmogorov-Smirnov test of time invariance $\Lambda_j(t) = t\Lambda_j(\tau)/\tau$ (Martinussen and Scheike 2006, p. 258). The following output suggests that, from the midpoint-imputed data, we cannot reject the hypothesis of time invariance $\Lambda_2(t) = t\Lambda_2(\tau)/\tau$ with a p-value of 75%.

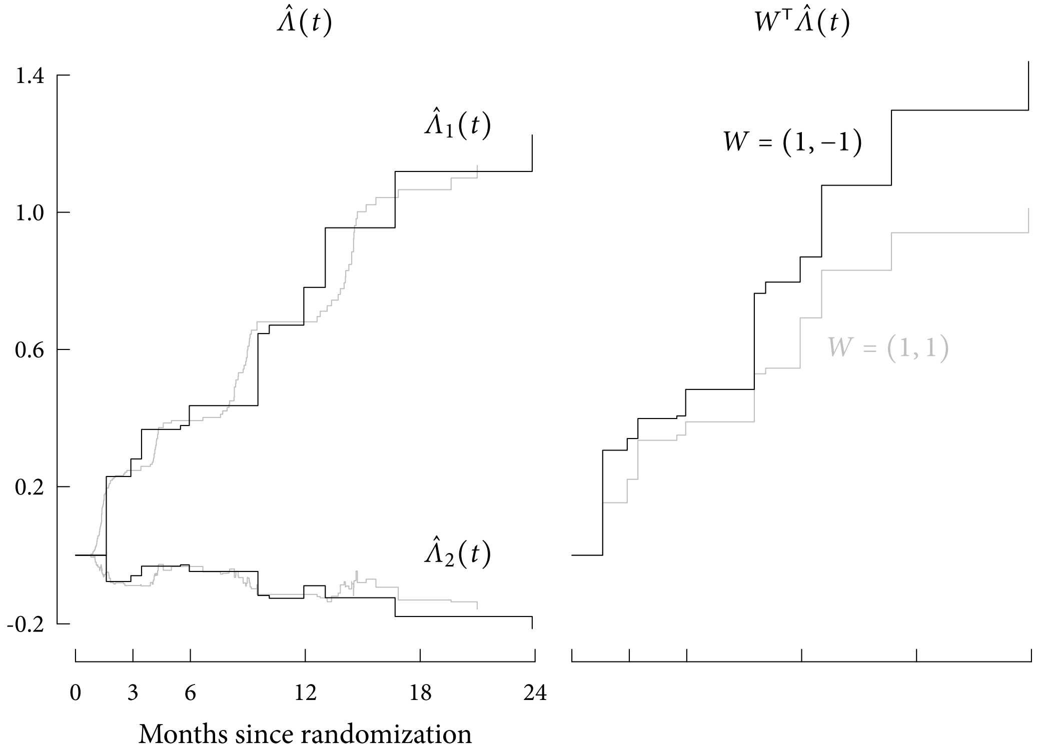


FIGURE 2.6 Left: midpoint-imputation (–) and SPMLE (–) cumulative baseline regression function (Λ_1) and cumulative coefficient of standardized age at study entry (Λ_2). Right: cumulative baseline intensity function for the oldest (–) and the youngest (–) individuals.

```
> rcfit <- cox.aalen(Surv(mid, !is.na(right), type = 'right')
                     ~ prop(trt) + age, data = p19)
> summary(rcfit)
```

```
Cox-Aalen Model

Test for Aalen terms
Test for nonparametric terms

Test for non-significant effects
            Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                          7.62               0.000
age                                  2.76               0.102

Test for time invariant effects
                  Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                        0.2040                       0.006
age                                0.0693                       0.750

Proportional Cox terms :
           Coef.    SE Robust SE D2log(L)^-1     z  P-val
prop(trt) -0.369 0.176     0.178       0.177 -2.08 0.0374
Test for Proportionality
          sup|  hat U(t) | p-value H_0
prop(trt)             4.45        0.466


  Call:
cox.aalen(Surv(mid, !is.na(right), type = "right") ~ prop(trt) +
    age, data = p19)
```

Replacing the icsurv model term age with prop(age) fits a two-covariate Cox model. This gives essentially the same age-adjusted hazard ratio for pamidronate as the Cox-Aalen model, suggesting that the effect of age is adequately described by a fixed parameter.

```
> icsurv(Surv(left, right, type = ’interval2’) ~ prop(trt) + prop(age),
         data = p19, eps = 1e-9, coef.typ = 1/2, coef.max = 2,
         rcsurv = list(Surv(mid, !is.na(right)) ~ .,
                       Surv(end, !is.na(right)) ~ .))
Call:
icsurv(formula = Surv(left, right, type = "interval2") ~ prop(trt) +
    prop(age), data = p19, rcsurv = list(Surv(mid, !is.na(right)) ~ .,
    Surv(end, !is.na(right)) ~ .), ... = list(eps = 1e-09,
    coef.typ = 1/2, coef.max = 2))

            coef se(coef)     z      p exp(coef)  2.5% 97.5%
prop(trt) -0.378   0.1755 -2.15 0.0310     0.685 0.486 0.967
prop(age) -0.223   0.0856 -2.61 0.0092     0.800 0.677 0.946

Based on n = 321
Initial log-likelihood: -352.546
Log-likelihood after 136 iterations: -326.995
```

```
               Left Interval Right
Censoring rate 0.19    0.224 0.586

Estimation from imputed data via timereg’s cox.aalen function

Formula:
Surv(mid, !is.na(right)) ~ prop(trt) + prop(age)

             coef se(coef)      z     p exp(coef)  2.5% 97.5%
prop(trt) -0.372   0.1737 -2.14 0.032     0.689 0.490 0.969
prop(age) -0.217   0.0817 -2.65 0.008     0.805 0.686 0.945

Formula:
Surv(end, !is.na(right)) ~ prop(trt) + prop(age)

             coef se(coef)      z     p exp(coef)  2.5% 97.5%
prop(trt) -0.382    0.174 -2.20 0.028     0.682 0.485 0.960
prop(age) -0.198    0.081 -2.44 0.015     0.821 0.700 0.962
```

The analysis presented here has some limitations. One is related to the fact that the primary endpoint of the trial considered the occurrence of the SREs during the first 12 months of follow-up. So efforts of the trialists to ensure balanced comparisons between the treatment groups may not carry over to the occurrence of bone lesions over a longer time period. Lesions were also not assessed in the 59 patients excluded from this analysis. The number excluded was equally-distributed in between the treatment groups, but the results here may still be subject to selection bias. Right-censoring preceding death is problematic; just over half of the sample analyzed died over the study period, with a median follow-up time of 13 months. The independent censoring assumption under this scenario is thus difficult to justify. This issue addressed in Section 4.5, where the occurrence of both new lesion and death are considered via the illness-death model.

CHAPTER 3

DOUBLY RIGHT-CENSORED DATA FROM AN ILLNESS-DEATH PROCESS

The terminal event in an illness-death process is often subject only to administrative censoring. Progression may be right-censored at an earlier time (Example 1.28), yielding what we call "doubly right-censored" data. Bebchuk and Betensky (2001, 2002) combine local likelihood and multiple imputation to estimate the marginal distribution of the event times. They further construct k-sample tests on the basis of a Markov Cox-type model with constant baseline transition intensities and time-dependent covariates (Bebchuk and Betensky 2005). Yuan et al. (2012) apply Bayesian methods to estimate covariate effect on the marginal distribution of event times within the family of parametric accelerated failure time models (1.19). Ke et al. (2011) devise a generalized Kaplan-Meier estimator for the subdistribution function of the exit time from the initial state.

This chapter considers the estimation of parameters specifying a Markov model with Cox-type transition intensities (Figure 3.1). Using the Fisher scoring algorithm proposed by Kalbfleisch and Lawless (1985), Jackson's (2011) msm package for R readily handles the time-homogeneous case, $\lambda_{hj}(t) = \lambda_{hj}$, and can extend to piecewise exponential intensities by way of time-dependent covariates. A relatively flexible piecewise exponential estimator is developed via sieve maximum likelihood.

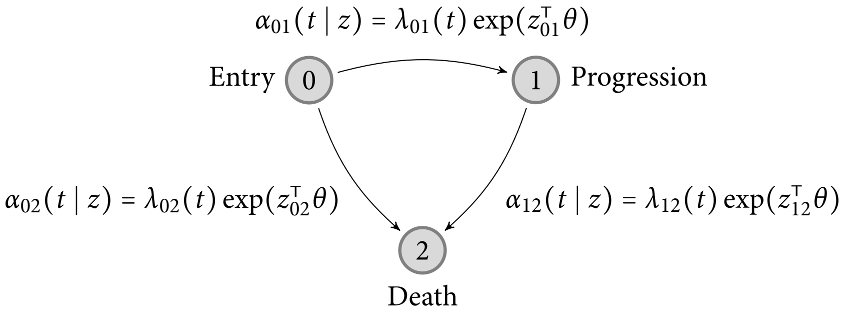

$\alpha_{01}(t \mid z) = \lambda_{01}(t)\exp(z_{01}^{\top}\theta)$
Entry
0
1
Progression
$\alpha_{02}(t \mid z) = \lambda_{02}(t)\exp(z_{02}^{\top}\theta)$
$\alpha_{12}(t \mid z) = \lambda_{12}(t)\exp(z_{12}^{\top}\theta)$
2
Death

FIGURE 3.1 A Markov progressive illness-death model.

3.1 MODEL AND OBSERVATION SCHEME

Consider a trivariate process $N = (N_{01}, N_{02}, N_{12})$ counting the 0 → 1, 0 → 2 and 1 → 2 state transitions in the Markov illness-death model. For states h and j, $h \neq j$, let $T_{hj} =$

$\inf\{t : N_{hj}(t) = 1\}$ denote the $h \to j$ transition time and $Y_h(t) = 1 - \sum_{j \neq h} N_{hj}(t-)$ be the $h \to j$ at-risk process. Assume that each N_{hj} has intensity process $Y_h \alpha_{hj}$ with

$$\alpha_{hj}(t \mid Z) = \lambda_{hj}(t) \exp(Z_{hj}^{\top} \theta), \tag{3.1}$$

where Z_{hj} is a transition-type–specific d_z-vector based on the fixed covariate Z, θ is a d_z-variate regression parameter and $\Lambda_{hj} = \int \lambda_{hj}$ is a nondecreasing baseline intensity function. The parameter θ is common to each of the transition intensities, but Z_{hj} can be suitably constructed from Z to give type-specific covariate effects (Andersen et al. 1993, pp. 478–80).

Over the finite interval $[0, \tau]$, suppose that observation of N is subject to the right-censoring times $0 < C \leq D \leq \tau$; N_{01} is known on $[0, C]$ and $N_{02} + N_{12}$ on $[0, D]$. Let $S = T_{01} \wedge T_{02}$ denote the exit time from the initial state and $T = T_{02} \wedge T_{12}$ the time of death. Put $U = S \wedge C$, $V = T \wedge D$, $\Delta_0 = 1(S < C)$ and $\Delta_2 = 1(T < D)$. Then the transition times (T_{01}, T_{02}, T_{12}) for an observation $X = (U, V, \Delta_0, \Delta_2, Z)$ are available when $\Delta_0 = \Delta_2 = 1$. In general the progression status $1(S < T)$ is known only if S occurs before C; that is, $\Delta_0 = 1$. Otherwise (U, V) is a "potential" censoring interval for T_{01}. Figure 3.2 illustrates this notation for two individuals, i and j. Exact data are available for subject j. The observation for i is doubly right-censored.

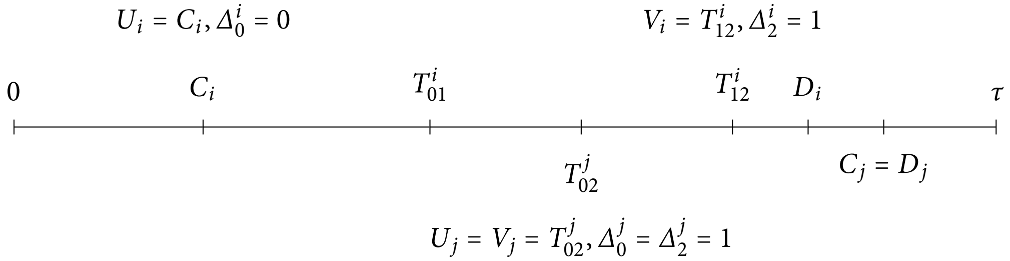

$U_i = C_i, \Delta_0^i = 0$
$V_i = T_{12}^i, \Delta_2^i = 1$
0
C_i
T_{01}^i
T_{12}^i
D_i
τ
T_{02}^j
$C_j = D_j$
$U_j = V_j = T_{02}^j, \Delta_0^j = \Delta_2^j = 1$

FIGURE 3.2 Observation of an illness-death process under double right censoring.

Let $A_{hj} \int \alpha_{hj}$ for $h \neq j$ and $A_{hh} = -\sum_{j \neq h} A_{hj}$. Then from Theorem 1.4 the transition probabilities are

$$P_{hh}(s, t \mid z) = \exp\left\{ \int_s^t A_{hh}(\mathrm{d}y \mid z) \right\}, \tag{3.2}$$

where

$$P_{01}(s, t \mid z) = \int_s^t P_{00}(s, y \mid z) A_{01}(\mathrm{d}y \mid z) P_{11}(y, t \mid z). \tag{3.3}$$

Assume the following basic condition.

B1 (T_{01}, T_{02}, T_{12}) is conditionally independent of (C, D) given Z.

Then the realization $X = x = (u, v, \delta_0, \delta_2, z)$ has density

$$\begin{aligned} p_{\theta,\Lambda}(x) = {} & P_{00}(0, u \mid z)[\alpha_{01}(u \mid z)P_{11}(u, v \mid z)\alpha_{12}(v \mid z)^{\delta_2}]^{\delta_0 1(u<v)} \\ & \times [P_{01}(u, v \mid z)\alpha_{12}(v \mid z)^{\delta_2} + P_{00}(u, v \mid z)\alpha_{02}(v \mid z)^{\delta_2}]^{1-\delta_0} \\ & \times \alpha_{02}(v \mid z)^{\delta_0(1-1(u<v))\delta_2} \end{aligned} \tag{3.4}$$

with respect to a dominating measure ν determined by the distribution of (C, D, Z).

3.1 REMARK. The expression in (3.4) is the same likelihood function obtained under (conditionally) independent censoring by C and D (Example 1.1). Such a mechanism permits dependence on the observed history. The stronger requirement in B1 simplifies the derivation of asymptotic properties. It is plausible under an intent-to-treat analysis in which C represents loss to follow-up for S, Z adequately explains variation in C, and D is an administrative censoring time for T. This precludes censoring individuals at the time of a change in treatment due to toxicity or need for additional therapies—a scenario that likely induces dependent censoring (Fleming et al. 2009). □

3.2 SIEVE MAXIMUM LIKELIHOOD ESTIMATION

Let $X_i = (U_i, V_i, \Delta_0^i, \Delta_2^i, Z_i)$, $i = 1, \ldots, n$, be n iid observations of X from (θ_0, Λ_0), $\Lambda_0 = (\Lambda_{hj}^0)$ for $h \neq j$. Note that $\mathbb{P}_n \log p_{\theta,\Lambda}$ *a priori* maximizes to infinity; with, say, Λ_{02} continuously differentiable $\alpha_{02}(V_i \mid Z_i)$ can be made arbitrarily large and $A_{00}(V_i \mid Z_i)$ close to zero at any $V_i = T_{02}^i < C_i$. The usual way out is to replace α_{hj} by the jump discontinuities ΔA_{hj}. However consider an individual i having unknown progression status ($\Delta_0^i = 1$) but known survival time ($\Delta_2^i = 1$):

$$U_i = C_i < V_i = S_i < D_i.$$

Suppose $(L, V_i]$ is a subinterval of $(U_i, V_i]$ containing no other observation times from the sample. Surely we need $\Lambda_{02}(V_i) + \Lambda_{12}(V_i) - \Lambda_{02}(L) - \Lambda_{12}(L) > 0$, but the available data are insufficient to *jointly* estimate $\Lambda_{02}(V_i) - \Lambda_{02}(L)$ and $\Lambda_{12}(V_i) - \Lambda_{12}(L)$. Thus no unbiased semiparametric estimator of (θ, Λ) exists. This problem is evaded by employing the method of sieves (Section 1.2.2) on the basis of the following assumptions.

B2 There exist $0 < \sigma < \tau$ and $0 < M < \infty$ such that $1/M < \Lambda_{hj}^0(\sigma-) < \Lambda_{hj}^0(\tau) < M$, $h \neq j$, and Λ_0 is continuously differentiable with bounded derivative λ_0 on $[\sigma, \tau]$.

B3 Let n_{hj} denote the number of individuals in the sample with $\{T_{hj} < \infty\}$ observed exactly, $h \neq j$. Then there exist $q_{hj} > 0$ such that $n_{hj}/n \to q_{hj}$ as $n \to \infty$.

Let $H = (H_{hj})$ denote the set of $\Lambda = (\Lambda_{hj})$ with each $\Lambda_{hj} : [0,\tau] \to [0, M]$ cadlag and nondecreasing. Any finite-dimensional approximation to H whose size increases with n is a *sieve*. Throughout consider the piecewise exponential sieve given by the set of piecewise linear interpolants of $\Lambda \in H$.

3.2 DEFINITION. For each $h \to j$, $h \neq j$, let $\mathcal{T}_{hj,n}$ be a set containing the $K_{hj,n} = O(n^\kappa)$, $0 < \kappa < 1$, points in $(0,\tau)$ from the partition

$$0 = t_0 < t_1 < \cdots < t_{K_{hj,n}} < t_{K_{hj,n}+1} = \tau$$

constructed so that every subinterval $[t_{k-1}, t_k)$ contains at least one exact $h \to j$ transition time observed in $X_1, \ldots, X_n$ and $\max_k(t_k - t_{k-1}) = O(n^{-\kappa})$. For every $\Lambda_{hj} \in H_{hj}$ let $\Lambda_{hj,n}$ denote the piecewise linear interpolant

$$\Lambda_{hj,n}(t) = \sum_{t_k \in \mathcal{T}_{hj,n}} I_k(t) \left\{ \left[1 - \frac{L_k(0,t)}{L_k(0,\tau)} \right] \Lambda_{hj}(t_{k-1}) + \left[\frac{L_k(0,t)}{L_k(0,\tau)} \right] \Lambda_{hj}(t_k) \right\}, \tag{3.5}$$

where $I_k(t) = 1_{[t_{k-1}, t_k)}(t)$ and $L_k(s,t)$ is the length of $[t_{k-1}, t_k) \cap [s,t)$. □

Let $A_{hj}(s, t \mid z) = A_{hj}(t \mid z) - A_{hj}(s \mid z)$. Then from B3, Λ_{02} and Λ_{12} are jointly estimable by maximizing

$$\begin{aligned}
&\log \operatorname{lik}_n(\theta, \Lambda) \\
&= \sum_{i=1}^n \log p_{\theta,\Lambda}(X_i) \\
&= -A_{01}(U_i \mid Z_i) - A_{02}(U_i \mid Z_i) \\
&\quad + \Delta_0^i 1(U_i < V_i)[\log \alpha_{01}(U_i \mid Z_i) - A_{12}(U_i, V_i \mid Z_i) + \Delta_2^i \log \alpha_{12}(V_i \mid Z_i)] \\
&\quad + (1 - \Delta_0^i) \log[P_{01}(U_i, V_i \mid Z_i) \alpha_{12}(V_i \mid Z_i)^{\Delta_2^i} + P_{00}(U_i, V_i \mid Z_i) \alpha_{02}(V_i \mid Z_i)^{\Delta_2^i}] \\
&\quad - \Delta_0^i (1 - 1(U_i < V_i)) \Delta_2^i \log \alpha_{02}(V_i \mid Z_i).
\end{aligned} \tag{3.6}$$

over the sieve $H_n = (H_{hj,n})$, $H_{hj,n} = \{\Lambda_{hj,n} : \Lambda_{hj} \in H_{hj}\}$. Let Θ denote the set of all possible θ. Then the piecewise exponential *sieve maximum likelihood estimator* (SMLE) satisfies

$$\log \operatorname{lik}_n(\hat\theta_n, \hat\Lambda_n) = \max_{\theta \in \Theta, \Lambda \in H_n} \log \operatorname{lik}_n(\theta, \Lambda). \tag{3.7}$$

This optimization problem is well-defined and has finite dimension. Its solution is characterized by the score equations

$$\frac{\partial}{\partial \theta} \log \operatorname{lik}_n(\hat\theta_n, \hat\Lambda_n) = 0,$$

$$\frac{\partial}{\partial \lambda_{hj}(t_k)} \log \mathrm{lik}_n(\hat{\theta}_n, \hat{\Lambda}_n) = 0, \qquad t_k \in \mathcal{T}_{hj,n},$$

which can be solved using a self-consistency algorithm. For fixed Λ it is straightforward to show that the $\log \mathrm{lik}_n(\theta, \Lambda)$ is strictly concave in θ, unless $Z_i = 0$, for every $i = 1, \dots, n$. Uniqueness of the SMLE for Λ is relatively difficult to establish. To safeguard against the potential for non-convexity or multiple stationary points in the objective function, standard methods such as the examination of different starting values and profile plots of the log-likelihood (e.g. Lawless 2003, p. 556) can be applied here. Further details on computation are deferred to Section 3.4.

3.3 ASYMPTOTIC PROPERTIES

Under some regularity conditions the sieve maximum likelihood estimator $(\hat{\theta}_n, \hat{\Lambda}_n)$ globally converges to the truth (θ_0, Λ_0) slower than the parametric rate $\sqrt{n}$, but $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). Proofs are constructed by adapting results from Section 2.3.

3.3.1 *Consistency*

The SMLE is asymptotically unbiased by application of Theorem 1.16. The conditions needed for this result can be verified along the same lines as Section 2.3.1, though some adjustments are needed to accommodate the sieve estimator. These easily follow by adaption of Y. Zhang et al.'s (2010) proof of consistency.

B4 θ_0 lies in the interior of Θ and Θ is a compact subset of $\mathbb{R}^{d_z}$.

B5 The distributions for C and D have support contained in $[\sigma, \tau]$ such that $\mathrm{P}(C = D = \tau \mid Z) > 0$, almost surely.

B6 The distribution of Z has support $\mathcal{Z} = \mathrm{supp}(F_Z)$ on a bounded subset of $\mathbb{R}^{d_z}$.

B7 For each $h \neq j$, $\mathrm{P}(Z_{hj}^{\top} a \neq c) > 0$ for every $a \in \mathbb{R}^{d_z}$ and $c \in \mathbb{R}$.

3.3 THEOREM. *Under the above conditions* $\|\hat{\theta}_n - \theta_0\| + \|\hat{\Lambda}_n - \Lambda_0\|_2 \overset{\text{as}}{\to} 0$, *where*

$$\|\hat{\Lambda}_n - \Lambda_0\|_2 = \sum_{h \neq j} \Big[\int_{\sigma}^{\tau} |\hat{\Lambda}_{hj,n} - \Lambda_{hj}^{0}|^2(u)\,\mathrm{d}u \Big]^{1/2}$$

is the L_2 distance between $\hat{\Lambda}_n$ and Λ_0 on (σ, τ).

Proof. We verify the conditions of Theorem 1.16 with criterion function

$$m_{\theta,\Lambda} = \log \frac{p_{\theta,\Lambda} + p_0}{2}.$$

From B2, H is assumed bounded on $[0,\tau]$. With B4, $\Theta \times H_n$ is a compact parametric class for each n with bracketing number

$$N_{[\,]}(\varepsilon, \Theta \times H_n, L_r(\mathbb{P}_n)) \lesssim (\operatorname{diam}\Theta/\varepsilon)^d (M/\varepsilon)^{K_n}, \tag{3.8}$$

where $K_n = \sum_{h \neq j} K_{hj,n}$. The corresponding bracketing integral converges and, by Theorem 1.12, $\Theta \times H_n$ is P-Donsker. Suppose that (θ_L, θ_R) and (Λ_L, Λ_R) is a bracket for (θ, Λ). By conditions B2 and B4 and the mean value theorem

$$\int_\sigma^\tau (\Lambda_{hj,L}(u)e^{z^\top \theta_L} - \Lambda_{hj,R}(u)e^{z^\top \theta_R})^2 \,\mathrm{d}u \lesssim \|\theta_L - \theta_R\| + \|\Lambda_{hj,L} - \Lambda_{hj,R}\|_2.$$

Since $m_{\theta,\Lambda}$ is pointwise Lipschitz in the transition intensities,

$$\int (m_{\theta_L,\Lambda_L} - m_{\theta_R,\Lambda_R})^2 \,\mathrm{d}\nu \lesssim \|\theta_L - \theta_R\| + \|\Lambda_L - \Lambda_R\|_2. \tag{3.9}$$

Thus $\{m_{\theta,\Lambda} : \theta \in \Theta, \Lambda \in H_n\}$ is also P-Donsker. Following (2.4), $P(m_{\theta,\Lambda} - m_{\theta_0,\Lambda_0}) \leq -d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) \leq 0$ with equality only if $(\theta,\Lambda) = (\theta_0,\Lambda_0)$ by Lemma 3.4 below. So it remains to show that $(\hat\theta_n, \hat\Lambda_n)$ is a near maximizer of $m_{\theta,\Lambda}$. Adapting the approach from Y. Zhang et al. (2010, pp. 352–53), let $\Lambda_{0,n} \in H_n$ be the piecewise linear interpolant of Λ_0 given by Definition 3.2. Since the logarithm is concave and the SMLE is a maximizer of the log-likelihood function $n\,\mathbb{P}_n \log p_{\theta,\Lambda}$ on the sieve $\Theta \times H_n$,

$$\begin{aligned}
\mathbb{P}_n(m_{\hat\theta,\hat\Lambda_n} - m_{\theta_0,\Lambda_0}) &= \mathbb{P}_n \log \frac{p_{\hat\theta,\hat\Lambda_n} + p_0}{2p_0} \\
&\geq \tfrac{1}{2}\,\mathbb{P}_n(\log p_{\hat\theta,\hat\Lambda_n} - \log p_0) \\
&\geq \tfrac{1}{2}\,\mathbb{P}_n(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0) \qquad (3.10)\\
&= \tfrac{1}{2}(\mathbb{P}_n - P)(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0) + \tfrac{1}{2}\,P(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0).
\end{aligned}$$

Elementary results from approximation theory (e.g. Dahlquist and Björck 1974, p. 10) show that $\|\Lambda_{0,n} - \Lambda_0\|_\infty$ has order

$$\max\{t_k - t_{k-1} : t_{k-1}, t_k \in \mathcal{T}_{hj,n}, h \neq j\} = O(\tau/n^\kappa),$$

under condition B2 and Definition 3.2. So $P(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0)^2 \lesssim n^{-2\kappa}$ and, by Theorem 1.19, $(\mathbb{P}_n - P)(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0) = o_P(n^{-1/2})$. From (1.25) the Kullback-Leibler distance between the measures under $(\theta_0, \Lambda_{0,n})$ and the truth is bounded above by zero. This gives $P(\log p_{\theta_0,\Lambda_{0,n}} - \log p_0) > -\,o(1)$. The above inequality then amounts to $\mathbb{P}_n\, m_{\hat\theta,\hat\Lambda_n} - \mathbb{P}_n\, m_{\theta_0,\Lambda_0} \geq o_P(n^{-1/2}) - o(1) = -\,o_P(1)$. ∎

3.4 LEMMA. *For every* $(\theta, \Lambda) \neq (\theta_0, \Lambda_0)$ *on* (σ, τ), $p_{\theta,\Lambda} \neq p_0$, *almost surely.*

Proof. Under conditions B2 and B4 to B6, each of $\mathrm{P}(S > \tau = C = D \mid Z)$, $\mathrm{P}(S = T < \tau = C = D \mid Z)$ and $\mathrm{P}(S < T < \tau = C = D \mid Z)$ are almost surely positive. Assume that $p_{\theta,\Lambda} = p_0$, almost surely. Then by Duhamel's equation (Theorem 1.5),

$$\begin{aligned} 0 &= |P_{00}(0,\tau \mid Z) - P_{00,0}(0,\tau \mid Z)| \\ &= \int_0^\tau P_{00}(0,u \mid Z)|A_{00} - A_{00,0}|(\mathrm{d}u \mid Z)P_{00,0}(u,\tau \mid Z), \end{aligned}$$

almost surely. This is satisfied only if A_{00} and $A_{00,0}$ are almost surely equal on $(0,\tau)$. Our assumption also implies that

$$\begin{aligned} P_{00}(0,t \mid Z)\alpha_{02}(t \mid Z) &= P_{00,0}(0,t \mid Z)\alpha_{02,0}(t \mid Z), \\ P_{00}(0,t \mid Z)\alpha_{01}(t \mid Z)P_{11}(t,\tau \mid Z) &= P_{00,0}(0,t \mid Z)\alpha_{01,0}(t \mid Z)P_{11,0}(t,\tau \mid Z), \end{aligned}$$

almost surely for each $\sigma < t < \tau$. Put $t^* = \inf\{\sigma \leq t < \tau : \lambda_{02,0}(t) > 0\}$. Then from B2 and the previous result, the first almost-sure identity above yields $e^{Z_{02}^\top(\theta-\theta_0)} = \lambda_{02}(t^*)/\lambda_{02,0}(t^*)$ and hence $Z_{02}^\top(\theta - \theta_0)$ is degenerate. Under B7 this implies that $\theta = \theta_0$, which in turn gives $\Lambda_{02} = \Lambda_{02,0}$ and $\Lambda_{01} = \Lambda_{01,0}$ on (σ,τ). With the second almost-sure identity, we obtain $\Lambda_{12} = \Lambda_{12,0}$ on (σ,τ). ▪

3.3.2 *Rate of convergence*

The rate at which the SMLE converges to the truth is obtained by application of Theorem 1.18. The requirements of this result can be verified by adapting the approach of Y. Zhang et al. (2010, pp. 352–53).

B8 For some $r \geq 1$, the rth derivative of Λ_0 continuous, positive and bounded on $[\sigma,\tau]$.

3.5 THEOREM. $\|\hat{\theta}_n - \theta_0\| + \|\hat{\Lambda}_n - \Lambda_0\|_2 = O_P(\max(n^{-(1-\kappa)/2}, n^{-r\kappa}))$ *under the conditions above.*

Proof. We apply Theorem 1.18 where, from the proof of Theorem 3.3 and Lemma 3.6 below, all but requirement (1.23) are verified. In particular for (1.22) the upper bound up to a constant is $-O(n^{-2rk})$ by condition B8. From (3.8)

$$J_{[\,]}(\delta, \{m_{\theta,\Lambda} : \theta \in \Theta, \Lambda \in H_n\}, L_r(\mathbb{P}_n)) \lesssim \delta\sqrt{d + K_n}.$$

By Lemma 1.22 the function $\varphi_n(\delta)$ in (1.23) then has order $\delta n^{\kappa/2} + n^{\kappa}n^{-1/2}$. The constraint $\varphi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ for every n is satisfied with equality by $\delta_n = n^{-(1-\kappa)/2}/2$. Since

$$n^{2r\kappa}\varphi_n(1/n^{r\kappa}) = n^{2r\kappa}\sqrt{n}(n^{-r\kappa}n^{(1-\kappa)/2} + n^{\kappa-1}) = \sqrt{n}(n^{r\kappa}n^{(1-\kappa)/2} + n^{2r\kappa}n^{-(\kappa-1)}),$$

we achieve a tighter bound of $n^{r\kappa}$ provided that $r\kappa \leq (1-\kappa)/2$. Thus the rate of convergence is $O_P(\min(n^{r\kappa}, n^{(1-\kappa)/2}))$. ∎

3.6 LEMMA. *Under the previous conditions* $d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) \gtrsim \|\theta - \theta_0\|^2 + \|\Lambda - \Lambda_0\|^2$.

Proof. Adapting the proof of Lemma 2.11,

$$d_{\mathrm{H}}^2(p_{\theta,\Lambda}, p_0) = \int \frac{(p_{\theta,\Lambda} - p_0)^2}{(\sqrt{p_{\theta,\Lambda}} + \sqrt{p_0})^2}\,\mathrm{d}\nu \gtrsim \int (p_{\theta,\Lambda} - p_0)^2\,\mathrm{d}\nu,$$

since $p_{\theta,\Lambda} + p_0$ can be uniformly bounded under B2, B4 and B6. By B3 and B8,

$$\begin{aligned}
&\int (p_{\theta,\Lambda} - p_0)^2\,\mathrm{d}\nu \\
&\quad\geq q_{02}\int_Z\int_0^\tau [P_{00}(0,s \mid z)\alpha_{02}(s \mid z) - P_{00,0}(0,s \mid z)\alpha_{02,0}(s \mid z)]^2\,\mathrm{d}s\,\mathrm{d}F_Z(z) \\
&\quad\geq q_{02}\int_Z\int_0^\tau [P_{00}(0,s \mid z) - P_{00,0}(0,s \mid z)]^2\alpha_{02,0}(s \mid z)^2\,\mathrm{d}s\,\mathrm{d}F_Z(z) \\
&\quad\gtrsim \int_Z\int_\sigma^\tau [P_{00}(0,s \mid z) - P_{00,0}(0,s \mid z)]^2\,\mathrm{d}s\,\mathrm{d}F_Z(z) \\
&\quad= \int_Z\int_0^\tau \left[\int_0^s P_{00}(0,s \mid z)(A_{00} - A_{00,0})(\mathrm{d}u \mid z)P_{00,0}(u,s \mid z)\right]^2\mathrm{d}s\,\mathrm{d}F_Z(z) \\
&\quad\gtrsim \int_Z\int_\sigma^\tau \left[\int_0^s (A_{00} - A_{00,0})(\mathrm{d}u \mid z)\right]^2\mathrm{d}s\,\mathrm{d}F_Z(z), \\
&\quad\geq \int_Z\int_\sigma^\tau (A_{0j} - A_{0j,0})^2(s \mid z)\,\mathrm{d}s\,\mathrm{d}F_Z(z), \qquad j = 1,2,
\end{aligned}$$

where the inequalities up to a constant holds because q_{02}, $\alpha_{02,0}$, P_{hh} and $P_{hh,0}$ are bounded away from zero on $[\sigma,\tau]$ and the equality follows from Duhamel's equation (Theorem 1.5). Similarly

$$\begin{aligned}
&\int (p_{\theta,\Lambda} - p_0)^2\,\mathrm{d}\nu \\
&\geq q_{12}\int_Z\int_0^\tau\int_s^\tau [P_{00}(0,s \mid z)\alpha_{01}(s \mid z)P_{11}(s,t \mid z)\alpha_{12}(t \mid z) \\
&\qquad\qquad - P_{00,0}(0,s \mid z)\alpha_{01,0}(s \mid z)P_{11,0}(s,t \mid z)\alpha_{12,0}(t \mid z)]^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}F_Z(z) \\
&\geq q_{12}\int_Z\int_0^\tau\int_s^\tau [P_{11}(s,t \mid z) - P_{11,0}(s,t \mid z)]^2
\end{aligned}$$

$$
\begin{aligned}
&\times [P_{00,0}(0,s \mid z)\alpha_{01,0}(s \mid z)\alpha_{12,0}(t \mid z)]^2 \,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}F_Z(z)\\
\gtrsim{}& \int_Z \int_\sigma^\tau \int_s^\tau [P_{11}(s,t \mid z) - P_{11,0}(s,t \mid z)]^2 \,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}F_Z(z)\\
={}& \int_Z \int_\sigma^\tau \int_s^\tau \left[\int_s^t P_{11}(s,u \mid z)(A_{11} - A_{11,0})(\mathrm{d}u \mid z)P_{11,0}(u,t \mid z)\right]^2 \mathrm{d}t\,\mathrm{d}s\,\mathrm{d}F_Z(z)\\
\gtrsim{}& \int_Z \int_\sigma^\tau \int_s^\tau \left[\int_s^t (A_{11} - A_{11,0})(\mathrm{d}u \mid z)\right]^2 \mathrm{d}t\,\mathrm{d}s\,\mathrm{d}F_Z(z)\\
\geq{}& \int_Z \int_\sigma^\tau (A_{12} - A_{12,0})^2(s \mid z)\,\mathrm{d}s\,\mathrm{d}F_Z(z).
\end{aligned}
$$

Let $\theta_t = t\theta + (1-t)\theta_0$, $\Lambda_{hj,t} = t\Lambda_{hj} + (1-t)\Lambda_{hj,0}$. From the mean value theorem, there is some $0 < t < 1$ depending on (y,z,h,j) such that

$$
\begin{aligned}
(A_{hj} - A_{hj,0})(y \mid z) &= \frac{\partial}{\partial t} A_{hj,t}(y \mid z)\\
&= t\exp(z_{hj}^\top \theta_t)\\
&\quad \times \{[1 + z_{hj}^\top(\theta - \theta_0)t](\Lambda_{hj} - \Lambda_{hj,0})(y) + z_{hj}^\top(\theta - \theta_0)\Lambda_{hj,0}(y)\}.
\end{aligned}
$$

For $(Y,Z) \sim \mu = 1_{[\sigma,\tau]} \times F_Z$, put $g_{hj,0}(z) = 1 + z_{hj}^\top(\theta - \theta_0)t$, $g_{hj,1}(y) = (\Lambda_{hj} - \Lambda_{hj,0})(y)$ and $g_{hj,2}(y,z) = z_{hj}^\top(\theta - \theta_0)\Lambda_{hj,0}(y)$. So $(A_{hj} - A_0)(Y \mid Z)$ is equal to $g_{hj,0}(Z)g_{hj,1}(Y) + g_{hj,2}(Y,Z)$ up to the factor $t\exp(Z_{hj}^\top \theta_t)$, which is bounded away from zero under A4, A5 and A7. By condition B7,

$$
[\mathrm{E}_\mu(g_{hj,1}g_{hj,2})]^2 < \mathrm{E}_\mu(g_{hj,1}^2)\,\mathrm{E}_\mu(g_{hj,2}^2).
$$

Since $g_{hj,0}(z)$ is uniformly close to 1 for θ close to θ_0 and $\Lambda_{hj,0}$ is bounded away from zero on $[\sigma,\tau]$,

$$
\begin{aligned}
\int (p_{\theta,\Lambda} - p_0)^2 \,\mathrm{d}\nu &\gtrsim \mu(g_{hj,0}g_{hj,1} + g_{hj,2})^2\\
&\gtrsim \mu g_{hj,2}^2 + \mu g_{hj,1}^2 \gtrsim \|\theta - \theta_0\|^2 + \|\Lambda - \Lambda_0\|^2,
\end{aligned}
$$

by Lemma 1.21. ■

3.3.3 *Asymptotic normality*

Here the SMLE is shown to be asymptotically efficient by application of Theorem 1.23 and Corollary 1.24. This result requires two further assumptions.

B9 Λ_0 has a second-order derivative that is uniformly bounded on $[\sigma,\tau]$.

3.7 THEOREM. *Let r be the order of the derivative of Λ_0 satisfying condition B8. If $1/(4r) < \kappa < 1/2$ then, under the above conditions, the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). In particular the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $\Sigma = \tilde{I}_0^{-1}$.*

Proof. We prove existence of a least favourable submodel meeting the requirements of Theorem 1.23 and Corollary 1.24 under the assumption that $d_z = 1$; the result for $d_z > 1$ follows by repeated application of the proof for this special case. The score function for $\theta \in \mathbb{R}$ is

$$\begin{aligned}\dot{\ell}_{\theta,\Lambda}(x) = &- z_{01}A_{01}(u) - z_{02}A_{02}(u) + \delta_0\delta_2 z_{02}\\ &+ \delta_0\, 1(u < v)\big[z_{01} - \delta_2 z_{02} + \delta_2 z_{12} - z_{12}A_{12}(u,v)\big]\\ &+ \frac{(1-\delta_0)}{p(u,v)}\Big\{\big[\delta_2 z_{02} - z_{01}A_{01}(u,v) - z_{02}A_{02}(u,v)\big]P_{00}(u,v)\alpha_{02}(v)^{\delta_2}\\ &\qquad - \textstyle\int_u^v [z_{01}\,\mathrm{d}A_{01}(y) + z_{02}\,\mathrm{d}A_{02}(y)]P_{01}(u,y)\alpha_{12}(v)^{\delta_2}\\ &\qquad + \big[z_{01}P_{01}(u,v) - \textstyle\int_u^v z_{12}\,\mathrm{d}A_{12}(y)P_{01}(u,y)P_{11}(y,v)\big]\alpha_{12}(v)^{\delta_2}\\ &\qquad + \delta_2 g_{12}(v)P_{01}(u,v)\alpha_{12}(v)^{\delta_2}\Big\},\end{aligned}$$

where conditionals on z are suppressed in α_{hj}, A_{hj} and P_{hj} for brevity and

$$p(u,v) = p(u,v \mid z) = P_{01}(u,v \mid z)\alpha_{12}(v \mid z)^{\delta_2} + P_{00}(u,v \mid z)\alpha_{02}(v \mid z)^{\delta_2}.$$

Since (3.1) is a multiplicative intensity model, the entries in Λ are essentially variation independent. A tangent set is then obtained by perturbing each Λ_{hj}. Consider a one-dimensional submodel $y \mapsto \Lambda_{hj,y}$ satisfying $\partial/\partial y_{|y=0}\, \mathrm{d}\Lambda_{hj,y} = g_{hj}\Lambda_{hj}$. Then the score for Λ is

$$\begin{aligned}&B_{\theta,\Lambda}g(x) = \sum_{h\neq j} B_{\theta,\Lambda}g_{hj}(x)\\ &= -\int_0^u g_{01}(y)\,\mathrm{d}A_{01}(y) - \int_0^u g_{02}(y)\,\mathrm{d}A_{02}(y) + \delta_0\delta_2 g_{02}(v)\\ &\quad + \delta_0\, 1(u < v)\big[g_{01}(u) - \delta_2 g_{02}(v) + \delta_2 g_{12}(v) - \textstyle\int_u^v g_{12}(y)\,\mathrm{d}A_{12}(y)\big]\\ &\quad + \frac{(1-\delta_0)}{p(u,v)}\Big\{\big[\delta_2 g_{02}(v) - \textstyle\int_u^v g_{01}(y)\,\mathrm{d}A_{01}(y) - \textstyle\int_u^v g_{02}(y)\,\mathrm{d}A_{02}(y)\big]P_{00}(u,v)\alpha_{02}(v)^{\delta_2}\\ &\qquad - \textstyle\int_u^v [g_{01}(y)\,\mathrm{d}A_{01}(y) + g_{02}(y)\,\mathrm{d}A_{02}(y)]P_{01}(y,v)\alpha_{12}(v)^{\delta_2}\\ &\qquad + \textstyle\int_u^v [g_{01}(y)\,\mathrm{d}A_{01}(y)P_{00}(u,y) - g_{12}(y)\,\mathrm{d}A_{12}(y)P_{01}(u,y)]P_{11}(y,v)\alpha_{12}(v)^{\delta_2}\\ &\qquad + \delta_2 g_{12}(v)P_{01}(u,v)\alpha_{12}(v)^{\delta_2}\Big\}.\end{aligned}$$

Note that the scores have similar form. Thus $\dot{\ell}_{\theta,\Lambda}(x)$ can be similarly written as a sum of terms indexed by z_{hj}. Denote these by $\dot{\ell}^{hj}_{\theta,\Lambda}(x)$. Since the distribution of $T_{0j} < \infty$ is specified by (θ, Λ_{0j}), we take the $0 \to j$ adjoint $B^*_{\theta,\Lambda}$ as the conditional expectation operator given $\{T_{0j} = t\}$ under (θ, Λ). The (positive) duration in state 1 is fully specified by (θ, Λ_{12}), so the $1 \to 2$ adjoint is the conditional expectation operator given $\{T_{12} - T_{01} = t\}$. Considering the $0 \to 1$ type-specific terms define, for an arbitrary function φ on $[0, \tau]$,

$$\begin{aligned} B_{01,1}(\varphi)(y) &= \varphi(y) e^{Z_{01}\theta} \lambda_{01}(y) 1_{(0,t)}(y)(1 - F_{C|Z}(y \mid Z)), \\ B_{01,2}(\varphi)(t) &= \varphi(t)(1 - F_{C,D|Z}(t, t \mid Z)), \end{aligned}$$

and

$$\begin{aligned} &B_{01,3}(\varphi)(t, y) = \varphi(y) e^{Z_{01}\theta} \lambda_{01}(y) \\ &\times \int_0^\tau \mathrm{E}_{C,D|Z}(Q(C, t_{12} \wedge D, 1(t_{12} < D), Z)(y) \mid Z) \exp(-A_{12}(t, t_{12} \mid Z)) A_{12}(\mathrm{d}t_{12} \mid Z), \end{aligned}$$

where

$$\begin{aligned} &Q(c, v, \delta_2, z)(y) = \frac{1_{(c,v)}(y)}{p(c, v \mid z)} \\ &\times \Big\{ \big[P_{00}(c, y \mid z) P_{11}(y, v \mid z) + P_{00}(y, v \mid z) \big] \alpha_{12}(v \mid z)^{\delta_2} + P_{00}(c, v \mid z) \alpha_{02}(v \mid z)^{\delta_2} \Big\}. \end{aligned}$$

Let $B_{01,k}(\cdot) \equiv B_{01,k}(1)(\cdot)$ and $b_{01,k}$ be the expectation of $B_{01,k}$ with respect to the distribution of Z. Then

$$\begin{aligned} B^*_{\theta,\Lambda} B_{\theta,\Lambda} g_{01}(t) &= -\int_0^\tau b_{01,1}(g_{01})(y)\,\mathrm{d}y + b_{01,2}(g_{01})(t) + \int_0^\tau b_{01,3}(g_{01})(t, y)\,\mathrm{d}y, \\ B^*_{\theta,\Lambda} \dot{\ell}^{01}_{\theta,\Lambda}(t) &= -\int_0^\tau b_{01,1}(Z_{01})(y)\,\mathrm{d}y + b_{01,2}(Z_{01})(t) + \int_0^\tau b_{01,3}(Z_{01})(t, y)\,\mathrm{d}y. \end{aligned}$$

If the least favourable direction $g^{01}_{\theta,\Lambda}$ exists, it is characterized by

$$B^*_{\theta,\Lambda} \dot{\ell}^{01}_{\theta,\Lambda}(t) = B^*_{\theta,\Lambda} B_{\theta,\Lambda} g^{01}_{\theta,\Lambda}(t).$$

This identity reduces to the Fredholm integral equation of the second kind

$$g^{01}_{\theta,\Lambda}(t) = f_{01}(t) + \int_0^\tau K_{01}(t, y) g^{01}_{\theta,\Lambda}(y)\,\mathrm{d}y, \tag{3.11}$$

where

$$f_{01}(t) = \frac{B^*_{\theta,\Lambda} \dot{\ell}^{01}_{\theta,\Lambda}(t)}{1 - \mathrm{E}_Z(F_{C,D|Z}(t, t \mid Z))} \quad \text{and} \quad K_{01}(t, y) = \frac{b_{01,1}(t, y) - b_{01,3}(t, y)}{1 - \mathrm{E}_Z(F_{C,D|Z}(t, t \mid Z))}.$$

At the truth (θ_0, Λ_0), f_{01} and K_{01} are bounded away from zero by B3 and from infinity by B5. From Fredholm's first theorem (e.g. Kanwal 1997, p. 48) there exists a unique solution g_0^{01} to (3.11) at (θ_0, Λ_0). Integral equations for $g_{\theta,\Lambda}^{02}$ and $g_{\theta,\Lambda}^{12}$ can be similarly derived. From condition B9, the least favourable directions at (θ_0, Λ_0) have bounded derivatives on $[\sigma, \tau]$. Let $g_{0,n}^{hj}$ denote the linear interpolant of g_0^{hj} under the same partition for $\hat{\Lambda}_n$. The least favourable submodel can then be defined as

$$\Lambda_{hj,y}(\theta, \Lambda) = \int (1 + (\theta - y) g_{0,n}^{hj}) \, \mathrm{d}\Lambda_{hj}, \quad h \neq j,$$

which is a cumulative baseline intensity function in H_n for y sufficiently close to θ. Equation (1.32) follows from Theorem 3.3 and the fact that θ and Λ are variation independent. Ordinary Taylor expansions yield an upper bound for $\dot{\ell}(\theta_0, \theta_0, \Lambda)$ proportional to $\|\Lambda - \Lambda_0\|^2$. By Theorem 3.5 and the restrictions on κ, the "bias" term $P_0\, \dot{\ell}(\theta_0, \hat{\theta}_n, \hat{\Lambda}_n)$ is faster than the required $O_P(n^{-1/4})$. This satisfies (1.33) via B8 and (1.35). The remaining structural requirements of Theorem 1.23 and Corollary 1.24 are met largely by assumption. ▪

Theorem 3.7 and Corollary 1.26 yield a consistent estimator for the profile information matrix of $\hat{\theta}_n$.

3.8 COROLLARY. *Let $e_1, \ldots, e_{d_z}$ be the unit vectors in $\mathbb{R}^{d_z}$ and ρ_n be a symmetric d_z-matrix whose entries ρ_{ij}, $i, j = 1, \ldots, d_z$, satisfy $(\sqrt{n}\rho_{ij})^{-1} = O_P(1)$. A consistent estimator for each entry in $\tilde{I}_0$ has the same form as* (2.12).

3.4 COMPUTATION

A computational algorithm for the SMLE can be specified using the same elements described in Section 1.2.4, though the objective function here is not convex. An EM-type approach is obtained by combining the Newton-Raphson method for θ and a self-consistency algorithm (Turnbull 1976) for Λ.

Let $\lambda_{hj} = (\lambda_{hj,0}, \ldots, \lambda_{hj,K_{hj,n}})^\top$ denote the vector representing the piecewise constant values of $\lambda_{hj}(t)$ over $[0, \tau)$, with $\lambda_{hj,k} = \lambda_{hj}(t_k)$ and $t_k \in \mathcal{T}_{hj,n}$. For brevity put $\lambda = (\lambda_{01}^\top, \lambda_{02}^\top, \lambda_{12}^\top)^\top$, $\phi = (\theta^\top, \lambda^\top)^\top$ and $\log \mathrm{lik}_n(\phi) \equiv \log \mathrm{lik}_n(\theta, \lambda)$. The SMLE is the solution to the score equation $\nabla_\phi \log \mathrm{lik}_n(\phi) = 0$, which can be solved numerically. A given value $\phi^{(r)} = (\theta^{(r)}, \lambda^{(r)})$ is updated via the candidate step $\eta^{(r)} = (\eta_\theta^{(r)}, \eta_\lambda^{(r)})$. Its first component $\eta_\theta^{(r)}$ is handled by the Newton-Raphson method

$$\eta_\theta^{(r)} = -\nabla_\theta^2 \log \mathrm{lik}_n(\phi^{(r)})^{-1} \nabla_\theta \log \mathrm{lik}_n(\phi^{(r)}). \tag{3.12}$$

For $\eta_\theta^{(r)}$, we incorporate the solution to the self-consistency equations, obtained by re-arranging $\nabla_\lambda \log \mathrm{lik}_n(\theta, \lambda) = 0$ to give a recursive expression for λ. In particular

$$\eta^{(r)}_{\lambda_{hj,k}} = r_{hj,n}(t_k; \theta^{(r)} + \eta^{(r)}_\theta, \lambda^{(r)}) - \lambda^{(r)}_{hj,k}, \quad t_k \in \mathcal{T}_{hj,n}, \tag{3.13}$$

$$r_{hj,n}(t_k; \theta, \lambda) = \frac{\sum_i \mathrm{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{d}N^i_{hj}(s) \mid X_i\right)}{\sum_i \exp(Z^{i\,\top}_{hj}\theta)\,\mathrm{E}\left(\int_{t_{k-1}}^{t_k} Y^i_h(s)\,\mathrm{d}s \mid X_i\right)}, \tag{3.14}$$

where the expectations evaluate to 1 if X_i provides exact information; otherwise, suppressing the dependence on Z,

$$\begin{aligned}
\mathrm{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{d}N_{01}(s) \mid X\right) &= \frac{1(U < V)}{p_{\theta,\Lambda}(X)} \int_L^R I_k(s) P_{00}(0,s)\alpha_{01}(s) P_{11}(s,V)\alpha_{12}(V)^{\Delta_2}\,\mathrm{d}s,\\
\mathrm{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{d}N_{02}(s) \mid X\right) &= \frac{1 - \Delta_0 1(U < V)}{p_{\theta,\Lambda}(X)} I_k(V) P_{00}(0,V)\alpha_{02}(V)^{\Delta_2},\\
\mathrm{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{d}N_{12}(s) \mid X\right) &= \frac{1(U < V)}{p_{\theta,\Lambda}(X)} I_k(V) P_{00}(0,L) P_{01}(L,R) P_{11}(R,V)\alpha_{12}(V)^{\Delta_2},\\
\mathrm{E}\left(\int_{t_{k-1}}^{t_k} Y_0(s)\,\mathrm{d}s \mid X\right) &= \frac{1(U < V)}{p_{\theta,\Lambda}(X)} \int_L^R L_k(0,s) P_{00}(0,s)\alpha_{01}(s) P_{11}(s,V)\alpha_{12}(V)^{\Delta_2}\,\mathrm{d}s\\
&\qquad + \frac{1 - \Delta_0 1(U < V)}{p_{\theta,\Lambda}(X)} L_k(0,V) P_{00}(0,V)\alpha_{02}(V)^{\Delta_2},\\
\mathrm{E}\left(\int_{t_{k-1}}^{t_k} Y_1(s)\,\mathrm{d}s \mid X\right) &= \frac{1(U < V)}{p_{\theta,\Lambda}(X)} \int_L^R L_k(s,V) P_{00}(0,s)\alpha_{01}(s) P_{11}(s,V)\alpha_{12}(V)^{\Delta_2}\,\mathrm{d}s,
\end{aligned}$$

with $[L, R]$ denoting the (potential) censoring interval for T_{01}. The notation in (3.14) draws a clear analogy to (1.11). Integrals over $[L, R]$ are straightforward to evaluate over a partition on which the model is time homogeneous; if $\alpha_{hj}(u \mid z) = \alpha_{hj}$ for $u \in [s, t)$, $0 \le s < t \le v \le \tau$, then

$$\begin{aligned}
&\int_s^t P_{00}(0, u \mid z)\alpha_{01}(u \mid z) P_{11}(u, v \mid z)\,\mathrm{d}u\\
&= P_{00}(0, s \mid z)\frac{\exp(-(t-s)\alpha_{12}) - \exp(-(t-s)(\alpha_{01} + \alpha_{02}))}{\alpha_{01} + \alpha_{02} - \alpha_{12}} P_{11}(t, u \mid z).
\end{aligned}$$

Overshoot is avoided by a simple step-halving procedure

$$\phi^{(r+1)} = \phi^{(r)} + \eta^{(r)}/2^j, \tag{3.15}$$

where j is the smallest nonnegative integer satisfying

$$\log \mathrm{lik}_n(\phi^{(r)}) \le \log \mathrm{lik}_n(\phi^{(r)} + \eta^{(r)}/2^j).$$

Putting this together gives the following algorithm, which can be proven to converge to a local maximizer of the log-likelihood on the basis of results from Wu (1983). The presence of a non-unique solution may be detected with different starting values, though experience thus far has not uncovered any instances of multiple maxima.

3.9 ALGORITHM. Let n_{hj} be the number of the exact $h \to j$ transition times observed from $X_1, \dots, X_n$. For given $C_{hj} > 0$ and $0 < \kappa < 1$, define $\mathcal{T}_{hj,n}$ as a partition of $[0, \tau)$ in which each subinterval contains $\lceil n_{hj}/(C_{hj} n^{\kappa}) \rceil$ exact $h \to j$ transition times. Set $r := 0$, $\theta^{(0)} = 0$ and $\lambda^{(0)} = 1$. Let $\eta^{(r)}$ be the candidate step with components given by (3.12) and (3.13) and $\phi^{(r+1)}$ be the result of the line search (3.15). If

$$\|\phi^{(r+1)} - \phi^{(r)}\|_{\infty} \leq \varepsilon,$$

for small positive value ε, then stop. Otherwise put $r := r + 1$. ◻

From condition B3 and Theorem 3.5 the optimal choice for κ is $1/(1+2k)$, where k is the order of the derivative on which we impose condition B8. Numerical studies of similar estimators suggest that larger partitions give better finite-sample results (e.g. Huang and Rossini 1997). Some broad comparisons in the choice of κ are provided in Section 3.5.

Holding $\theta^{(r)}$ fixed in Algorithm 3.9 evaluates the profile likelihood needed for variance estimation under Corollary 3.8. To approximate entries from the profile information matrix we follow the same procedure outlined in Section 2.4.2.

3.5 SIMULATION STUDY

Numerical properties of the SMLE were investigated for variants of the censoring scheme described in Remark 3.1. Each of these considered the same transition intensity model

$$A_{hj}(t \mid Z) = \Lambda_{hj}(t) \exp(\theta_1 Z_{01} + \theta_2 Z_{02} + \theta_3 Z_{12}), \tag{3.16}$$

where $\Lambda_{01}(t) = t^{4/5}$, $\Lambda_{02}(t) = 3t/4$, $\Lambda_{12}(t) = (3t/2)^{5/4}$, Z is uniform on $\{0,1\}$, Z_{hj} is the product of Z and the $h \to j$ transition type indicator, $\theta_1 = -\log(2)$, $\theta_2 = -\log(2)$ and $\theta_3 = 0$. Note that Z influences only the risk of transition out of the initial state 0 and its effect is the same for each transition type, but we did not assume either of these properties in estimating θ. Transition times were right-censored by $D = \tau = 2$ and $C = D$ with probability $1 - p(Z)$, where

$$\operatorname{logit}(p(Z)) = \beta_0 + \beta_1 Z.$$

In the event that $C \neq D$, $C = Y \wedge \tau$ where Y is exponentially-distributed with mean $\tau/2 = 1$. Here D can be thought of as an administrative censoring time and C, when $C < D$, a random dropout time occurring earlier in the observation period.

Three thousand Monte Carlo replicates with sample sizes $n = 500, 1000$ and 2000 were generated under three scenarios:

- independent censoring with $\beta_0 = \log(1/3)$ and $\beta_1 = 0$,
- independent censoring with $\beta_0 = \log(9)$ and $\beta_1 = 0$, and
- conditionally independent censoring $\beta_0 = \log(1/3)$ and $\beta_1 = \log(3/2)$.

With $\exp(\beta) = (1/3, 1)$, the probability of $\{C < D\}$ is $p(Z) = 1/4$. Under $\exp(\beta) = (9, 1)$ this increases to 9/10. Under $\exp(\beta) = (1/3, 3/2)$ the probability is 1/4 for subjects with $Z = 0$. For those with $Z = 1$, $p = 1/3$.

Estimates for each sample were obtained using a C implementation of Algorithm 2.15 relying on routines from the LAPACK library (Anderson et al. 1999) to carry out matrix inversion. Tuning parameters were set to $C_{hj} = 1$, $\kappa = 2/5$, $\varepsilon = 10^{-7}$, $\operatorname{typ}\theta = 1$ and $\sup\theta = 10$. This ensured convergence within a reasonable number of iterations over all scenarios and sample sizes. Under the first scenario estimates were re-evaluated with the alternative sieve sizes $\kappa = 4/15$ and $\kappa = 2/5$.

The SPMLE based on "singly" right-censored data (Example 1.6) was also considered for:

- the underlying or "latent" transition times (T_{01}, T_{02}, T_{12}) right-censored by D and
- the observed transition times (T_{01}, T_{02}, T_{12}) right-censored by C.

Estimates were obtained using Therneau's (2012) `coxph` routine from the `survival` package for R. The first case above corresponds to the latent data, observable only if C was always equal to D. The second approach is essentially right-censors all transition times at C, ignoring any observed $T = T_{02} \wedge T_{12}$ with $C < T < D$. The resulting SPMLE for the latter is consistent under both condition B1 and the weaker assumption of independent right-censoring at C (Example 1.1). In practice, however, the observed data are often reduced further to obtain the composite endpoint $S = T_{01} \wedge T_{02}$, commonly known as progression-free survival (PFS). Some ad hoc methods for PFS mentioned by Ke et al. (2011) call for survival analysis of

- the observed time $U = S \wedge C$ with event status $\Delta_0 = 1(S \leq C)$,
- U with any available status $\max(\Delta_0, \Delta_2)$, and
- $\Delta_0 U + (1 - \Delta_0) V$ with $\max(\Delta_0, \Delta_2)$.

The first approach is a reasonable simplification of the three-state model when the $0 \to 1$ and $0 \to 2$ intensity functions are similar. The last two methods involve systematic imputation and are thus prone to bias. From data under each of these cases we fit a Cox model with a single covariate Z.

Under the fixed transition intensity model parameters roughly 50% of subjects in the sample progressed by τ ($S < T \leq \tau$), 12% were event-free ($S > \tau$), and 16% survived to study closure ($T > \tau$). Censoring times generated from $\beta_0 = \log(1/3)$, gave a low proportion of doubly right-censored observations; less than 15%. With $\beta_0 = \log(9)$, the doubly right-censoring rate increases to almost 40% (Table 3.1). Censoring rates were roughly comparable between the positive and negative progression status groups.

			Censoring rate	
		Exact observation	Singly right-censored	Doubly right-censored
e^{β_0}	e^{β_1}	$T < C$	$C = D < T$ or $S < C < T$	$C < D$ and $C < S$
1/3	1	0.704	0.153	0.142
9	1	0.473	0.142	0.385
1/3	3/2	0.722	0.150	0.128

TABLE 3.1 Average censoring rates over 3000 replicates with $n = 2000$.

The simulation results for the SMLE $\hat{\theta}_n$ summarized by Table 3.2 are compatible with the asymptotic properties derived in Section 3.3. Bias becomes negligible with larger sample size. Monte Carlo sample standard deviations for $\hat{\theta}_n$ decrease with larger n and is reasonably approximated by the standard error estimates. Empirical coverage probabilities of the 95% confidence intervals are close to the nominal level. The SPMLE arising from singly right-censored data at C has a higher degree of variability, but achieves lower average bias in some scenarios. The difference in finite-sample bias between the two estimators becomes small in larger samples with higher rates of censoring. Since the covariate effect on the exit time from the initial state is invariant to the transition type, analysis of progression-free survival right-censored at C also performs reasonably well. However the imputation variants are heavily biased (Table 3.3).

Table 3.4 suggests that we should generally prefer larger sieves to mitigate finite-sample bias in $\hat{\theta}_n$. Under the independent, low-rate censoring scenario, $e^{\beta} = (1/3, 1)$, the average bias for the SMLE with a sieve growing at the rate $n^{4/15}$ was up to eleven times larger than the bias achieved by an $n^{2/5}$-sieve estimator. The SMLE under the asymptotically optimal rate of $n^{1/3}$ also fared no better than the $n^{2/5}$-sieve, with mean bias ratios up to 4.5. The corresponding standard deviation ratios are close to 1, so the reduction in bias came at little to no loss in precision. The improvement achieved by a larger sieve appears to diminish with increasing sample size for the estimates specific to the terminal event types. The scenarios examined here generate exactly observed 0 → 1 times more frequently, so in practice selection of the sieve constants

TABLE 3.2 Bias, standard deviation (SD), average standard error estimate (ASE) and empirical coverage probability of 95% confidence intervals (CP) for the SMLE of θ over 3000 replicates. The same results for the SPMLE based on transition times and PFS, both right-censored at C, are provided for comparison.

e^{β_0},				SMLE			SPMLE censored by C		PFS
e^{β_1}	n		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ
1/3,	500	Bias	-0.0054	-0.0007	0.0025	-0.0033	-0.0019	0.0011	-0.0022
1		SD	0.1334	0.1508	0.1430	0.1347	0.1572	0.1442	0.1031
		ASE	0.1352	0.1538	0.1401	0.1374	0.1611	0.1413	0.1047
		CP	0.9477	0.9557	0.9427	0.9517	0.9550	0.9453	0.9527
	1000	Bias	-0.0005	0.0017	0.0010	0.0004	0.0005	-0.0004	0.0007
		SD	0.0961	0.1109	0.0988	0.0970	0.1148	0.1000	0.0741
		ASE	0.0956	0.1082	0.0990	0.0972	0.1135	0.1000	0.0739
		CP	0.9540	0.9437	0.9507	0.9547	0.9473	0.9487	0.9490
	2000	Bias	-0.0008	0.0014	-0.0012	-0.0003	0.0007	-0.0020	0.0002
		SD	0.0663	0.0756	0.0703	0.0670	0.0787	0.0708	0.0504
		ASE	0.0675	0.0765	0.0699	0.0687	0.0802	0.0707	0.0522
		CP	0.9510	0.9563	0.9457	0.9547	0.9567	0.9523	0.9607
9,	500	Bias	-0.0067	-0.0021	0.0039	-0.0050	-0.0039	0.0021	-0.0035
1		SD	0.1497	0.1618	0.1565	0.1552	0.1866	0.1604	0.1203
		ASE	0.1499	0.1632	0.1545	0.1560	0.1884	0.1589	0.1203
		CP	0.9483	0.9517	0.9433	0.9487	0.9497	0.9437	0.9523
	1000	Bias	-0.0002	0.0017	-0.0003	0.0019	0.0025	-0.0009	0.0026
		SD	0.1065	0.1170	0.1093	0.1098	0.1334	0.1122	0.0849
		ASE	0.1060	0.1146	0.1091	0.1102	0.1325	0.1123	0.0848
		CP	0.9523	0.9497	0.9553	0.9550	0.9503	0.9547	0.9517
	2000	Bias	-0.0011	0.0004	-0.0002	0.0002	0.0005	-0.0005	0.0006
		SD	0.0737	0.0799	0.0776	0.0767	0.0912	0.0796	0.0583
		ASE	0.0749	0.0810	0.0770	0.0778	0.0937	0.0793	0.0599
		CP	0.9523	0.9540	0.9493	0.9583	0.9530	0.9493	0.9630
1/3,	500	Bias	-0.0053	-0.0007	0.0024	-0.0033	-0.0019	0.0011	-0.0021
3/2		SD	0.1321	0.1503	0.1425	0.1332	0.1562	0.1435	0.1020
		ASE	0.1341	0.1532	0.1394	0.1363	0.1595	0.1404	0.1037
		CP	0.9520	0.9563	0.9420	0.9550	0.9553	0.9443	0.9510
	1000	Bias	-0.0001	0.0015	0.0009	0.0009	0.0007	-0.0002	0.0012
		SD	0.0953	0.1103	0.0985	0.0962	0.1137	0.0993	0.0737
		ASE	0.0949	0.1078	0.0985	0.0964	0.1123	0.0994	0.0732
		CP	0.9510	0.9460	0.9517	0.9503	0.9493	0.9470	0.9500
	2000	Bias	-0.0006	0.0014	-0.0012	-0.0002	0.0007	-0.0020	0.0004
		SD	0.0657	0.0754	0.0700	0.0664	0.0781	0.0703	0.0500
		ASE	0.0670	0.0762	0.0696	0.0681	0.0794	0.0703	0.0517
		CP	0.9513	0.9543	0.9473	0.9530	0.9553	0.9520	0.9570

		U-imputed PFS	V-imputed PFS	Latent SPMLE		
n		θ	θ	θ_1	θ_2	θ_3
500	Bias	0.0473	0.1471	-0.0034	-0.0020	0.0004
	SD	0.0964	0.0872	0.1265	0.1469	0.1359
	ASE	0.0982	0.0774	0.1292	0.1498	0.1334
	CP	0.9253	0.5103	0.9487	0.9513	0.9450
1000	Bias	0.0491	0.1491	0.0016	0.0008	-0.0008
	SD	0.0705	0.0631	0.0906	0.1075	0.0947
	ASE	0.0693	0.0546	0.0914	0.1055	0.0945
	CP	0.8797	0.2463	0.9493	0.9463	0.9517
2000	Bias	0.0488	0.1485	-0.0005	0.0012	-0.0022
	SD	0.0467	0.0422	0.0632	0.0735	0.0669
	ASE	0.0490	0.0386	0.0646	0.0746	0.0668
	CP	0.8383	0.0447	0.9550	0.9580	0.9500

TABLE 3.3 Simulation results for estimators of θ based on imputation variants of PFS and the latent data over 3000 replicates from $e^{\beta} = (1/3, 1)$.

C_{hj} should probably consider relative frequencies of the exact event times.

		$\kappa = 4/15$			$\kappa = 1/3$		
n		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
500	Relative bias	2.401	2.184	2.279	1.601	1.497	1.506
	Relative precision	1.002	0.998	0.992	1.001	0.999	0.998
1000	Relative bias	10.949	0.736	3.673	4.478	0.920	1.810
	Relative precision	1.003	0.998	0.995	1.000	0.999	0.999
2000	Relative bias	6.795	0.737	-0.835	2.450	0.864	0.515
	Relative precision	1.001	0.998	0.995	1.000	0.999	0.999

TABLE 3.4 Ratio of bias (relative bias) and SD (relative precision) for $\hat{\theta}_n$ between the specified κ and $\kappa = 2/5$, based on 3000 replicates from $k = 8$ and $e^{\beta} = (1/3, 1)$.

Under the scenarios $e^{\beta} = (9, 1)$ and $e^{\beta} = (1/3, 3/2)$, pointwise means and percentiles for SMLE of the cumulative baseline intensity functions are depicted in Figure 3.3. Pointwise estimates for Λ_{01} and Λ_{02} appear unbiased. Early in the observation period the SMLE for λ_{12} tends to be larger than the truth, giving a general overestimate for the cumulative function Λ_{12}. This bias becomes negligible under larger samples and lower rates of double right-censoring. Variability in each component of $\hat{\Lambda}_n$ is low and also decreases with increasing n and smaller β_0. In smaller sample sizes, bias appears more influenced by the size of the sieve rather than the rate of double censoring (Figure 3.4). The level of variability was similar between the largest and smallest sieve sizes examined.

One apparent trade-off in a larger sieve is longer time to convergence in terms of both the number of iterations required by Algorithm 3.9 and the computing time for

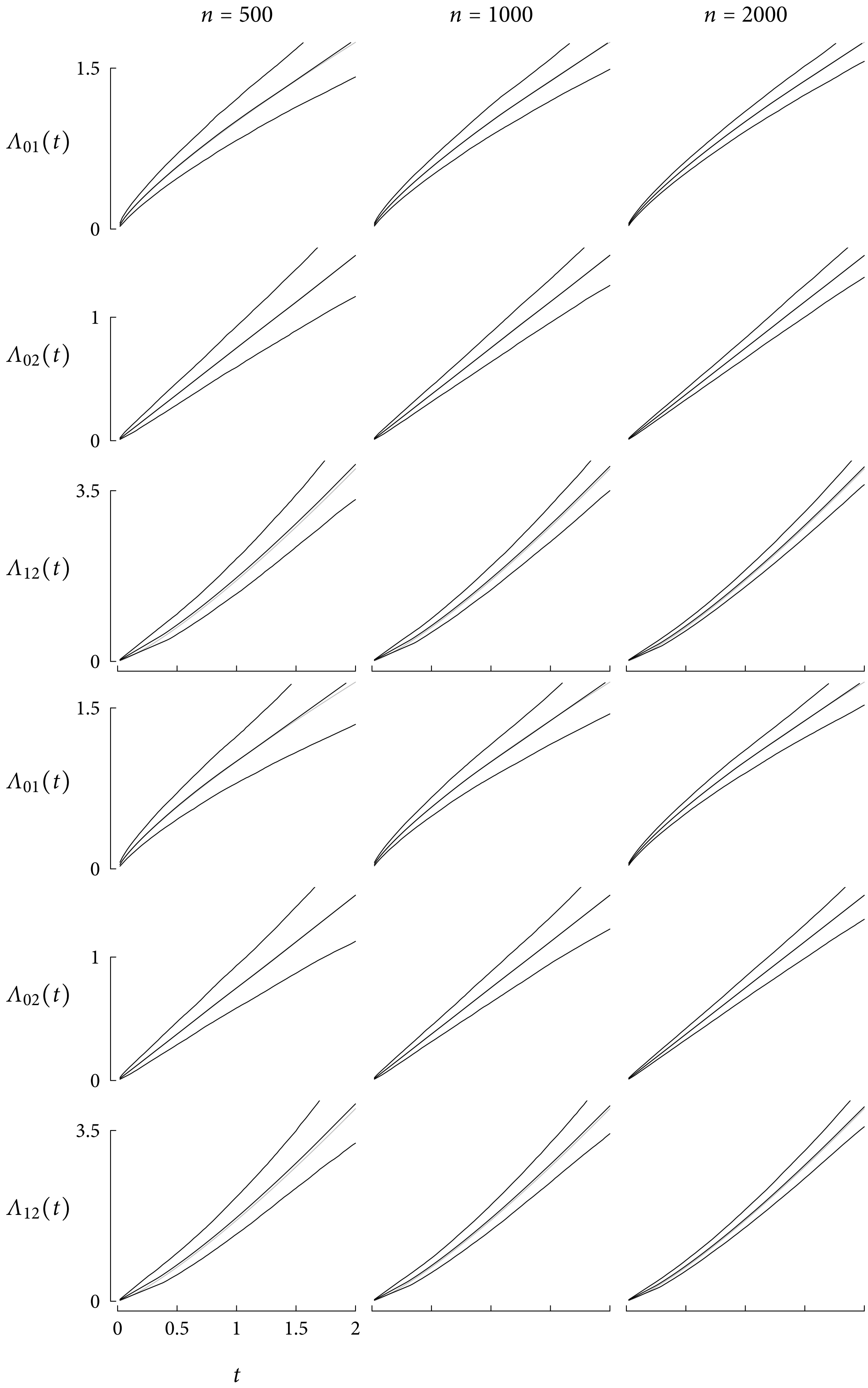


figure 3.3 True values for Λ (–) depicted with pointwise lower and upper 2.5th percentiles and pointwise means for the spmle of Λ with $\kappa = 2/5$ based on 3000 replicates under $e^{\beta} = (1/3, 3/2)$ (top) and $e^{\beta} = (9, 1)$ (bottom).

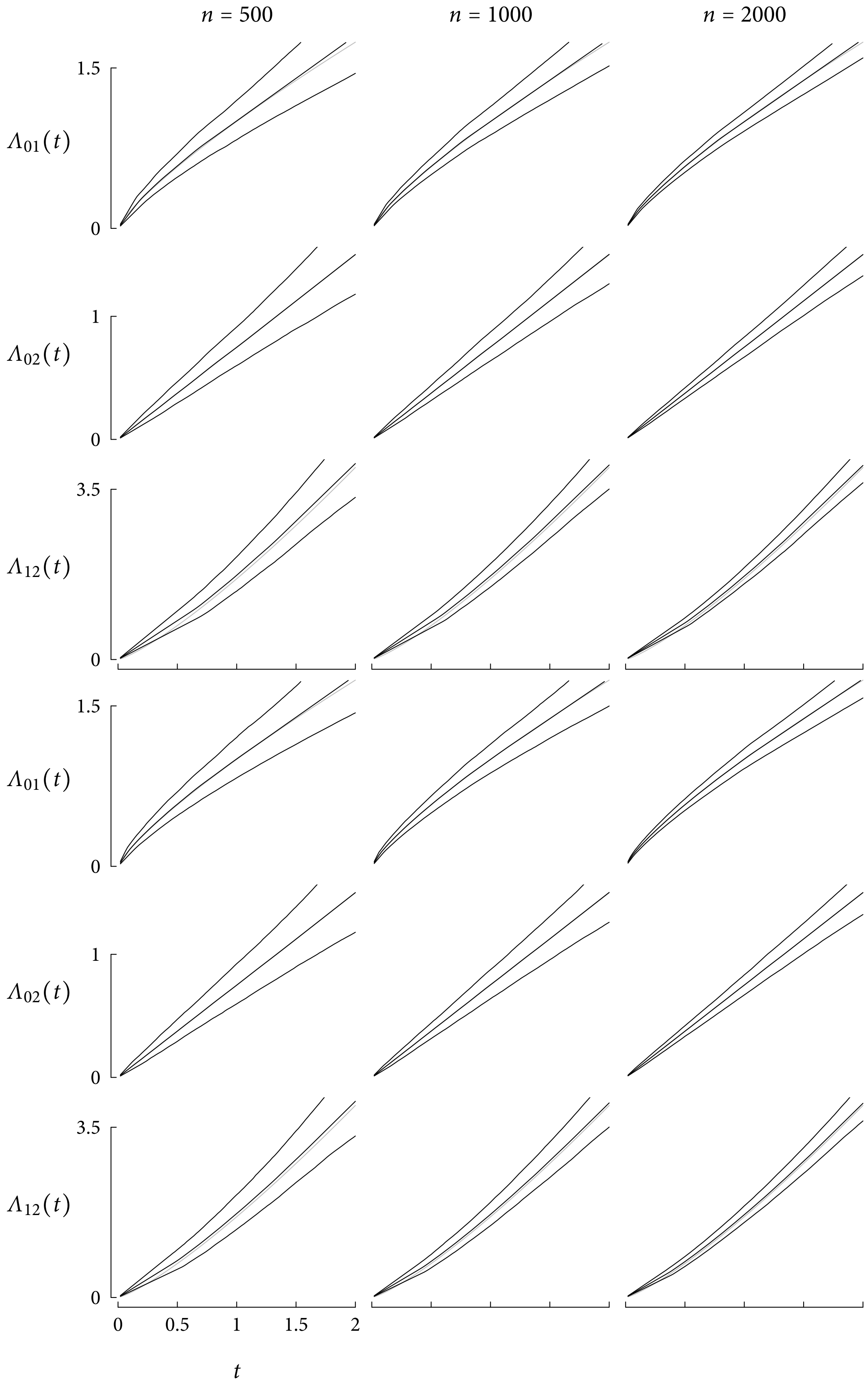


FIGURE 3.4 True values for Λ (–) depicted with pointwise lower and upper 2.5th percentiles and pointwise means for the SPMLE of Λ with $\kappa = 4/15$ (top) and $\kappa = 1/3$ (bottom) based on 3000 replicates under $e^{\beta} = (1/3, 1)$.

TABLE 3.5 Average time to convergence for Algorithm 3.9. The CPU time, given in seconds, covers variance estimation.

	$e^{\beta_0} = 1/3$						$e^{\beta_0} = 9$	
	$\kappa = 4/15$		$\kappa = 1/3$		$\kappa = 2/5$		$\kappa = 2/5$	
n	CPU	Iterations	CPU	Iterations	CPU	Iterations	CPU	Iterations
500	0.23	30	0.32	30	0.41	30	1.33	48
1000	0.45	30	0.65	30	0.98	30	3.36	49
2000	1.06	29	1.63	30	2.87	29	9.89	50

both parameter and variance estimation (Table 3.5). The latter was however below 10 seconds on average in all of the scenarios and sample sizes considered.

3.6 APPLICATION

In 2001 zoledronic acid, a later-generation bisphosphonate, was reported to demonstrate at least equivalent efficacy and safety versus pamidronate, the standard treatment in the prevention of bone lesion complications at the time (e.g. Major et al. 2001; Rosen et al. 2001). Rosen et al. (2001) evaluated an international phase III double-blinded comparative trial of the two bisphosphonates in 1,648 patients with breast cancer or multiple myeloma and at least one bone lesion. Skeletal-related events (SREs), including fractures, spinal cord compression, radiation to bone and bone surgery, were recorded up to 30 months following randomization (Figure 3.5). Since the primary endpoint for the trial considered the occurrence of at least one SRE over 13 months, follow-up for SREs ended before 30 months for many subjects. The

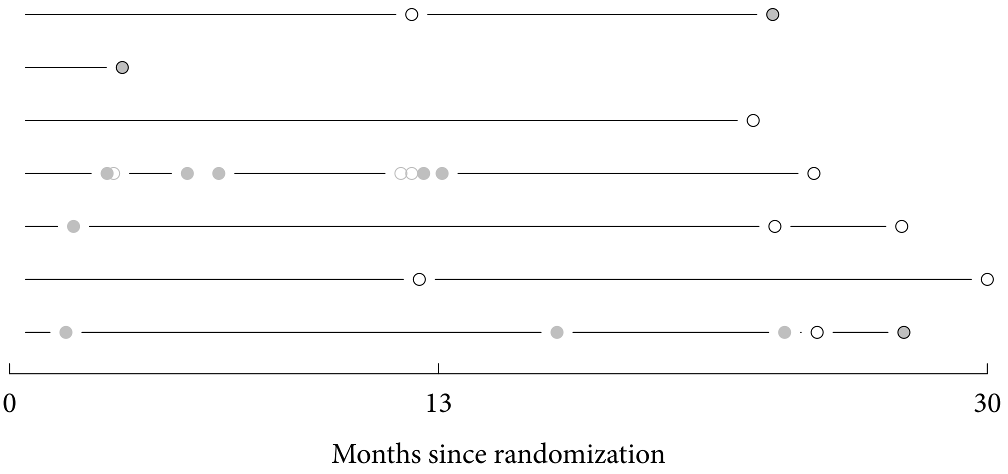


FIGURE 3.5 Occurrence of skeletal-related events (○, ●), right-censoring (○) and death (●). Only the first radiation or surgery to bone (●) is considered in the analysis.

data over the longer observation period are thus doubly right-censored. This section illustrates application of the SMLE to compare the effect of the bisphosphonates on both time to the first bone intervention and survival in a subgroup of 777 breast cancer patients from the North American contingent of the trial. "Progression" here corresponds to the composite event coinciding with the earliest radiation to bone or bone surgery.

A new R function, called `dcprog`, was devised to give a user interface to the C estimation routine described in Section 3.5. Regression models are specified in a manner similar to Therneau's (2012) `survival` package for R. Observations depicted in Figure 3.5 are represented using the R data frame

```
   id start stop from to status z z01 z02 z12
1   1     0  360    0  1      0 0   0   0   0
2   1     0  360    0  2      0 0   0   0   0
3   1   360  690   NA  2      1 0   0   0   0
4   2     0   90    0  1      0 0   0   0   0
5   2     0   90    0  2      1 0   0   0   0
6   3     0  660    0  1      0 0   0   0   0
7   3     0  660    0  2      0 0   0   0   0
8   4     0   90    0  1      1 0   0   0   0
9   4     0   90    0  2      0 0   0   0   0
10  4    90  720    1  2      0 0   0   0   0
11  5     0   30    0  1      1 1   1   0   0
12  5     0   30    0  2      0 1   0   1   0
13  5    30  810    1  2      0 1   0   0   1
14  6     0  360    0  1      0 0   0   0   0
15  6     0  360    0  2      0 0   0   0   0
16  6   360  900   NA  2      0 0   0   0   0
17  7     0   30    0  1      1 0   0   0   0
18  7     0   30    0  2      0 0   0   0   0
19  7    30  810    1  2      1 0   0   0   0
```

where the transition types are indicated in the variables `from` and `to`. Left- and right-endpoints of the at-risk time intervals for a given type are measured in days by the variables `start` and `stop`, respectively. The variable `status` indicates the occurrence of a transition at `stop`. Terminal events having unknown transition type are recorded in the same manner, but with the `from` variable set to the missing value `NA`. Any type-specific covariates for these cases should provide values corresponding to the $1 \rightarrow 2$ transition. This data format ensures that no modification is needed to fit the SPMLE from singly right-censored data at C via the well-known `survival` function `coxph` under R's default `NA` action, `na.omit`. Note that the cases shown here do not represent an excerpt from the actual data set; values have been rounded or randomized for confidentiality.

In the code fragment below the `start`, `stop` and `status` variables are combined into a response using a `counting`-type `Surv` object.

```
> fit <- dcprog(Surv(start, stop, status) ~ cluster(id)
               + trans(from, to) + I(z * (to==1))
               + I(z * (from %in% 0 & to==2))
               + I(z * (from %in% c(NA,1) & to==2)), data = p10,
               sieve.const = 1, sieve.rate = 2/5, eps = 1e-9,
               coef.typ = 1/2, coef.max = 2)
```

Two special terms should appear in any model: `cluster` and `trans`. The variable passed to the `survival` function `cluster` should uniquely identify individuals in the sample. The new function `trans` is akin to the `survival` function `strata` in that `trans` indicates transition type terms, but it also extracts information used to determine the state structure and perform data checks. The remaining model terms specify the covariates. Calculation of $0 \to 1$, $0 \to 2$ and $1 \to 2$ type-specific copies of the zoledronic acid treatment indicator `z` are explicitly shown here using the as-is function `I`.

Additional option settings specify values for tuning parameters. A sieve growing at the rate $C_{hj}n^{\kappa} = n^{2/5}$ is obtained with `sieve.const = 1` and `sieve.rate = 2/5`. The remaining parameters, fully described in Section 2.6, control variance estimation. By default `dcprog` also fits the same model to singly right-censored data at C. This is essentially done by passing the same model formula and data frame to `coxph`, but with the `trans` term replaced by `strata` and all rows in the data frame having `from` equal to `NA` excluded. The `dcprog` function returns an object of the type `dcprog`. Its `print` routine reproduces the function call, summarizes regression coefficients, reports the initial and final log-likelihood values and gives the rates at which exact transition times (S, T) and the progression status $1(S < T)$ were observed in the sample.

The output below shows that the log-likelihood Equation (3.6) at the initial parameter value $\phi^{(0)}$ is -4656. After $r = 48$ iterations the log-likelihood at the final parameter value $\phi^{(r)}$ is -4533. The exact transition times (S, T) are observed for only 33% of the 777 individuals in the sample. Incomplete transition times but known progression status $1(S < T)$ are available from 29%. Observations for the remaining 38% are doubly right-censored; both exact times and progression status are unavailable. The $0 \to 1$, $0 \to 2$ and $1 \to 2$ type-specific regression coefficients $\hat{\theta}_n$ measuring the increase in the risk of transition associated with receiving zoledronic acid versus pamidronate are -0.173, 0.189 and -0.101, respectively. None of these differences are statistically significant, with p-values for the two-sided test $\theta_j = 0$ no smaller than 18%.

```
> fit
Call:
dcprog(formula = Surv(start, stop, status) ~ cluster(id)
    + trans(from, to) + I(z * (to == 1)) + I(z * (from %in% 0 & to == 2))
    + I(z * (from %in% c(NA, 1) & to == 2)), data = p10,
    ... = list(sieve.const = 1, sieve.rate = 2/5, eps = 1e-09,
               coef.typ = 1/2, coef.max = 2))

                                        coef se(coef)      z    p
I(z * (to == 1))                      -0.173    0.148 -1.166 0.24
I(z * (from %in% 0 & to == 2))         0.189    0.141  1.344 0.18
I(z * (from %in% c(NA, 1) & to == 2)) -0.101    0.199 -0.507 0.61
                                      exp(coef)  2.5% 97.5%
I(z * (to == 1))                          0.841 0.629  1.12
I(z * (from %in% 0 & to == 2))            1.209 0.917  1.59
I(z * (from %in% c(NA, 1) & to == 2))     0.904 0.612  1.34

Based on n = 777 subjects contributing 2231 observation times

Initial log-likelihood: -4655.97
Log-likelihood after 48 iterations: -4533

                 (S, T) 1(S < T) Neither
Observation rate  0.329     0.29   0.381

Estimation from right-censored data via survival's coxph function

Formula:
Surv(start, stop, status) ~ cluster(id) + strata(from, to) +
    I(z * (to == 1)) + I(z * (from %in% 0 & to == 2)) + I(z *
    (from %in% c(NA, 1) & to == 2))

                                           coef se(coef)       z
I(z * (to == 1))                      -0.133455    0.148 -0.90090
I(z * (from %in% 0 & to == 2))         0.156815    0.283  0.55368
I(z * (from %in% c(NA, 1) & to == 2)) -0.000385    0.146 -0.00264
                                         p exp(coef)  2.5% 97.5%
I(z * (to == 1))                      0.37     0.875 0.655  1.17
I(z * (from %in% 0 & to == 2))        0.58     1.170 0.671  2.04
I(z * (from %in% c(NA, 1) & to == 2)) 1.00     1.000 0.751  1.33

Based on 1935 observation times (296 deleted due to missingness)
```

The `print` function also provides the same summary of the regression coefficients obtained from `coxph` fit to the singly right-censored data at C. The estimate for the effect on the 1 → 2 transition intensity is markedly different from the SMLE, but none of the coefficients were found to be significantly different from zero. Additional output indicates that 296 of the 2231 rows in data frame have a missing `from` value and were thus excluded from analysis. This corresponds to 38% of the sample having doubly right-censored data.

Values for the sieve estimates can be accessed directly from the `dcprog` object's list arguments `coef` and `bhaz`. The SPMLE from singly right-censored data at C are similarly represented, but the corresponding `coef` vector and `bhaz` data frame are nested in the `dcprog` list argument `rcfit`.

```
> fit$coef
                     I(z * (to == 1))
                           -0.1728308
       I(z * (from %in% 0 & to == 2))
                            0.1894571
I(z * (from %in% c(NA, 1) & to == 2))
                           -0.1009796
> fit$bhaz[1:3, ]
      hazard time        trans
1 0.00000000    0 from=0, to=1
2 0.05814021   31 from=0, to=1
3 0.08540514   57 from=0, to=1

> fit$rcfit[[1]]$coef
                     I(z * (to == 1))
                        -0.1334553657
       I(z * (from %in% 0 & to == 2))
                         0.1568147185
I(z * (from %in% c(NA, 1) & to == 2))
                        -0.0003853867

> fit$rcfit[[1]]$bhaz[1:3, ]
       hazard time        trans
1 0.000000000 0.01 from=0, to=1
2 0.007047897 1.00 from=0, to=1
3 0.009880123 2.00 from=0, to=1
```

Results with alternative sieve sizes can be investigated by refitting the model with different parameters. The above listing shows that similar estimates are achieved with $C_{hj}n^{\kappa} = n^{1/3}$.

```
> dcprog(Surv(start, stop, status) ~ cluster(id)
         + trans(from, to) + I(z * (to==1))
         + I(z * (from %in% 0 & to==2))
         + I(z * (from %in% c(NA,1) & to==2)), data = p10,
         sieve.const = 1, sieve.rate = 1/3, eps = 1e-9,
         coef.typ = 1/2, coef.max = 2)
Call:
dcprog(formula = Surv(start, stop, status) ~ cluster(id)
    + trans(from, to) + I(z * (to == 1)) + I(z * (from %in% 0 & to == 2))
    + I(z * (from %in% c(NA, 1) & to == 2)), data = p10,
    ... = list(sieve.const = 1, sieve.rate = 1/3, eps = 1e-09,
               coef.typ = 1/2, coef.max = 2))

                                         coef se(coef)      z    p
I(z * (to == 1))                       -0.177    0.149 -1.190 0.23
I(z * (from %in% 0 & to == 2))          0.194    0.143  1.360 0.17
I(z * (from %in% c(NA, 1) & to == 2)) -0.111    0.200 -0.556 0.58
                                      exp(coef)  2.5% 97.5%
I(z * (to == 1))                          0.838 0.626  1.12
I(z * (from %in% 0 & to == 2))            1.214 0.918  1.61
I(z * (from %in% c(NA, 1) & to == 2))     0.895 0.605  1.32

Based on n = 777 subjects contributing 2231 observation times

Initial log-likelihood: -4655.98
Log-likelihood after 46 iterations: -4545.54

                 (S, T) 1(S < T) Neither
Observation rate  0.329     0.29   0.381

Estimation from right-censored data via survival’s coxph function

Formula:
Surv(start, stop, status) ~ cluster(id) + strata(from, to) +
    I(z * (to == 1)) + I(z * (from %in% 0 & to == 2)) + I(z *
    (from %in% c(NA, 1) & to == 2))

                                           coef se(coef)        z
I(z * (to == 1))                      -0.133455    0.148 -0.90090
I(z * (from %in% 0 & to == 2))         0.156815    0.283  0.55368
I(z * (from %in% c(NA, 1) & to == 2)) -0.000385    0.146 -0.00264
                                         p exp(coef)  2.5% 97.5%
I(z * (to == 1))                      0.37     0.875 0.655  1.17
I(z * (from %in% 0 & to == 2))        0.58     1.170 0.671  2.04
I(z * (from %in% c(NA, 1) & to == 2)) 1.00     1.000 0.751  1.33

Based on 1935 observation times (296 deleted due to missingness)
```

To examine estimates achieved with starting values different from the ones described in Algorithm 3.9, the dcprog function can accept alternative initial parameter values via the input argument init. Initial values can also be obtained from the SPMLE under singly right-censored data using the option setting rcinit = TRUE. With the latter approach the difference in the resulting sieve estimates is of the order 10^{-7}.

```
> fit.rcinit <- dcprog(Surv(start, stop, status) ~ cluster(id)
                  + trans(from, to) + I(z * (to==1))
                  + I(z * (from %in% 0 & to==2))
```

```
                    + I(z * (from %in% c(NA,1) & to==2)), data = p10,
                    rcinit = TRUE, sieve.const = 1, sieve.rate = 2/5,
                    eps = 1e-9, coef.typ = 1/2, coef.max = 2)
> max(abs(fit.rcinit$coef - fit$coef))
[1] 3.344081e-07
> max(abs(fit.rcinit$bhaz - fit$bhaz))
[1] 1.584767e-07
```

Estimates for the cumulative baseline intensity functions are fully depicted in Figure 3.6. The SPMLE ignoring any terminal event following C gives smaller estimates for both Λ_{02} and Λ_{12} than the SMLE. Although both estimators are asymptotically consistent we should expect to see differences between the two in finite samples, particularly under high rates of doubly right-censored data.

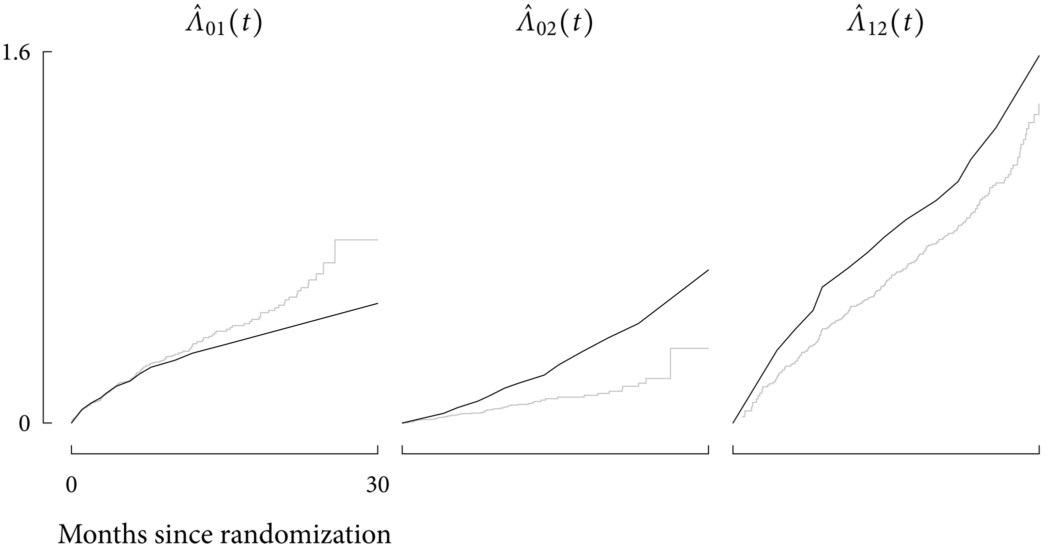


FIGURE 3.6 The SPMLE based on data right-censored at C (–) and SMLE (–) for the cumulative baseline transition intensity functions.

Interpretation of the results here is limited by the fact that the primary endpoint considered only the first 13 months of follow-up, so attrition later in the follow-up period may be related to the transition times. Another consideration is the validity of two assumptions: proportional hazards and the Markov property. With the availability of singly right-censored data, these properties can be examined by existing methods. The `survival` function `cox.zph` implements Grambsch and Therneau's (1994) test of the proportional hazards assumption based on the correlation between the event times, under some specified transformation, and the scaled Schoenfeld residuals. The test with log-transformed times

```
> cox.zph(fit$rcfit[[1]],transform='log')
```

```
                                           rho chisq     p
I(z * (to == 1))                        0.0413 0.781 0.377
I(z * (from %in% 0 & to == 2))         -0.0409 0.745 0.388
I(z * (from %in% c(NA, 1) & to == 2))  0.0163 0.115 0.734
GLOBAL                                      NA 1.598 0.660
```

indicates that the correlation is not significantly different from zero. The same test under alternative transformations gives similar results. Following Andersen et al. (2000), one method of assessing departures from Markovity is to include the duration in the intermediate state 1 as a time-dependent covariate. This can be achieved by calculating the total duration in a new variable `wait` and refitting the SPMLE with total duration as a time-transformed term `tt(wait)`.

```
> p10$wait <- with(p10, (from == 1) * (stop - start))
> coxph(Surv(start, stop, status) ~ cluster(id)
             + strata(from, to) + tt(wait)
             + I(z * (to==1)) + I(z * (from == 0 & to==2))
             + I(z * (from == 1 & to==2)), data = p10)
Call:
coxph(formula = Surv(start, stop, status) ~ cluster(id) + strata(from,
    to) + tt(wait) + I(z * (to == 1)) + I(z * (from == 0 & to ==
    2)) + I(z * (from == 1 & to == 2)), data = p10)


                                coef exp(coef) se(coef) robust se
tt(wait)                    -0.79645     0.451    0.111     0.448
I(z * (to == 1))            -0.13346     0.875    0.148     0.148
I(z * (from == 0 & to == 2))  0.15681     1.170    0.285     0.283
I(z * (from == 1 & to == 2)) -0.00205     0.998    0.150     0.157
                                 z     p
tt(wait)                    -1.779 0.075
I(z * (to == 1))            -0.901 0.370
I(z * (from == 0 & to == 2))  0.554 0.580
I(z * (from == 1 & to == 2)) -0.013 0.990

Likelihood ratio test=147  on 4 df, p=0  n= 1935, number of events= 451
   (296 observations deleted due to missingness)
```

From the output it is apparent the influence of duration on mortality is large, with an increased risk of death soon after the first bone intervention. The statistical significance of this effect is weak, with a p-value of 7.5%, but we should not be confident that Markov assumption is met. Since the observed data likelihood (3.6) is constructed on the basis of Theorem 1.4, the SMLE may not be robust to departures from the Markov property. In cases where this condition cannot be reasonably justified, duration-dependent models based on the right-censored data at C offer a practical alternative.

CHAPTER 4

INTERVAL-CENSORED DATA FROM AN ILLNESS-DEATH PROCESS

Use of the illness-death model in dealing with both interval-censored progression times and right-censored terminal events was first proposed by Frydman (1995). Progression status was assumed known, a requirement that can imposed on the available data by carrying the last-observed progression status forward to death or right-censoring. This form of imputation generally known as “last observation carried forward” (LOCF) is of course prone to misclassification when the last negative inspection occurs long before the end of follow-up. Such a scenario is addressed in later works. Joly et al. (2002) construct spline-based penalized maximum likelihood estimators for Markov cumulative transition intensity functions in the general case where progression status remains unknown unless it is confirmed by a positive inspection (Example 1.30). Frydman and Szarek (2009) estimate the subdistribution functions F_{01} and F_{02} and the cumulative intensity Λ_{12} (Figure 4.1) also under Example 1.30, but allow for the possibility that negative progression status can sometimes be confirmed at right-censoring or death.

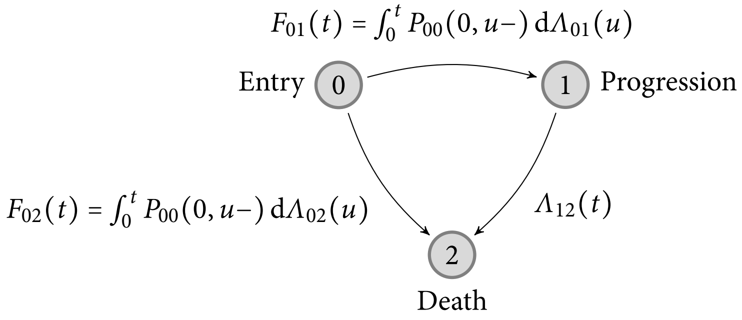

$F_{01}(t) = \int_0^t P_{00}(0, u-)\, \mathrm{d}\Lambda_{01}(u)$
Entry
0
1
Progression
$F_{02}(t) = \int_0^t P_{00}(0, u-)\, \mathrm{d}\Lambda_{02}(u)$
$\Lambda_{12}(t)$
2
Death

FIGURE 4.1 Subdistribution functions in the Markov illness-death model.

This final chapter examines estimation of a Cox-type Markov model under a variant of Frydman and Szarek’s (2009) scheme that assumes negative progression status is confirmed at terminal events for some positive proportion of the sample. Estimates in the case with constant transition intensities and time-dependent covariates are easily obtained from Kalbfleisch and Lawless’s (1985) Fisher scoring algorithm, implemented by `msm` package for R (Jackson 2011). Here the results of Chapter 3 are extended to obtain a flexible piecewise exponential sieve estimator.

4.1 MODEL AND OBSERVATION SCHEME

Considering the same semiparametric model from Chapter 3, assume that each component in the Markov illness-death process $N = (N_{01}, N_{02}, N_{12})$ has intensity process $Y_h \alpha_{hj}$ with

$$\alpha_{hj}(t \mid Z) = \lambda_{hj}(t) \exp(Z_{hj}^{\top} \theta), \tag{4.1}$$

where Z_{hj} is a type-specific d-vector based on the fixed covariate Z, θ is a d-variate regression parameter and $\Lambda_{hj} = \int \lambda_{hj}$ is a nondecreasing baseline intensity function.Improvement in bias

Let $T_{hj} = \inf\{t : N_{hj}(t) = 1\}$ denote the $h \to j$ transition time. Suppose that the time of death $T = T_{02} \wedge T_{12}$ is observed exactly whenever it precedes the right-censoring time C, $0 < \sigma \leq C \leq \tau$, but the progression status $1(S < T)$, $S = T_{01} \wedge T_{02}$, is detected by the inspection times $\mathbf{Y}_K = (Y_{K,1}, \ldots, Y_{K,K})$ where

$$Y_{K,0} \equiv 0 < \sigma < Y_{K,1} < \cdots < Y_{K,K} < \infty \equiv Y_{K,K+1}$$

and $1 \leq K < \infty$. A complete set of inspections at $\mathbf{Y}_K$ gives $\mathbf{\Delta}_K = (\Delta_{K,1}, \ldots, \Delta_{K,K})$ with $\Delta_{K,j} = 1_{(Y_{K,j}, Y_{K,j+1}]}(S)$. Any inspections following $V = T \wedge C$ are presumed unavailable, though additional information about S may be assessed at V. Denote this by $(\Delta, \Delta_1, \Delta_2)$, where $\Delta_2 = 1(T \leq C)$, $\Delta_1 = 1(S < V)$ if $\Delta = 1$ and $\Delta_1 = 0$ when $\Delta = 0$. Under this observation scheme the progression status $1(S < T)$ is known from the observation X provided that either a positive inspection precedes V or the status is assessed, $\Delta = 1$, at V (Figure 4.2). Otherwise the progression status is unknown over the time interval from the last negative inspection preceding V up to V (Figure 4.2, bottom).

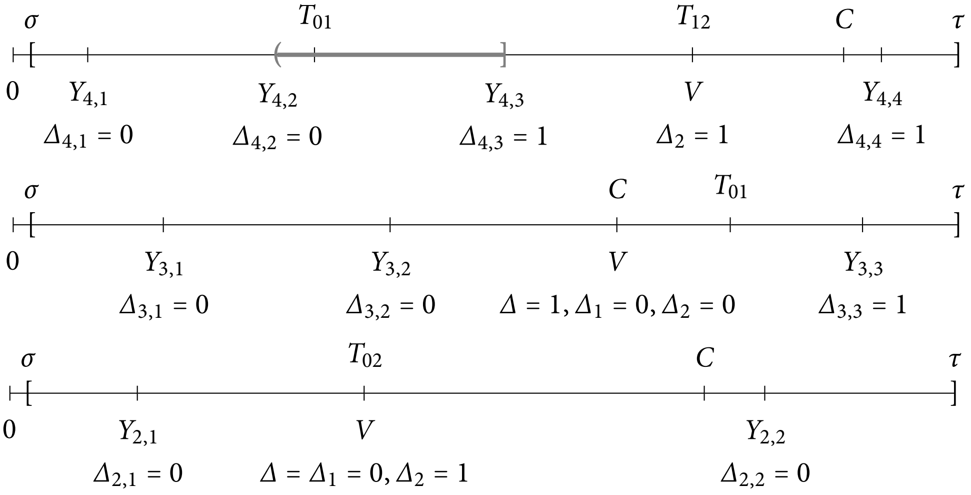


FIGURE 4.2 Under intermittent inspection of an illness-death process, the progression status $1(S < T)$ may be known (top, middle) or unavailable (bottom) from the observed data.

Let $(\mathbf{Y}_K(V), \boldsymbol{\Delta}_K(V))$ be the observed part of $(\mathbf{Y}_K, \boldsymbol{\Delta}_K)$, $X = (K, \mathbf{Y}_K(V), \boldsymbol{\Delta}_K(V), V, \Delta, \Delta_1, \Delta_2, Z)$ and $\mathbf{Y}$ denote the triangular array of "potential" inspection times $\{Y_{k,j} : j = 1, \ldots, k, k = 1, 2, \ldots\}$. Assume the following:

C1 Suppose that $\Delta = g(X, \Gamma)$ where g is a known function and Γ is a random variable such that (T_{01}, T_{02}, T_{12}) is conditionally independent of $(K, \mathbf{Y}, \Gamma, C)$ given Z.

Then under Theorem 1.4 and Λ absolutely continuous, the realization $X = x = (k, \mathbf{y}_k(v), \boldsymbol{\delta}_k(v), v, \delta, \delta_1, \delta_2, z)$ has density

$$\begin{aligned}
&p_{\theta,\Lambda}(x) \\
&= \prod_{j=1}^{k} \Big[P_{00}(0, y_{k,j} \mid z) P_{01}(y_{k,j}, y_{k,j+1} \wedge v \mid z) P_{11}(y_{k,j+1} \wedge v, v \mid z) \\
&\qquad\quad \times \alpha_{12}(v \mid z)^{\delta_2}\Big]^{\delta_{k,j} 1_{(0,v)}(y_{k,j+1}) \vee \delta\delta_1 1_{(y_{k,j}, y_{k,j+1}]}(v)} \\
&\quad \times \prod_{j=1}^{k} \Big[P_{00}(0, y_{k,j} \mid z) P_{01}(y_{k,j}, y_{k,j+1} \wedge v \mid z) P_{11}(y_{k,j+1} \wedge v, v \mid z) \alpha_{12}(v \mid z)^{\delta_2} \\
&\qquad\qquad + P_{00}(0, v \mid z) \alpha_{02}(v \mid z)^{\delta_2}\Big]^{(1-\delta)(1-\delta_{k,j}) 1_{(y_{k,j}, y_{k,j+1}]}(v)} \\
&\quad \times \big[P_{00}(0, v \mid z) \alpha_{02}(v \mid z)^{\delta_2}\big]^{\delta(1-\delta_1)},
\end{aligned} \tag{4.2}$$

with respect to a dominating measure ν determined by the distribution of $(K, \mathbf{Y}, \Gamma, C)$. The same notation from Section 3.1 is used here: $A_{hh} = -\sum_{j \neq h} A_{hj}$, $A_{hj} = \int \alpha_{hj}$, P_{hh} is defined by (3.2) and P_{01} by (3.3).

4.1 REMARK. The expression in (4.2) can be rewritten as a product of transition probabilities accumulating over the subintervals in $\mathbf{y}_k(v)$. This is a valid likelihood under a noninformative censoring mechanism similar to the one considered in Definition 1.32. The stronger requirement in C1 simplifies the derivation of asymptotic properties. It may be motivated by the setting in which progression status before V is inspected according to a predetermined schedule, with the completion and exact timing of assessments determined by some random process related to (S, T) only via Z. Whether or not additional inspection occurs at V is essentially known from X with any extra uncertainty following some mechanism that depends on (S, T) only through $(K, \mathbf{Y}, C, Z)$. For example, the form $\Delta = (1 - \Delta_2)\Gamma_C + \Delta_2 \Gamma_T$ and $\Gamma = (\Gamma_C, \Gamma_T)$ allows for different rates of inspection depending on the terminal event status at V. ▫

4.2 SIEVE MAXIMUM LIKELIHOOD ESTIMATION

Let $X_i = (K_i, \mathbf{Y}^i_K(V_i), \boldsymbol{\Delta}^i_K(V_i), V_i, \Delta_i, \Delta^i_1, \Delta^i_2, Z_i)$, $i = 1, \ldots, n$, be n iid observations of X from (θ_0, Λ_0), $\Lambda_0 = (\Lambda^0_{hj})$ for $h \neq j$. As with the density in Chapter 3, $\mathbb{P}_n \log p_{\theta,\Lambda}$

maximizes to infinity. An empirical-type likelihood is equally problematic as the increment in Λ_{02} and Λ_{12} cannot be jointly estimated at any $V_i = T_i$ for which the progression status is unknown, $\max\{\mathbf{\Delta}_K^i(V_i), \Delta_i\} = 0$. So unless the progression status is observed at every terminal event, no unbiased non- or semiparametric maximimum likelihood estimators exist. Under at least moderate levels of censoring one might expect to underestimate Λ_{01} and Λ_{12}. This seems apparent for Λ_{12} based on numerical results reported by Frydman and Szarek (2009, Table 1) and Szarek (2008, p. 125). Joly et al.'s (2002) smooth estimator may mitigate this problem, but their simulation study is void of replication and thus inconclusive. Bias can be fully eliminated by smoothing over the survival times with *known* progression status. For this task a more restrictive observation scheme is imposed in order to adapt the piecewise exponential sieve maximum likelihood estimator from Chapter 3.

C2 There exist $0 < \sigma < \tau$ and $0 < M < \infty$ such that $1/M < \Lambda_{hj}^0(\sigma-) < \Lambda_{hj}^0(\tau) < M$, $h \neq j$, and Λ_0 is continuously differentiable on $[\sigma, \tau]$.

C3 Let n_{h2} denote the number of individuals in the sample with $T_{h2} < \infty$ observed exactly, $h = 0, 1$. Also let n_{01} be the number for whom $T_{01} < \infty$ is left- or interval-censored; that is, $n_{01} = \sum_i \max\{\mathbf{\Delta}_K^i(V_i), \Delta_1^i\}$. Then there exist $q_{hj} > 0$ such that $n_{hj}/n \to q_{hj}$ as $n \to \infty$.

C4 The interval (σ, τ) is a subset of the combined support of the inspection times.

4.2 DEFINITION. Under condition C3 consider the SPMLE based on the observed data from the $\sum_{h \neq j} n_{hj}$ individuals for whom the progression status is known. Let $\mathcal{I} = \{(L_j, R_j] : j = 1, \ldots, n_{01}\}$ denote the set of censoring intervals among the n_{01} individuals known to have progressed; that is,

$$\begin{aligned}\mathcal{I} &= \{(L_j, R_j] : j = 1, \ldots, n_{01}\} \\ &= \{(Y_{K_i,j}^i, Y_{K_i,j+1}^i \wedge V_i] : \Delta_{K_i,j}^i 1_{(0,V_i)}(Y_{K_i,j+1}) \vee \Delta_i \Delta_1^i 1_{(Y_{K_i,j}, Y_{K_i,j+1}]}(V_i) = 1\}.\end{aligned}$$

From a straight-forward adaptation of Definition 2.4 and Proposition 2.5 the $h \to j$ component of this estimator increases on the set $\mathcal{U}_{hj}$ given by one of

$$\begin{aligned}&\mathcal{U}_{01} \subseteq \{R_j : R_j \in \mathcal{I}\}, \\ &\mathcal{U}_{02} = \{V_i : \max\{\mathbf{\Delta}_K^i(V_i)\} = 0, \Delta_i = 1, \Delta_1^i = 0, \Delta_2^i = 1\}, \\ &\mathcal{U}_{12} = \{V_i : \max\{\mathbf{\Delta}_K^i(V_i), \Delta_1^i\} = 1.\Delta_2^i = 1\},\end{aligned}$$

where the superset for $\mathcal{U}_{01}$ can be further reduced to the right-endpoints of the maximal intersections of $\mathcal{I}$. □

Let $H = (H_{hj})$, $h \neq j$, denote the set of $\Lambda = (\Lambda_{hj})$ with each $\Lambda_{hj} : [0, \tau] \to [0, M]$ cadlag and nondecreasing. Consider the following finite-dimensional approximation to H.

4.3 DEFINITION. For each $h \to j$, $h \neq j$, let $\mathcal{T}_{hj,n}$ be a set containing the $K_{hj,n} = O(n^{\kappa})$ points in $(0, \tau)$ from the partition

$$0 = t_0 < t_1 < \cdots < t_{K_{hj,n}} < t_{K_{hj,n}+1} = \tau,$$

defined so that every subinterval $[t_{k-1}, t_k)$ contains at least one element from $\mathcal{U}_{hj}$ and $\max_k(t_k - t_{k-1}) = O(n^{-\kappa})$. For every $\Lambda_{hj} \in H_{hj}$ let $\Lambda_{hj,n}$ denote its piecewise linear interpolant given by (3.5). ▫

4.4 REMARK. This partition essentially corresponds to the support of the SPMLE based on the observations with known progression status, previously characterized by Frydman (1995). A requisite of Definition 4.3 is that the inspection times become dense (σ, τ) as $n \to \infty$, a condition implied by C4. ▫

From C3, Λ_{02} and Λ_{12} are jointly estimable by maximizing

$$\begin{aligned} \log \operatorname{lik}_n(\theta, \Lambda) &= \sum_{i=1}^{n} \log p_{\theta,\Lambda}(X_i) \\ = \sum_{j=1}^{k} &[\Delta_{k,j}^i 1_{(0,V_i)}(Y_{k,j+1}^i) \vee \Delta_i \Delta_1^i 1_{(Y_{k,j}^i, Y_{k,j+1}^i]}(V_i)][A_{00}(0, Y_{k,j}^i \mid Z_i) \\ &\quad + \log P_{01}(Y_{k,j}^i, Y_{k,j+1}^i \wedge V_i \mid Z_i) - A_{12}(Y_{k,j+1}^i \wedge V_i, V_i \mid Z_i) + \Delta_2^i \log \alpha_{12}(V_i)] \\ &+ \sum_{j=1}^{k} (1 - \Delta_i)(1 - \Delta_{k,j}^i) 1_{(Y_{k,j}^i, Y_{k,j+1}^i]}(V_i)[A_{00}(0, Y_{k,j}^i \mid Z_i) \\ &\qquad + \log P_{01}(Y_{k,j}^i, Y_{k,j+1}^i \wedge V_i \mid Z_i) - A_{12}(Y_{k,j+1} \wedge V_i, V_i \mid Z_i) + \Delta_2^i \log \alpha_{12}(V_i \mid Z_i) \\ &\qquad + A_{00}(V_i \mid Z_i) + \Delta_2^i \log \alpha_{02}(V_i \mid Z_i)] \\ &+ \Delta_i (1 - \Delta_1^i)[A_{00}(V_i \mid Z_i) + \Delta_2^i \log \alpha_{02}(V_i \mid Z_i)]. \end{aligned} \tag{4.3}$$

over the sieve $H_n = (H_{hj,n})$, $H_{hj,n} = \{\Lambda_{hj,n} : \Lambda \in H_{hj}\}$. Let Θ denote the set of all possible θ. Then the piecewise exponential *sieve maximum likelihood estimator* (SMLE) satisfies

$$\log \operatorname{lik}_n(\hat{\theta}_n, \hat{\Lambda}_n) = \max_{\theta \in \Theta, \Lambda \in H_n} \log \operatorname{lik}_n(\theta, \Lambda). \tag{4.4}$$

This optimization problem is well-defined and has finite dimension. As with Chapter 3 its solution is characterized by the score equations, which can be solved by the following variant of Algorithm 3.9.

4.5 ALGORITHM. Let n_{hj} be the size of $\mathcal{U}_{hj}$. For given $C_{hj} > 0$ and $0 < \kappa < 1$, define $\mathcal{T}_{hj,n}$ as a partition of $[0, \tau)$ in which each subinterval contains $\lceil n_{hj}/(C_{hj} n^{\kappa}) \rceil$ elements from $\mathcal{U}_{hj}$. Set $r := 0$, $\theta^{(0)} = 0$ and $\lambda^{(0)} = 1$. Let $\eta^{(r)}$ be the candidate step with components given by (3.12) and (3.13) and $\phi^{(r+1)}$ be the result of the line search (3.15). If

$$\|\phi^{(r+1)} - \phi^{(r)}\|_{\infty} \leq \varepsilon,$$

for small positive value ε, then stop. Otherwise put $r := r + 1$. ▫

4.3 ASYMPTOTIC PROPERTIES

Under some regularity conditions the sieve maximum likelihood estimator $(\hat{\theta}_n, \hat{\Lambda}_n)$ globally converges to the truth (θ_0, Λ_0) slower than the parametric rate $n^{1/2}$, but $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). These results are derived largely by adapting the proofs from Sections 2.3 and 3.3.

Following Section 2.3.1, define for any $A \in \mathcal{B}[\sigma, \tau]$ and $B \in \mathcal{B}(\mathbb{R}^{d_z})$

$$\mu_{y,z}(A \times B) = \int_B \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid Z = z) \sum_{j=1}^{k} \mathrm{P}(Y_{k,j} \in A \mid Z = z)\, \mathrm{d}F_Z(z),$$
$$\tilde{\mu}_{y,z}(A \times B) = \int_B \sum_{k=1}^{\infty} \mathrm{P}(K = k \mid Z = z) \frac{1}{k} \sum_{j=1}^{k} \mathrm{P}(Y_{k,j} \in A \mid Z = z)\, \mathrm{d}F_Z(z),$$

$\mu_1(A) = \mu_{y,z}(A \times \mathbb{R}^{d_z})$ and μ_2 be the Lebesgue measure on $[\sigma, \tau]$.

C5 θ_0 lies in the interior of Θ and Θ is a compact subset of $\mathbb{R}^{d_z}$.

C6 The distribution for C has support contained in $[\sigma, \tau]$ such that $\mathrm{P}(C = \tau \mid Z) > 0$, almost surely.

C7 The distribution of Z has support $\mathcal{Z} = \mathrm{supp}(F_Z)$ on a bounded subset of $\mathbb{R}^{d_z}$.

C8 For each $h \neq j$, $\mathrm{P}(Z_{hj}^{\top} a \neq c) > 0$ for every $a \in \mathbb{R}^{d_z}$ with $a \neq 0$ and $c \in \mathbb{R}$.

The same arguments used to derive Theorem 3.3 give the following result.

4.6 THEOREM. *Under the above conditions* $\|\hat{\theta}_n - \theta_0\| + \|\hat{\Lambda}_n - \Lambda_0\|_{\mu,2} \overset{\text{as}}{\to} 0$, *where*

$$\|\hat{\Lambda}_n - \Lambda_0\|_{\mu,2} = \sum_{h \neq j} \int_{\sigma}^{\tau} (\hat{\Lambda}_{hj,n} - \Lambda_{hj})^2(u)\, \mathrm{d}\mu_j,$$

is the L_2 distance between $\hat{\Lambda}_n$ and Λ_0 on $\mathrm{supp}(\mu_1) \times (\sigma, \tau)^2$.

The rate at which the SMLE converges to the truth essentially follows by the approach used to prove Theorem 3.5. Some adaptation is needed to allow for convergence in the measure $\mu_1 \times \mu_2 \times \mu_2$. For this we defer to the proof of Lemma 2.11.

C9 For $(Y, Z) \sim \mu_1/\mu_{y,z}([\sigma,\tau] \times \mathcal{Z})$ there exists $0 < \rho < 1$ such that $a^\top \operatorname{Var}(Z_{01})a \leq \rho a^\top \operatorname{E}(Z_{01}Z_{01}^\top)a$, almost surely, for all $a \in \mathbb{R}^{d_z}$.

C10 For some $r \geq 1$, the rth derivative of Λ_0 continuous, positive and bounded on $[\sigma, \tau]$.

4.7 THEOREM. *$\|\hat{\theta}_n - \theta_0\| + \|\hat{\Lambda}_n - \Lambda_0\|_2 = O_P(\max(n^{-(1-\kappa)/2}, n^{-r\kappa}))$ under the conditions above.*

Asymptotic normality is derived by combining the proofs for Theorems 2.13 and 3.7 under the one-dimensional submodels $y \mapsto \Lambda_{01,y}$ and $y \mapsto \Lambda_{h2,y}$ satisfying $g_{01} = \partial/\partial y_{|y=0}\Lambda_{01}$ and $\partial/\partial y_{|y=0}\, \mathrm{d}\Lambda_{h2,y} = g_{h2}\Lambda_{h2}$, respectively.

C11 Λ_0 has a second-order derivative that is uniformly bounded on $[\sigma, \tau]$.

C12 There is a constant $y_0 > 0$ such that $P(Y_{K,j} - Y_{K,j-1} \geq y_0 : j = 1, \ldots, K, Z) = 1$, almost surely.

C13 For $k = 1, 2, \ldots, j = 2, \ldots, k$, the conditional density functions $f_{Y_{k,1}|Z}$, $f_{Y_{k,j}|Z}$ and $f_{Y_{k,j-1},Y_{k,j}|Z}$ exist. Moreover the partial derivatives of the conditional expectations $\operatorname{E}_{K|Z}(\sum_{j=1}^{K} f_{Y_{K,j}|Z}(u \mid z))$ and $\operatorname{E}_{K|Z}(\sum_{j=2}^{K} f_{Y_{K,j-1},Y_{K,j}|Z}(u, v \mid z))$ with respect to u and v are uniformly bounded in z.

4.8 THEOREM. *Let r be the order of the derivative of Λ_0 satisfying condition C10. If $1/(4r) < \kappa < 1/2$ then, under the above conditions, the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at (θ_0, Λ_0). In particular the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $\Sigma = \tilde{I}_0^{-1}$.*

4.9 COROLLARY. *Let $e_1, \ldots, e_{d_z}$ be the unit vectors in $\mathbb{R}^{d_z}$ and ρ_n be a symmetric d_z-matrix whose entries ρ_{ij}, $i, j = 1, \ldots, d_z$, satisfy $(\sqrt{n}\rho_{ij})^{-1} = O_P(1)$. A consistent estimator for each entry in $\tilde{I}_0$ has the same form as (2.12).*

4.4 SIMULATION STUDY

Numerical properties of the SMLE were investigated for variants of the censoring scheme described in Remark 4.1. Each of these considered the transition intensity model as the previous chapter, given by Equation (3.16). Transition times were right-censored by $C = \tau = 2$. Inspection times were generated in the same manner as Section 2.5. Recall that this scheme considered k "scheduled" inspections. Here every visit after the first was missed with probability $p(Z)$ such that

$$\operatorname{logit} p(Z) = \beta_0 + \beta_1 Z.$$

The last observation time $V = T \wedge C$ offered one further inspection of progression status with a fixed probability of 0.2.

Three thousand Monte Carlo replicates with sample sizes $n = 250$, 500 and 1000 were generated under three scenarios:

- an independent inspection process with $k = 8$, $\beta_0 = \log(1/9)$ and $\beta_1 = 0$;
- an independent inspection process with $k = 4$, $\beta_0 = \log(1/9)$ and $\beta_1 = 0$; and
- a conditionally independent inspection process with $k = 8$, $\beta_0 = \log(1/4)$ and $\beta_1 = \log(4/9)$.

Estimates for each sample were obtained using the same C routine devised for Algorithm 3.9. Changes needed to address differences in the data format and sieve construction were handled by an additional R front end. Tuning parameters were set to the following values: $C_{hj} = 1$, $\kappa = 1/3$, $\varepsilon = 10^{-7}$, $\operatorname{typ}\theta = 1$ and $\sup\theta = 10$. This ensured convergence within a reasonable number of iterations over all scenarios and sample sizes. Under the first scenario estimates were re-evaluated using the alternative sieve sizes $\kappa = 4/15$ and $\kappa = 2/5$. In addition to the SMLE, the same or roughly equivalent Cox models were fit to right-censored variants of the data (Example 1.6) via Therneau's (2012) `coxph` routine from the `survival` package for R. Estimates were based on four different right-censored data sets:

- underlying or "latent" transition times right-censored only by $C = \tau$,
- left- and interval-censored progression times imputed to the midpoint of the censoring interval with last negative inspections carried forward (midpoint-imputed),
- time to the first positive inspection or death (PFS), and
- a variant of PFS obtained by right-censoring times at the (negative) inspection preceding two missed visits (PFS FDA).

The last form of imputed data directly follows the FDA guideline presumably devised to mitigate bias in the analysis of progression-free survival times (FDA 2007, Table A, 2011, Table X).

TABLE 4.1 Average censoring rates over 3000 replicates with $n = 1000$.

			Rate of known progression status at V			
k	e^{β_0}	e^{β_1}	Overall	$S > C$	$S < V = T \wedge C$	$S = T = V < C$
8	1/9	1	0.498	0.199	0.789	0.201
4	1/9	1	0.431	0.199	0.656	0.201
8	1/4	4/9	0.493	0.199	0.780	0.201

Under the fixed transition intensity model parameters, roughly 50% of subjects in the sample progressed ($S < T$), 12% were event-free ($S > \tau$), and 16% survived to study closure ($T > \tau$). The overall observation rate of progression status ranged from 43% to 50% (Table 4.1). This rate was held fixed at 20% among progression-free subjects. The status was known more often among subjects who progressed. Under $k = 8$ the rate was close to 80%. With fewer potential inspections, $k = 4$, this decreased to 66%.

The simulation results for the SMLE $\hat{\theta}_n$ summarized by Table 4.2 are compatible with the asymptotic properties in Section 4.3. Bias generally becomes negligible with larger sample size. Monte Carlo sample standard deviations for $\hat{\theta}_n$ also decrease with larger n and are reasonably approximated by the standard error estimates. Empirical coverage probabilities of the 95% confidence intervals are close to the nominal level. The SPMLE based on midpoint-imputed progression times and PFS had, on average, larger finite-sample bias. Bias in both of these estimators generally did not diminish with increasing sample size. The omission of inspections following two missed visits offered a small reduction in the bias of PFS under the independent censoring schemes, but under conditionally independent censoring this strategy achieved average biases 60 to 80% larger than PFS (Table 4.3).

Table 4.4 suggests that the asymptotically-optimal sieve rate $n^{1/3}$ performs reasonably well in finite samples compared to smaller ($n^{4/15}$) and larger ($n^{2/5}$) alternatives. Under the independent censoring scenario with frequent inspections ($k = 9, e^{\beta} = (1/9, 1)$), the average bias for the SMLE with a sieve growing at the rate $n^{4/15}$ was up to three and a half times larger than the bias achieved by an $n^{1/3}$-sieve estimator. Smaller average bias was achieved by increasing the rate from $n^{1/3}$ to $n^{2/5}$, but this was not entirely the case for coefficients specific to the terminal event times. Since the standard deviation ratios averaged close to 1, any reduction in bias came at little to no loss in precision. Instances of negative relative bias were encountered in the

k,			SMLE			Midpoint-imputed SPMLE			PFS
e^{β_1}	n		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ
8,	250	Bias	-0.0075	-0.0060	0.0078	0.0133	-0.0081	-0.0387	0.0058
1		SD	0.2130	0.2525	0.2131	0.2090	0.1934	0.2059	0.1434
		ASE	0.2069	0.2489	0.2077	0.2059	0.1893	0.1977	0.1398
		CP	0.9477	0.9527	0.9410	0.9480	0.9493	0.9367	0.9440
	500	Bias	-0.0028	0.0065	-0.0030	0.0162	0.0010	-0.0463	0.0114
		SD	0.1474	0.1748	0.1456	0.1455	0.1344	0.1382	0.0984
		ASE	0.1459	0.1722	0.1460	0.1453	0.1336	0.1396	0.0985
		CP	0.9487	0.9450	0.9507	0.9447	0.9483	0.9417	0.9527
	1000	Bias	-0.0036	-0.0003	0.0004	0.0119	-0.0035	-0.0441	0.0062
		SD	0.1029	0.1211	0.1037	0.1016	0.0951	0.0991	0.0699
		ASE	0.1031	0.1205	0.1032	0.1027	0.0944	0.0986	0.0696
		CP	0.9483	0.9477	0.9483	0.9460	0.9487	0.9287	0.9533
4,	250	Bias	-0.0089	-0.0045	0.0063	0.0357	-0.0052	-0.0602	0.0171
1		SD	0.2323	0.2785	0.2283	0.2257	0.1831	0.2215	0.1434
		ASE	0.2267	0.2724	0.2209	0.2257	0.1777	0.2134	0.1401
		CP	0.9500	0.9503	0.9430	0.9527	0.9453	0.9357	0.9450
	500	Bias	-0.0027	0.0062	-0.0028	0.0414	0.0016	-0.0653	0.0223
		SD	0.1639	0.1927	0.1572	0.1599	0.1265	0.1520	0.0980
		ASE	0.1598	0.1867	0.1552	0.1591	0.1255	0.1504	0.0988
		CP	0.9487	0.9410	0.9430	0.9427	0.9490	0.9323	0.9490
	1000	Bias	-0.0061	0.0013	0.0000	0.0362	-0.0024	-0.0637	0.0173
		SD	0.1146	0.1314	0.1119	0.1119	0.0883	0.1081	0.0703
		ASE	0.1128	0.1300	0.1095	0.1124	0.0887	0.1060	0.0698
		CP	0.9500	0.9503	0.9417	0.9410	0.9507	0.9073	0.9467
8,	250	Bias	-0.0075	-0.0054	0.0076	0.0431	-0.0229	-0.0280	0.0154
4/9		SD	0.2145	0.2543	0.2136	0.2104	0.1929	0.2066	0.1431
		ASE	0.2084	0.2509	0.2084	0.2068	0.1885	0.1982	0.1397
		CP	0.9483	0.9530	0.9433	0.9383	0.9463	0.9383	0.9437
	500	Bias	-0.0033	0.0071	-0.0031	0.0451	-0.0135	-0.0357	0.0208
		SD	0.1484	0.1757	0.1466	0.1460	0.1339	0.1390	0.0982
		ASE	0.1468	0.1733	0.1465	0.1459	0.1330	0.1399	0.0985
		CP	0.9497	0.9493	0.9510	0.9370	0.9463	0.9443	0.9457
	1000	Bias	-0.0035	-0.0006	0.0007	0.0414	-0.0183	-0.0332	0.0157
		SD	0.1035	0.1217	0.1037	0.1021	0.0946	0.0992	0.0698
		ASE	0.1037	0.1212	0.1034	0.1031	0.0940	0.0988	0.0696
		CP	0.9507	0.9503	0.9500	0.9310	0.9417	0.9387	0.9487

TABLE 4.2
Bias, standard deviation (SD), average standard error estimate (ASE) and empirical coverage probabilities of 95% confidence intervals (CP) for the SMLE of θ over 3000 replicates. Results for the SPMLE based on midpoint-imputed progression times and PFS are provided for comparison.

TABLE 4.3 Simulation results for estimators of θ based on the FDA-recommended variant of PFS and the latent data over 3000 replicates.

		PFS FDA θ			Latent SPMLE		
		$k = 8$	$k = 4$	$k = 8$			
n		$e^{\beta_1} = 1$	$e^{\beta_1} = 1$	$e^{\beta_1} = 4/9$	θ_1	θ_2	θ_3
250	Bias	0.0045	0.0168	0.0276	0.0009	-0.0067	0.0021
	SD	0.1441	0.1435	0.1447	0.1871	0.2170	0.1963
	ASE	0.1404	0.1402	0.1411	0.1831	0.2119	0.1886
	CP	0.9423	0.9453	0.9410	0.9487	0.9463	0.9363
500	Bias	0.0101	0.0221	0.0333	0.0036	0.0035	-0.0039
	SD	0.0989	0.0980	0.0991	0.1293	0.1505	0.1319
	ASE	0.0990	0.0988	0.0995	0.1293	0.1494	0.1335
	CP	0.9527	0.9493	0.9370	0.9510	0.9493	0.9540
1000	Bias	0.0049	0.0171	0.0282	-0.0008	-0.0018	-0.0010
	SD	0.0702	0.0704	0.0705	0.0909	0.1068	0.0951
	ASE	0.0699	0.0698	0.0703	0.0914	0.1056	0.0944
	CP	0.9523	0.9480	0.9303	0.9550	0.9453	0.9460

largest sample size, implying that the size of the sieve can affect the average direction of the bias. However from Table 4.2 magnitudes of the associated average biases were no greater than 0.0004.

TABLE 4.4 Ratio of bias (relative bias) and SD (relative precision) for $\hat{\theta}_n$ between the specified κ and $\kappa = 1/3$, based on 3000 replicates under $k = 8$ and $e^{\beta} = (1/9, 1)$.

		$\kappa = 4/15$			$\kappa = 2/5$		
n		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
250	Relative bias	1.816	0.768	1.227	0.656	1.110	0.916
	Relative precision	1.001	0.997	0.986	1.003	1.004	1.008
500	Relative bias	2.485	1.175	0.542	0.596	1.059	1.166
	Relative precision	0.998	0.998	0.992	1.004	1.002	1.007
1000	Relative bias	1.706	-1.303	3.640	0.833	0.569	-0.839
	Relative precision	1.000	1.002	0.996	1.004	1.000	1.005

Under the independent censoring scenarios with $k = 4$ and $k = 8$, pointwise means and percentiles for SMLE of the cumulative baseline intensity functions are depicted in Figure 4.3. Pointwise estimates for Λ_{02} appear unbiased. Late in the observation the SMLE for Λ_{01} tends to provide overestimates. The same is true in estimating Λ_{12}, but bias appears early in the observation period the SMLE. Both forms of bias diminish over larger sample sizes and more frequent inspections. Variability in each component of $\hat{\Lambda}_n$ is low and also decreases as n increases. Similar features arise in $\hat{\Lambda}_n$ under the smallest $\kappa = 4/15$ and the largest $\kappa = 2/5$ sieve sizes (Figure 4.4). Small

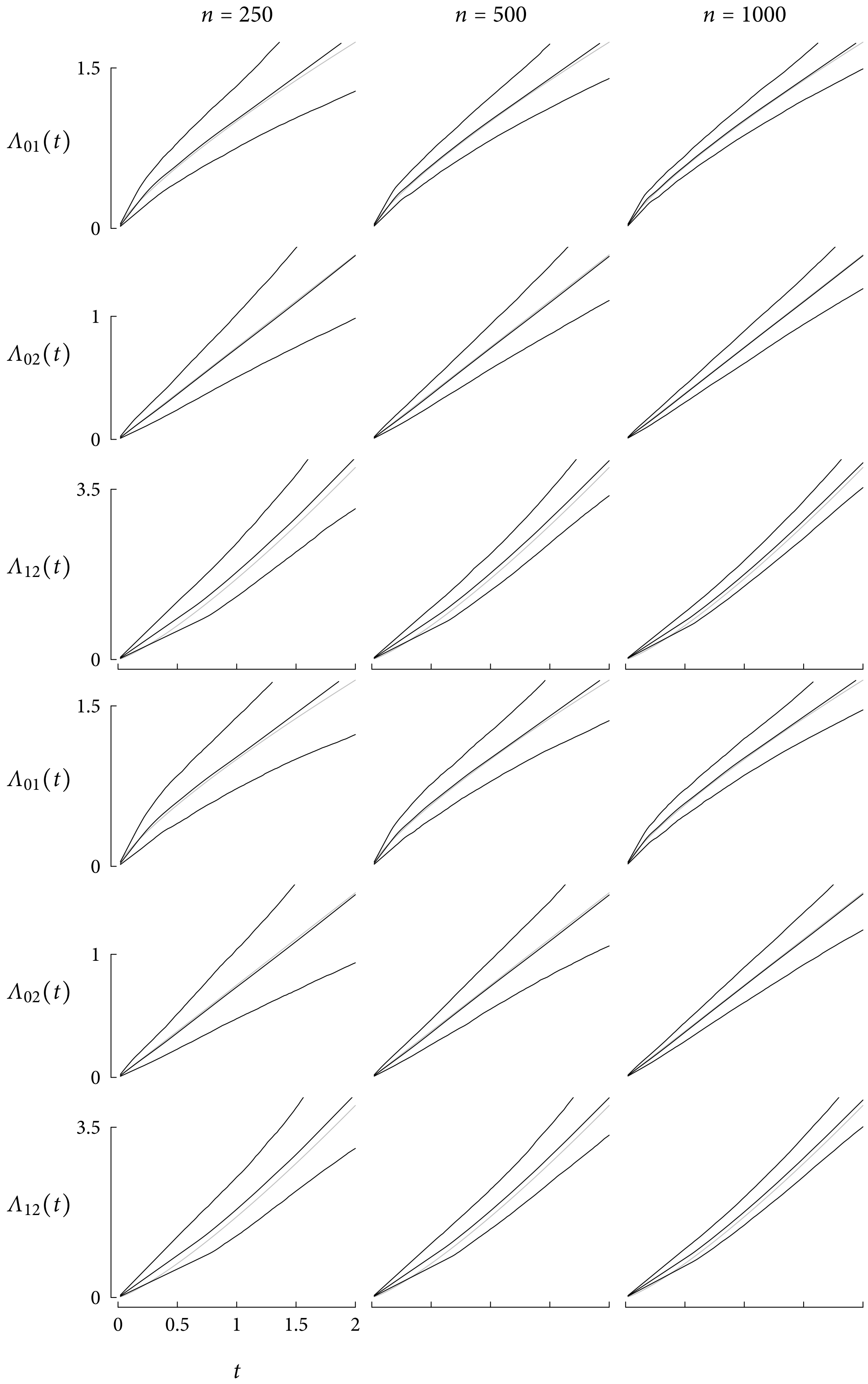


FIGURE 4.3
True values for Λ (–) depicted with pointwise lower and upper 2.5th percentiles and pointwise means for the SPMLE of Λ with $\kappa = 1/3$ based on 3000 replicates under $e^{\beta} = (1/9, 1)$ and $k = 8$ (top) and $k = 4$ (bottom).

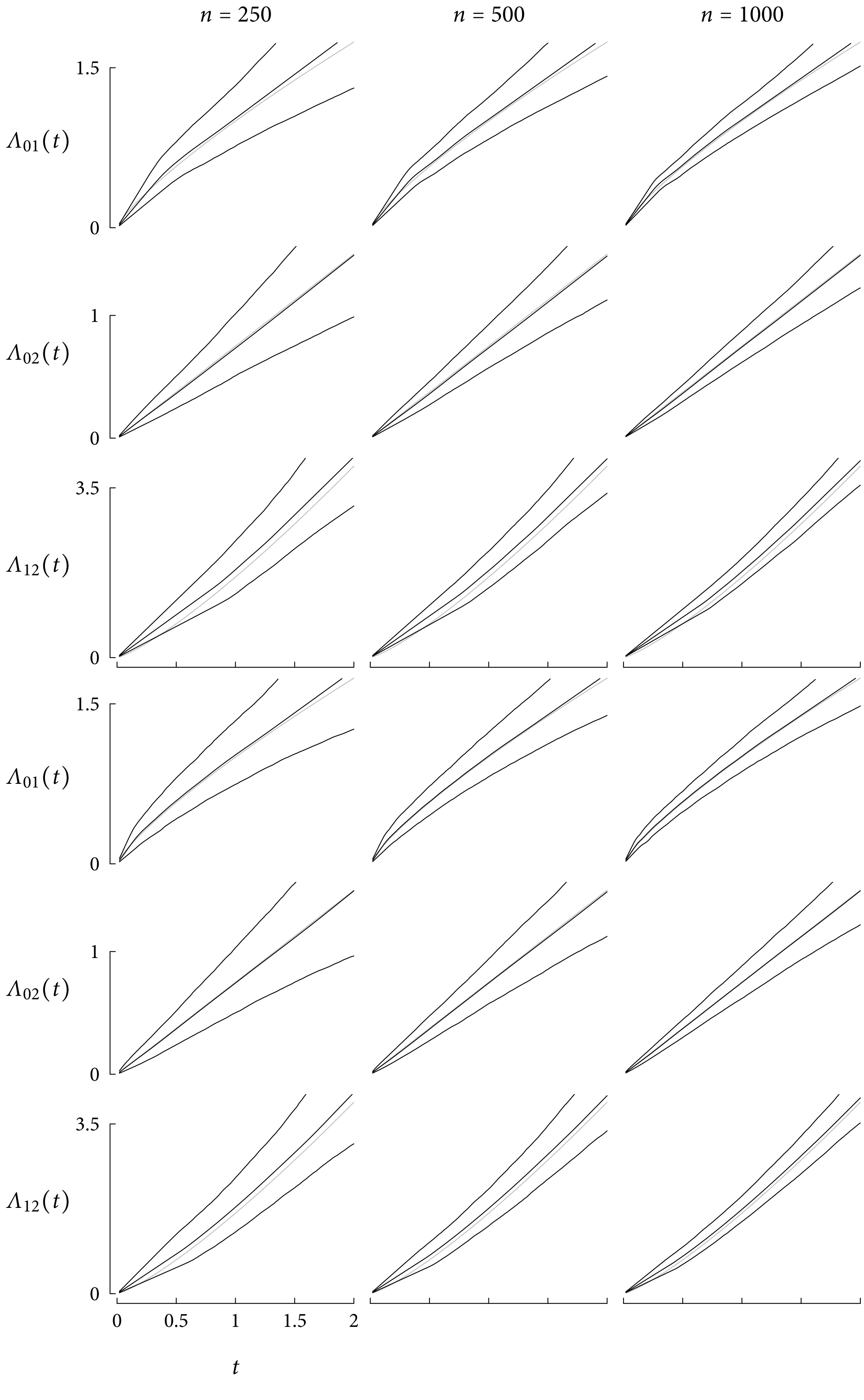


FIGURE 4.4 True values for $\varLambda$ (–) depicted with pointwise lower and upper 2.5th percentiles and pointwise means for the SPMLE of $\varLambda$ with $\kappa = 4/15$ (top) and $\kappa = 2/5$ (bottom) based on 3000 replicates under $k = 8$ and $e^{\beta} = (1/9, 1)$.

improvements in pointwise bias are achieved with a larger sieve, with no obvious increase in variability.

From Table 4.5, it is apparent that larger sieves need much more time to convergence in terms of both the number of iterations carried out by Algorithm 4.5 and the computing time for both parameter and variance estimation. Algorithm 4.5 does not appear to scale particularly well in sieve and sample size, with the CPU ranging between a fraction of a second up to just over 2 minutes. Longer processing times are also needed when inspections are infrequent.

	$k = 8$						$k = 4$	
	$\kappa = 4/15$		$\kappa = 1/3$		$\kappa = 2/5$		$\kappa = 1/3$	
n	CPU	Iterations	CPU	Iterations	CPU	Iterations	CPU	Iterations
250	0.23	48	0.78	154	4.57	793	3.33	666
500	0.62	58	4.48	388	25.85	1675	17.98	1390
1000	1.60	69	22.76	803	129.54	3311	82.08	2576

TABLE 4.5 Average time to convergence for Algorithm 4.5. The CPU time, given in seconds, covers variance estimation.

4.5 APPLICATION

In this section we return to the bone lesion data examined in Chapter 2. Both time to the first new lesion and death are considered via the Markov illness-death model, so all 380 subjects in the sample are included in the analysis. The observed data are not entirely compatible with the required assumptions of the SMLE, since the progression status is not available at right-censoring or terminal events. As convenient workaround, a narrow form of LOCF was applied to the last negative inspection within six weeks of right-censoring or death.

A new R function, called icprog, was constructed to repurpose the C routine described in Section 3.5. The observations depicted in Figure 2.5 are represented in the following R data frame.

```
  id start stop from to status mid.start mid.stop mid.status
1  1    90  180    0  1      3         0      135          1
2  1   Inf   NA    0  2      0         0      135          0
3  1   870   NA    1  2      0       135      870          0
4  2   180  360    0  1      3         0      270          0
5  2   360   NA    0  2      1         0      270          1
6  2   360   NA    1  2      1        NA       NA         NA
7  3   330  630    0  1      3         0      480          0
8  3   630   NA    0  2      1         0      480          1
9  3   630   NA    1  2      1        NA       NA         NA
```

```
10  4     0   60    0  1      3         0       30          1
11  4   Inf   NA    0  2      0         0       30          0
12  4   480   NA    1  2      1        30      480          1
13  5     0   60    0  1      3         0       30          1
14  5   Inf   NA    0  2      0         0       30          0
15  5   870   NA    1  2      0        30      870          0
16  6   Inf   NA    0  1      0        60       60          0
17  6    60   NA    0  2      1        60       60          1
   pfs.time pfs.event trt trt01 trt02 trt12
1       180         1   0     0     0     0
2        NA        NA   0     0     0     0
3        NA        NA   0     0     0     0
4       360         1   0     0     0     0
5        NA        NA   0     0     0     0
6        NA        NA   0     0     0     0
7       630         1   0     0     0     0
8        NA        NA   0     0     0     0
9        NA        NA   0     0     0     0
10       60         1   1     1     0     0
11       NA        NA   1     0     1     0
12       NA        NA   1     0     0     1
13       60         1   1     1     0     0
14       NA        NA   1     0     1     0
15       NA        NA   1     0     0     1
16       60         1   1     1     0     0
17       NA        NA   1     0     1     0
```

The transition types are indicated in the variables `from` and `to`. For the 0 → 1 type, the variables `start` and `stop` specify the potential censoring interval for the progression time in days. The event `status` for these cases are given the `survival` interval-censoring value of `3`. If the progression status is observed to be negative, `stop` is set to the missing value `NA` and `start` to the infinite value `Inf`. When the progression status is known to be positive, the 0 → 2 type is similarly coded. Otherwise 0 → 2 times are recorded in `start`, `stop` is set to `NA`, and `status` indicates the occurrence of events. The remaining 1 → 2 type is similarly represented, but is necessary only for subjects whose 0 → 1 `status` is `3`. Midpoint-imputed progression times with last negative status carried forward are given by the additional columns `mid.start`, `mid.stop` and `mid.status`, with values corresponding to the extra 1 → 2 types needed by the observed data set to missing (`NA`). This format follows the usual counting process representation of multistate data (e.g. Therneau and Grambsch 2000, Section 8.6). Data corresponding to PFS are provided in the standard time-to-event format with the variables `pfs.time` and `pfs.event`. The values can be recorded in one of the mandatory 0 → 1 or 0 → 2 types, with the remaining types set to missing (`NA`).

```
> fit <- icprog(Surv(start, stop, status, type="interval") ~ cluster(id)
+ trans(from, to) + I(trt * (to==1))
+ I(trt * (from==0 & to==2)) + I(trt * (from==1)),
data = p19.state, sieve.const = 1, sieve.rate = 1/3,
eps = 1e-9, coef.typ = 1/2, coef.max = 2,
rcprog = list(Surv(mid.start, mid.stop, mid.status) ~ .,
              Surv(pfs.time, pfs.event) ~ trt))
```

These variables are collected into an `interval`-type `Surv` object to give the model response. The remaining model terms follow the same format described in Section 3.6. Models to be fit using `coxph` are specified using the optional list argument `rcprog`. The code fragment above fits two additional models based on midpoint-imputed progression times and PFS, with the former having the same predictor terms already provided in the first argument to `icprog`. The `icprog` function returns an `icprog`-type object. Its `print` routine reproduces the function call, summarizes regression coefficients, reports the initial and final log-likelihood values and gives the rates at which the terminal event T and progression status $1(S < T)$ were observed in the sample.

From the output below the log-likelihood Equation (4.3) at the initial parameter value is -2532. After 1075 iterations this increases to -2479. Both the terminal event T and the progression status $1(S < T)$ were observed for 28% of the 380 individuals in the sample. Only the progression status was observed in 13%. In the remaining 59% the terminal event was right-censored and the progression status was unknown. The $0 \to 1$, $0 \to 2$ and $1 \to 2$ type-specific regression coefficients $\hat{\theta}_n$ measuring the increase in the risk of transition associated with receiving pamidronate are -0.3905, -0.0388 and -0.0461, respectively. Only the $0 \to 1$ effect is significant, with a p-value for the two-sided test $\theta_1 = 0$ of 3.4%. It is estimated that an individual treated with pamidronate has 0.677 times the rate of new lesion (95% confidence interval 0.472 – 0.971) versus another patient receiving placebo. This result is in general agreement with the analysis of TTP presented in Section 2.6.

```
> fit
Call:
icprog(formula = Surv(start, stop, status, type = "interval") ~
    cluster(id) + trans(from, to) + I(trt * (to == 1)) + I(trt *
        (from == 0 & to == 2)) + I(trt * (from == 1)), data = p19.state,
    rcprog = list(Surv(mid.start, mid.stop, mid.status) ~ .,
        Surv(pfs.time, pfs.event) ~ trt),
    ... = list(sieve.const = 1, sieve.rate = 1/3, eps = 1e-09,
               coef.typ = 1/2, coef.max = 2))

                                 coef se(coef)      z     p
I(trt * (to == 1))            -0.3905    0.184 -2.120 0.034
```

```
I(trt * (from == 0 & to == 2)) -0.0388    0.209 -0.185 0.850
I(trt * (from == 1))           -0.0461    0.198 -0.233 0.820
                               exp(coef)  2.5% 97.5%
I(trt * (to == 1))                 0.677 0.472 0.971
I(trt * (from == 0 & to == 2))     0.962 0.638 1.450
I(trt * (from == 1))               0.955 0.648 1.410

Based on n = 380 subjects contributing 1117 observation times

Initial log-likelihood: -2532.27
Log-likelihood after 1075 iterations: -2478.7

                 T and 1(S < T) Only 1(S < T) Neither
Observation rate          0.284         0.126   0.589

Estimation from imputed data via survival's coxph function

Formula:
Surv(mid.start, mid.stop, mid.status) ~ cluster(id) + strata(from,
    to) + I(trt * (to == 1)) + I(trt * (from == 0 & to == 2)) +
    I(trt * (from == 1))

                                  coef se(coef)      z    p exp(coef)
I(trt * (to == 1))             -0.2251    0.174 -1.292 0.20     0.798
I(trt * (from == 0 & to == 2)) -0.1145    0.141 -0.812 0.42     0.892
I(trt * (from == 1))           -0.0315    0.207 -0.152 0.88     0.969
                                2.5% 97.5%
I(trt * (to == 1))             0.567  1.12
I(trt * (from == 0 & to == 2)) 0.676  1.18
I(trt * (from == 1))           0.646  1.45

Based on 893 observation times (224 deleted due to missingness)

Formula:
Surv(pfs.time, pfs.event) ~ trt

      coef se(coef)     z    p exp(coef)  2.5% 97.5%
trt -0.148     0.11 -1.35 0.18     0.862 0.695  1.07

Based on 380 observation times (737 deleted due to missingness)
```

The `print` function also provides the same summary of regression coefficients obtained from the `coxph` fit to the models specified by the `rcprog` argument. The estimate for θ based on midpoint-imputed progression times are somewhat similar to the SMLE. None of the differences measured are significantly different from zero, with p-values for the two-sided test $\theta_j = 0$ no smaller than 20%. Both estimators based on the illness-death model suggest that pamidronate has no influence on mortality, which makes analysis via PFS problematic. The hazard ratio for the earliest of time to new lesion and death associated with pamidronate is 0.862 (95% confidence interval 0.695 – 1.07). This is not significantly different from 1, with a p-value of 18%. The output also provides the number of cases included by `coxph`.

The SMLE can be accessed directly from the `icprog` object list arguments `coef` and `bhaz`. Estimates based on imputed data are similarly represented, but nested in the `icprog` list argument `rcfit`.

```
> fit$coef
                     I(z * (to == 1))
                           -0.1728308
       I(z * (from %in% 0 & to == 2))
                            0.1894571
I(z * (from %in% c(NA, 1) & to == 2))
                           -0.1009796
> fit$bhaz[1:3, ]
      hazard time        trans
1 0.00000000    0 from=0, to=1
2 0.05814021   31 from=0, to=1
3 0.08540514   57 from=0, to=1

> fit$rcfit[[1]]$coef
                     I(z * (to == 1))
                        -0.1334553657
       I(z * (from %in% 0 & to == 2))
                         0.1568147185
I(z * (from %in% c(NA, 1) & to == 2))
                        -0.0003853867

> fit$rcfit[[1]]$bhaz[1:3, ]
       hazard time        trans
1 0.000000000 0.01 from=0, to=1
2 0.007047897 1.00 from=0, to=1
3 0.009880123 2.00 from=0, to=1
```

Alternative sieve sizes can be examined by fitting the same model with different parameters. The listing below shows that similar estimates are achieved with $C_{hj}n^{\kappa} = n^{4/15}$, but the treatment effect measured here is stronger. Since the reduction in sieve size did not result in a large decrease in the likelihood, we may prefer a smaller sieve for this data set.

```
> icprog(Surv(start, stop, status, type='interval') ~ cluster(id)
         + trans(from, to) + I(trt * (to==1))
         + I(trt * (from==0 & to==2)) + I(trt * (from==1)),
          data = p10, sieve.const = 1, sieve.rate = 4/15,
          eps = 1e-9, coef.typ = 1/2, coef.max = 2)
Call:
icprog(formula = Surv(start, stop, status, type = "interval") ~
    cluster(id) + trans(from, to) + I(trt * (to == 1)) + I(trt *
        (from == 0 & to == 2)) + I(trt * (from == 1)), data = p19.state,
    ... = list(sieve.const = 1, sieve.rate = 4/15, eps = 1e-09,
               coef.typ = 1/2, coef.max = 2))

                                  coef se(coef)       z     p
I(trt * (to == 1))             -0.4339    0.183 -2.3667 0.018
I(trt * (from == 0 & to == 2))  0.0100    0.212  0.0473 0.960
I(trt * (from == 1))           -0.0857    0.197 -0.4359 0.660
                               exp(coef)  2.5% 97.5%
I(trt * (to == 1))                 0.648 0.452 0.928
I(trt * (from == 0 & to == 2))     1.010 0.666 1.530
I(trt * (from == 1))               0.918 0.624 1.350

Based on n = 380 subjects contributing 1117 observation times

Initial log-likelihood: -2532.27
Log-likelihood after 80 iterations: -2481.53

                 T and 1(S < T) Only 1(S < T) Neither
Observation rate          0.284         0.126   0.589
```

To check results achieved with starting values different from the ones described in Algorithm 4.5, the `icprog` function can accept alternative initial parameter values via the input argument `init`. Initial values can also be obtained from estimates based on imputed right-censored data using the option setting `rcinit = TRUE`. Under the latter approach the difference in the resulting sieve estimates is of the order 10^{-6}.

```
> fit.rcinit <- icprog(Surv(start, stop, status, type='interval')
                 ~ cluster(id) + trans(from, to) + I(trt * (to==1))
                 + I(trt * (from==0 & to==2))
                 + I(trt * (from==1)), data = p19.state,
                 rcinit = TRUE, sieve.const = 1, sieve.rate = 1/3,
                 eps = 1e-9, coef.typ = 1/2, coef.max = 2
                 rcprog = list(Surv(mid.start, mid.stop, mid.status) ~ .))
> max(abs(fit.rcinit$coef - fit$coef))
[1] 2.37378e-07
> max(abs(fit.rcinit$bhaz - fit$bhaz))
[1] 2.587994e-06
```

Estimates for the baseline cumulative intensity functions are fully represented in Figure 4.5. The SPMLE based on midpoint-imputed progression times with last negative inspection carried forward gives smaller pointwise estimates for Λ_{01} and Λ_{12}, and larger estimates for Λ_{02} compared to the SMLE. This difference is probably due to the fact that negative status was carried forward for almost 60% of the sample. Restricting LOCF to within six weeks reduces the rate of imputation to less than

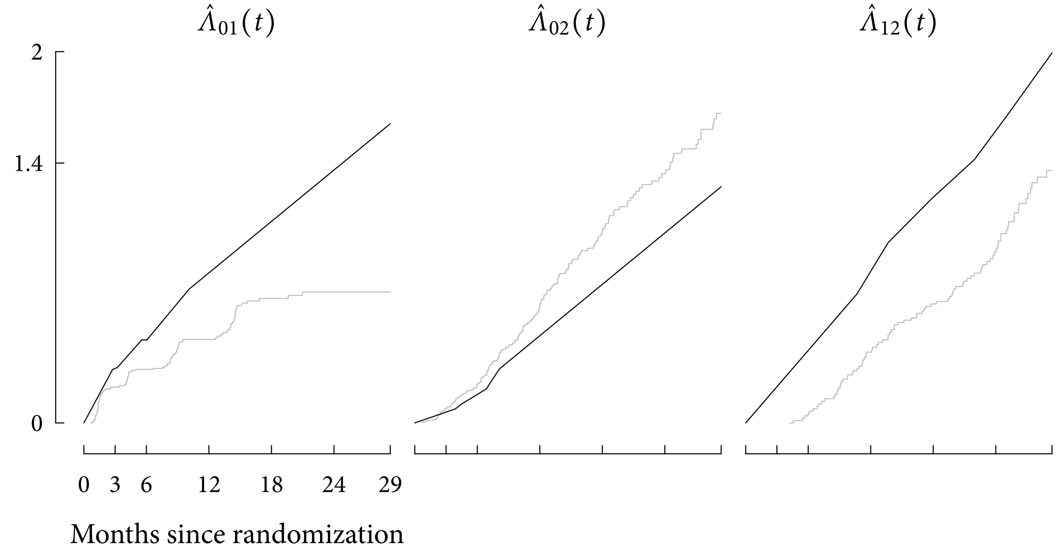


FIGURE 4.5
The SPMLE from midpoint-imputed data (–) and the SMLE (–) for the cumulative baseline transition intensity functions.

seven percent. With very few known 0 → 2 transition times meeting the criteria for restricted LOCF, the SMLE is nearly linear.

Even in this restricted form, LOCF imputation is subject to bias resulting from misclassification of progression status, so this limits the interpretation of the analysis based on the SMLE. Following Section 3.6 it would be useful to examine the proportional hazards and Markov assumptions of this model, but existing methods do not account for the presence of interval-censored data. We could consider the midpoint-imputed data, but since the estimates under this imputation scheme are markedly different the result of any inference tests would be difficult to extend to the SMLE.

CHAPTER 5

DISCUSSION

Chapter 2 of this thesis examined maximum likelihood estimation of the Cox-Aalen hazard model (Figure 5.1) from mixed case interval-censored data with fixed covariates. A simulation study showed that resulting estimator is preferable to imputation-based alternatives, particularly under larger samples with infrequent inspections. The new methods also mark the first extension of any Aalen model variant to interval-censored data. The following two chapters constructed sieve maximum likelihood estimators for a Markov illness-death process having Cox-type transition intensities (Figure 5.2) under two censoring schemes. The first considered "double" right censoring arising when progression is right-censored earlier than survival. The second examined interval-censored progression times. Both sieve estimators demonstrated superior empirical properties compared to imputation-based alternatives encountered in practice. Together the new estimators pose a variety of open problems. These include:

FIGURE 5.1
The Cox-Aalen model considered in Chapter 2.

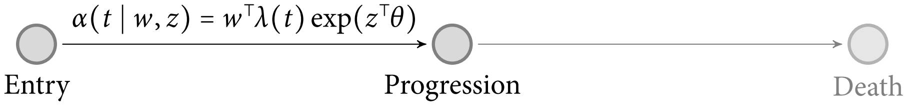
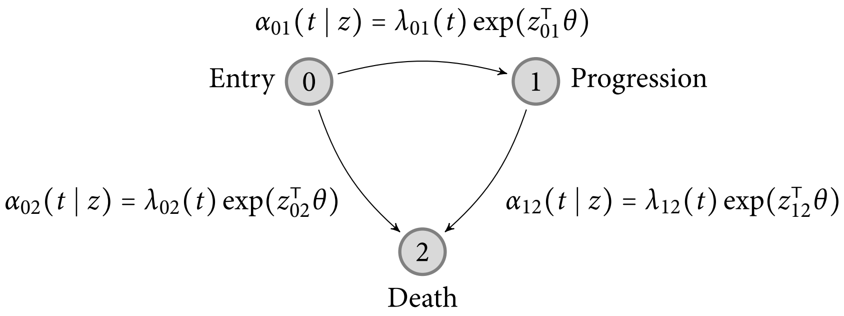

$\alpha(t \mid w, z) = w^{\top}\lambda(t)\exp(z^{\top}\theta)$
Entry
Progression
Death

FIGURE 5.2
The extended Cox model considered in Chapters 3 and 4.


$\alpha_{01}(t \mid z) = \lambda_{01}(t)\exp(z_{01}^{\top}\theta)$
Entry
0
1
Progression
$\alpha_{02}(t \mid z) = \lambda_{02}(t)\exp(z_{02}^{\top}\theta)$
$\alpha_{12}(t \mid z) = \lambda_{12}(t)\exp(z_{12}^{\top}\theta)$
2
Death

- criteria for the selection of tuning parameters,
- the limiting distribution for $n^{1/3}|\hat{\Lambda}_n - \Lambda_0|(t)$,
- estimation of functionals such as the conditional survivor distribution,
- methods to assess goodness-of-fit, and
- extension to time-dependent covariates.

The first issue largely refers to the sieve constant C_{hj} and growth rate κ for the estimators constructed in Chapters 3 and 4. From the relative precision reported in Tables 3.4 and 4.4, C_{hj} should compensate for large discrepancies between n^κ and the number of exact $h \to j$ transition times available. Detailed simulation study may prove useful in devising more concrete criteria for selecting the sieve size. Additional tuning parameters arise in variance estimation via profile likelihood. Section 2.4.2 reduces the choice to specifying the typical and maximum values among the entries of θ. This is motivated by numerical methods for evaluating derivatives, but some empirical evidence endorsing such an approach would be valuable.

Estimating variance and functionals of the infinite-dimensional parameter Λ is a persistent open problem encountered with most semiparametric and sieve maximum likelihood estimators from interval-censored data. The estimators proposed here are no exception. Addressing this limitation will likely require an extension of the theory described in Section 1.2.

New tools for model assessment are also needed. For the Cox-Aalen model considered in Chapter 2, Martinussen and Scheike (2006, pp. 255–58) offer inference tests of time-invariance and proportionality for covariate effects. These are limited to right-censored data. In practice there may be no obvious separation of covariates into additive (W) and multiplicative (Z) components, so similar tests for interval-censored data would be useful. An extension of Grambsch and Therneau's (1994) test for proportional hazards to interval-censored data would be relevant in checking the functional form of the sieve estimator proposed in Chapter 4. In the application illustrated in Section 3.6, the Markov assumption under the illness-death model proved too restrictive. Extension of both the sieve estimators to time-dependent covariates would enable departures from the Markov property.

A few more limitations are found by considering the estimators individually. In particular the computational demands of the semiparametric estimator devised in Chapter 2 are prohibitive in practical settings. Adaptation of candidate support reduction methods (e.g. Dümbgen et al. 2006; Y. Wang 2008) may offer some improvement processing time.

Finite-sample efficiency gains achieved by the sieve estimator of Chapter 3 over the simple alternative obtained by ignoring all events following the earlier censoring time are modest. More detailed simulation study is needed to investigate settings in which the proposed estimator may offer a clear advantage. Some informal experimentation suggest that scenarios having moderate-to-high levels of censoring with sufficient variation in the earlier right-censoring time deserve examination.

The application presented in Section 4.5 employed a restricted variant of the last observation carried forward approach to enforce the data requirements of the proposed sieve estimator. Since all of the new methods are devised as better alternatives to imputation, this is less than ideal. The construction of unbiased support- or sieve-finding methods would enable application of the estimator to more general observation schemes similar to the one encountered in the application.

Aside from their practical shortcomings, the new estimators show potential for some interesting extensions. These include the adaptation of the Cox-Aalen estimator to panel count data via the Wellner and Y. Zhang's (2007) Poisson process framework. Spline-based variants of the sieve estimators should easily follow from the results of Y. Zhang et al. (2010). The Markov assumption imposed on the progressive illness-death model greatly simplifies the construction of estimators, but is difficult to justify in practice. A restricted form of duration dependence may be possible through an extended Cox model and some variant of the observation schemes considered in Chapters 3 and 4.

REFERENCES

Aalen, O. O. (1975). "Statistical Inference for a Family of Counting Processes." PhD thesis. University of California, Berkeley (cit. on p. 6).

Aalen, O. O. (1978). "Nonparametric inference for a family of counting processes." *Annals of Statistics* 6(4): 701–26 (cit. on p. 6).

Aalen, O. O. (1980). "A model for nonparametric regression analysis of counting processes." In *Mathematical Statistics and Probability Theory: Proceedings, Sixth International Conference, Wisa (Poland) 1978*. Ed. by K. Witold, A. Kozek and J. Rosinski. New York: Springer, 1–25 (cit. on p. 8).

Aalen, O. O. (1989). "A linear regression model for the analysis of life times." *Statistics in Medicine* 8(8): 907–25 (cit. on pp. 8, 29).

Aalen, O. O. and S. Johansen (1978). "An empirical transition matrix for non-homogeneous Markov chains based on censored observations." *Scandinavian Journal of Statistics* 5(3): 141–50 (cit. on p. 6).

Alioum, A. and D. Commenges (1996). "A proportional hazards model for arbitrarily censored and truncated data." *Biometrics* 52(2): 512–24 (cit. on p. 33).

Andersen, P. K., Ø. Borgan, R. D. Gill and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer (cit. on pp. 3, 5, 59).

Andersen, P. K., S. Esbjerg and T. I. A. Sørensen (2000). "Multi-state models for bleeding episodes and mortality in liver cirrhosis." *Statistics in Medicine* 19(4): 587–99 (cit. on p. 85).

Andersen, P. K. and R. D. Gill (1982). "Cox's regression model for counting processes: A large sample study." *Annals of Statistics* 10(4): 1100–1120 (cit. on p. 7).

Andersen, P. K., J. P. Klein and S. Rosthoj (2003). "Generalised linear models for correlated pseudo-observations, with applications to multi-state models." *Biometrika* 90(1): 15–27 (cit. on p. 26).

Andersen, P. K. and B. B. Ronn (1995). "A nonparametric test for comparing two samples where all observations are either left- or right-censored." *Biometrics* 51 (1): 323–29 (cit. on p. 24).

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney and D. Sorensen (1999). *LAPACK Users' Guide*. Third. Philadelphia, PA: Society for Industrial and Applied Mathematics (cit. on pp. 48, 72).

Armijo, L. (1966). “Minimization of functions having Lipschitz continuous first partial derivatives.” *Pacific Journal of Mathematics* 16(1): 1–3 (cit. on p. 45).

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid and E. Silverman (1955). “An empirical distribution function for sampling with incomplete information.” *The Annals of Mathematical Statistics* 26(4): 641–47 (cit. on p. 22).

Bebchuk, J. D. and R. A. Betensky (2000). “Multiple imputation for simple estimation of the hazard function based on interval censored data.” *Statistics in Medicine* 19(3): 405–19 (cit. on p. 25).

Bebchuk, J. D. and R. A. Betensky (2001). “Local likelihood analysis of survival data with censored intermediate events.” *Journal of the American Statistical Association* 96: 449–57 (cit. on pp. 20, 27, 58).

Bebchuk, J. D. and R. A. Betensky (2002). “Local likelihood analysis of the latency distribution with interval censored intermediate events.” *Statistics in Medicine* 21 (22): 3475–91 (cit. on pp. 27, 58).

Bebchuk, J. D. and R. A. Betensky (2005). “Tests for treatment group differences in the hazards for survival, before and after the occurrence of an intermediate event.” *Statistics in Medicine* 24(3): 359–78 (cit. on pp. 27, 58).

Betensky, R. A., J. C. Lindsey, L. M. Ryan and M. P. Wand (1999). “Local EM estimation of the hazard function for interval-censored data.” *Biometrics* 55(1): 238–45 (cit. on p. 25).

Bickel, P. J., C. A. Klaassen, J. Ritov and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press (cit. on p. 24).

Breslow, N. (1972). Contribution to the discussion of “Regression models and life-tables” by D. R. Cox. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2): 216–17 (cit. on p. 7).

Cai, T. and R. A. Betensky (2003). “Hazard regression for interval-censored data with penalized spline.” *Biometrics* 59(3): 570–79 (cit. on p. 25).

Calle, M. L. and G. Gómez (2001). “Nonparametric bayesian estimation from interval-censored data using Monte Carlo methods.” *Journal of Statistical Planning and Inference* 98(1-2): 73–87 (cit. on p. 26).

Chen, B. and X.-H. Zhou (2011). “Non-homogeneous Markov process models with informative observations with an application to Alzheimer’s disease.” *Biometrical Journal* 53(3): 444–63 (cit. on p. 26).

Cheng, G., Y. Zhang and L. Lu (2011). “Efficient algorithms for computing the non and semi-parametric maximum likelihood estimates with panel count data.” *Journal of Nonparametric Statistics* 23(2): 567–79 (cit. on p. 44).

Cox, D. R. (1972). "Regression models and life-tables." (With discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2): 187–220 (cit. on pp. 7, 25).

Dahlquist, G. and Å. Björck (1974). *Numerical Methods*. Courier Dover Publications (cit. on p. 63).

Datta, S., L. Lan and R. Sundaram (2009). "Nonparametric estimation of waiting time distributions in a markov model based on current status data." *Journal of Statistical Planning and Inference* 139(9): 2885–97 (cit. on p. 23).

Datta, S. and R. Sundaram (2006). "Nonparametric estimation of stage occupation probabilities in a multistage model with current status data." *Biometrics* 62(3): 829–37 (cit. on p. 23).

Day, R., J. Bryant and M. Lefkopoulou (1997). "Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators." *Biometrika* 84(1): 45–56 (cit. on p. 27).

De Gruttola, V. and S. W. Lagakos (1989). "Analysis of doubly-censored survival data, with application to AIDS." *Biometrics* 45(1): 1–11 (cit. on p. 23).

Dehghan, M. and T. Duchesne (2011). "A generalization of Turnbull's estimator for nonparametric estimation of the conditional survival function with interval-censored data." *Lifetime Data Analysis* 17(2): 234–55 (cit. on p. 23).

Dejardin, D., E. Lesaffre and G. Verbeke (2010). "Joint modeling of progression-free survival and death in advanced cancer clinical trials." *Statistics in Medicine* 29 (16): 1724–34 (cit. on p. 27).

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1): 1–38 (cit. on p. 22).

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM (cit. on p. 46).

Dümbgen, L., S. Freitag-Wolf and G. Jongbloed (2006). "Estimating a unimodal distribution from interval-censored data." *Journal of the American Statistical Association* 101(475): 1094–106 (cit. on pp. 19, 23, 24, 44, 45, 107).

Dümbgen, L. and K. Rufibach (2009). "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency." *Bernoulli* 15(1): 40–68 (cit. on p. 23).

Efron, B. (1965). "The two sample problem with censored data." In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. Le Cam and J. Neyman. Berkeley: University of California Press, 831–53 (cit. on p. 22).

Fang, H.-B., J. Sun and M.-L. T. Lee (2002). "Nonparametric survival comparisons for interval-censored continuous data." *Statistica Sinica* 12(4): 1073–83 (cit. on p. 24).

Farrington, C. P. (1996). "Interval censored survival data: A generalized linear modelling approach." *Statistics in Medicine* 15(3): 283–92 (cit. on p. 26).

FDA (2007). *Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics*. Clinical/Medical Guidances. U.S. Department of Health and Human Services (cit. on pp. 1, 26, 27, 48, 94).

FDA (2011). *Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics (Draft Guidance)*. Clinical/Medical Guidances. U.S. Department of Health and Human Services (cit. on p. 94).

Fine, J. P., H. Jiang and R. Chappell (2001). "On semi-competing risks data." *Biometrika* 88(4): 907–19 (cit. on p. 27).

Finkelstein, D. M. (1986). "A proportional hazards model for interval-censored failure time data." *Biometrics* 42(4): 845–54 (cit. on p. 24).

Fleming, T. R., M. D. Rothmann and H. L. Lu (2009). "Issues in using progression-free survival when evaluating oncology products." *Journal of Clinical Oncology* 27(17): 2874–80 (cit. on pp. 1, 26, 60).

Fletcher, R. (1987). *Practical methods of optimization*. 2nd ed. Wiley (cit. on p. 45).

Frydman, H. (1995). "Nonparametric estimation of a Markov 'illness-death' process from interval-censored observations, with application to diabetes survival data." *Biometrika* 82(4): 773–89 (cit. on pp. 23, 27, 86, 90).

Frydman, H. and J. Liu (2013). "Nonparametric estimation of the cumulative intensities in an interval censored competing risks model." *Lifetime Data Analysis* 19(1): 79–99 (cit. on p. 23).

Frydman, H. and M. Szarek (2009). "Nonparametric estimation in a Markov 'illness-death' process from interval censored observations with missing intermediate transition status." *Biometrics* 65(1): 143–51 (cit. on pp. 23, 27, 28, 86, 89).

van de Geer, S. (1993). "Hellinger-consistency of certain nonparametric maximum likelihood estimators." *The Annals of Statistics* 21(1): 14–44 (cit. on p. 23).

Geman, S. and C.-R. Hwang (1982). "Nonparametric maximum likelihood estimation by the method of sieves." *The Annals of Statistics* 10(2): 401–14 (cit. on p. 14).

Gill, R. D. and S. Johansen (1990). "A survey of product-integration with a view toward application in survival analysis." *The Annals of Statistics* 18(4): 1501–55 (cit. on pp. 3, 6).

Goggins, W. B., D. M. Finkelstein, D. A. Schoenfeld and A. M. Zaslavsky (1998). "A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model." *Biometrics* 54(4): 1498–507 (cit. on p. 25).

Gómez, G., M. L. Calle and R. Oller (2004). "Frequentist and bayesian approaches for interval-censored data." *Statistical Papers* 45(2): 139–73 (cit. on p. 26).

Grambsch, P. M. and T. M. Therneau (1994). "Proportional hazards tests and diagnostics based on weighted residuals." *Biometrika* 81(3): 515–26 (cit. on pp. 84, 107).

Greenwood, P. E. and W. Wefelmeyer (1991). "Efficient estimating equations for nonparametric filtered models." In *Statistical Inference in Stochastic Processes*. Ed. by N. U. Prabhu and I. V. Basawa. Vol. 6. New York: Marcel Dekker, 107–41 (cit. on p. 9).

Grenander, U. (1981). *Abstract inference*. Wiley (cit. on p. 14).

Griffin, B. and S. Lagakos (2010). "Nonparametric inference and uniqueness for periodically observed progressive disease models." *Lifetime Data Analysis* (cit. on p. 23).

Groeneboom, P. (1991). *Nonparametric Estimators Maximum Likelihood Estimators for Interval Censoring and Deconvolution*. Technical Report 378. Department of Statistics, Stanford University (cit. on pp. 22, 23).

Groeneboom, P. (1996). "Lectures on inverse problems." In *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXIV – 1994*. Ed. by P. Bernard. Berlin: Springer-Verlag, 67–164 (cit. on p. 19).

Groeneboom, P. (2012). "Likelihood ratio type two-sample tests for current status data." *Scandinavian Journal of Statistics* (cit. on p. 24).

Groeneboom, P., G. Jongbloed and B. I. Witte (2010). "Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model." *Annals of Statistics* 38(1): 352–87 (cit. on p. 25).

Groeneboom, P., M. H. Maathuis and J. A. Wellner (2008). "Current status data with competing risks: Limiting distribution of the MLE." *Annals of Statistics* 36(3): 1064–89 (cit. on p. 23).

Groeneboom, P. and J. A. Wellner (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser (cit. on pp. 22, 32).

Grüger, J., R. Kay and M. Schumacher (1991). "The validity of inferences based on incomplete observations in disease state models." *Biometrics* 47(2): 595–605 (cit. on pp. 21, 22, 29).

Gu, M. G. and C.-H. Zhang (1993). "Asymptotic properties of self-consistent estimators based on doubly censored data." *The Annals of Statistics* 21(2): 611–24 (cit. on p. 23).

Han, S., A.-C. Andrei and K.-W. Tsui (January 15, 2013). "A semiparametric regression method for interval-censored data." *Communications in Statistics - Simulation and Computation*. Advance online publication (cit. on pp. 26, 29).

Heller, G. (2011). "Proportional hazards regression with interval censored data using an inverse probability weight." *Lifetime Data Analysis* 17(3): 373–85 (cit. on p. 25).

Hortobagyi, G. N., R. L. Theriault, L. Porter, D. Blayney, A. Lipton, C. Sinoff, H. Wheeler, J. F. Simeone, J. Seaman, R. D. Knight, M. Heffernan, D. J. Reitsma, I. Kennedy, S. G. Allan, K. Mellars and the Protocol 19 Aredia Breast Cancer Study Group (1996). "Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases." *New England Journal of Medicine* 335(24): 1785–92 (cit. on p. 52).

Hu, X. J., J. Sun and L.-J. Wei (2003). "Regression parameter estimation from panel counts." *Scandinavian Journal of Statistics* 30(1): 25–43 (cit. on p. 29).

Huang, J. (1994). "Estimation in Regression Models with Interval Censoring." PhD thesis. Department of Statistics: University of Washington (cit. on p. 24).

Huang, J. (1996). "Efficient estimation for the proportional hazards model with interval censoring." *Annals of Statistics* 24(2): 540–68 (cit. on pp. 24, 42).

Huang, J. (1999). "Asymptotic properties of nonparametric estimation based on partly interval-censored data." *Statistica Sinica* 9(2): 501–19 (cit. on p. 23).

Huang, J. and A. J. Rossini (1997). "Sieve estimation for the proportional-odds failure-time regression model with interval censoring." *Journal of the American Statistical Association* 92(439): 960–67 (cit. on pp. 25, 71).

Huang, J. and J. A. Wellner (1995). *Efficient Estimation for the Proportional Hazards Model with 'Case 2' Interval Censoring*. Technical Report 290. Seattle: Department of Statistics, University of Washington (cit. on pp. 24, 25, 34, 40).

Hudgens, M. G., G. A. Satten and I. M. Longini Jr. (2001). "Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation." *Biometrics* 57(1): 74–80 (cit. on p. 23).

Huffer, F. W. and I. W. McKeague (1991). "Weighted least squares estimation for Aalen's additive risk model." *Journal of the American Statistical Association* 86 (413): 114–29 (cit. on pp. 8, 9, 29).

IBM (2012). *ILOG CPLEX Optimization Studio*. Version 12.5 (cit. on pp. 45, 48).

Jackson, C. (2011). "Multi-state models for panel data: The msm package for R." *Journal of Statistical Software* 38(8) (cit. on pp. 58, 86).

Jacobsen, M. (2006). *Point Process Theory and Applications*. Boston: Birkhäuser (cit. on p. 4).

Jewell, N. P., M. van der Laan and T. Henneman (2003). "Nonparametric estimation from current status data with competing risks." *Biometrika* 90(1): 183–97 (cit. on p. 23).

Johansen, S. (1983). "An extension of Cox's regression model." *International Statistical Review* 51(2): 165–74 (cit. on p. 7).

Joly, P., D. Commenges, C. Helmer and L. Letenneur (2002). "A penalized likelihood approach for an illness-death model with interval-censored data: Application to age-specific incidence of dementia." *Biostatistics* 3(3): 433–43 (cit. on pp. 21, 25, 86, 89).

Joly, P., D. Commenges and L. Letenneur (1998). "A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia." *Biometrics* 54(1): 185–94 (cit. on p. 25).

Jongbloed, G. (1998). "The iterative convex minorant algorithm for nonparametric estimation." *Journal of Computational and Graphical Statistics* 7(3): 310–21 (cit. on pp. 22, 44, 45).

Kalbfleisch, J. D. and J. F. Lawless (1985). "The analysis of panel data under a Markov assumption." *Journal of the American Statistical Association* 80(392): 863–71 (cit. on pp. 26, 58, 86).

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley (cit. on p. 5).

Kanwal, R. P. (1997). *Linear Integral Equations*. 2nd ed. Boston: Birkhäuser (cit. on pp. 42, 69).

Ke, C., B. Ding, J. Wang and Q. Jiang (2011). "Analysis of a Composite Endpoint Under Different Censoring Schemes For Component Events." Joint Statistical Meetings contributed paper presentation (cit. on pp. 58, 72).

Kim, J. S. (2003). "Maximum likelihood estimation for the proportional hazards model with partly interval-censored data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2): 489–502 (cit. on p. 24).

van der Laan, M. J. and A. Hubbard (1997). "Estimation with interval censored data and covariates." *Lifetime Data Analysis* 3(1): 77–91 (cit. on p. 23).

van der Laan, M. J. and J. M. Robins (1998). "Locally efficient estimation with current status data and time-dependent covariates." *Journal of the American Statistical Association* 93(442): 693–701 (cit. on p. 23).

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. 2nd ed. New York: Wiley (cit. on pp. 26, 29, 31, 39, 62).

Lin, D. Y., D. Oakes and Z. Ying (1998). “Additive hazards regression with current status data.” *Biometrika* 85(2): 289–98 (cit. on p. 24).

Lin, D. Y. and Z. Ying (1994). “Semiparametric analysis of the additive risk model.” *Biometrika* 81(1): 61–71 (cit. on pp. 8, 24, 25).

Lindsey, J. C. and L. M. Ryan (1998). “Methods for interval-censored data.” *Statistics in Medicine* 17(2): 219–38 (cit. on p. 26).

Maathuis, M. H. (2006). “Nonparametric estimation for current status data with competing risks.” PhD thesis. University of Washington (cit. on p. 23).

Major, P., A. Lortholary, J. Hon, E. Abdi, G. Mills, H. D. Menssen, F. Yunus, R. Bell, J. Body, E. Quebe-Fehling and J. Seaman (2001). “Zoledronic acid is superior to pamidronate in the treatment of hypercalcemia of malignancy: A pooled analysis of two randomized, controlled clinical trials.” *Journal of Clinical Oncology* 19(2): 558–67 (cit. on p. 78).

Martinussen, T. and T. H. Scheike (2002). “Efficient estimation in additive hazards regression with current status data.” *Biometrika* 89(3): 649–58 (cit. on p. 24).

Martinussen, T. and T. H. Scheike (2006). *Dynamic Regression Models for Survival Data*. New York: Springer (cit. on pp. 2, 9, 48, 53, 55, 107).

McKeague, I. W. and P. D. Sasieni (1994). “A partly parametric additive risk model.” *Biometrika* 81(3): 501–14 (cit. on p. 8).

Murphy, S. A. and A. W. van der Vaart (1997). “Semiparametric likelihood ratio inference.” *The Annals of Statistics* 25(4): 1471–509 (cit. on pp. 15, 34, 37).

Murphy, S. A. and A. W. van der Vaart (2000). “On profile likelihood.” *Journal of the American Statistical Association* 95(450): 449–65 (cit. on pp. 17, 18, 25, 43).

Murphy, S. A., A. W. van der Vaart and J. A. Wellner (1999). “Current status regression.” *Mathematical Methods of Statistics* 8(3): 407–25 (cit. on p. 25).

Pan, W. (1999). “Extending the iterative convex minorant algorithm to the Cox model for interval-censored data.” *Journal of Computational and Graphical Statistics* 8(1): 109–20 (cit. on p. 44).

Peng, L. and J. P. Fine (2007). “Regression modeling of semicompeting risks data.” *Biometrics* 63(1): 96–108 (cit. on p. 27).

Petroni, G. R. and R. A. Wolfe (1994). “A two-sample test for stochastic ordering with interval-censored data.” *Biometrics* 50(1): 77–87 (cit. on p. 24).

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing*. 3rd ed. Cambridge University Press (cit. on p. 46).

Rabinowitz, D., A. A. Tsiatis and J. Aragón (1995). “Regression with interval-censored data.” *Biometrika* 82(3): 501–13 (cit. on p. 24).

Rebolledo, R. (1980). “Central limit theorems for local martingales.” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 51(3): 269–86 (cit. on p. 4).

Robins, J. M. and A. Rotnitzky (1992). “Recovery of information and adjustment for dependent censoring using surrogate markers.” In *AIDS Epidemiology: Methodological Issues*. Ed. by N. Jewell, K. Dietz and V. Farewell. Birkhäuser, 297–331 (cit. on p. 23).

Rosen, L. S., D. Gordon, M. Kaminski, A. Howell, A. Belch, J. Mackey, J. Apffelstaedt, M. Hussein, R. E. Coleman, D. J. Reitsma, J. J. Seaman, B. L. Chen and Y. Ambros (2001). “Zoledronic acid versus pamidronate in the treatment of skeletal metastases in patients with breast cancer or osteolytic lesions of multiple myeloma: A phase III, double-blind, comparative trial.” *Cancer Journal* 7(5): 377–87 (cit. on p. 78).

Rossini, A. J. and A. A. Tsiatis (1996). “A semiparametric proportional odds regression model for the analysis of current status data.” *Journal of the American Statistical Association* 91(434): 713–21 (cit. on p. 24).

Sasieni, P. D. (1992). “Information bounds for the additive and multiplicative intensity models.” In *Survival Analysis: State of the Art*. Ed. by J. P. Klein and P. K. Goel. Dordrecht: Kluwer Academic, 249–65 (cit. on pp. 8, 9).

Satten, G. A. (1996). “Rank-based inference in the proportional hazards model for interval censored data.” *Biometrika* 83(2): 355–70 (cit. on p. 25).

Satten, G. A. and M. R. Sternberg (1999). “Fitting semi-Markov models to interval-censored data with unknown initiation times.” *Biometrics* 55(2): 507–13 (cit. on p. 25).

Scheike, T. H. and M.-J. Zhang (2002). “An additive-multiplicative Cox-Aalen regression model.” *Scandinavian Journal of Statistics* 29(1): 75–88 (cit. on pp. 9, 29).

Schick, A. and Q. Yu (2000). “Consistency of the GMLE with mixed case interval-censored data.” *Scandinavian Journal of Statistics* 27(1): 45–55 (cit. on pp. 21, 22, 29).

Shen, X. (2000). “Linear regression with current status data.” *Journal of the American Statistical Association* 95(451): 842–52 (cit. on p. 25).

Sternberg, M. R. and G. A. Satten (1999). “Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation.” *Biometrics* 55(2): 514–22 (cit. on p. 23).

Sun, J. (1997). “Regression analysis of interval-censored failure time data.” *Statistics in Medicine* 16(5): 497–504 (cit. on p. 26).

Sun, J. (1999). "A nonparametric test for current status data with unequal censoring." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(1): 243–50 (cit. on p. 24).

Sun, J. and J. D. Kalbfleisch (1993). "The analysis of current status data on point processes." *Journal of the American Statistical Association* 88(424): 1449–54 (cit. on p. 24).

Sun, J. and J. S. Shen (2009). "Efficient estimation for the proportional hazards model with competing risks and current status data." *Canadian Journal of Statistics* 37(4): 592–606 (cit. on p. 24).

Sun, J. and L. Sun (2005). "Semiparametric linear transformation models for current status data." *Canadian Journal of Statistics* 33(1): 85–96 (cit. on p. 24).

Sun, J., X. Tong and X. He (2007). "Regression analysis of panel count data with dependent observation times." *Biometrics* 63(4): 1053–59 (cit. on p. 29).

Sun, J. and L. J. Wei (2000). "Regression analysis of panel count data with covariate-dependent observation and censoring times." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62(2): 293–302 (cit. on p. 29).

Sun, J., Q. Zhao and X. Zhao (2005). "Generalized log-rank tests for interval-censored failure time data." *Scandinavian Journal of Statistics* 32(1): 49–57 (cit. on p. 24).

Sutradhar, R. and R. J. Cook (2008). "Analysis of interval-censored data from clustered multistate processes: application to joint damage in psoriatic arthritis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(5): 553–66 (cit. on p. 26).

Szarek, M. (2008). "Estimation of the distribution of time to first event in a composite endpoint from interval censored observations with incomplete non-fatal event status." PhD thesis. New York University (cit. on p. 89).

Therneau, T. (2012). *A Package for Survival Analysis in S*. R package version 2.37-2 (cit. on pp. 52, 72, 79, 93).

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag (cit. on p. 100).

Titman, A. C. (2011). "Flexible nonhomogeneous Markov models for panel observed data." *Biometrics* 67(3): 780–87 (cit. on p. 26).

Tolusso, D. and R. J. Cook (2009). "Robust estimation of state occupancy probabilities for interval-censored multistate data: An application involving spondylitis in psoriatic arthritis." *Communications in Statistics - Theory and Methods* 38(18): 3307–25 (cit. on p. 25).

Turnbull, B. W. (1976). "The empirical distribution function with arbitrarily grouped, censored and truncated data." *Journal of the Royal Statistical Society: Series B (Methodological)* 38(3): 290–95 (cit. on pp. 22, 23, 32, 69).

van der Vaart, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. Vol. 44. CWI Tracts (cit. on p. 24).

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press (cit. on pp. 10, 12, 13, 16, 42).

van der Vaart, A. (2002). "Semiparametric statistics." In *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXIX - 1999*. Ed. by P. Bernard. Lecture Notes in Mathematics. Saint-Flour: Springer (cit. on p. 10).

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer (cit. on pp. 10, 12–15, 24).

van der Vaart, A. and J. A. Wellner (2000). "Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes." In *High Dimensional Probability II*. Ed. by E. Giné, D. M. Mason and J. A. Wellner. Boston: Birkhäuser, 115–33 (cit. on pp. 12, 22, 29, 34, 36).

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate." *The Annals of Mathematical Statistics* 20(4): 595–601 (cit. on p. 13).

Wang, W. (2003). "Estimating the association parameter for copula models under dependent censoring." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(1): 257–73 (cit. on p. 27).

Wang, Y. (2008). "Dimension-reduced nonparametric maximum likelihood computation for interval-censored data." *Computational Statistics & Data Analysis* 52(5): 2388–402 (cit. on p. 107).

Wellner, J. A. and Y. Zhang (2007). "Two likelihood-based semiparametric estimation methods for panel count data with covariates." *Annals of Statistics* 35 (5): 2106–42 (cit. on pp. 24, 29, 34, 37, 39, 44, 108).

Wen, C.-C. (2012). "Cox regression for mixed case interval-censored data with covariate errors." *Lifetime Data Analysis* 18(3): 321–38 (cit. on pp. 24, 29).

Wu, C. F. J. (1983). "On the convergence properties of the EM algorithm." *The Annals of Statistics* 11(1): 95–103 (cit. on pp. 22, 70).

Xu, J., J. D. Kalbfleisch and B. Tai (2010). "Statistical analysis of illness-death processes and semicompeting risks data." *Biometrics* 66(3): 716–25 (cit. on p. 27).

Yashin, A. and E. Arjas (1988). "A note on random intensities and conditional survival functions." *Journal of Applied Probability* 25(3): 630–35 (cit. on p. 5).

Yu, Q., A. Schick, L. Li and G. Y. C. Wong (1998). "Asymptotic properties of the GMLE with case 2 interval-censored data." *Statistics & Probability Letters* 37(3): 223–28 (cit. on p. 23).

Yuan, Y., P. F. Thall and J. E. Wolff (2012). "Estimating progression-free survival in paediatric brain tumour patients when some progression statuses are unknown." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61(1): 135–49 (cit. on pp. 28, 58).

Yuen, K.-C., J. Shi and L. Zhu (2006). "A *k*-sample test with interval censored data." *Biometrika* 93(2): 315–28 (cit. on p. 24).

Zeng, D., J. Cai and Y. Shen (2006). "Semiparametric additive risks model for interval-censored data." *Statistica Sinica* 16(1): 287–302 (cit. on pp. 24, 46).

Zhang, Y., L. Hua and J. Huang (2010). "A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data." *Scandinavian Journal of Statistics* 37(2): 338–54 (cit. on pp. 25, 62–64, 108).

Zhang, Z., L. Sun, X. Zhao and J. Sun (2005). "Regression analysis of interval-censored failure time data with linear transformation models." *Canadian Journal of Statistics* 33(1): 61–70 (cit. on p. 24).

Zhao, Q. and J. Sun (2004). "Generalized log-rank test for mixed interval-censored failure time data." *Statistics in Medicine* 23(10): 1621–29 (cit. on p. 24).

Zhao, X., Q. Zhao, J. Sun and J. S. Kim (2008). "Generalized log-rank tests for partly interval-censored failure time data." *Biometrical Journal* 50(3): 375–85 (cit. on p. 24).