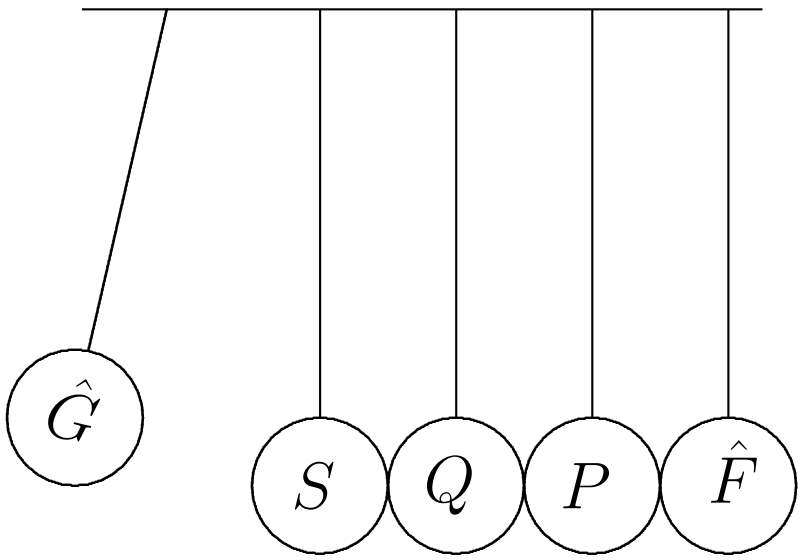
The Analysis of doubly censored Survival Data

An Application to Data collected from the Amsterdam Cohort Studies on HIV Infection and AIDS



Longhow Lam

Contents

List of Figures

Acknowledgements

This report presents the results of a one-year project performed at the division of Public Health and Environment of the Municipal Health Service in Amsterdam. This project completes my two-year post-M.Sc. mathematical design engineering program at the department Technical Mathematics and Informatics at the Delft University of Technology (in Dutch: postdoctorale opleiding Wiskundige Beheers- en Beleidsmodellen). The examination board was formed by Dr. R.B. Geskus, Dr. I.P.M. Keet (Municipal Health Service Amsterdam), Dr. H.P. Lopuhaä, Prof. Dr. P. Groeneboom, Prof. Dr. R.M. Cooke (Delft University of Technology).

I would like to thank the Municipal Health Service for offering me the opportunity to work for one year on the project and providing me with an excellent working environment. At the Municipal Health Service I would like to thank the following people involved in the project. Dr. Ronald Geskus for his supervision of the project, his great support, criticism and feedback on both mathematical and non mathematical issues was an import factor for the completion of the project. Dr. René Keet for his support on epidemiological and HIV related issues. Drs. Ethel Boucher and Nel Albrecht-van Lent for providing data in a format suitable for analyses.

At the Delft University of Technology I would like to thank the following people involved in the project. Prof. Dr. Piet Groeneboom for his mathematical support and programming work. Dr. Rik Lopuhaä and Ir. Bert van Zomeren for their mathematical support. Peter van der Wijden for his programming work.

Prof. André Tits of the Institute for Systems Research at the University of Maryland is acknowledged for making the C routine 'cfsqp' available.

All the participants of the Amsterdam cohort studies are acknowledged for their contribution to AIDS research. As a mathematician I didn't see any of those participants. However, I realize that the simple expression $(u_i, v_i, z_i, \delta_i)$ $i = 1, \ldots, N$ I am using in this report represents the tragic history of a lot of persons.

Last but not least, I would like to thank my roommates and all my colleagues at the Municipal Health Service for making the stay at the Municipal Health service such a pleasant one. Special thanks must be given to Nel, Maria, Marja and the secretariat. Without their good cups of coffee, tea and water I certainly wouldn't have survived my long computer sessions.

Amsterdam, August 1997 Longhow Lam

Chapter 1

Introduction

1.1 The doubly censored Incubation Time of AIDS

The incubation time of a disease is defined as the time interval between time of infection and time of manifestation of clinical disease. In case of AIDS, many studies can not identify the date of infection with the human immunodeficiency virus (HIV), the virus which causes AIDS. Then the incubation time of AIDS (= time to AIDS) is defined as the time interval between the time of seroconversion, the appearance of antibodies against HIV, and the time of diagnosis of AIDS. The incubation time of AIDS is of interest for several reasons. It is important to find significant covariates of the incubation time, *i.e.* factors that do have an influence on the duration of the incubation time. Furthermore, the incubation time is of interest for its key role in the method of backcalculating for estimating the number of HIV positive individuals in the population (see [5]), and in the development of mathematical models of the epidemic.

For statistical analysis it would be ideal to know exactly the time of seroconversion and the time of diagnosis for all participants in a cohort study on AIDS. However, due to the periodic screening and limited follow-up time of participants both time points are likely to be censored. In fact, in every study a person's seroconversion moment is only known to have occurred after time u of the last seronegative test and before time v of the first seropositive test. The time interval $(u, v]$ is called the seroconversion interval. For some cases the last negative test is not available, then for time u we may take a lowerbound for the moment of seroconversion. The moment of AIDS diagnosis z is known exactly or is right censored if the diagnosis has not yet been observed. This kind of censoring is often referred to as double censoring, since both the begin point and the end point of the incubation time can be censored, see figure 1.1. Therefore, with double censoring two time scales play a role in statistical inferences: the seroconversion time scale, usually the calendar time scale on which the seroconversion intervals are situated, and the incubation time scale, the time scale on which the incubation time lives, which ranges from 0 to 20 years, or more.

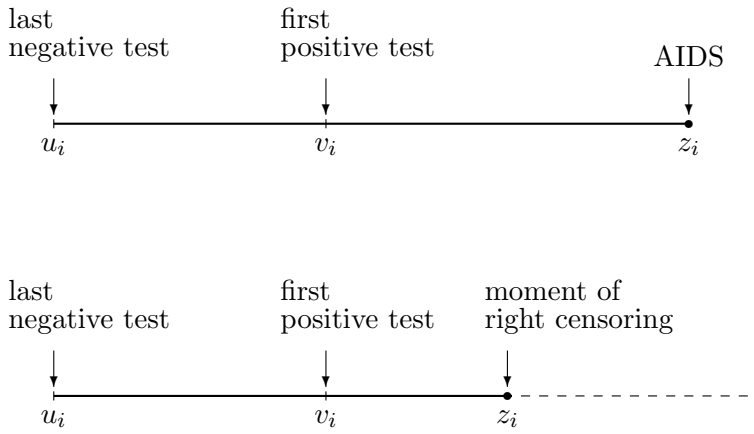


Figure 1.1: Double censoring: Moment of seroconversion is interval censored and moment of AIDS is possibly right censored

Outline of this report

The remaining part of this introduction gives an overview of the available data in Amsterdam. Chapter 2 deals with the methods that are already used in the analysis of the incubation time. Chapter 3 is devoted to the nonparametric approach of double censoring and the calculation of the estimators. In chapter 4 we will discuss how to include seroprevalent cases into the analysis of the incubation time. Chapter 5 will discuss how to use parametric models for doubly censored data and the summary will summarize the main results.

1.2 The Amsterdam Cohort Studies

In order to study the incubation time and several other aspects of the Human Immunodeficiency Virus (HIV) infection and AIDS, several cohort studies were started in Amsterdam in the early eighties. A detailed description of the Amsterdam cohort studies can be found in [2, 25, 24]

The HOM-study

The HOM-study consists of homosexual men living mainly in and around the city of Amsterdam. The study was started in October 1984. Men were recruited trough announcements in the gay press, advertisements and by word of mouth. Only those who were free of AIDS symptoms were allowed to enter the study. Until April 1985 748 men entered the study of whom about one third was already seropositive; these men are called seroprevalent cases. We only know that the seroconversion of these men took place between the start of the HIV epidemic and their entrance into the study (between October 1984 and April 1985). For the

start of the HIV epidemic we chose 1980. This is based on the study in [27], where it is shown that the number of HIV infections before 1980 is very low. Therefore the seroprevalent cases result in wide seroconversion intervals of more than 4 years. Between April '85 and February '88 only seronegative men could enter the study. Together with persons who where negative at the start of the study this resulted in a so-called seroconverter group of over 100 men with narrow seroconversion intervals of a few months width. After 1988 both seropositive and seronegative men could enter the study. Seropositive men who entered the study at this stage have very long seroconversion intervals, since we only now that seroconversion took place between 1980 and sometime after 1988. Figure 1.2 shows the seroconversion intervals for the HOM-study ordered by the left endpoints of the intervals, it clearly shows the three different groups: the seroprevalent cases entering in 84/85, the seropositives entering after 1988 and the seroconverters. Seropositives were seen every 3 months. Clinical, epidemiological and social scientific data are collected with standardized questionnaires and by physical examination. Blood samples are drawn and cryo preserved for virological and immunological tests.

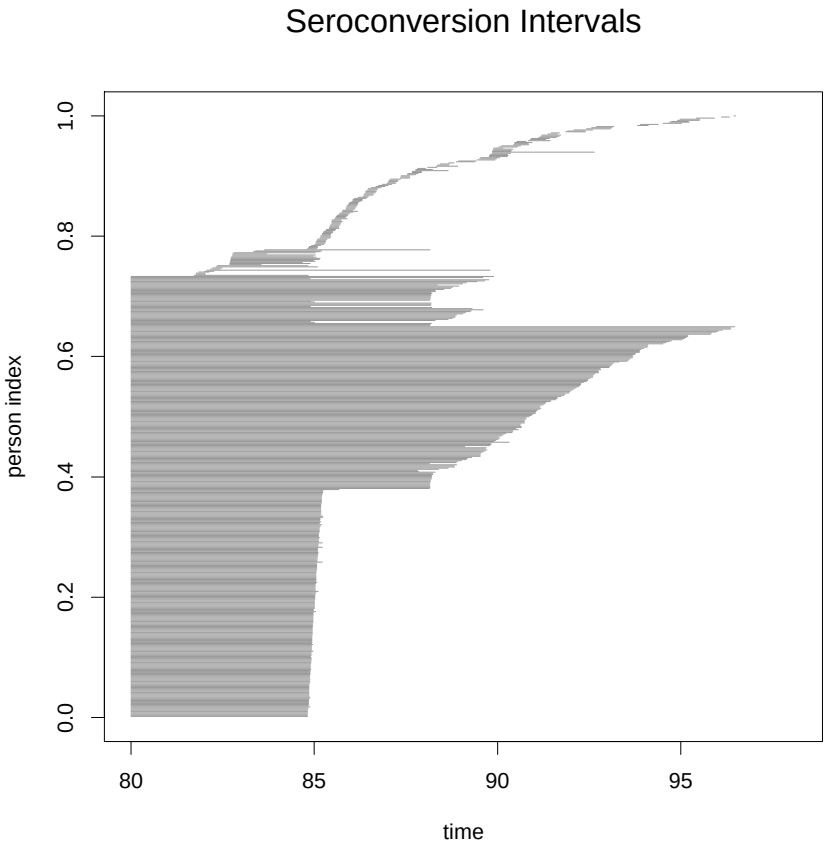


Figure 1.2: Seroconversion intervals for the HOM-study

The HBvac-study

The HBvac-study consists of homosexual men who have participated in a clinical trial to test the efficacy of a hepatitis B vaccine, the so-called HBV studies [26]. A total of 800 men were enrolled, 120 of these 800 men entered the HOM-study in 1984. In the HBV studies blood samples were taken and stored at a regular basis. In 1990 part of the men who participated in the HBV studies were asked to participate in an HIV-1 follow-up study and HIV-1 antibody testing of specimens collected and stored between 1982 and 1990. The seroconversion intervals of 89 men were found via stored blood samples or via follow-up. Figure 1.3 shows these intervals ordered by the left endpoint of the intervals. We see that the majority of the individuals were already seropositive in 1990. For some of those individuals a last negative test could only be restored from the blood samples, resulting in long seroconversion intervals

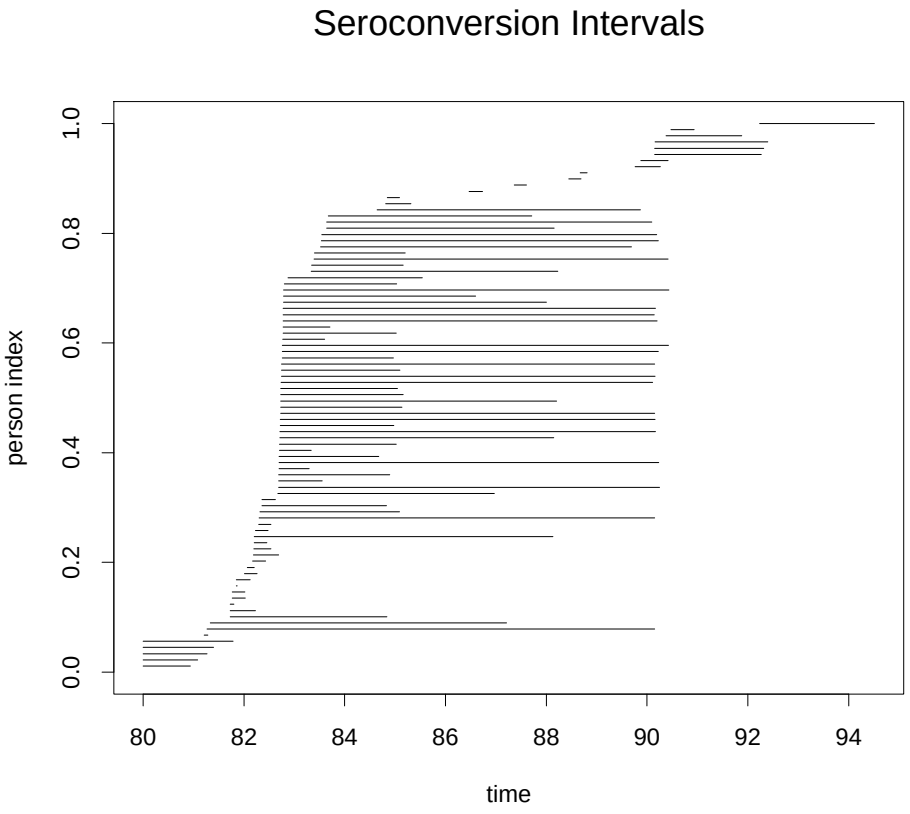


Figure 1.3: Seroconversion intervals for the HBvac-study

The IDU-study

The IDU-study consists of injecting drug users and was started at the end of 1985. At that time only one drug user had been reported with AIDS. However, studies in the US [10] showed that HIV could spread very fast among injecting drug users. The cohort study among drug users is an open study in which new participants have been enrolled continuously. Participants were recruited at methadone outposts and at sexually transmitted disease clinics for drug-using prostitutes. Some epidemiological results of the cohort study among drug users can be found in [19]. Figure 1.4 shows the seroconversion intervals of 99 participants of the IDU-study ordered by the left endpoint of the intervals.

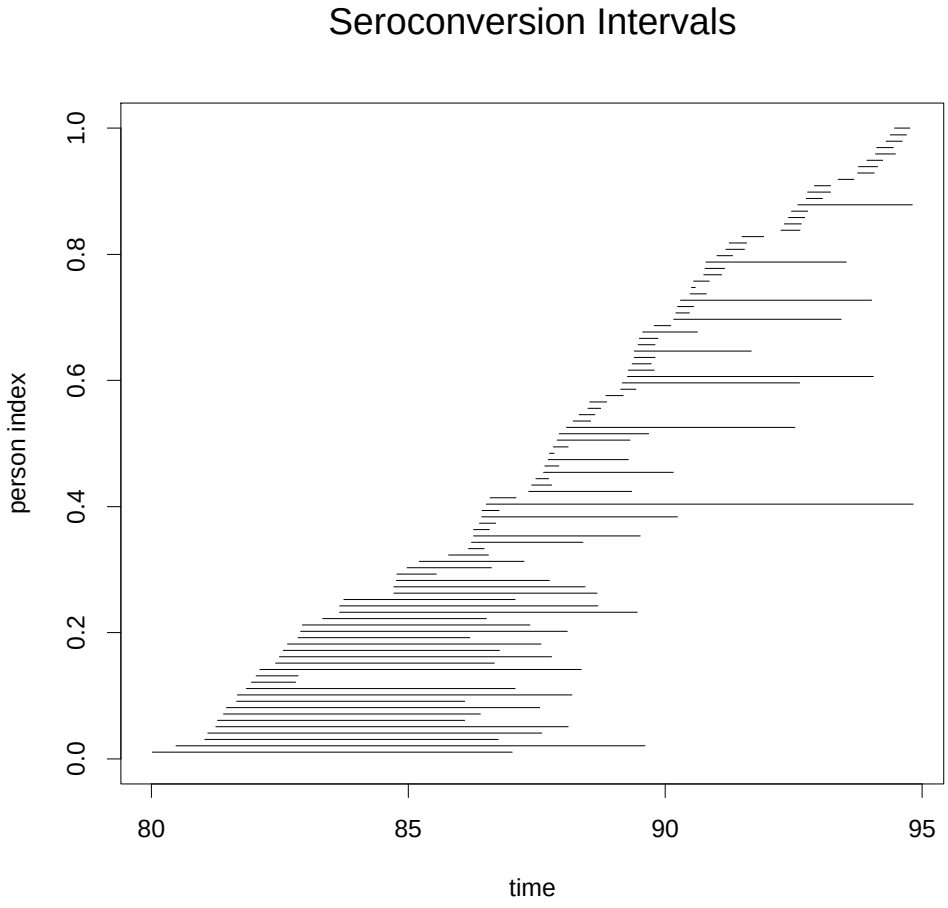


Figure 1.4: Seroconversion intervals for the IDU-study

Chapter 2

Already used Methods in the Amsterdam Cohort Studies

In this chapter we describe two methods that have already been used to estimate the AIDS incubation time of the Amsterdam cohorts, the Kaplan-Meier method and the interval transformation method. These methods are univariate methods, *i.e.* the doubly censored nature of the data is not taken fully account of.

2.1 The Kaplan-Meier Estimator

2.1.1 The Likelihood

The Kaplan-Meier estimator (KME) is the most widely used estimator in the analysis of the HIV incubation time. Just take a random volume of the American Journal of Epidemiology and the Kaplan-Meier plots will emerge. The KME is suitable for estimating the distribution function F of a random variable T if one has a sample that consists of directly observable realizations and right-censored observations of T. If we assume that the censoring mechanism is independent of T, one can write down the log-likelihood as follows:

$$\log L \;=\; \sum_{i}^{n} \delta_i \log(f(t_i)) + (1-\delta_i)\log(1-F(t_i^{\mathrm{rc}})) \tag{2.1}$$

where $\delta_i = 1$ if observation i is observed directly and $\delta_i = 0$ if observation i is right censored.

Maximizing (2.1) with respect to F under the restriction that $F(t_2) \geq F(t_1)$ if $t_2 \geq t_1$, results in the maximum likelihood estimator $\hat{F}_n$. One can show (see [4]) that $\hat{F}_n$ satisfies

$$1 - \hat{F}_n(t) = \prod_{i:\ t_i<t} \frac{n_i - d_i}{n_i} \tag{2.2}$$

where n_i is the number of individuals at risk just prior to t_i and d_i is the number of uncensored individuals who have an AIDS diagnosis at t_i. So from (2.2) we see that the survival curve

$S(t) = 1 - F(t)$ is a step function and has steps at the directly observable realizations t_i. Confidence intervals for the survivor function can be obtained by Greenwood's formula (see [4]) for the asymptotic variance of $1 - \hat{F}(t)$:

$$\widehat{\text{var}}\,[1 - \hat{F}(t)] = (1 - \hat{F}(t))^2 \sum_{i:\ t_i<t} \frac{d_i}{n_i(n_i - d_i)}.$$

2.1.2 Midpoint Imputation

In the cohort studies we have no directly observable realizations of the incubation time, since we only know that the moment of seroconversion of individual i lies in the interval (u_i, v_i). An ad hoc approach is to impute the moment of seroconversion as the midpoint of the interval $m_i = (u_i + v_i)/2$. For individuals who have an AIDS diagnosis at time z_i, the incubation time $t_i = z_i - m_i$ is observed and for individuals who did not have AIDS at time z_i the incubation time is right censored with value $t_i^{\text{rc}} = z_i - m_i$. With these imputed incubation times we can use the Kaplan-Meier estimator. Figure 2.1 shows an example of a Kaplan-Meier estimator together with the 95% confidence intervals for the data of the seroconverter group of the HOM-study. The strategy of imputing midpoints will not induce large biases in the estimator of the incubation time distribution as long as the majority of the seroconversion intervals is small. As a rule of thumb, if the width of the seroconversion intervals is smaller than 2 years, the bias in the estimator of the incubation time is small, see [5]. For the HOM-study this means that we have to leave out a lot of seroprevalent cases, since these cases have (very) wide seroconversion intervals, see figure 1.3.

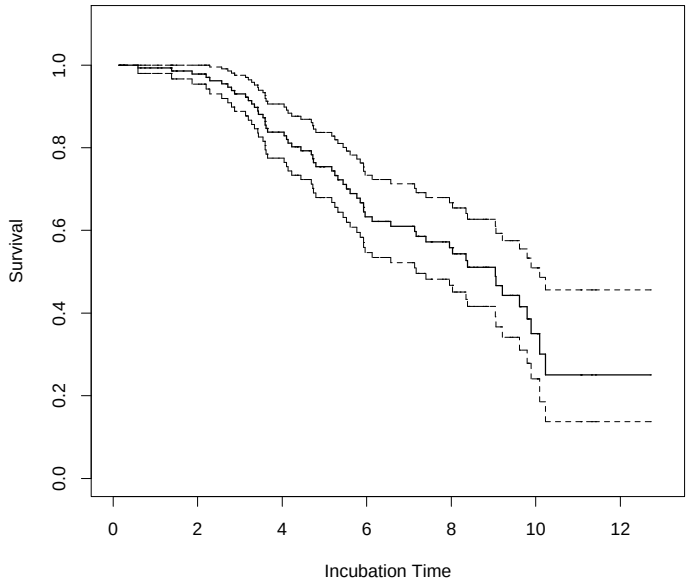


Figure 2.1: Kaplan-Meier estimate of the incubation time of the seroconverters, based on midpoint imputation.

2.1.3 Expected Seroconversion Dates

A better approach, certainly for wider seroconversion intervals, is to estimate the date of seroconversion by the expected date of seroconversion, given that seroconversion occurred between u_i and v_i. To do this we need an estimate $\hat{G}$ of the seroconversion distribution, this estimate is based on the observed seroconversion intervals. An nonparametric estimator for G is the nonparametric maximum likelihood estimator, described in the next section. Given this seroconversion distribution $\hat{G}$ and the seroconversion interval (u_i, v_i) for person i we can calculate the corresponding expected date of seroconversion s_i,

$$s_i = \frac{\int_{u_i}^{v_i} s d\hat{G}(s)}{\hat{G}(v_i) - \hat{G}(u_i)}. \tag{2.3}$$

The Kaplan-Meier is then used on the estimated incubation times $t_i = z_i - s_i$ to estimate the incubation time distribution. For the HBvac-study this strategy is applied since this cohort contains some very long seroconversion intervals. First an estimation of the seroconversion distribution G is calculated based on the seroconversion intervals shown in figure 1.3; then instead of using midpoints the expected seroconversion dates are calculated based on G. Figure 2.2 shows the expected seroconversion dates. For the longer intervals these dates are shifted one or two years to the left compared with the midpoints. Because of these earlier seroconversion the survival of the total group is slightly better, compared with the survival based on midpoints as illustrated in figure 2.3.

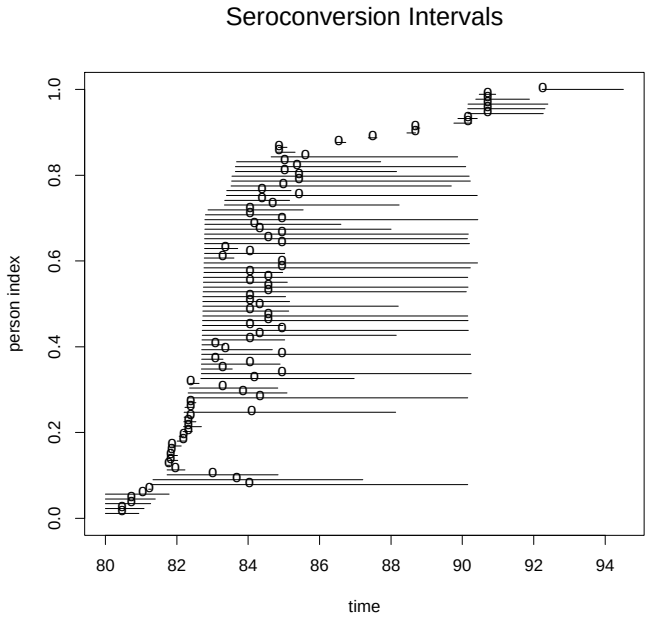


Figure 2.2: Seroconversion intervals of the HBvac-study, together with the expected seroconversion dates (o).

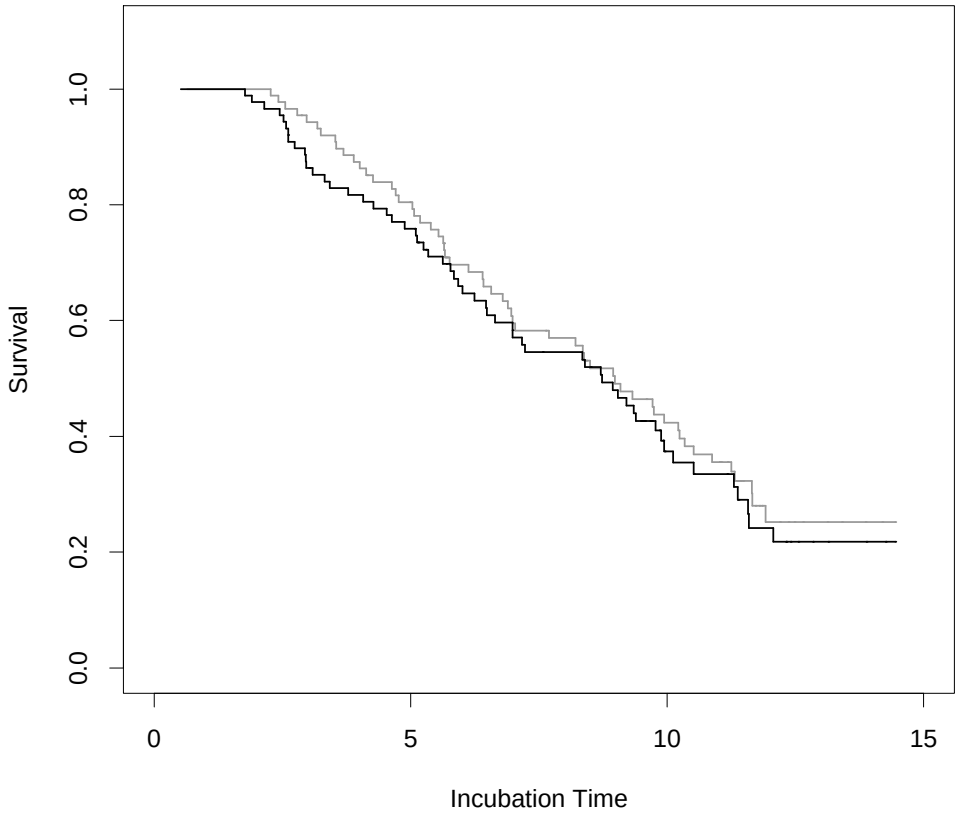


Figure 2.3: dark curve: estimation based on midpoints, light curve: estimation based on expected seroconversion dates

2.2 Interval Transformation

If seroconversion intervals are wide then choosing a seroconversion date in that interval by the midpoint method can lead to biases in the estimate of the incubation time. Instead of using partially right censored data as in the Kaplan-Meier case, we can take account of the interval censored nature of the data. For a person with an observed AIDS diagnosis we can transform the seroconversion interval into an interval in the incubation time scale (figure 2.4 illustrates this). For a person with a right censored AIDS diagnosis the transformation results in a minimal incubation time. The incubation interval is unbounded as illustrated in figure 2.5. After this transformation of the data we can apply a nonparametric maximum likelihood method for interval censored data, which is described extensively in [1]. The estimator for the incubation time distribution F is obtained by maximizing the following log-likelihood:

$$\log L \;=\; \sum_{i=1}^{n} \delta_i \log(F(r_i) - F(l_i)) \;+\; (1-\delta_i)\log(1-F(l_i)) \tag{2.4}$$

under order restrictions for F, *i.e.* F must satisfy the conditions of a distribution function. Here $\delta_i = 1$ if individual i has an AIDS diagnosis, and $(l_i, r_i]$ is the incubation interval obtained after transformation and contains the unobserved incubation time. If $\delta_i = 0$ then

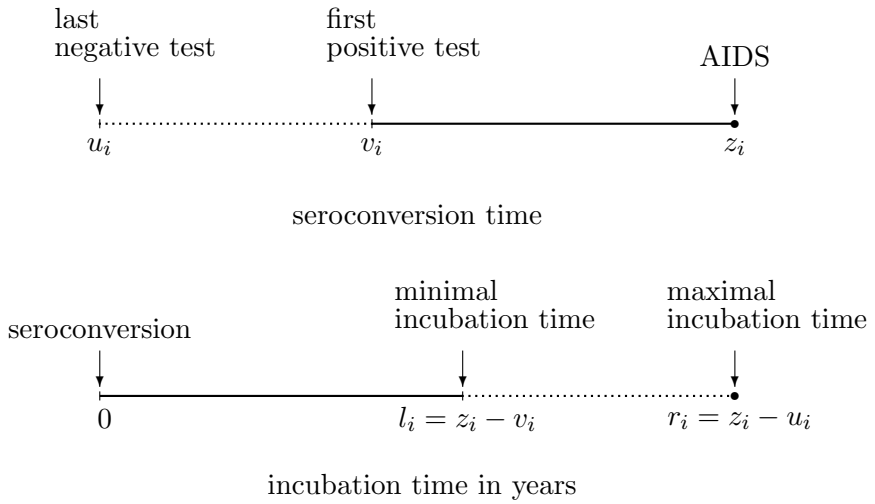


Figure 2.4: Transformation from seroconversion time scale to incubation time scale

individual i has a right censored AIDS diagnosis and we only know that the incubation time is longer than l_i. Optimization of (2.4) is not as straightforward as in the Kaplan-Meier case, an iterative search algorithm is needed to find the solution. The iterative convex minorant (ICM) algorithm is a simple and fast algorithm to find the solution in case of interval censored data.

We note that likelihood approach (2.4) is only an approximation of the real likelihood, since the method described in [1] assumes that the observation times are independent of the incubation time. However, for interval transformation this is not the case. For example, take an individual i with an observed AIDS diagnosis. His observed data is $(u_i, v_i, z_i, \delta_i = 1)$. If we transform this to interval censored data for the incubation time we get: $(l_i = z_i - v_i, r_i = z_i - u_i, \delta_i = 1)$, the probability of this event is given by $\mathbb{P}(L = l_i, R = r_i, l_i \leq T \leq r_i)$. So if F is the distribution function of the incubation time T and F is independent of the observation time distribution (L, R) then we would get the term corresponding with $\delta = 1$ in formula (2.4). However this not the case since l_i and r_i contain z_i, which is the sum of seroconversion period and the incubation time period.

2.2.1 The ICM Algorithm

Maximization of (2.4) boils down to minimization of a convex function ϕ over the cone $C = \{\beta \in \mathbb{R}^n : 0 \leq \beta_1 \leq \ldots \leq \beta_n \leq 1\}$. In case of interval transformation ϕ has the form

$$\phi(\beta) \;=\; -\sum_{i=1}^{N} \delta_i \log(\beta_{j_i} - \beta_{k_i}) + (1 - \delta_i) \log(1 - \beta_{j_i}) \;,$$

where j_i and k_i are indices indicating which β's correspond with $F(r_i)$ and $F(l_i)$ respectively. The ICM algorithm is especially designed for such problems, see [14]. In the k-th iteration

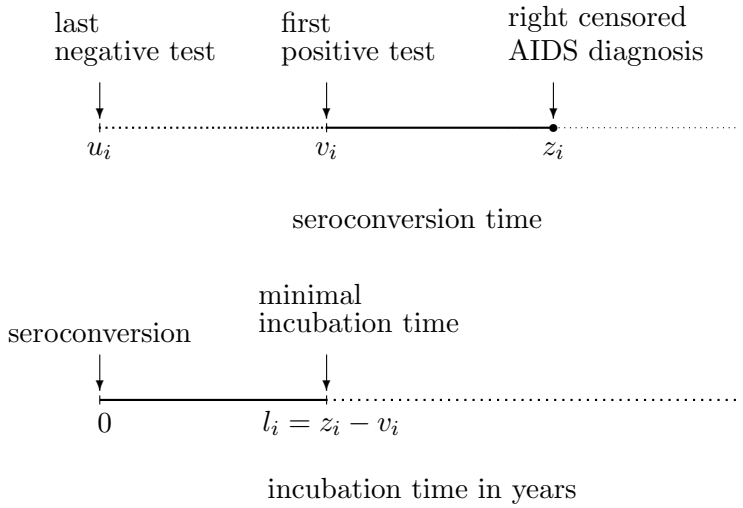


Figure 2.5: Transformation from seroconversion time scale to incubation time scale

of the algorithm, ϕ is locally approximated at $\beta^{(k)}$ with a special Taylor expansion,

$$q(\beta) = \phi(\beta^{(k)}) + (\beta - y^{(k)})' W(\beta^{(k)})(\beta - y^{(k)}) \ ,$$

where the Hessian matrix $W(\beta^{(k)})$ evaluated at $\beta^{(k)}$ only contains the diagonal elements,

$$\begin{aligned} W(\beta^{(k)}) &= \operatorname{diag}(w_i), \ w_i = \frac{\partial^2}{\partial\beta_i \partial\beta_i} \ \phi(\beta^{(k)}) \\ y^{(k)} &= \beta^{(k)} - W^{-1}(\beta^{(k)}) \nabla\phi(\beta^{(k)}). \end{aligned}$$

The quadratic programming problem

$$\hat{\beta} = \underset{\beta \in C}{\operatorname{argmin}}\, q(\beta)$$

can then be solved directly, $\hat{\beta}_i$ is the left derivative of the convex minorant of the cumulative sum diagram consisting of the points $P_0 = (0, 0)$ and

$$P_j = \left(\sum_{l=1}^{j} w_l \ , \sum_{l=1}^{j} w_l y_l \right) \quad j = 1, \ldots, n$$

evaluated at P_i. This result is given in [14] and illustrated in figure 2.6.

We get the following algorithm:

- step 1. $k = 0$. Choose a starting value $\beta^{(k)}$; one can take $\beta_i^{(k)} = i/n, \ i = 1, \ldots, n$.

- step 2. Calculate $\beta^{(k+1)}$ via the cumulative sum diagram:

$$\beta_*^{(k+1)} = \underset{\beta \in C}{\operatorname{argmin}} \ (\beta - y^{(k)})' W(\beta^{(k)})(\beta - y^{(k)})$$

- step 3. If convergence is reached: stop, else: proceed with step 4.

- step 4. If $\beta^{(k+1)}$ leads to a direction of sufficient descent:

$$\phi(\beta^{(k+1)}) < \phi(\beta^{(k)}) + \epsilon \nabla \phi(\beta)'(\beta^{(k+1)} - \beta^{(k)})$$

then $k = k + 1$; proceed with step 2.
Else perform a line search to create a direction of descent $\beta^{(k+1)}$, *i.e.*

$$\beta^{(k+1)} = \underset{0 \le \lambda \le 1}{\operatorname{argmin}} \, \phi(\beta^{(k)} + \lambda(\beta^{(k+1)} - \beta^{(k)})).$$

Let $k = k + 1$; proceed with step 2.

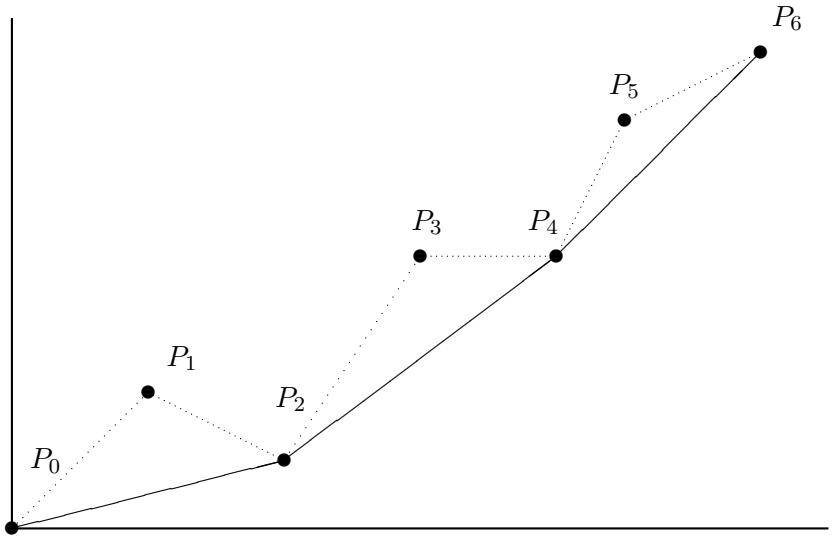


Figure 2.6: solid line: convex minorant, dashed line: cumulative sum diagram

In step 3 we have to check for convergence of the algorithm, one general way to test convergence is to check whether or not the decrease in function value is fractionally smaller than some tolerance. For minimizing convex functions on cones there is an alternative approach, convergence of the algorithm can be checked by the so-called Fenchel conditions (see [15, 14]). Without loss of generality we assume that the optimal solution $\hat{\beta}$ is blockwise constant and define $k = (k_1, \ldots, k_J)$ by

$$0 \le \hat{\beta}_1 = \ldots = \hat{\beta}_{k_1} < \hat{\beta}_{k_1+1} = \ldots = \hat{\beta}_{k_2} < \ldots < \hat{\beta}_{k_J} < \hat{\beta}_{k_J+1} = \ldots = \hat{\beta}_n \le 1 \ .$$

A convex function ϕ has its minimum $\hat{\beta}$ on the cone C if and only if:

$$\sum_{i=k_j+1}^{k_{j+1}} \hat{\beta}_i \frac{\partial \phi}{\partial \beta_i}(\hat{\beta}) \;\; = \;\; 0 \;\; j = 1, \ldots, J-1 \;\; ,$$

where the indices j also contains $j = J$ if $\beta_{K_J+1}, \ldots, \beta_n < 1$.
It is proven in [14] that the line search in the algorithm will lead to a converging algorithm. We use Armijo's rule to perform a so called inexact line search, *i.e.* we do not search for the exact location of the minimum of ϕ along the line segment

$$\left\{ y = \beta^{(k)} + \lambda(\beta^{(k+1)} - \beta^{(k)}) \mid \lambda \in [0,1] \right\} \;\; ,$$

it suffices to find an approximate line optimization. See [20] for a complete description.

In fact there is an even faster algorithm that works well for interval censoring. It is the so-called hybrid algorithm which is a combination of the ICM and the EM algorithm (see section 3.3.1). After each calculation of $\beta^{(k+1)}$ we add an extra EM step to refine $\beta^{(k+1)}$ before calculating a new iterate with the quadratic approximation. For likelihood (2.4) it is an empirical finding that this hybrid algorithm converges faster than the ICM algorithm alone.

2.2.2 Results from the Cohort Studies

Figure 2.7 shows an estimate (dark line) of the incubation time based on interval transformation. The data are from the seroconverter group of the HOM-study. For small seroconversion intervals there is not much difference in the location of the survival curves between the Kaplan-Meier method and the interval transformation method. However, the smaller number of jumps in the survival curve with the interval transformation method better reflects the uncertainty in the seroconversion dates. This is more clearly illustrated if we use the data from the HBvac-study, where the uncertainty in the seroconversion dates is larger. The difference between interval transformation and Kaplan-Meier is illustrated in figure 2.8.

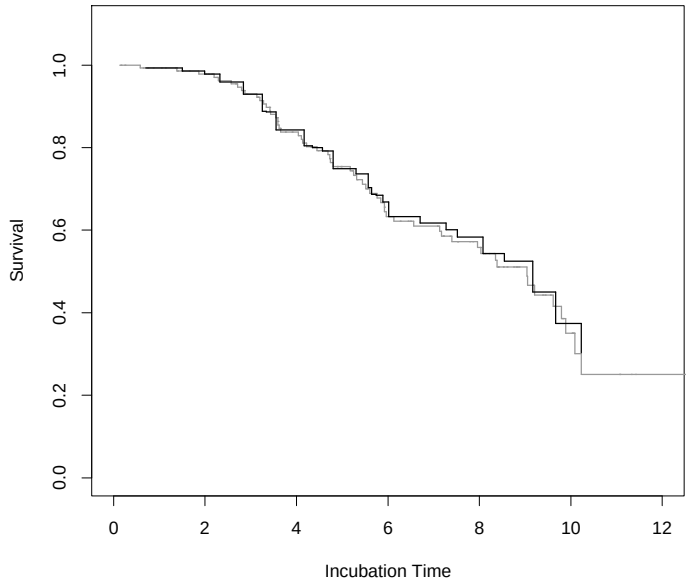


Figure 2.7: Estimates of the incubation time distribution of the seroconverters: interval transformation (dark curve) and Kaplan-Meier (light curve).

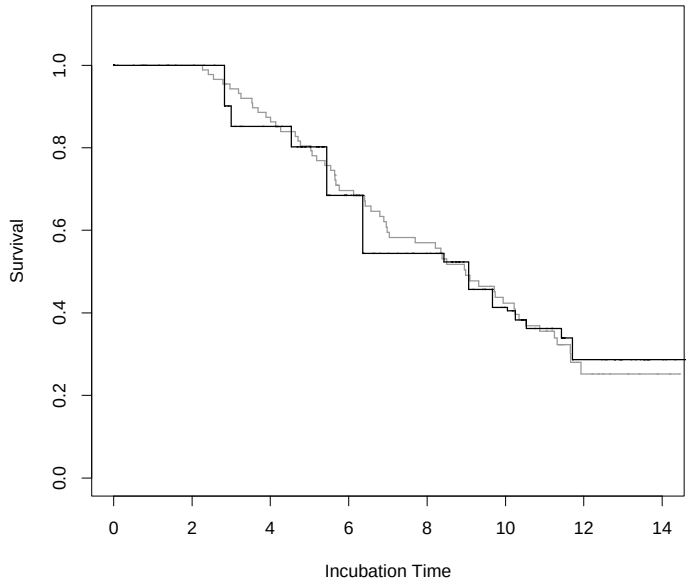


Figure 2.8: Incubation time estimates of the HBvac-study: interval transformation (dark curve) and Kaplan-Meier (light curve) based on midpoints

Chapter 3

Double Censoring

The methods of the following chapters differ from the ones of the preceding chapter. We now take account of the interval censored nature of the seroconversion dates and right censored moment of AIDS diagnosis simultaneously, a so-called bivariate approach.

3.1 The Nonparametric Model

The available data are $\{u_i, v_i, z_i, \delta_i\}$ for $i = 1, \ldots, N$, where $(u_i, v_i]$ is the seroconversion interval, z_i, the time of right censoring or diagnosis and δ_i an indicator indicating whether the time of diagnosis is right-censored ($\delta_i = 0$) or observed directly ($\delta_i = 1$). The incubation interval is defined as $(z_i - v_i, z_i - u_i)$. Let Y be the date of seroconversion with distribution function G and density g, Z date of AIDS diagnosis and $T = Z - Y$ the incubation time with distribution function F. We must make three assumptions that simplify the construction of the likelihood (see [5]).

1. A cohort of uninfected individuals is assembled at a fixed calendar time, say $s = 0$.
2. The date of seroconversion is independent of the incubation time.
3. The censoring times u_i, v_i and possibly z_i are generated by a point process that is independent of both seroconversion time and incubation time, *i.e.* if we are only interested in the estimation of the seroconversion and incubation time distribution we can leave out the terms in the likelihood which contain the probability of the observation times.

The three assumptions do not always hold in practice. For example assumption one is violated if a cohort contains seroprevalent cases. These seroprevalent cases will induce a truncation effect and the likelihood must be corrected for this effect. However, if these seroprevalent cases enter in the beginning of the AIDS epidemic, the truncation effect can be ignored without inducing a large bias.

We can express the log-likelihood for the data as follows:

$$\log L = \sum_{i=1}^{N} \delta_i \log \left[\int_{z_i - v_i}^{z_i - u_i} g(z_i - s) dF(s) \right] + (1 - \delta_i) \log \left[\int_{z_i - v_i}^{z_i - u_i} g(z_i - s)[1 - F(s)] ds \right] \quad (3.1)$$

The terms in the likelihood corresponding to $\delta_i = 1$ arise as follows. Of all the incubation times s, only those who occurred between $z_i - v_i$ and $z_i - u_i$ are admissible for individual i. Once we have assumed that the incubation time for individual i is a certain s we know that the seroconversion time is $z_i - s$. Integrating over all possible s we get the term corresponding with $\delta_i = 1$ in the likelihood. For an individual i with a right-censored AIDS diagnosis, we know that every seroconversion time between u_i and v_i is possible for this person. If we assume that the seroconversion time is a certain s, then the incubation time is at least $z_i - s$ Integrating over all possible seroconversion times s we get the term corresponding with $\delta_i = 0$ in the likelihood.

Log-Likelihood (3.1) can be seen as a generalization of the log-likelihood corresponding to the Kaplan-Meier and interval transformation. If we take for g in each likelihood term an indicator function which equals 1 at the midpoint of the seroconversion intervals and zero elsewhere, then the integrals in the likelihood function reduce to probabilities of a single incubation time if $\delta_i = 1$ or a single censoring time ($\delta_i = 0$). This is equivalent to the Kaplan-Meier log-likelihood. The interval transformation log-likelihood can be obtained by taking g constant for the terms where $\delta_i = 1$, the seroconversion density g then disappears from the integral and the term reduces to a term which is equivalent to interval transformation. For terms where $\delta_i = 0$, we take for the seroconversion density g an indicator function, reducing the integral to a term which is the survival of a single incubation time.

With a completely parametric analysis one assumes that the seroconversion time distribution belongs to a certain class of distributions, $\left\{ G_{\theta_1} : \theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1} \right\}$, and the incubation time distribution belongs to a certain class of distributions $\left\{ F_{\theta_2} : \theta_2 \in \Theta_2 \subset \mathbb{R}^{d_2} \right\}$. Then maximization of (3.1) with respect to θ_1 and θ_2 results in the estimates of θ_1 and θ_2, see chapter 5. With a nonparametric analysis no assumptions for F and G are made, we try to estimate the 'complete' distributions F and G by maximization of (3.1) with respect to F and G. However, we can not allow just 'everything' for G and F. The following example shows that in that case the log-likelihood can be made infinite.

If we take for the seroconversion time density $g(x) = c/\sqrt{-x + v_i}$, and for the incubation time density $f(x) = c'/\sqrt{x - z_i + v_i}$, then (3.1) becomes infinite, so to make this maximization sensible we must restrict G and F to a certain class of functions. For example, the class of functions with bounded derivative would exclude the previous example. However, the usual approach is to take the distributions F and G discrete and let them live on a finite grid. If we let F live on the grid $0 < t_1 < \ldots < t_s$, then G lives on the grid $0 < y_1 < \ldots < y_r$, which will be induced by the grid of F in such a way that y_j is one of the points of $z_i - t_k$ as long as $z_i - t_k$ is in the seroconversion interval $(u_i, v_i]$.

3.1.1 The Grid

The number of gridpoints (t_k) in the incubation time scale determines the 'difficulty' of the estimation problem. If this number is large then this will induce an even larger number of grid points ($z_i - t_k$) in the seroconversion time. So a fine grid can lead to time consuming calculations. However, one should not take a too coarse grid to save time. A too coarse grid may not reveal the subtle patterns in the shapes of the distributions. There are several ways to define a grid:

- An evenly spaced grid. Divide the time scale in periods of the same length, three months, say, and let the incubation time distribution live on that grid and the seroconversion time distribution on the grid induced by the incubation time grid. Or let the seroconversion time distribution also live on a evenly spaced grid, this is what De Gruttola and Lagakos [6] used in their illustration. In fact they need a grouping in both time scales of the data to avoid the problem of inducing a new gridpoint.
- A more data-adaptive method is to use a right-endpoint grid or left-endpoint grid. For each person, one determines the right endpoint or left endpoint of his incubation interval as a gridpoint. This means that in periods where there are many incubation intervals, there are more gridpoints. The number of gridpoints in the incubation time scale equals the number of individuals. For the seroconversion time scale the number of gridpoints is much larger, seven times the number of data, say.

Example of induced seroconversion grid

If we have the following data, notated as $(u_i; v_i; z_i)$: (0.5; 4; 8), (2; 5; 10.5) and (1.5; 3; 10) then the left endpoints of the incubation intervals form the grid $t_1 = 8-4 = 4$, $t_2 = 10.5-5 = 5.5$ and $t_3 = 10 - 3 = 7$. Each grid point in the incubation time scale induces one or more points in the seroconversion scale. Incubation time $t_1 = 4$ will lead to a seroconversion time of $z_1 - t_1 = 8 - 4 = 4$, incubation time 5.5 will lead to 8-5.5=2.5 and 10.5-5.5=5, and incubation time 7 will lead to 8-7=1, 10.5-7=3.5 and 10-7=3. So the seroconversion grid is $y_1 = 1, y_2 = 2.5, y_3 = 3, y_4 = 3.5, y_5 = 4$ and $y_6 = 5$.

3.1.2 The Likelihood

The likelihood is a discrete version of (3.1). If we assume that the seroconversion intervals are open to the left and closed to the right we get the following formula:

$$\sum_{i=1}^{N} \delta_i \log \sum_{z_i - v_i \le t_k < z_i - u_i} [F(t_k) - F(t_{k-1})](G(z_i - t_k) - G(z_i - t_k)^-) \; + \\ + \; (1-\delta_i) \log \sum_{z_i - v_i \le t_k < z_i - u_i} [1 - F(t_{k-1})][G(z_i - t_k) - G(z_i - t_{k+1})] \; , \qquad (3.2)$$

where $G(z_i - t_k)^-$ means the value of G at the largest gridpoint smaller than $z_i - t_k$ and $t_0 = 0$. If we take the seroconversion intervals open to right and closed to the left we get a somewhat different formula.

Define the following variables:

$$\begin{aligned} \beta_k &= F(t_k) \\ \alpha_{ik} &= G(z_i - t_k) \\ K_i &= \{k \mid z_i - v_i \leq t_k < z_i - u_i\}, \quad i = 1, \ldots, N \end{aligned}$$

Maximization of (3.2) with respect to the distributions F and G is equivalent to the following constrained optimization problem:

$$\begin{aligned} \min_{\alpha,\beta} \; -\log L \;=\; & -\sum_{i=1}^{N} \delta_i \log \sum_{k \in K_i} (\beta_k - \beta_{k-1})(\alpha_{ik} - \alpha_{ik}^{-}) \; + \\ & + (1 - \delta_i) \log \sum_{k \in K_i} (1 - \beta_{k-1})(\alpha_{i,k} - \alpha_{i,k+1}) \end{aligned} \tag{3.3}$$

under the restrictions

$$\begin{aligned} & (\alpha_{ik}) \text{ a distribution function} \\ & 0 \leq \beta_1 \leq \ldots \leq \beta_s \leq 1. \end{aligned}$$

The log-likelihood (3.2) has already been discussed by De Gruttola and Lagakos [6]. They use the observed Fisher information matrix to estimate the variances of $\hat{F}(t_k)$ and $\hat{G}(y_j)$. This only makes sense if the grid (number of parameters) is fixed as the number of observations tends to infinity, since then one can expect that the usual maximum likelihood theory holds. Groeneboom and Wellner [1] showed that in interval censoring the NPMLE of $F(t)$ for fixed t converges to a distribution which is not normal and that the rate of convergence differs from the usual $\sqrt{n}$. Since we can consider double censoring as a generalization of interval censoring, we can also expect non normal limiting distributions for $\hat{F}(t)$ and $\hat{G}(t)$. Precise asymptotic properties have not been derived yet.

Example

We have observed two individuals, one with seroconversion interval $[1, 3]$ and AIDS diagnosis at 7, the other with seroconversion interval $[3, 4]$ and AIDS diagnosis 9. We take as grid for the incubation time scale $0, 4, 5$ and for the seroconversion time scale $0, 2, 3, 4$, the 2 is taken in the seroconversion grid since 7-5=2 lies in the seroconversion interval of person 1. Then the corresponding constrained maximization with respect to F and G is given by:

$$\begin{aligned} \max \quad & \log[\, (F(4) - F(0))(G(3) - G(2)) + (F(5) - F(4))(G(2) - G(0)) \,] \\ & + \; \log[(F(5) - F(4))(G(4) - G(3))] \end{aligned}$$

restricted to

$$\begin{aligned} & 0 \leq G(0) \leq G(2) \leq G(3) \leq G(4) \leq 1 \\ & 0 \leq F(0) \leq F(4) \leq F(5) \leq 1 \end{aligned}$$

3.2 Piecewise uniform Distributions

The nonparametric approach described in this chapter is closely related to a weakly structured parametric model (see [6]) for the incubation and seroconversion time distribution. Until now we assumed a discrete time scale, where the probability masses were placed on the gridpoints. It is also possible to have a slightly different view on this; choose a grid $t_0 < t_1 < \ldots < t_s$ for the incubation time scale and a grid $y_0 < y_1 < \ldots < y_r$ for the seroconversion time scale, and then assume that the incubation time and seroconversion time are piecewise uniformly distributed, i.e. the heights of the distribution functions $F(t)$ and $G(t)$ in a gridpoint are given by:

$$\begin{aligned} F(t_k) &= \beta_k \quad k = 1, \ldots, s \\ G(t_k) &= \alpha_j \quad j = 1, \ldots, r \end{aligned}$$

and between two gridpoint the distribution function is linearly interpolated. The α_j's and β_k's must satisfy the order restrictions $0 \leq \alpha_1 \leq \ldots \leq \alpha_r \leq 1$ and $0 \leq \beta_1 \leq \ldots \leq \beta_s \leq 1$, making G and F (defective) distribution functions.

Define K_i as the set of indices which indicates which t_k's are admissible for the i-th incubation interval,

$$K_i = \{k | z_i - v_i \leq t_k \leq z_i - u_i\} \ .$$

Define J_{ik} as the set of indices which indicates which y_j's correspond with t_k for individual i.

$$J_{ik} = \{j | z_i - t_k \leq y_j \leq z_i - t_{k-1}\}$$

Then the (pseudo) log-likelihood is given by:

$$\begin{aligned} \sum_{i=1}^{N} \delta_i \log \left[\sum_{k \in K_i} \left(\theta_{ik}^{(1)} (\beta_k - \beta_{k-1}) \sum_{j \in J_{ik}} \theta_{ij}^{(2)} (\alpha_j - \alpha_{j-1}) \right) \right] \ + \\ + (1 - \delta_i) \log \left[\sum_{k \in K_i} \left((1 - \beta_{k-1}) \sum_{j \in J_{ik}} \theta_{ij}^{(2)} (\alpha_j - \alpha_{j-1}) \right) \right] \end{aligned} \tag{3.4}$$

where $\theta_{ik}^{(1)} \in (0, 1)$ is a fraction which indicates how much of the interval $(t_{k-1}, t_k]$ is covered by the incubation time interval $(z_i - v_i, z_i - u_i]$, and $\theta_{ij}^{(2)} \in (0, 1)$ is a fraction which indicates how much of a part the interval $(y_{j-1}, y_j]$ is covered by $(z_i - t_k, z_i - t_{k-1}]$. An advantage of this approach is that we don't have to use an induced seroconversion grid. In the previous section, if t_k was a gridpoint in the incubation time scale then $z_i - t_k$ was a gridpoint in the seroconversion time scale, since it could be a moment of seroconversion for individual i. With the piecewise uniform approach we only indicate between which two gridpoints seroconversion could take place.

Formula (3.4) needs some words of explanation. First, it is not the log-likelihood itself, the expression $(\alpha_k - \alpha_{k-1})$ in the terms corresponding with $\delta_i = 1$ should contain an extra constant indicating the distance between two successive gridpoints. However this constant has no influence on the maximization of the likelihood. Second, the terms corresponding with $\delta_i = 0$ are only approximations of the real likelihood term corresponding with the piecewise uniform approach. Another approach is to let F be piecewise constant (a step function), which will lead to a likelihood as in 3.4.

Example

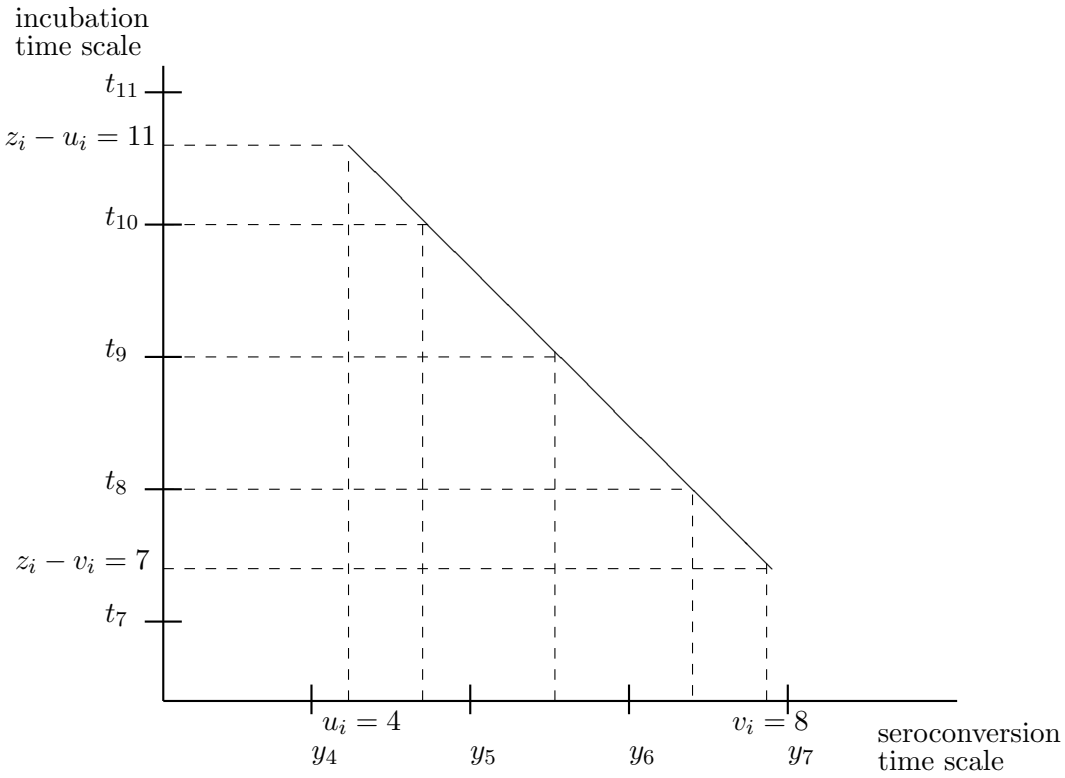


Figure 3.1: Incubation interval and seroconversion interval

Figure 3.1 shows an individual with seroconversion interval $(4, 8)$ and an incubation interval $(7, 11)$. The numbers along the axis are the indices of the corresponding gridpoints, we can see that only a part of $\beta_{11} - \beta_{10}$ is admissible, the corresponding part in the seroconversion time scale is only a part of $\alpha_5 - \alpha_4$. The log-likelihood term for this individual is given by:

$$\begin{aligned} \log \quad & [\ 0.6(\beta_{11} - \beta_{10})0.4(\alpha_5 - \alpha_4) \\ + \quad & (\beta_{10} - \beta_9)(0.3(\alpha_5 - \alpha_4) + 0.5(\alpha_6 - \alpha_5)) \\ + \quad & (\beta_9 - \beta_8)(0.6(\alpha_6 - \alpha_5) + 0.15(\alpha_7 - \alpha_6)) \\ + \quad & 0.7(\beta_8 - \beta_7)(0.35(\alpha_7 - \alpha_6))\ ] \end{aligned}$$

3.3 Optimization of the Likelihood

We first note that the optimization problems arising from double censoring are well defined, *i.e.* there is an optimum for (3.4) since the likelihood function is continuous on a bounded parameter space. The optimization of the likelihood can be done in several ways. One simple algorithm is the Expectation Maximization (EM) algorithm; another more advanced algorithm is the Sequential Quadratic Programming (SQP) algorithm. Both are described in the following sections.

3.3.1 The EM Algorithm

To solve the maximization problems (3.3) and (3.4) we can use the EM algorithm. The EM algorithm is widely used for missing data problems. For doubly-censored data we use a generalization of the self-consistency algorithm proposed by Turnbull [8], described by De Gruttola and Lagakos [6]. To use the EM algorithm it is necessary to reparameterize problems (3.3) and (3.4). Rearrange α_{ik} into a vector α_j so that $\alpha_j \leq \alpha_{j+1}$, and define the following variables

$$\begin{aligned} f_k &= \beta_k - \beta_{k-1} \\ w_j &= \alpha_j - \alpha_{j-1} \\ \alpha^i_{jk} &= \begin{cases} 1 & \text{if } \delta_i = 1 \text{ and } u_i \leq y_j \leq v_i \text{ and } z_i - v_i \leq t_k \leq z_i - u_i \\ 1 & \text{if } \delta_i = 0 \text{ and } u_i \leq y_j \leq v_i \text{ and } z_i - v_i \leq t_k \leq z_i - u_i \\ 0 & \text{else} \end{cases} \end{aligned}$$

the α^i_{jk}'s indicate whether or not a specific combination of w_j and f_k are admissible for observation i. So instead of using α_j and β_k, which represents the height of the distributions we use the probabilities w_j and f_k. For example, problem (3.3) can now be reformulated as follows,

$$\max_{w,f} \quad \log L = \sum_{i=1}^{N} \log \sum_j \sum_k \alpha^i_{jk} f_k w_j$$

under the restrictions

$$\sum_j w_j = 1$$

$$\sum_k f_k = 1$$

The density f always sums up to one, if we take a point larger than the largest observed incubation time and place the remaining probability mass on that point. The same can be done for the seroconversion time distribution.

To remove the equality restrictions we can use the Lagrange multiplier method, to form the Lagrange function Φ:

$$\Phi = \sum_{i=1}^{N} \log \sum_j \sum_k \alpha^i_{jk} f_k w_j + \lambda_1 (\sum_j w_j - 1) + \lambda_2 (\sum_k f_k - 1)$$

We now have an unconstrained maximization problem which we can solve by setting the partial derivatives to zero:

$$\begin{aligned}
\frac{\partial \Phi}{\partial w_j} &= \sum_{i=1}^{N} \frac{\sum_k \alpha_{jk}^i f_k}{\sum_j \sum_k \alpha_{jk}^i w_j f_k} + \lambda_1 = 0 \\
\frac{\partial \Phi}{\partial f_k} &= \sum_{i=1}^{N} \frac{\sum_j \alpha_{jk}^i w_j}{\sum_j \sum_k \alpha_{jk}^i w_j f_k} + \lambda_2 = 0 \\
\frac{\partial \Phi}{\partial \lambda_1} &= \sum_j w_j - 1 = 0 \\
\frac{\partial \Phi}{\partial \lambda_2} &= \sum_k f_k - 1 = 0
\end{aligned}$$

If we multiply the equations $\frac{\partial \Phi}{\partial w_j} = 0$ by w_j and sum them up

$$\sum_j w_j \frac{\partial \Phi}{\partial w_j} = N + \lambda_1 \sum_j w_j = 0,$$

we get the optimal value for λ_1, which equals $-N$. The same trick is used to get the optimal value for λ_2, which also equals $-N$. Using these values for λ_1 and λ_2 we get the following equations for w_j and f_k:

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \frac{w_j \sum_k \alpha_{jk}^i f_k}{\sum_j \sum_k \alpha_{jk}^i w_j f_k} &= w_j, \quad j = 1, \ldots, r \\
\frac{1}{N} \sum_{i=1}^{N} \frac{f_k \sum_j \alpha_{jk}^i w_j}{\sum_j \sum_k \alpha_{jk}^i w_j f_k} &= f_k, \quad k = 1, \ldots, s
\end{aligned}$$

These equations are called the self-consistency equations (see [1]).

From these equations we can derive the following algorithm to find the unknown w_j and f_k:

1. $l = 0$
2. choose starting values for $w_j^{(l)}$, $f_k^{(l)}$, one can take $w_j^{(l)} = 1/r$ and $f_k^{(l)} = 1/s$, where r is the number of gridpoints in the seroconversion timescale and s the number of gridpoints in the incubation time scale.
3. compute

$$\mu_{jk}^i = \frac{\alpha_{jk}^i w_j^{(l)} f_k^{(l)}}{\sum_j \sum_k \alpha_{jk}^i w_j^{(l)} f_k^{(l)}}$$

4. refine the values for $w_j^{(l)}$ and $f_k^{(l)}$ by

$$w_j^{(l+1)} = \frac{\sum_{i,k} \mu_{jk}^i}{n} \quad , \quad f_k^{(l+1)} = \frac{\sum_{i,j} \mu_{jk}^i}{n}$$

5. $l = l + 1$. Go to step 3 until convergence

The EM algorithm is a relatively simple algorithm, it only uses the structure of the derivative equations of the Lagrange function (self-consistency equations). It is clear from the algorithm, that once a value of w_j or f_k is put to zero it remains zero, until convergence. If the starting distributions put zero masses at some points t_k or y_j, the EM algorithm would converge to a solution of the self-consistency equations, but this solution would not necessarily maximize the likelihood. Furthermore, it is the general empirical finding that the number of iteration steps will increase with the sample size. In [6] it is shown that the solution found by the EM algorithm is either a saddle point or a local optimum of the likelihood.

3.3.2 Sequential Quadratic Programming

Optimization problems of the form (3.3) and (3.4) are just special cases of a general nonlinear programming problem (P), where (P) is defined as

$$\begin{aligned} (P) \qquad & \min_{x \in \mathbb{R}^n} f(x) \\ \text{subject to} \quad & \\ & g_j(x) \leq 0, \ j = 1, \ldots, m_1 \\ & h_j(x) = 0, \ j = 1, \ldots, m_2 \end{aligned}$$

In our case, the likelihood function plays the role of f, α and β together compose x, and we have no equality constraints, so $m_2 = 0$, and only linear inequality constraints of the form $\alpha_j - \alpha_{j+1} \leq 0$ and $\beta_j - \beta_{j+1} \leq 0$. For problems like (P) several algorithms have been developed over the past decades. One of those is the Sequential Quadratic Programming (SQP) algorithm, which we describe briefly (see [18] for a detailed description).

The necessary conditions for optimality in (P) without equality constraints are given by the so-called Kuhn-Tucker (KT) relations. Define the Lagrange function L as

$$L(x, u) = f(x) + \sum_{j=1}^{m_1} u_j g_j(x)$$

then the (KT) relations are given by:

$$\begin{aligned} (KT) \quad \nabla_x L(x, u) &= \nabla f(x) + \sum_{j=1}^{m_1} u_j \nabla g_j(x) = 0 \\ & u_j g_j(x) = 0, \ j = 1, \ldots, m_1 \\ & u_j \geq 0 \end{aligned}$$

The computations start at an arbitrary point $(x^{(0)}, u^{(0)})$, $u^{(0)} \geq 0$. In the k-th iteration we solve $(x^{(k+1)}, u^{(k+1)})$ from the first order approximations with respect to x and u of (KT):

$$\nabla f(x^{(k)}) + \nabla^2_{xx} L(x^{(k)}, u^{(k)})(x - x^{(k)}) + \sum_{j=1}^{m_1} u_j \nabla g_j(x^{(k)}) \;=\; 0 \tag{3.5}$$

$$u_j g_j(x^{(k)}) + u_j^{(k)} \nabla g_j(x^{(k)})(x - x^{(k)}) \;=\; 0,\; j = 1, \ldots, m_1 \tag{3.6}$$

$$u_j \;\geq\; 0,\; j = 1, \ldots, m_1 \tag{3.7}$$

This linear system in (x, u) almost coincides with the Kuhn-Tucker relations of the following quadratic programming problem:

$$\begin{aligned} (QP) \quad \min_x \quad & \nabla f(x^{(k)})'(x - x^{(k)}) + 1/2(x - x^{(k)})' \nabla^2_{xx} L(x^{(k)}, u^{(k)})(x - x^{(k)}) \\ \text{subject to} \quad & \\ & g_j(x^{(k)}) + \nabla g_j(x^{(k)})'(x - x^{(k)}) \geq 0 \end{aligned}$$

Instead of (3.6) we find

$$u_j[g_i(x^k) + \nabla g_i(x^k)'(x - x^k)] = 0,\; j = 1, \ldots, m_1$$

It can be proven, however, that any limit point (x^*, u^*) of the sequence $\left\{(x^{(k)}, u^{(k)})\right\}$, whether it is generated by (3.5) - (3.7) or (QP), satisfies the Kuhn-Tucker relations (KT).

Two refinements can be made to improve the method. First, a line search can be included, instead of solving (QP) directly we generate a search direction $s^{(k)}$ from the problem

$$\begin{aligned} \min_s \quad & \nabla f(x^{(k)})' s + 1/2 s' \nabla^2_{xx} L(x^{(k)}, u^{(k)}) s \\ \text{subject to} \quad & \\ & g_j(x^{(k)}) + \nabla g_j(x^{(k)})' s \geq 0 \end{aligned}$$

after which we explore $f(x^{(k)} + \lambda s^{(k)})$, subject to the additional requirement that we have to satisfy the inequality restrictions on x, to find a new iterate $x^{(k+1)}$. Second, we do not need the exact Hessian matrix $\nabla^2_{xx} L(x^{(k)}, u^{(k)})$ of the Lagrange function at $(x^{(k)}, u^{(k)})$. The idea is to replace it with variable-metric approximations (see [7]). Under mild conditions the sequence $\left\{(x^{(k)}, u^{(k)})\right\}$ converges to a Kuhn-Tucker point (x^*, u^*).

The SQP algorithm can be seen as a general case of the ICM algorithm, as described in section 2.2.1. First, instead of only approximating the objective function with a quadratic form with diagonal elements, the SQP algorithm also uses off-diagonal elements in the approximation. Second, the SQP algorithm can take account of more general restrictions than the ICM algorithm. For example, with double censoring we have two sets of order restrictions, which ICM can't handle. However, these two generalizations make the SQP algorithm

too slow for interval tranformation. The ICM algorithm is tailor-made for situations with interval censored data, since the diagonal approximation and the simple order restrictions allow for a very fast cumulative sum diagram solution, whereas the SQP algorithm solves in each step a complete quadratic programming problem.

3.3.3 Multiple optimal Solutions

It is possible to create doubly censored data problems with multiple (locally) optimal solutions, in principle any algorithm that is used in multiple optimal problems can converge to a local optimum only. Take the following trivial example. Suppose there is just one observation $(5, 9, 15, 1)$, so the seroconversion interval is $(5, 9]$ and the incubation time interval is $(6, 10]$. If the grids in both time scales are $1, 2, 3, \ldots$, then the likelihood is simply

$$\log(w_6 f_{10} + w_7 f_9 + w_8 f_8 + w_9 f_7).$$

Maximizing this likelihood under the restrictions that the w's and f's must sum up to one we get several optimal solutions: $w_6 = f_{10} = 1$, $w_7 = f_9 = 1$, $w_8 = f_8 = 1$ and $w_9 = f_7 = 1$ are all optimal solutions. It is clear that this problem of multiple solutions is merely the result of a too fine grid or too few observations. For example if we add an extra observation $(7, 11, 17, 1)$ to our example then we only have two optimal solutions: $w_8 = f_{10} = 1$ and $w_9 = f_9 = 1$. So to avoid non-unique maximum likelihood estimates the grids should not be chosen too fine. From the example above it is clear that a seroconversion or incubation interval should not contain too many gridpoints, unless there are other intervals who have a part of the gridpoints in common.

3.3.4 The Grid

A too fine grid will not lead to a sensible maximum likelihood estimator. To see this we take as grid for the incubation time scale $\epsilon_1, 2\epsilon_1, \ldots$ and for the seroconversion time scale $\epsilon_2, 2\epsilon_2, \ldots$. The incubation density is defined as $f(s) = f_k$ if $s \in (k\epsilon_1, (k+1)\epsilon_1]$ and the seroconversion density is defined as $g(s) = g_j$ if $s \in (j\epsilon_2, (j+1)\epsilon_2]$, where $\sum f_k \epsilon_1 = 1$ and $\sum g_j \epsilon_2 = 1$. Likelihood 3.1 can then be rewritten as follows. For terms with $\delta_i = 1$ we get:

$$\begin{aligned}
\int_{z_i - v_i}^{z_i - u_i} g(z_i - s) f(s) ds &= \sum_{k \in K_i} \int_{k\epsilon_1}^{(k+1)\epsilon_1} g(z_i - s) f(s) ds \\
&= \sum_{k \in K_i} f_k \int_{k\epsilon_1}^{(k+1)\epsilon_1} g(z_i - s) ds \\
&= \sum_{k \in K_i} f_k \sum_{j \in J_{ik}} \int_{j\epsilon_2}^{(j+1)\epsilon_2} g(z_i - s) ds \\
&= \epsilon_2 \sum_{k \in K_i} \sum_{j \in J_{ik}} f_k g_j \ .
\end{aligned}$$

Where K_i and J_{ik} are defined on page 21.

For terms with $\delta_i = 0$ we get:

$$\int_{z_i-v_i}^{z_i-u_i} g(z_i - s)[1 - F(s)]ds \approx \sum_{k\in K_i} \sum_{j\in J_{ik}} f_k g_j \epsilon_1 \epsilon_2 \ .$$

The log-likelihood is then given by:

$$\log L = \sum_{i}^{N} \left[\delta_i \sum_{k\in K_i} \sum_{j\in J_{ik}} f_k g_j \epsilon_2 + (1-\delta_i) \sum_{k\in K_i} \sum_{j\in J_{ik}} f_k g_j \epsilon_1 \epsilon_2 \right].$$

By taking $f_k = c/\epsilon_1$ and $g_j = c/\epsilon_2$ the likelihood will tend to infinity as $\epsilon_1 \downarrow 0$ and $\epsilon_2 \downarrow 0$.

3.4 Extensions of double Censoring

3.4.1 Truncation Effects

The effect of truncation occurs when sample data are drawn from a non representative subset of a larger population of interest. In the previous sections we assumed that a cohort of uninfected persons was assembled at calendar time $s = 0$. However this is not always the case; for example the HOM-study contains seroprevalent cases. If we include these cases into the study then we may introduce a bias, since these cases could enter the study because they were free from AIDS defining illnesses at study entry. So the effect of this sampling scheme is to selectively exclude cases with very short incubation periods, since those who developed AIDS before 1984 were not included into the cohort study. We can correct for this bias by using a truncated density in the analysis. For example, suppose we have an uncensored sample $X_1, \ldots, X_n$ from a ditribution with dustribution function F and we know that truncation occurs at a_i, for $i = 1, \ldots, n$. Then we must use the truncated densities

$$f(X_i | X_i > a_i) = \frac{f(X_i)}{1 - F(a_i)}$$

in the likelihood.

For a censored sample, the estimation of distribution functions via Kaplan-Meier, interval censoring or double censoring can also be adapted to account for truncation effects (see [17, 16, 8, 28]). For example, suppose a prevalent case i enters the study at v_i. Then we use the truncated density

$$g_i(s) = \frac{g(s)}{\int_0^{v_i} g(v_i - s)[1 - F(s)]ds + (1 - G(v_i))} \tag{3.8}$$

instead of g in likelihood (3.1). The denominator of (3.8) is the probability of being free of AIDS diagnosis, this probability is the sum of the probability of being infected and AIDS free and the probability of not being seroconverted before study entry v_i.

3.4.2 Interval censored AIDS Diagnoses

In the Amsterdam cohort studies the moments of AIDS diagnosis are either known or right censored. However, in other studies it may occur that (some) moments of AIDS diagnosis are interval censored, this occurs when an individual leaves the study at a moment z_1 before he has an AIDS diagnosis and dies of AIDS at moment z_2 and no further information is available. The AIDS diagnosis is then interval censored by the interval (z_1, z_2), see figure 3.2. The moment z_2 can be obtained via the death registries, where moment and cause of death are stored for the entire population of a certain area.

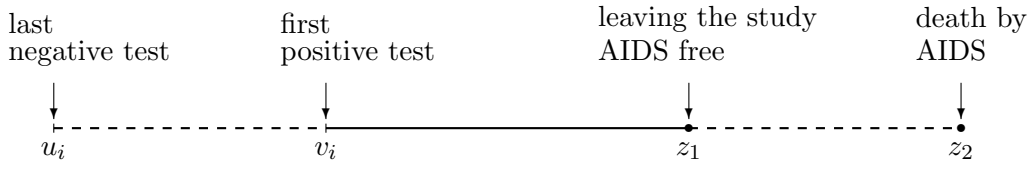


Figure 3.2: Double censoring with interval censored AIDS diagnosis

Individuals with an interval censored AIDS diagnosis contribute the following term to the likelihood

$$\log \int_{u_i}^{v_i} g(s)[F(z_{2,i} - s) - F(z_{1,i} - s)]\ ds\ .$$

So the likelihood can consist of three different kinds of terms, terms for known AIDS diagnosis, right censored AIDS diagnosis and interval censored AIDS diagnosis.

3.4.3 The Distinction of Subgroups

In some cohort studies subgroups can be identified. For example, there may be an indication that men follow a different seroconversion pattern than women. However, it is possible that there is no reason to assume a different incubation time distribution. In the modelling of double censoring we can stratify into these two groups. The likelihood then contains an extra seroconversion distribution which has to be estimated and is given by:

$$\sum_{i=1}^{N} \delta_{i,m} \log \left[\int_{z_i - v_i}^{z_i - u_i} g_m(z_i - s) dF(s) \right]^{\delta_i} + \delta_{i,m} \log \left[\int_{z_i - v_i}^{z_i - u_i} g_m(z_i - s)[1 - F(s)] ds \right]^{1-\delta_i} +$$
$$(1 - \delta_{i,m}) \log \left[\int_{z_i - v_i}^{z_i - u_i} g_v(z_i - s) dF(s) \right]^{\delta_i} + (1 - \delta_{i,m}) \log \left[\int_{z_i - v_i}^{z_i - u_i} g_v(z_i - s)[1 - F(s)] ds \right]^{1-\delta_i}$$

where $\delta_{i,m} = 1$ if case i is an man and $\delta_{i,m} = 0$ if case i is a woman. The densities g_m and g_v are the corresponding seroconversion densities.

It can also occur that the group containing information on the seroconversion pattern is larger than the group containing information on both seroconversion pattern and incubation time. For example, in the IDU-study described in section 1.2 there are 113 cases which have been selected for the analysis of the incubation time. However, there are more cases but these cases only give us only some information on the seroconversion pattern. The likelihood can then be extended with terms which only contains information on the seroconversion pattern. These terms indirectly contribute to the estimation of the incubation time. We get the following log-likelihood function:

$$\sum_{i=1}^{N} \delta_{i,G} \log \left[\int_{z_i - v_i}^{z_i - u_i} g(z_i - s) dF(s) \right]^{\delta_i} + \delta_{i,G} \log \left[\int_{z_i - v_i}^{z_i - u_i} g(z_i - s)[1 - F(s)] ds \right]^{1-\delta_i} + (1 - \delta_{i,G}) \log \left[\int_{u_i}^{v_i} g(s) ds \right]^{\delta_i} + (1 - \delta_{i,G}) \log(1 - G(v_i))^{1-\delta_i} \quad (3.9)$$

where $\delta_{i,G} = 1$ if a case contains information on both seroconversion and incubation. The last two terms in (3.9) correspond to cases which only contain interval censored information and right censored information on the seroconversion time, respectively.

3.5 Results

Simulations

To check the correctness of the implementation of the algorithm we first simulate a data set and then try to estimate the seroconversion distribution and the incubation time distribution. The seroconversion times x_i are drawn from a Weibull(1.75, 5) and the incubation times t_i are drawn from a Weibull(2,6). We can now form the times of aids $z_i = x_i + t_i$. To construct the seroconversion interval for case i we form a random grid $u_1, u_2, u_3, \ldots$, where $u_i = u_1 + u_2 + \ldots + u_{i-1} + B$ and B is a realization of a uniform random variable $U(0, 2)$. The interval $(u_{j-1}, u_j]$ is taken as seroconversion interval if $x_i \in (u_{j-1}, u_j]$. Finally we randomly right-censor the time z_i, by taking $z_i = \min(C_i, z_i)$ where C_i is a realization of a uniform random variable $U(z_i - 3, z_i + 3)$. So z_i is right censored ($\delta_i = 0$) if $C_i < z_i$ and z_i is observed directly ($\delta_i = 1$) if $z_i > C_i$. The SQP algorithm results in the estimates of the seroconversion distribution and the incubation time distribution which are depicted in figure 3.3 and 3.4. With the simulated data it is also possible to look at the influences of different number of parameters (grid sizes) on the estimate of the incubation time. It appears that the influences are small. Figure 3.5 shows the plot of the estimates of the incubation time distribution of the simulated data. The number of gridpoints varies from 15 to 70 in the incubation time scale.

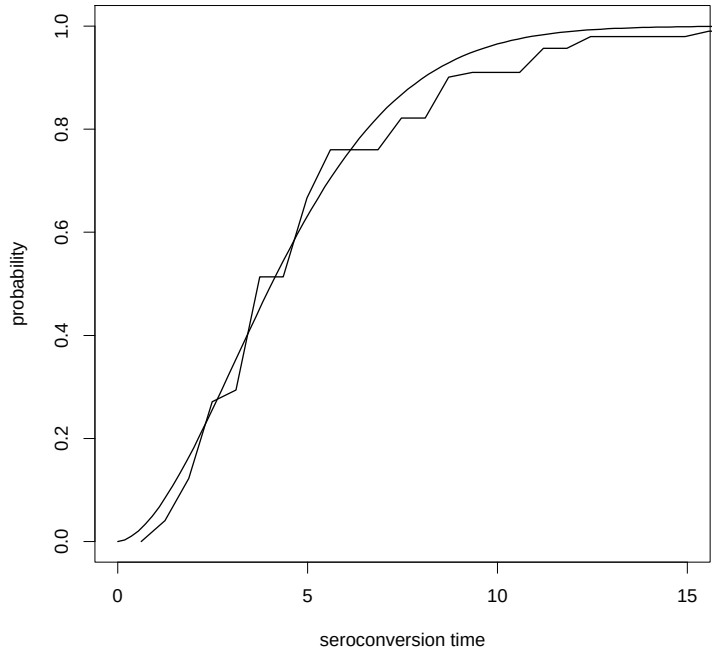


Figure 3.3: Step function is the estimated seroconversion distribution of the simulated dataset, smooth function is the 'real' distribution function

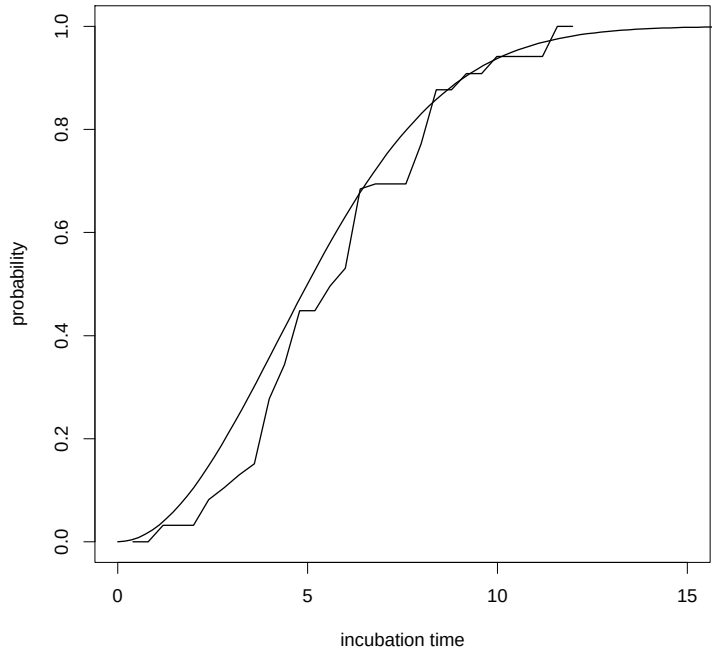


Figure 3.4: Step function is the estimated seroconversion distribution of the simulated dataset, smooth function is the 'real' distribution function

The HBvac study

Figure 3.6 shows the survival curve based on double censoring for the incubation time of the HBvac-study, together with a Kaplan-Meier estimate of the incubation time based on the ex-

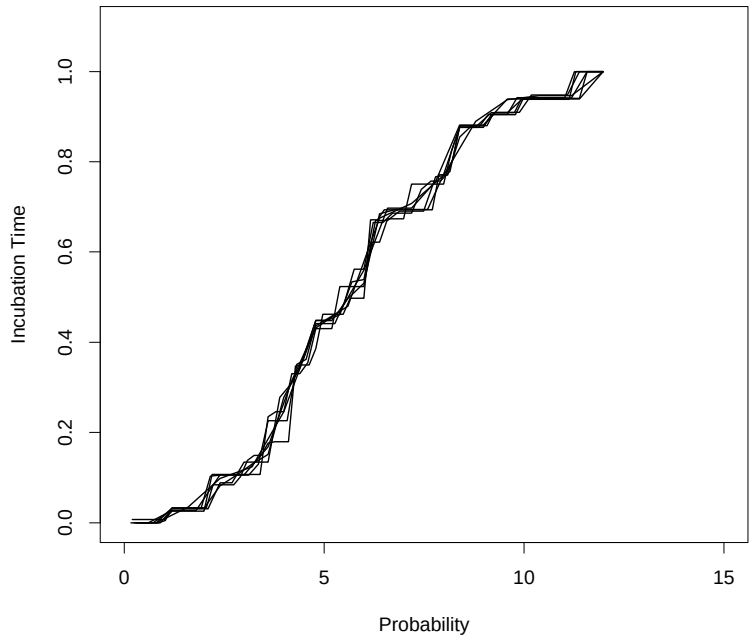


Figure 3.5: Estimates of the incubation time distribution of the simulated data set with different grid sizes.

pected seroconversion dates. We see that the double censoring estimate has less larger pieces where the survival is constant than the Kaplan-Meier. This reflects the uncertainty of the seroconversion period, whereas the Kaplan-Meier method assumes a known seroconversion time.

The HOM-study

The HOM-study among homosexual men is a much larger cohort than the HBvac and contains seroprevalent cases. Two groups of prevalent cases can be distinguished, the first group is the group entering at the beginning of the cohort study (1985) and the second group consists of persons entering after 1988. We leave out the second prevalent group, since truncation effects are too severe to be modelled by likelihood (3.1) and these cases may lead to a bias in the estimate of the incubation time. The estimated seroconversion distribution is plotted in figure 3.7 and the estimated incubation time survival curve is plotted in figure 3.8. The huge jump in figure 3.7 around 1983 reflects the large amount of uncertainty with respect to the date of seroconversion for the seroprevalent cases. The Kaplan-Meier based on midpoints is also plotted in figure 3.8, we see that the survival estimate based on the Kaplan-Meier is better. This is caused by the fact that the midpoints of the seroprevalent cases (ca. June 1982) is earlier compared to the estimated seroconversion pattern of the HOM-study. Table 3.1 numerically summarizes the results of the several analyses of the incubation time of the HBvac-study and the HOM-study.

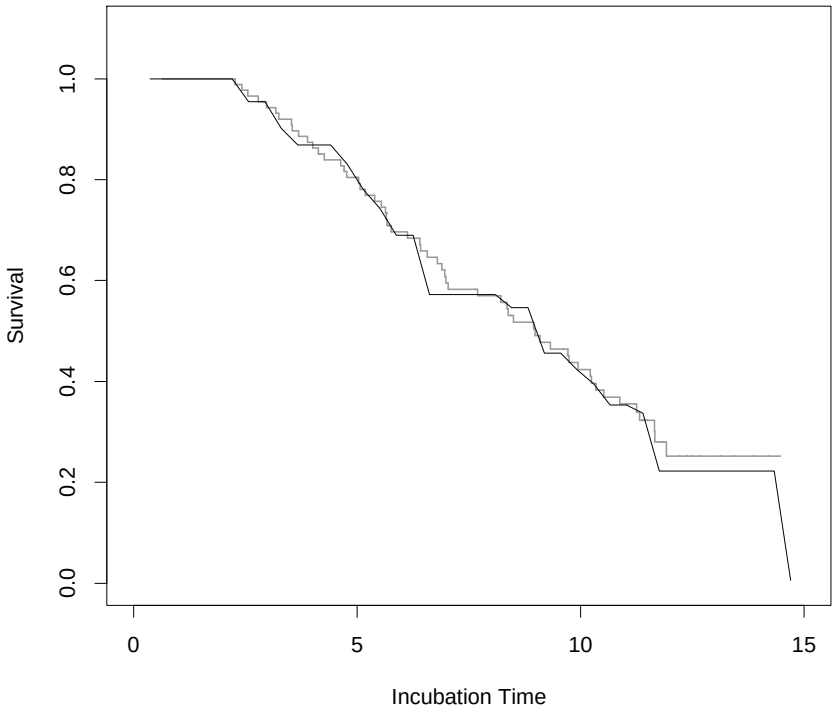


Figure 3.6: Survival curves of the HBvac-study. dark: double censoring light: Kaplan-Meier based on expected date of seroconversion

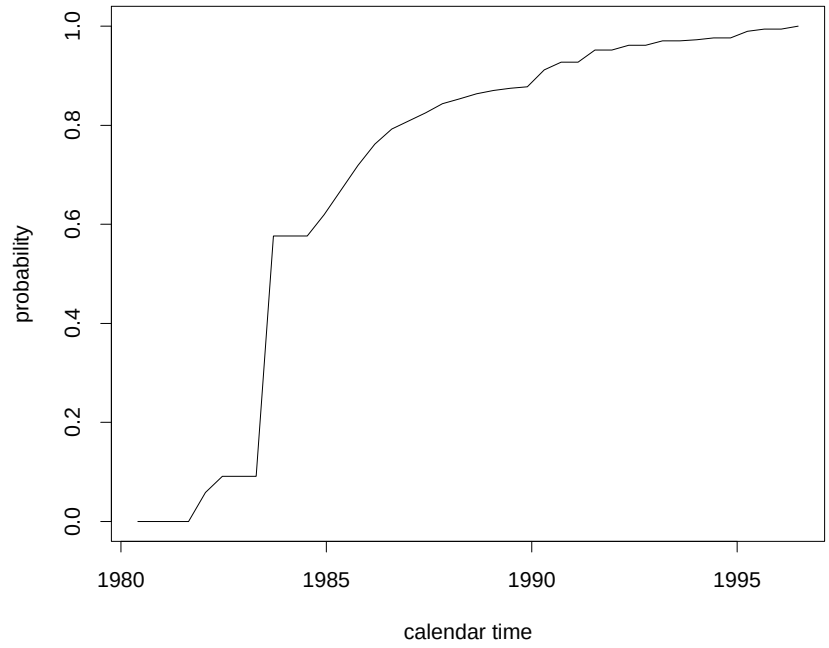


Figure 3.7: Estimated seroconversion distribution of the HOM-study calculated via the double censoring method

Percentiles in years for the incubation time			
HOM-study	25%	50%	75%
KM-mid	5.92	9.49	14.30
double	5.53	9.14	13.60
HBvac-study	25%	50%	75%
KM-expd	5.47	9.01	11.90
double	5.51	9.00	11.76

Table 3.1: Numerical summary of the incubation time of the HOM-study and the HBvac-study.

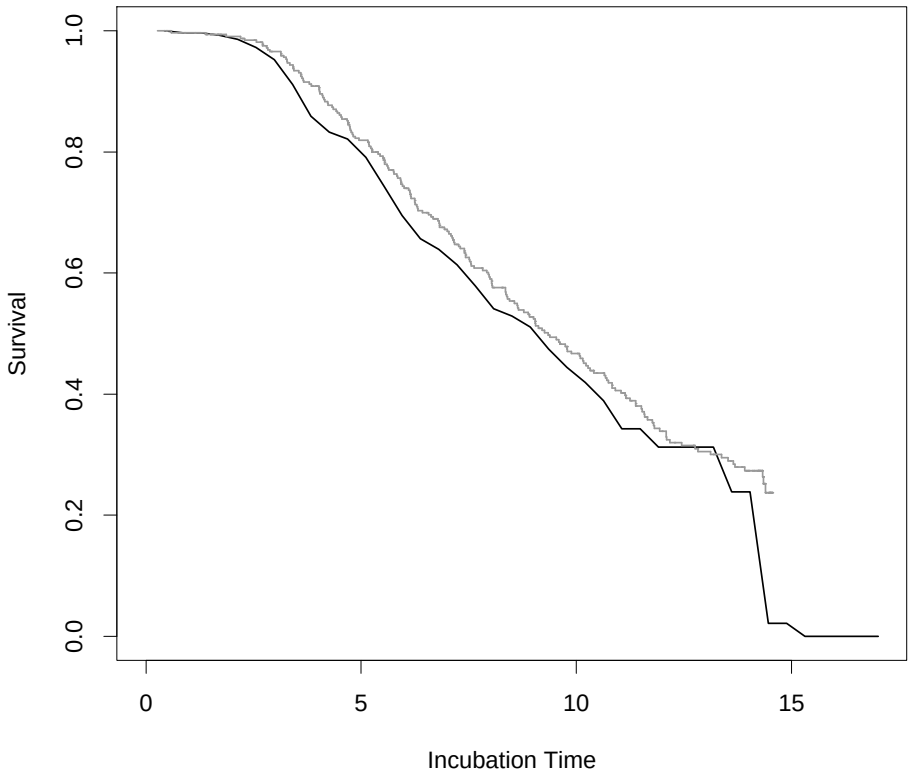


Figure 3.8: Estimated survival curves from the HOM-study: dark = double censoring, light = Kaplan-Meier based on midpoints

The IDU-study

As mentioned in section 3.4.3 the IDU-study contains more cases with information on the seroconversion pattern than cases with information on both seroconversion and incubation pattern. First, we analysed 113 cases with information on both seroconversion and incubation pattern. Figure 3.9 shows two approaches, curve 1 is based on Kaplan-Meier using midpoints and curve 2 is based on double censoring (likelihood (3.4)). Second, we added 899 cases to our first analysis, these cases only contain information on the seroconversion pattern, so we

have to use likelihood (3.9). Curve 3 is the resulting survival curve of this approach. It turns out that for the first 10 years the three approaches don't differ too much.

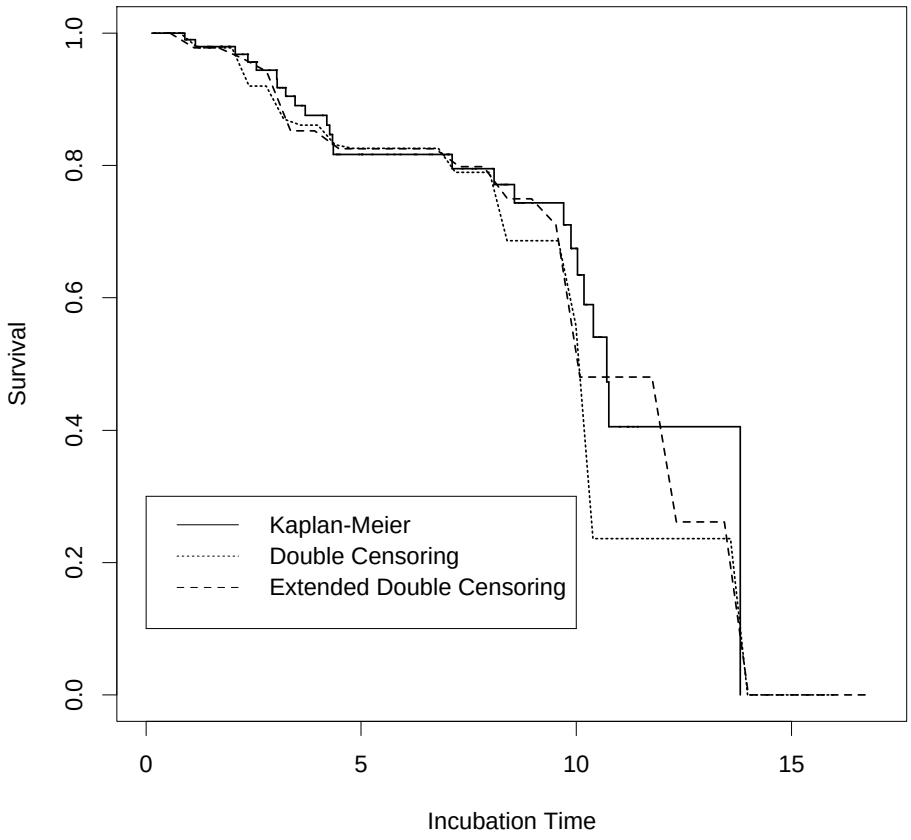


Figure 3.9: Estimated survival curves of the IDU-study

Chapter 4

The seroprevalent Cases

Until now we analysed the seroconversion distribution and the incubation time distribution nonparametrically, as described in the previous chapter. However this approach may lead to some problems due to the seroprevalent cases. To deal with these cases we slightly adapt our approach in the previous chapter. We look at semiparametric models and individual seroconversion distributions based on marker values.

4.1 Semiparametric Models

4.1.1 The Likelihood

As mentioned in chapter 1, the HOM-study contains seroprevalent cases. These cases hardly contain any information on the seroconversion time distribution before 1984. A nonparametric approach in this situation will lead to large jumps in the estimate of the distribution in the time period 1980-1985 (see figure 3.7).

One method to deal with this problem is to assume a parametric form G_θ for the seroconversion distribution and leave the incubation time distribution piecewise uniform as described in chapter 3. In fact, by choosing a parametric form we force a seroconversion structure in the period 1980-1985. We get the following log-likelihood,

$$\begin{aligned}\sum_{i=1}^{N} \delta_i \log \sum_{z_i-v_i \le t_k < z_i-u_i} [F(t_k) - F(t_{k-1})][G_\theta(z_i - t_{k-1}) - G_\theta(z_i - t_k)] \\ + \ (1-\delta_i) \log \sum_{z_i-v_i \le t_k < z_i-u_i} [1 - F(t_{k-1})][G_\theta(z_i - t_{k-1}) - G_\theta(z_i - t_k)] \ . \end{aligned} \tag{4.1}$$

Where the second term is comparable with the second term in formula (3.4) and is also an approximation of the real log-likelihood term.

To estimate θ and F we maximize (4.1). The corresponding optimization problem is given by

$$\begin{aligned}\min_{\theta,\beta} \log L &= -\sum_{i=1}^{N} \delta_i \log \sum_{k\in K_i} (\beta_k - \beta_{k-1})[G_\theta(z_i - t_{k-1}) - G_\theta(z_i - t_k)] + \\ &+ (1-\delta_i) \log \sum_{k\in K_i} (1-\beta_{k-1})[G_\theta(z_i - t_{k-1}) - G_\theta(z_i - t_k)] \qquad (4.2)\end{aligned}$$

subject to

$$\theta \in \mathbb{R}^k$$
$$0 \le \beta_1 \le \ldots \le \beta_s \le 1.$$

which is almost identical to (3.3), but differs in the kind of parameters we are dealing with. Here we have a parametric part for the seroconversion time distribution and a nonparametric part for the incubation time distribution, a so-called semiparametric form. This results in a different approach to solve (4.2).

4.1.2 Optimization of the Likelihood

To maximize the log-likelihood (4.1) we could use an EM algorithm for both θ and β. However, there are no order restrictions for θ, so given β we can use 'conventional' optimization algorithms, such as the Newton method or the conjugate gradient method (see [7]), to find an optimal value for θ. For a given θ the likelihood is concave in β, so we can use the ICM algorithm, which we described in chapter 2. So we can optimize likelihood (4.1) by alternating between the Newton method and the ICM method.

An alternative approach is to use the SQP algorithm, as described in chapter 3. With this approach we don't have to switch between the two parameters θ and β. Instead, we solve the problem in the total parameter space (θ, β), with the order restrictions on β.

4.1.3 Results from the HOM-study

We now apply the semiparametric method to the HOM-study. This cohort contains a lot of seroprevalent cases. We look at two approaches. First, we assume that the seroconversion distribution G_θ is Weibull, and estimate the incubation time distribution F nonparametrically. Second, we assume that the seroconversion distribution G_θ resembles epidemic curve, and estimate the incubation time distribution nonparametrically. An epidemic curve arises from the fact that the number of new infections exponentially increases in the first period of an epidemic and then slowly decreases in the last period. In the HIV epidemic, there was an exponential increase in the first four or five years, followed by a slow decrease.

For the seroconversion time scale we get the results as plotted in figure 4.1. It shows the seroconversion intervals, the fitted Weibull distribution and a sort of epidemic curve. We see the drawback of the Weibull approach, the seroconversion distribution behaves almost uniform in the first five years, indicating that seroconversions in the period 1980-1985 are equally

likely to occur, which is not probable according to previous epidemiologic data [27] . With the epidemic curve approach we force a higher probability of seroconversion in the period 1983-1985 than in 1980-1983. This is more probable, but it can never be checked without the help of additional data. For the epidemic curve we have looked at two possibilities. First the distribution curve follows an exponential form for the first four years and then a Weibull piece is used for the remaining part. So the seroconversion distribution $G(s)$ is modeled as follows:

$$G(s) = \begin{cases} pe^{cs} & \text{for } s \leq 4 \\ (1-p)(1-\exp[-(\lambda s)^{\alpha}]) + pe^{4c} & \text{for } s > 4 \end{cases} .$$

were p is the proportion of seroprevalent cases.

Second, logistic growth curves are used to simulate the effect of an exponential growth in the first part of the seroconversion time scale and then a slow decrease. We choose for a mixture of logistic curves, since a one parameter logistic curve was unable to describe the data properly. We choose the following expression for $G(s)$:

$$G(s) = p\frac{0.01}{0.01 + 0.99e^{-K_1 s}} + (1-p)\frac{0.01}{0.01 + 0.99e^{-K_2 s}} .$$

Figure 4.1 shows the results obtained by using the several choices for the seroconversion distribution.

For the incubation time scale we get the results plotted in figure 4.2. It shows the nonparametric estimates of the incubation time distributions corresponding to a Weibull seroconversion distribution, and the two epidemic curve approaches for the seroconversion distribution. The better survival for the Weibull approach arises from the fact that the Weibull seroconversion distribution gives a higher probability for early seroconversions than the epidemic curve approaches, resulting in longer incubation times. However the differences are small as illustrated in table 4.1 and figure 4.2. It has to be noted that the two choices for the epidemic curves are merely choices to illustrate how much influence a non-uniform seroconversion pattern in the first four or five years of the HIV epidemic has on the incubation time. The problem is that we don't now with certainty whether or not the parametric form holds for the seroprevalent group. For the seroprevalent cases it may be better to use marker values to reveal some information on their seroconversion pattern as described in the next section.

Percentiles in years of the incubation time			
seroconversion pattern	25%	50%	75%
Weibull	5.70	9.12	11.96
Exp + Weibull	5.42	8.22	12.33
Mixed logistic	5.53	9.10	14.1

Table 4.1: Numerical summary of the incubation time distribution estimates for the HOM-study including prevalent cases.

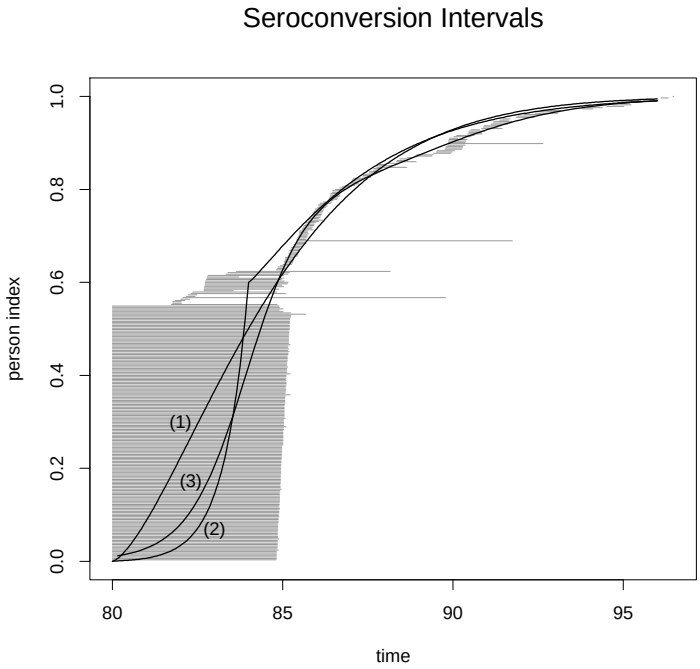


Figure 4.1: Estimated distribution functions of the seroconversion time : (1) = Weibull, (2) = exp + Weibull, (3) = logistic

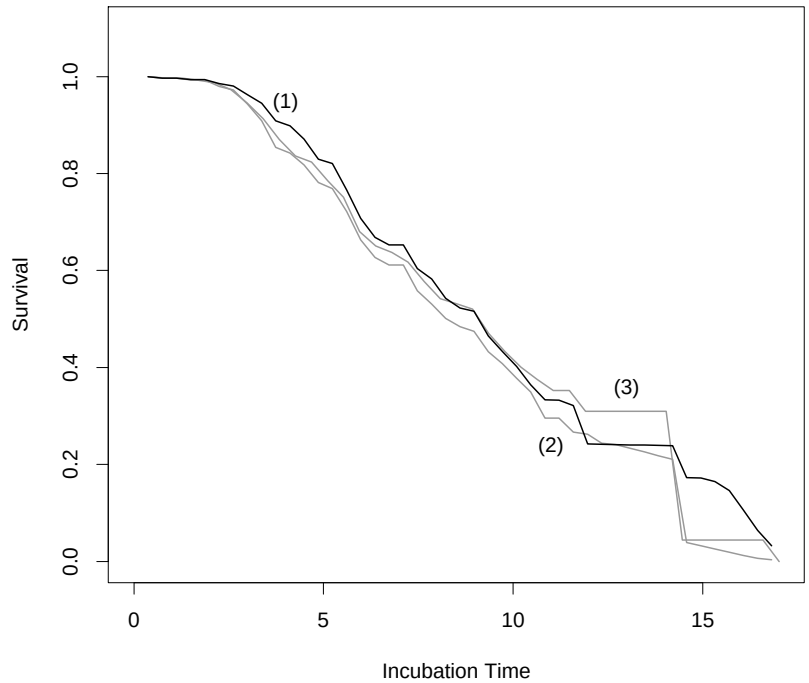


Figure 4.2: Survival functions of the incubation time based on: (1) = Weibull, (2) = Exp + Weibull (3) = logistic

4.2 The Use of Marker Values

Markers are variables that track the progression of HIV infection. The course of the marker values are consequences of disease progression. Three classes of markers have emerged (see

[5]). Immunological markers, for example numbers of $CD4^+$ T cells, measure of T cell function and levels of serum neopterin. Virological markers, for example the number of HIV RNA copies in serum or plasma (the viral load), presence or absence of detectable p24 antigen, and the presence or absence of syncytium inducing variants of HIV (SI/NSI). Clinical markers, are for example weight loss, candidiasis, and persistent fever.

One way to reveal information on the seroconversion times of a seroprevalent group is to build a model for marker values over time for a seroconverter group. This model is then used for the seroprevalent group. The marker values at study entry of each person in the seroprevalent group are used to calculate a time of seroconversion. However, due to the large intra- and outerindividual variance it is not possible to calculate one moment of seroconversion and treat that moment as the seroconversion moment of a seroprevalent person. We have to settle for a distribution which indicates how likely it is a person converted t years before study entry.

4.2.1 A Parametric Approach

The relation between time after seroconversion T and a marker value X is modeled by a regression model of the following form:

$$g(T_k) \;=\; \alpha + \beta\ f(X_k) + \epsilon_k \tag{4.3}$$

where g and f are known transformations and ϵ_k's are error terms. For example, Muñoz et al (see [11]) applied this model to the $CD4^+$ T cells marker. He specified:

$$\log\ T_k \;=\; \alpha + \beta\ \mathrm{CD4_k} + \epsilon_\mathrm{k} \tag{4.4}$$

where the ϵ_k's are i.i.d. extreme value distributed (see [4]). So T has a Weibull distribution with a location that depends on the CD4 count. Standard regression analysis can be used to check the validity of the model, for example scatterplots to check the linearity, residual plots to check the error distribution assumptions etc. The model can also be extended to allow heteroscedasticity and truncation effects.

From the seroconverter group of the HOM-study, we have measurements of CD4 numbers. The vast majority of the seroconverters visited the Municipal Health Service at intervals of 3 months. At these visits the CD4 number was measured. These data do not really suggest that the model (4.4) as used by Muñoz is applicable, see figure 4.3. Instead, after some transforming on the time variable T and the CD4 variable we find that the following linear model looks better for our cohort data

$$\left(\frac{T_k}{\mathrm{CD4_k}}\right)^{1/4} \;=\; \alpha + \beta\ \log\ \mathrm{CD4_k}\ +\ \epsilon_\mathrm{k}\ , \tag{4.5}$$

where ϵ_k are i.i.d. normal distributed with mean zero and variance σ^2, see figure 4.3.
This means that T has a somewhat strange distribution, a normal distribution to the power four scaled by CD4,

$$T \quad\sim\quad \mathrm{CD4}\ \mathcal{N}(\alpha + \beta\ \log\ \mathrm{CD4}\ , \sigma^2)^4 \tag{4.6}$$

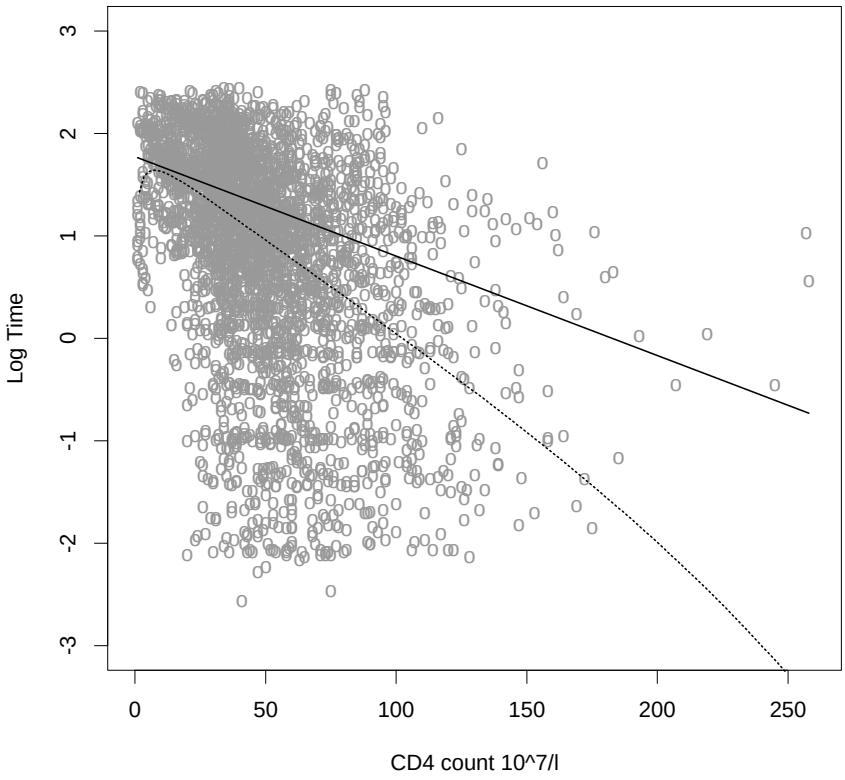


Figure 4.3: Scatterplots and predicted values based on formula (4.4) (solid line) and (4.5) (dashed line)

With this model we can calculate the seroconversion distribution of a prevalent person, by inserting his CD4 count at study entry into equation (4.6).

Example
Suppose we have a seroprevalent person entering the study at the beginning of 1985. At that moment his CD4 count is 70 $10^7/l$. From the seroconverter group we estimated $\alpha = 1.38$, $\beta = -0.23$ and $\sigma^2 = 0.011$ in (4.6). Thus, the distribution of T, the time after seroconversion with a particular CD4 count, can be calculated. This distribution must be truncated on the interval (0,5) because seroconversions before 1980 are very unlikely. Figure 4.4 shows the seroconversion distribution over calendar time. This figure shows that it is more likely that this person converted after 1983 than before.
So for all the seroprevalent cases who have a CD4 measurement at study entry we can calculate his seroconversion distribution G_i given his CD4 number. Consequently we can substitute this distribution in the semiparametric likelihood (4.1) in order to get an estimate of the incubation time distribution.

It is possible that every data set has its own transformation and error distribution assumptions. Moreover, often it is not really clear whether or not a certain model fits the data well. For example, if we look at the plot of model (4.5) in figure 4.5 then we see a strange

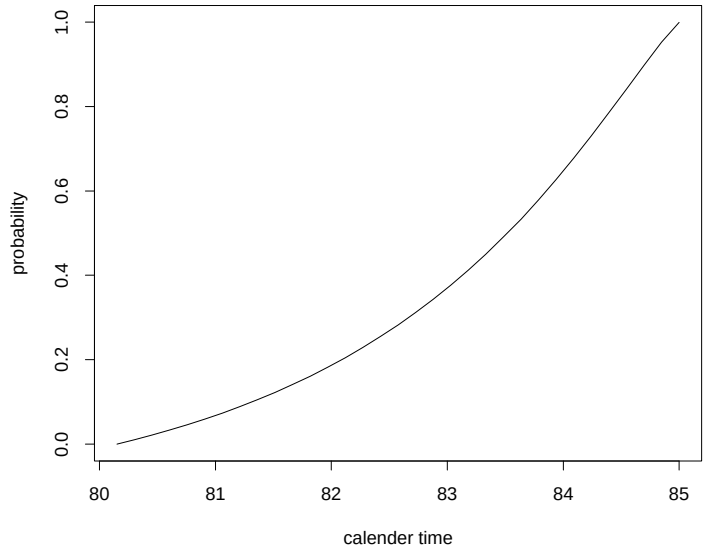


Figure 4.4: Distribution function of the seroconversion time for CD4 number 70 $10^7/l$

behaviour at the lower CD4 numbers of the time after seroconversion distribution. Therefore we look at a nonparametric approach in the next section.

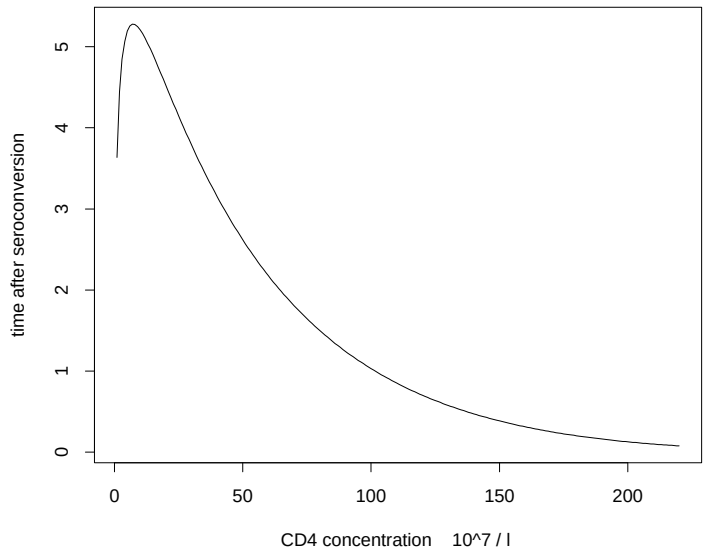


Figure 4.5: Plot of predicted time after seroconversion against CD4, according to model (4.5)

4.2.2 A Nonparametric Approach

A nonparametric approach will let the data speak more for themselves, instead of forcing them into a model. We take the raw data of marker values of the seroconverter group and estimate, for a given marker value, the time after seroconversion distribution. Let the data

consist of points $(T_i, X_i),\ i = 1, \ldots, N$, where T_i is the time after seroconversion and X_i the corresponding marker value. For a give marker value m define $d_i = |X_i - m|\ , i = 1, \ldots, N$, and let $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N)}$. Define for some $k < N$

$$L_m = \left\{ T_i : |X_i - m| \leq d_{(k)} \right\}\ .$$

For a given marker value m we use the times $T_i \in L_m$ to estimate the time after seroconversion distribution G_m, given the marker value m. So L_m is the set of k T_i's corresponding to the k nearest-neighbour marker values of m, for some $k \leq N$. The set L_m consists of non-censored observations, hence it is possible to use the empirical cumulative distribution function (ECDF) as an estimate of G_m. The ECDF puts a probability mass $1/k$ on every $T_i \in L_m$, where k is the number of $T_i \in L_m$. However, the set L_m can contain several times belonging to the same person, which may lead to a bias in the estimation of G_m. For example, the data set may contain a person who has a stable CD4 number over time, say 70 $10^7/l$. If we look in the neighbourhood of 70 we take all his observations. It is then better to use a weighted empirical cumulative distribution function (WCDF).

The WCDF

Let the times T_i in L_m belong to l persons and $n = n_1 + n_2 + \ldots + n_l$, where n_j is the number of times of person j. Thus the times T_i can be regrouped by person: $T_{j,i}\ j = 1, \ldots, l,\ i = 1, \ldots n_j$. If $l = n$ then we have the situation that every T_i in L_i belongs to a different person. The WCDF puts mass $1/n_j l$ on every $T_{j,i}$, thus the $T_{j,i}$'s that belong to persons with many observations receive less probability mass.

Example of the WCDF

The following somewhat extreme example shows the effect of the WCDF. Suppose we have a sample $x_1, \ldots, x_{100}$ from a standard normal distribution, the ECDF of this sample is plotted in figure 4.6. It looks normal. However, suppose that for some reason the 30 largest observations are from one person, then these observations receive less probability mass than they normally would, since these observation are only from one person. The WCDF for this example is illustrated in figure 4.6.

4.2.3 The nonparametric Approach for the HOM-study

To illustrate the nonparametric approach we use the measured CD4 numbers of the seroconverter group of the HOM-study for the analysis of the seroprevalent group.

Estimation of individual seroconversion curves of prevalent cases

We estimate the distributions of the time after seroconversion of seroprevalent cases with their measured CD4 counts at study entry in 1985. To deal with the random fluctuation of CD4 measurements we took the average of the first four CD4 measurements, if available, for each prevalent case. For each seroprevalent case with average CD4 number m at study entry,

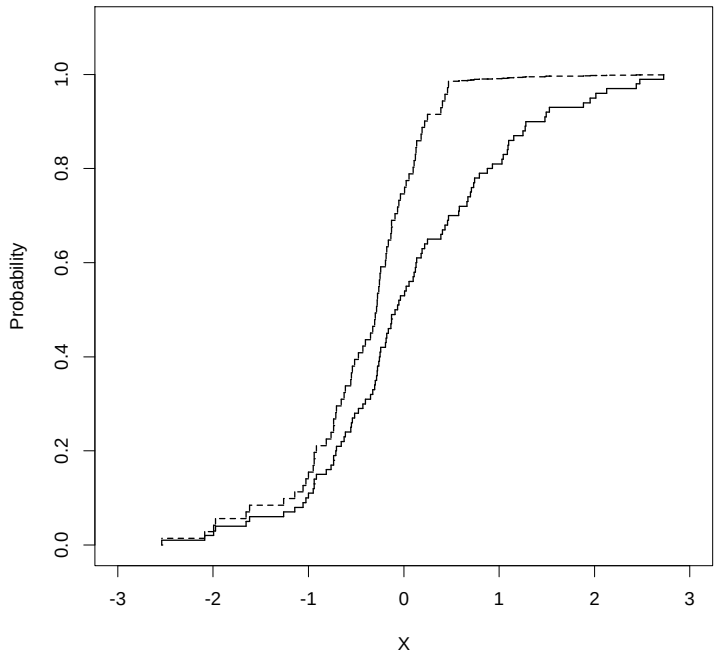


Figure 4.6: solid: the ECDF, dashed: the WCDF, based on a normal sample.

we determine the set L_m by taking the 150 CD4 measurements of the seroconverters closest to m. Using the WCDF we estimate his distribution of the time after seroconversion. With this distribution at hand we can calculate the seroconversion distribution over calendar time. For most of the seroprevalent cases we know that seroconversion took place between 1980 and study entry which is between October 1984 and April 1985. Hence, we condition the seroconversion distribution on these intervals. Figure 4.7 shows the conditioned distribution functions of all the seroprevalent cases of the HOM-study entering around 1985, together with the average curve. We see that the majority has an exponentially looking curve, indicating that the probability of seroconversions increases over the period 1980-1985.

The analysis of the incubation time of prevalent cases

With all the calculated individual seroconversion curves we can use two methods to analyse the incubation time. First, the double censoring method which uses each individual's seroconversion distribution. The term of a seroprevalent case i in likelihood (3.1) is replaced with a term that contains the individual seroconversion density of case i g_i, instead of a general seroconversion density g. With this method we directly take account of the interval censored nature of the date of seroconversion. However, the computation of the incubation time distribution is time consuming. Figure 4.10 shows the estimate (dashed curve) of the incubation time.

Second, imputation methods calculate one date of seroconversion for each prevalent case. Once a date of seroconversion has been imputed for each case we can use the Kaplan-Meier estimator for the estimation of the incubation time. One way to impute a date is to calculate

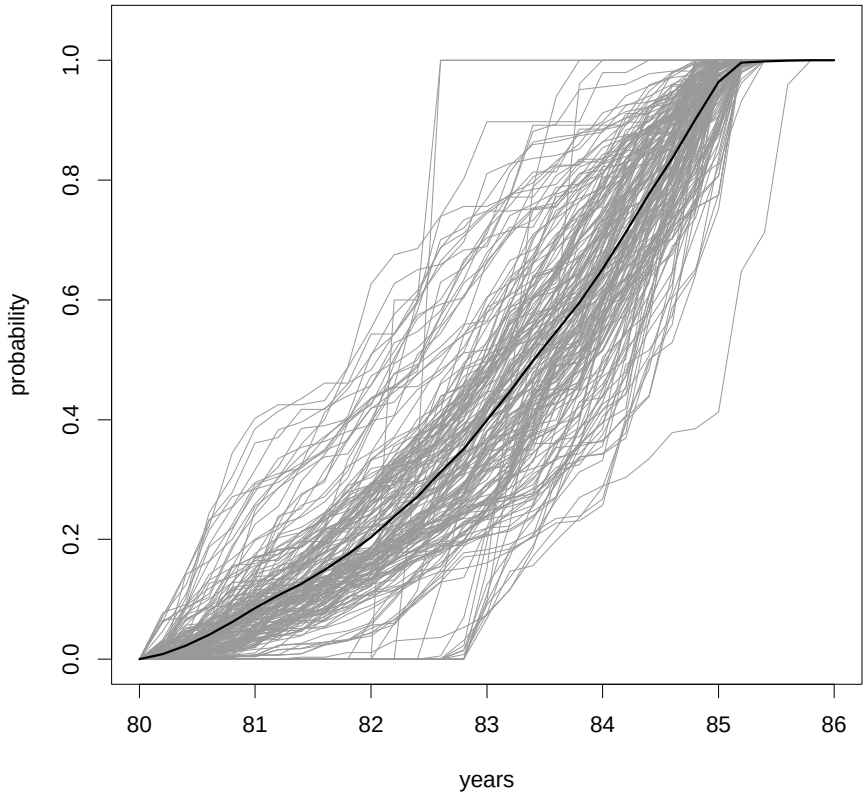


Figure 4.7: The estimated seroconversion distribution functions of the prevalent cases (light curves) and the average curve (black curve)

the expected seroconversion date, in a similar way as in section 2.1.3. Instead of using (2.3) with one seroconversion curve for the total population we use the individual seroconversion curve based on CD4 numbers. Figure 4.8 shows the empirical distribution (dark line) of the expected dates of seroconversion for all the seroprevalent cases over the calendar time scale. The estimate of the incubation time distribution based on the expected dates of seroconversion is shown in figure 4.10 for the total population. Instead of imputing the expected date of seroconversion for a prevalent case, we can also impute a date of seroconversion by randomly drawing one date of seroconversion for each prevalent case based on his individual seroconversion curve. Figure 4.8 shows the empirical distribution (light line) of one sample of randomly drawn dates. The difference in the incubation time estimates between expected seroconversion dates and randomly drawn is illustrated in figure 4.9. To take account of the fluctuation of randomly drawn seroconversion dates we took the average of ten incubation time survival curves based on ten randomly drawn dates of seroconversion. In appendix B we performed a small simulation study to see the difference of randomly drawn and expected seroconversion dates on the incubation time. It turns out that the expected seroconversion dates perform slightly better, *i.e.* the resulting estimated incubation time distribution resembles the 'true' incubation time distribution more than with randomly drawn seroconversion dates.

Estimating the incubation time via double censoring or via imputed expected dates of sero-

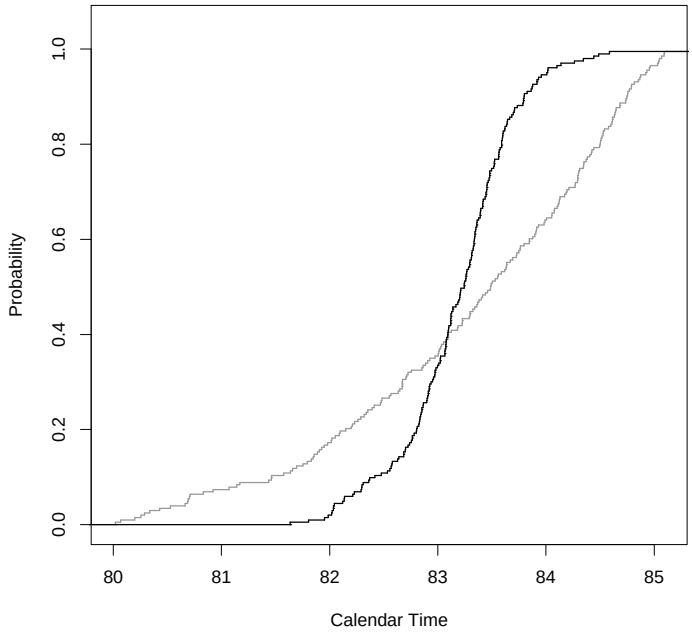


Figure 4.8: Emperical distribution of dates of seroconversion. dark: expected dates of seroconversion, light: randomly drawn dates

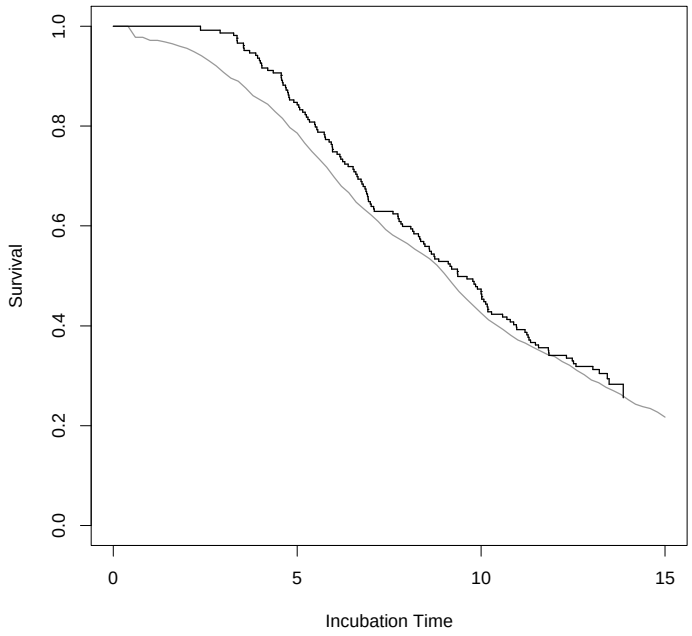


Figure 4.9: Kaplan-Meier based on expected seroconversion dates (dark) and Kaplan-Meier based on randomly drawn seroconversion dates (light).

conversion could lead to two resembling curves. This is caused by the uniform behaviour of the incubation time distribution. Let SC_i be the seroconversion distribution of case i. If we look at a likelihood term of an observed AIDS diagnosis in the double censoring approach we

get the following derivation:

$$\begin{aligned}\int_{z_i-v_i}^{z_i-u_i} g_i(z_i-s)f(s)ds &= \mathbb{E}(f(z_i-SC_i)) \\ &\approx f(\mathbb{E}(z_i-SC_i)) \text{ if } f \text{ is almost linear or constant.} \qquad (4.7)\end{aligned}$$

For a right censored AIDS diagnosis we get:

$$\begin{aligned}\int_{z_i-v_i}^{z_i-u_i} g_i(z_i-s)[1-F(s)]ds &= \mathbb{E}(1-F(z_i-SC_i)) \\ &\approx 1-F(\mathbb{E}(z_i-SC_i)) \text{ if } F \text{ is almost linear or constant.} \quad (4.8)\end{aligned}$$

The terms (4.7) and (4.8) are just the terms in the Kaplan-Meier likelihood.

Truncation effects

If we estimate the incubation time distribution with double censoring or imputation methods, we have to deal with truncation effects. It is possible that individuals with (very) short incubation times are selectively excluded from the study, since the HOM-study only followed individuals who were AIDS-free at study entry. For example, a person who seroconverted in 1982 and had AIDS in 1984 was not selected for the HOM-study. So we have to correct for this by using the method in section 3.4.1. For the imputation methods this is easy to perform but not completely correct, for the double censoring method this is more complicated.

The start of the AIDS epidemic

Until now we took 1980 as the beginning of the AIDS epidemic. This is plausible for the epidemic in Amsterdam. For the nonparametric approach as described in chapter 3 it didn't matter what we took as the beginning of the AIDS epidemic. However, with the individual seroconversion curves approach we assumed that seroconversions before 1980 were not possible and conditioned the individual curves on the interval 1980-1985. We can also look at the effect of taking 1978 as the beginning of the AIDS epidemic. Instead of conditioning the individual seroconversion curves of the seroprevalent cases on the interval 1980-1985 we have to condition them on the interval 1978-1985. Figure 4.10 shows the corresponding incubation time distribution (light curve). There is not much difference compared with taking 1980 as the beginning of the AIDS epidemic. This is explained by the fact that in the period 1978-1982 few infections took place, so the impact on the incubation time distribution is small.

4.2.4 Comparison of prevalent Cases and Seroconverters

Using the individual seroconversion patterns for the prevalent cases we can compare the incubation time of the prevalent cases (infected before April 1985) and seroconverters (infected after October 1984). So we can get an indication whether or not the incubation time of people infected early in the epidemic is on average longer than the incubation time of people infected later in the epidemic. For both 1978 and 1980 as starting date of the AIDS epidemic

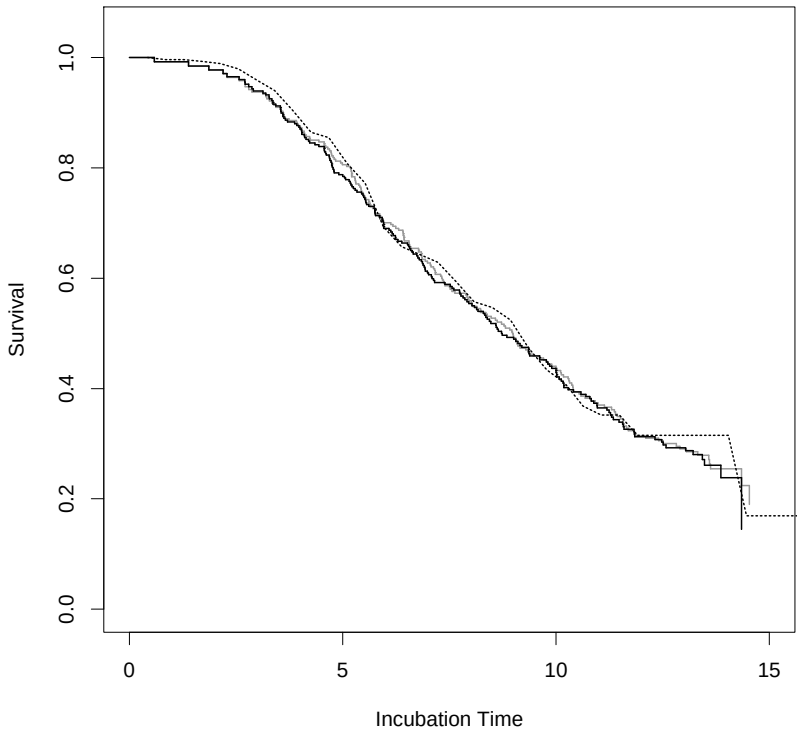


Figure 4.10: Incubation time distributions based on CD4 number, solid = Kaplan-Meier dashed = double censoring, light = Kaplan-Meier based on 1978 as begin of AIDS epidemic.

we looked at the differences in survival. The analysis is quite simple but has to be interpreted carefully. We split the data into a prevalent group ($n = 203$) and a seroconverter group ($n = 126$). For a prevalent case we take as date of seroconversion the expected seroconversion date given his seroconversion pattern and for a seroconverter we take the midpoint of the seroconversion interval as seroconversion date. The three resulting Kaplan-Meier estimates are plotted in figure 4.11. We see that the prevalent group has a longer time to AIDS than the seroconverter group. The log-rank test statistic for testing the difference between two survival curves doesn't give a significant result. For the 1978 curve the test statistic has a value of 3.01 with a corresponding p-value of 0.083 and for the 1980 curve the value is 1.91 with a corresponding p-value of 0.16.

Percentiles in years of the incubation time and 95% confidence intervals			
	25%	50%	75%
converters	4.73 (4.04, 5.76)	8.03 (7.15, 10.80)	11.59 (10.83, ∞)
prevalent cases	5.96 (5.48, 6.76)	9.36 (8.44, 10.30)	13.90 (13.00, ∞)

Table 4.2: Numerical summary of the prevalent group and converter group.

The log-rank test statistic should be used with caution in this case. It doesn't take account of the uncertainty in the dates of seroconversion. So taking the expected dates of seroconversion

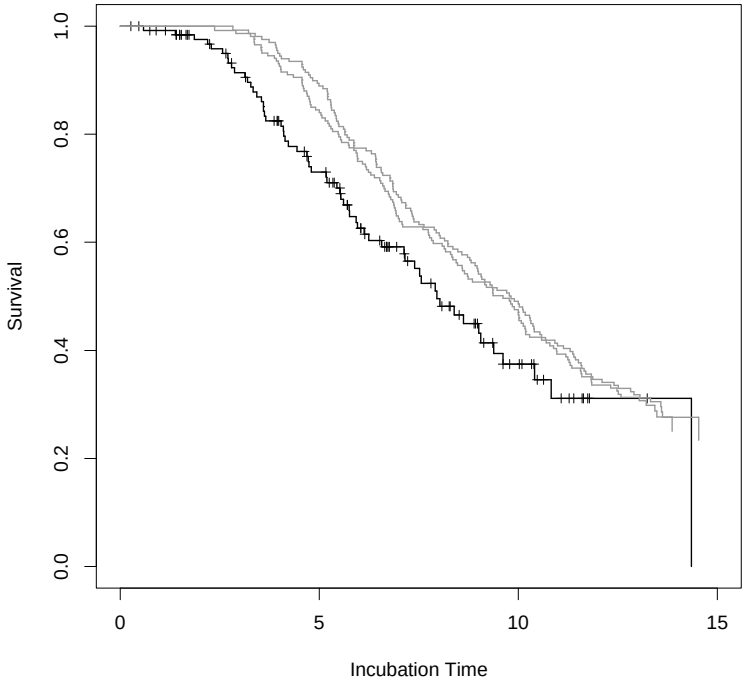


Figure 4.11: Incubation time estimates of prevalent cases: 1978 as start (upper curve), 1980 as start (middle curve). Incubation time estimate of seroconverters (lower curve)

and using them for further analysis will lead to an underestimation of the true p-value of the log-rank test statistic. So in any analysis where expected dates of seroconversion are used we should validate a borderline significant p-value. This can be done by a bootstrap procedure. A bootstrap procedure can be used to reveal information on the distribution of a test statistic by means of computer simulations. Under the assumption that the seroprevalent group and the seroconverters have the same incubation time distribution we reconstruct the data for the two groups by simulation and calculate the log-rank test statistic, in similar way as we did with the original data. By repeating this procedure a large number of times we can get an idea of the randomness of the test statistic. We used the following bootstrap procedure:

- Step 1. Use the individual seroconversion curves of the seroprevalent cases to simulate an individual seroconversion date SC_i for each prevalent case i.

- Step 2. For case i let $Z_i = SC_i + T_i$, where T_i a randomly drawn (T_i, δ_i) incubation time or censoring time from the seroprevalent group based on the expected seroconversion dates $\mathbb{E}SC_i$. These times may be right censored or not. Let the simulated incubation time or censoring time be $T_i' = Z_i - \mathbb{E}SC_i$.

- Step 3. Randomly sample from the incubation times T_i' obtained from step 2 to simulate the variability in the Kaplan-Meier. So if n_1 is the number of prevalent cases we n_1 times draw with probability $1/n_1$ from our incubation/censoring times with replacement, obtaining a data set of n_1 cases.

- Step 4. Calculate the Kaplan-Meier of the data set obtained in step 3, call it KMprev.
- Step 5. Randomly sample form the original incubation/censoring times of the seroprevalent group based on the expected seroconversion dates to form a simulated seroconverter group. So if n_2 is the number of seroconverter cases, we n_2 times draw with probability $1/n_2$ from the prevalent cases with replacement, obtaining a data set of n_2 cases.
- Step 6. Calculate the Kaplan-Meier of the data set obtained in step 5, call it KMconv.
- Step 7. Use KMprev and KMconv to calculate the log-rank test statistic T^*

Repeat this procedure N times to get N log rank test statistics $T_1^*, T_2^*, \ldots, T_N^*$. Let T be the test statistic obtained by comparing the original seroconverter group and the seroprevalent group based on expected dates of seroconversion. The p-value of the test statistic can then be calculated as the proportion of T^* values greater than T. Figure 4.12 shows the histogram of 300 simulated log-rank test statistics T^* together with the χ_1^2 density. We see that the simulated distribution has a sligthly heavier tail than the χ_1^2. So the p-values corresponding with the test statistics 3.01 and 1.91, found earlier are 0.11 and 0.22, instead of 0.083 and 0.16

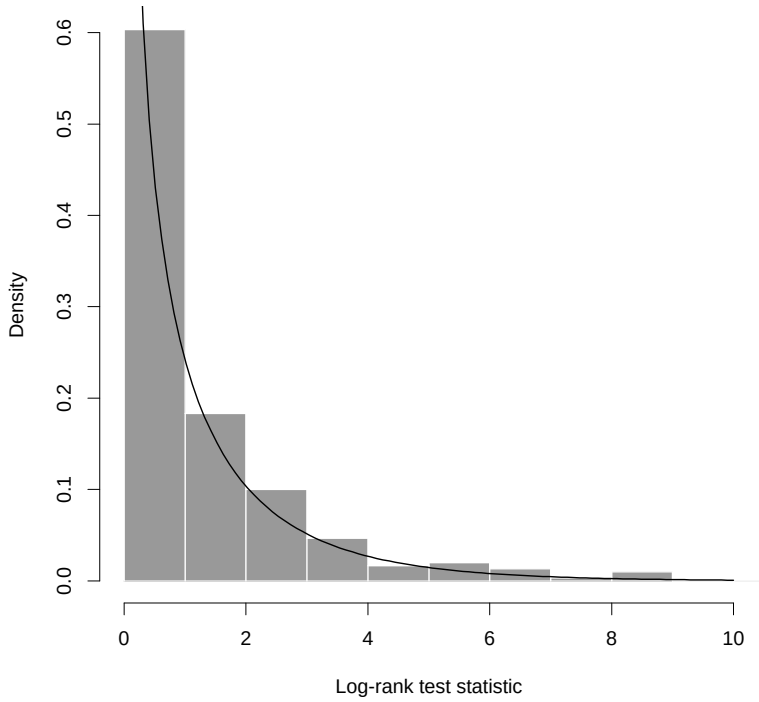


Figure 4.12: Histogram of 300 simulated log-rank test statistics together with the χ_1^2 density

Chapter 5

Parametric Models and Covariates

In this chapter we describe a completely parametric approach of double censoring and the inclusion of covariates.

5.1 Parametric Models

5.1.1 The Likelihood

In a completely parametric model, we assume that the seroconversion time and the incubation time have parametric forms g_{θ_1} and f_{θ_2}, respectively. The log-likelihood $\log L(\theta)$ for the data set $(u_i, v_i, z_i, \delta_i),\ i = 1, \ldots, N$ is given by:

$$\sum_{i=1}^{N} \log \left[\int_{z_i - v_i}^{z_i - u_i} g_{\theta_1}(z_i - s) f_{\theta_2}(s) ds \right]^{\delta_i} + \log \left[\int_{z_i - v_i}^{z_i - u_i} g_{\theta_1}(z_i - s)[1 - F_{\theta_2}(s)] ds \right]^{1-\delta_i} \tag{5.1}$$

where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^d$, and d is the dimension of the joint parameter space of the seroconversion and incubation time.

The maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$ is obtained by maximizing $\log L$ with respect to θ. The advantage of using completely parametric models is that there is a 'standard' maximum likelihood theory, which holds under regularity conditions. This theory gives us the following (asymptotic) properties of $\hat{\theta}_{\mathrm{ML}}$

- $\hat{\theta}_{\mathrm{ML}}$ is consistent: $\hat{\theta}_{\mathrm{ML}}$ converges in probability to the true value θ_0 of θ when N tends to infinity.
- It is asymptotically normally distributed:

  $$\sqrt{N}(\hat{\theta}_{\mathrm{ML}} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1})$$

  when N tends to infinity.

The variance of $\hat{\theta}$ is, for large values of N, approximately equal to $\mathcal{I}(\theta)^{-1}/N$ and will vanish for $N \to \infty$. $\mathcal{I}(\theta)$ is known as the Fisher Information matrix. It is defined as the $d \times d$ matrix, with i,j-th element $\mathcal{I}_{ij}(\theta)$ given by

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log L(\theta)}{\partial\theta_i \partial\theta_j}\right]$$

which can be evaluated at $\hat{\theta}_{\mathrm{ML}}$ to estimate the covariance matrix for the MLE. However the exact expected value will rarely be available in practice, because of the complicated non-linear second derivatives. Instead we may estimate the (i,j)-th element of $\mathcal{I}(\theta_0)$ by

$$\hat{\mathcal{I}}_{ij}(\theta_0) = \left[-\frac{\partial^2 \log L(\theta)}{\partial\theta_i \partial\theta_j}\right]_{\theta=\hat{\theta}_{\mathrm{ML}}}$$

This is computed simply by evaluating the second derivative matrix of the log-likelihood function at the maximum likelihood estimate. It is the so-called 'observed information matrix', not the expected.

The estimation of θ is often not of interest, one is more interested in the estimation of a function of θ, $h(\theta)$. The maximum likelihood estimate of $h(\theta)$ is simply $h(\hat{\theta}_{\mathrm{ML}})$ if h is differentaible, and as a consequence of the Delta-method, the asymptotic variance can be obtained by:

$$\sqrt{N}\left(h(\hat{\theta}_{\mathrm{ML}}) - h(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0,\ \nabla h(\theta)\mathcal{I}(\theta)^{-1}\nabla h(\theta)^t\right)$$

when N tends to infinity. For example, the mean of a log normal incubation time with parameters λ and α is given by $h(\lambda,\alpha) = \exp(\lambda + \alpha^2/2)$ and is estimated by $h(\hat{\lambda},\hat{\alpha})$. An estimate of the variance of the mean is given by $\nabla h\ \mathcal{I}^{-1}\ \nabla h^t$.

For the optimization of (5.1) with respect to θ we can use the conjugate gradient method (see [7]) with numerical derivatives, since analytical derivatives are algebraically messy. To compute the integrals in each likelihood term of (5.1) we use Romberg integration (see [7]).

5.1.2 Parametric Models for Distributions

Although estimators in parametric models have nice asymptotic properties, and are in models with censored data easier to calculate than nonparametric estimates, they must be used with caution. A misspecified model of the incubation time or seroconversion time distribution can lead to parameter estimates which are useless and confidence interval estimations which are unrealistic small. So a validation of the specified parametric forms is absolutely necessary. This validation is usually based on nonparametric estimates. We give a brief description of some parametric models (see [4]) that can be used and we will show how to validate them. Although we did not check wheter or not the regularity conditions hold for the following models, we used the asymptotic properties of the maximum likelihood estimators.

The Weibull Model

The density f of the Weibull model with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ is given by

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1} \exp[-(\lambda t)^{\alpha}],\ t > 0.$$

The distribution function F is given by

$$F(t) = 1 - \exp[-(\lambda t)^{\alpha}],\ t > 0.$$

For $\lambda < 1$ the hazard function of the Weibull distribution is strictly decreasing, and for $\lambda > 1$ the hazard function is strictly increasing. The Weibull distribution can be developed as the limiting distribution of the minimum of a sample from a continuous distribution with support on $[0, u)$ for some $0 < u < \infty$.

The Log-normal Model

If a random variable T has a normal (Guassian) distribution with parameters μ and $\sigma > 0$ then $Y = \exp(T)$ has a log-normal distribution. The density f of Y is given by

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp[-(\log(t) - \mu)^2/2\sigma^2],\ t > 0\ ,$$

and the distribution function is given by

$$F(t) = \Phi\left(\frac{\log(t) - \mu}{\sigma}\right),\ t > 0, \tag{5.2}$$

where $\Phi(x)$ is the standard normal distribution function. The hazard function of the log-normal distribution is first increasing and then decreases to zero as $t \to \infty$, and thus the hazard is only decreasing in the long-life range.

The Log-Logistic Model

The log-logistic model is only used occasionally in modeling life time distributions. It has the advantage of having a simple algebraic expression for the distribution function F. The log-logistic distribution with parameters λ and p is given by

$$F(t) = 1 - \frac{1}{1 + (\lambda t)^p},\ t > 0.$$

The Generalized Gamma Model

The generalized gamma model with parameters λ, p and k has a somewhat complicated density function f, given by

$$f(t) = \frac{\lambda p(\lambda t)^{pk-1} \exp[-(\lambda t)^p]}{\Gamma(k)},\ t > 0$$

where Γ is the gamma function $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$. The generalized gamma includes as special cases of the parameters the Weibull ($k = 1$) and the log-normal ($k \to \infty$) distribution. So it permits their evaluation relative to each other and to a more general model.

Validation

One simple way to validate a parametric form is to estimate the parametric form and then plot its distribution function $F_\theta(t)$ together with the nonparametric estimate $F(t)$ of the distribution function. This approach has two minor drawbacks. First, the parametric form has to be estimated first and second the s-shaped curves of the distribution function may be misleading, it can 'hide' some subtle differences between the two curves. A better approach is to find a transformation ψ such that $\psi(F_\theta(t))$ is linear in $\log t$ or t. Applying this transformation to the nonparametric estimate $F(t)$ and plotting it against $\log t$ or t should give a fairly straight line if the underlying distribution is indeed F_θ.

5.1.3 Results for the HOM-study

We validate whether or not parametric models give a reasonable fit for the HOM-study. For a log-normal model the distribution is given by (5.2), so if we take $\psi(x) = \Phi^{-1}(x)$ then $\psi(F_{\mu,\sigma}(t)) = \sigma^{-1}\log(t) - \mu/\sigma$. Applying ψ to the nonparametric estimate $\hat{F}$ of the incubation time, for example based on double censoring from section 3.5, and plotted against $\log t$ results in the graph shown in figure 5.1. We see a reasonable straight line with three outliers, moreover the slope and intercept can serve as rough estimates of μ and σ. The Weibull family can be validated by taking $\psi(x) = \log(-\log(1-x))$, yielding $\psi(F_{\lambda,\alpha}(t)) = \alpha\log(\lambda) + \log(t)$. Figure 5.1 shows the plot of $\psi(\hat{F})$ against $\log t$ for the seroconversion distribution. There is some deviation from the straight line, indicating the poor fit of a Weibull model.
If we exclude the prevalent cases and only look at the seroconverter group of the HOM-study we get a much better validation. There is an indication that both the seroconversion and incubation time distribution follow a Weibull distribution, see figure 5.3. We used the conjugate gradient method to estimate the parameters and the observed Fisher information matrix to derive the 95% confidence intervals. Table 5.1 shows the results.

Weibull seroconversion time $\theta_1 = (\lambda_1, \alpha_1)$		
	λ_1	α_1
estimate	0.281	1.105
95 % conf. int.	(0.234, 0.328)	(0.951, 1.259)
Weibull incubation time $\theta_2 = (\lambda_2, \alpha_2)$		
	λ_2	α_2
estimate	0.104	1.974
95 % conf. int.	(0.090, 0.119)	(1.557, 2.391)

Table 5.1: Parameter estimates for the Weibull model for the seroconverter group

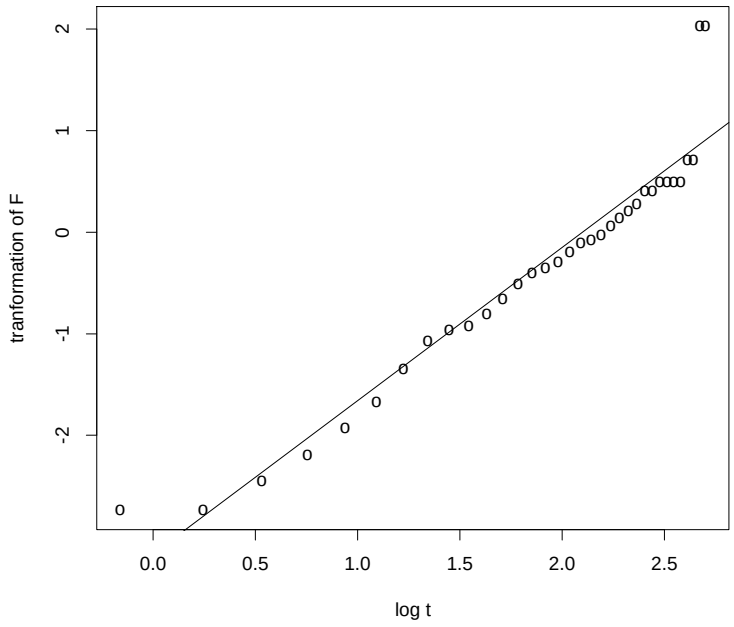


Figure 5.1: Validation of the log normal form of the incubation time distribution of the HOM-study.

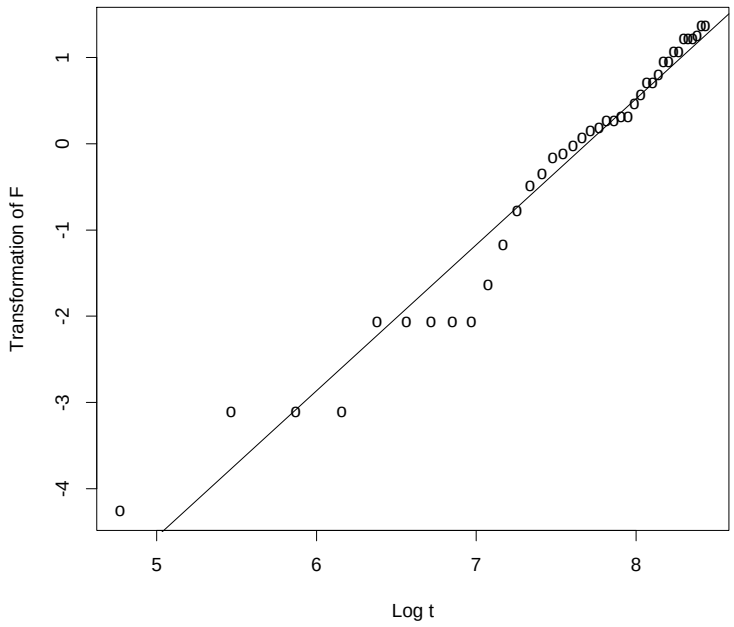


Figure 5.2: Validation of the Weibull form of the seroconversion time distribution of the HOM-study.

Figure 5.3 shows the estimated Weibull seroconversion distribution together with the seroconversion intervals of the seroconverter group, and figure 5.4 shows the estimated Weibull survival function together with the Kaplan Meier estimate. The esitimated median incuba-

tion time is 7.95 years with a 95% confidence interval of (6.87, 9.00). The 25% and 75% percentiles are 5.09 and 11.30 years with 95% confidence intervals of (4.27, 5.89) and (9.48, 13.09) respectively. We also compared the parametric double censoring approach with the parametric right censoring approach using midpoints of the seroconversion intervals as the date of seroconversion. The parameter estimates of the incubation time where almost identical, which we would expect with small seroconversion intervals.

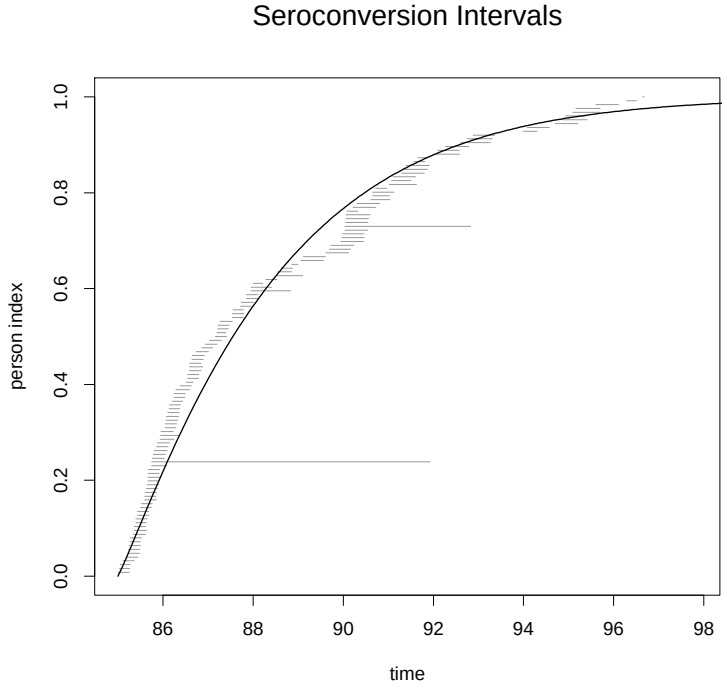


Figure 5.3: Seroconversion intervals of the seroconverter group of the HOM-study and fitted Weibull distribution.

5.2 Covariates and the Incubation Time

This section describes the analysis of covariates x of the incubation time T. Covariates are variables that affect the duration of the incubation time. They explain why some individuals progress to AIDS faster than others. It therefore becomes of interest to collect information on covariates of the incubation time. In several studies different covariates have been identified as not significant and some as significant. For example, in [21] no relation of sexual behavior, history of sexual transmitted diseases or use of alcohol, tobacco and recreational drugs with rates of disease progression could be demonstrated. However, younger age at seroconversion and use of prophylaxis were significantly related to a slower progression from seroconversion to death [21]. Also it is demonstrated [22] that certain host genetics have influence on disease progression.

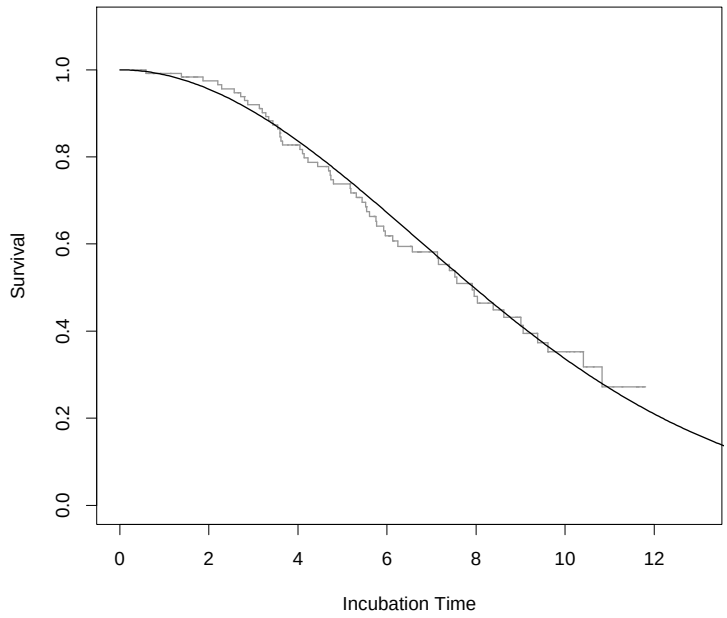


Figure 5.4: Survival functions of the incubation time: Weibull and Kaplan Meier

Two models that take account on information of covariates are widely used in survival analyses, the proportional hazard model (Cox regression) and the accelerated failure time model, see [4].

5.2.1 The Proportional Hazards Model

The proportional hazards model specifies a multiplicative effect of the covariates on the hazard. Let $\lambda(t;x)$ be the hazard at time t for an individual with covariates x, then the model specifies that

$$\lambda(t;x) \quad = \quad \lambda_0(t)e^{x\beta} \tag{5.3}$$

where $\lambda_0(t)$ is an arbitrary base-line hazard function for T. To see the impact of the covariates on the distribution function F of T in the proportional hazard model we rewrite (5.3) in terms of distribution functions. We get

$$F(t;x) \quad = \quad 1-[S_0(t)]^{\exp(x\beta)}\ , \tag{5.4}$$

where $S_0(t)$ is the base-line survivor function $S_0(t) = \exp(-\int_0^t \lambda_0(u)\ du)$

5.2.2 The Accelerated Failure Time Model

The accelerated failure time model is a log-linear model for the incubation time T. If we define $Y = \log T$ then the model assumes a linear relationship between Y and the covariates x,

$$Y \quad = \quad x\beta + W$$

where W is an error variable. In terms of T we get

$$T = \exp(x\beta)T_0 \ , \ \text{where} \ \ T_0 = \exp(W)$$

T_0 can be regarded as a base-line incubation time. So depending on β the role of x is to accelerate or decelerate the incubation time. The distribution function of T with covariates x is given by

$$F(t;x) \ = \ F_0(t\exp(-x\beta)) \tag{5.5}$$

5.2.3 Plugging the Covariates into double Censoring

Estimation of the Cox regression model and the accelerated failure time model has only been done in samples with right censored data, since most software packages are only equiped with such routines. For interval and double censoring no general software is available. However, estimation can be done by plugging the distribution functions (5.4) or (5.5) into the log-likelihood (2.4) or (3.1) and then optimizing the log-likelihood with respect to the seroconversion time distribution G, incubation time distribution F and the regression parameters β.

For example, if we take the Cox regression model and interval censoring we get the following log-likelihood

$$\begin{aligned} \log L \ &= \ \sum_{i=1}^{n} \delta_i \log([1-F_0(l_i)]^{\exp(x\beta)} - [1-F_0(r_i)]^{\exp(x\beta)}) \\ &\quad + \ (1-\delta_i)\log([1-F_0(l_i)]^{\exp(x\beta)}) \end{aligned} \tag{5.6}$$

For double censoring and the accelerated failure time model we get the following log-likelihood

$$\begin{aligned} \log L(F_0,\beta) \ &= \ \sum_{i=1}^{N} \log\left[\int_{z_i-v_i}^{z_i-u_i} g(z_i-s)dF_0(s\exp(-x\beta))\right]^{\delta_i} \\ &\quad + \log\left[\int_{z_i-v_i}^{z_i-u_i} g(z_i-s)[1-F_0(s\exp(-x\beta))]ds\right]^{1-\delta_i} \end{aligned} \tag{5.7}$$

For log-likelihood (5.6) some results have been given in [15]. Under appropriate regularity conditions the maximum likelihood estimator for the regression parameter β is shown to be asymptotically normal and efficient. To estimate F_0 and β a profile likelihood approach can be used. In this approach the log-likelihood is first maximized over F_0 for fixed values of β to obtain $\hat{F}_0(\beta)$, then the profile log likelihood function $\log(\hat{F}_0(\beta),\beta)$ is maximized over β to find $\hat{\beta}$. As mentioned in [15] this method is computationally feasible for low dimensional β. For higher dimensional β such an exhaustive search is computationally infeasible, iterative search methods like the SQP algorithm are needed. However, thses search methods can not guarantee to find a global maximum.

For double censoring, no proof of asymptotic normality and efficiency is available yet. However, for a fixed small grid compared to the sample size, one might consider the model as completely parametric and use asymptotic normality and Fisher Information matrix, see [23]. A profile likelihood approach in the double censoring case would probably be too slow, since the computation of the distributions F_0 and G_0 takes a lot more time than the computation of F_0 in the interval censoring case.

Bibliography

[1] Groeneboom P. Wellner J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* Birkhäuser Verlag.

[2] GGGD, UvA, CLB (1996). *The amsterdam cohort studies on HIV infection and AIDS.*

[3] Bindels P.J.E. (1996). *Surveillance and survival studies on HIV/AIDS in Amsterdam.* Ponsen & Looijen.

[4] Kalbfleisch J. Prentice R.L. (1980). *The statistical analyses of failure time data.* Wiley, New York.

[5] Brookmeyer R. Gail M.H. (1994). *AIDS Epidemiology: A quantitative approach.* Oxford university press.

[6] De Gruttola V. Lagakos S.W. (1989). *Analysis of doubly-censored survival data, with application to AIDS.* Biometrics 45, p. 1-11.

[7] Press W.H. Teukolsky S.A. Vetterling W.T. Flannery B.P. (1992). *Numerical recipes in C.* Cambridge university press.

[8] Turnbull B.W. (1976). *The emperical distribution function with arbitrarily grouped, censored, and truncated data.* Journal of the Royal Statistical Society, Series B **38,** 290-295.

[9] Munoz A. Xu J. (1996). *Models for the incubation period of AIDS and variations according to age and period.* To appear in Statistics in Medicine (1996).

[10] Peterman T.A. Drotman D.P. Curran J.W. (1985). *Epidemiology of the acquired immunodeficiency syndrome (AIDS).* Epidemiologic Reviews 1985, Vol. 7.

[11] Taylor J.M.G. et al (1990). *Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation.* Statistics in Medicine 1990, Vol. 9, 505-514.

[12] Munoz A. et al (1992). *Estimation of time since exposure for a prevalent cohort.* Statistics in Medicine 1992, Vol. 11, 939-952.

[13] Wand M.P. Jones M.C. (1995) *Kernel smoothing.* Monographs on Statistics and Applied Probability 60, Chapman & Hall

[14] Jongbloed G. (1995). *The iterative convex minorant algorithm for nonparametric estimation.* Technical Report 95-105, Delft University of Technology, submitted. **ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1995**

[15] Huang J. Wellner J.A. (1995b). *Efficient Estimation for the proportional hazards model with "Case 2" interval censoring.* Technical Report 290, University of Washington, Seattle. **http://www.stat.washington.edu:80/jaw/jaw.research.available.html**

[16] Tu X.M. (1995). *Nonparametric estimation of survival distribution with censored initiating time and censored and truncated terminating time.* Appl. Statist. **44**, No.1, pp 3-16.

[17] Sun J. (1995). *Emperical estimation of a distribution function with truncated ad doubly interval-censored data and its application to AIDS studies.* Biometrics 51, 1096-1104.

[18] Lawrence C. Zhou J.L. Tits A.L. (1997). *User's guide for CFSQP version 2.5: A C code for solving (large scale) constrained nonlinear optimization problems, generating iterates satisfying all inequality constraints.* University of Maryland, College Park, MD 20742. **http://www.isr.umd.edu/Labs/CACSE/FSQP/fsqp.html**

[19] Haastrecht H.J.A. van (1997). *HIV infection and drug use in the Netherlands: The course of the epidemic.* Journal of Drug Issues 27(1), 57-72 1997.

[20] Bazaraa M.S. Sherali H.D. Shetti C.M. (1993). *Nonlinear programming, theory and algorithms.* Wiley, New York.

[21] Veugelers P.J. Page K.A. et al (1994). *Determinants of HIV disease progression among homosexual men registered in the tricontinental seroconverter study.* American Journal of Epidemiology 1994; 140:747-758.

[22] Keet I.P.M. Klein M.R. et al (1996).*The rolr of host genetics in the natural history of HIV-1 infection: needles in the haystack.* AIDS 1996, **10** (supl A): s59-s67.

[23] Kim M.Y. De Gruttola V.G. Lagakos S.W. (1993). *Analyzing double censored data with covariates, with application to AIDS.* Biometrics 49, March 1993.

[24] I.P.M. Keet and others *Predictors of rapid progression to AIDS in HIV-1 seroconverters.* AIDS volume 7,pages 51-57, 1993.

[25] De Wolf F and Lange JMA and others. *Numbers of CD4+ cells and the levels of core antigens and of antibodies to the human immunodeficiency virus as predictors of AIDS among seropositive homosexual men.* Journal Infect Dis volume = 158, pages 615-622, 1988.

[26] R. A. Coutinho and P. N. Lelie and P. Albrecht-van Lent and E. E. Reerink-Brongers and L. Stoutjesdijk and P. Dees and J. Nivard and J. Huisman and H. W. Reesink. *Efficacy of heat inactivated hepatitis B vaccine in male homosexuals: outcome of a placebo controlled double blind trial.* Br Med J, volume 286, pages 1305-1308, 1983.

[27] G.J.P.van Griensven and E.M.M. de Vroome and J. Goudsmit and R.A. Coutinho. *Changes in sexual behaviour and the fall in incidence of HIV infection among homosexual men.* Br Med J, volume = 298, pages 218-221, year 1989.

[28] Wang, M.-C. and Jewel, N.P. and Tsai, W.-Y. *Asymptotic properties of the product limit estimator under random truncation.* Ann. Statist., volume 14, pages = 1597-1605, year 1986.

Appendix A

Density Estimation

With the estimate F of the distribution function at hand, we can derive an estimate of the density f corresponding to F. This density is assumed to be absolute continuous. However the NPMLE of F is usually a step function corresponding to a purely discrete probability distribution. Hence, some kind of smoothing is needed.

A frequently used estimator is the so called kernel estimator (see [13]). An estimate of the density f is given by:

$$\hat{f}(x) \;=\; h^{-1} \int K_x((a-t)/h)\; d\hat{F}(t)$$

where K_x is a so-called boundary kernel and h the bandwidth. For the kernel K_x, there are several forms: normal, triangle, triweight, etc. The choice of the kernel is not really significant for the estimate $\hat{f}$, the estimate is mainly determined by the choice of the bandwidth h. Usually h is selected via a bootstrap or cross-validation method.

Once we have an estimate $\hat{f}$ of the density we can use it for smoothing the NPLME of the distribution function and estimation of the hazard rate. A smoothed version $\hat{F}_s$ of $\hat{F}$ can be obtained by integrating the estimated density $\hat{f}$:

$$\hat{F}_s(t) \;=\; \int_0^t \hat{f}(t)\; dt$$

and the hazard rate $\lambda(t)$ is estimated by

$$\hat{\lambda}(t) \;=\; \frac{\hat{f}(t)}{1-\hat{F}_s(t)}\;.$$

Example
Figure A.1 shows an example of an NPMLE and a smoothed version based on a bandwidth of one year, and figure A.2 shows the effect of three different bandwidths on the density estimation. A smaller bandwidth leads to a more peaked version of the estimate.

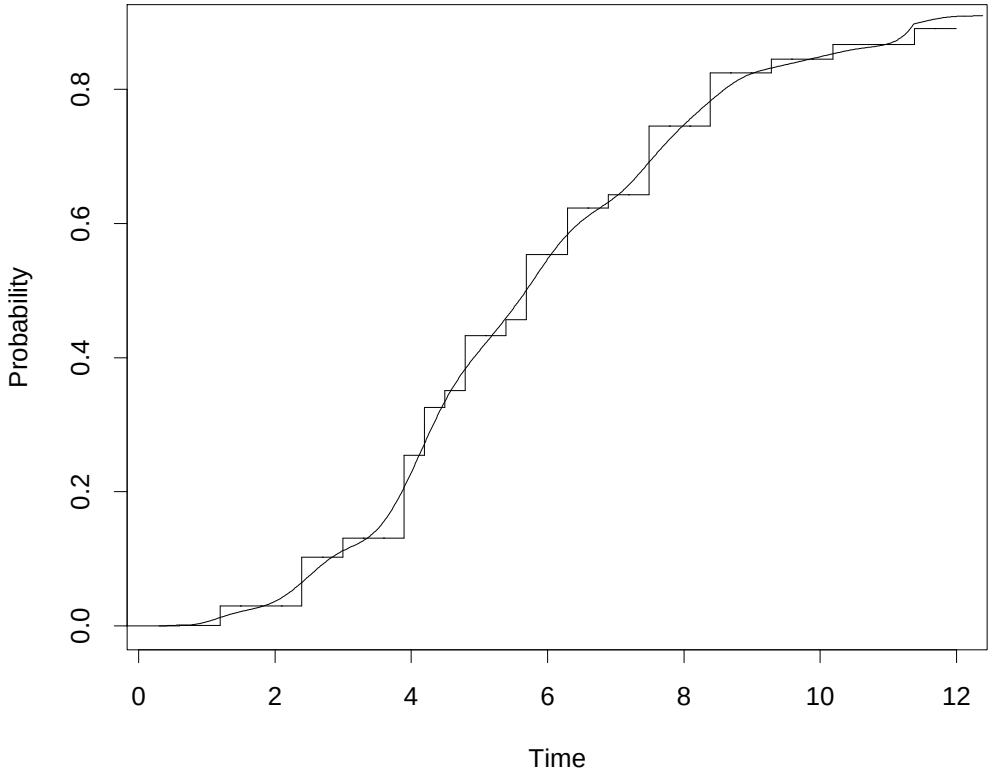


Figure A.1: NPMLE and smoothed version of an illustration data set

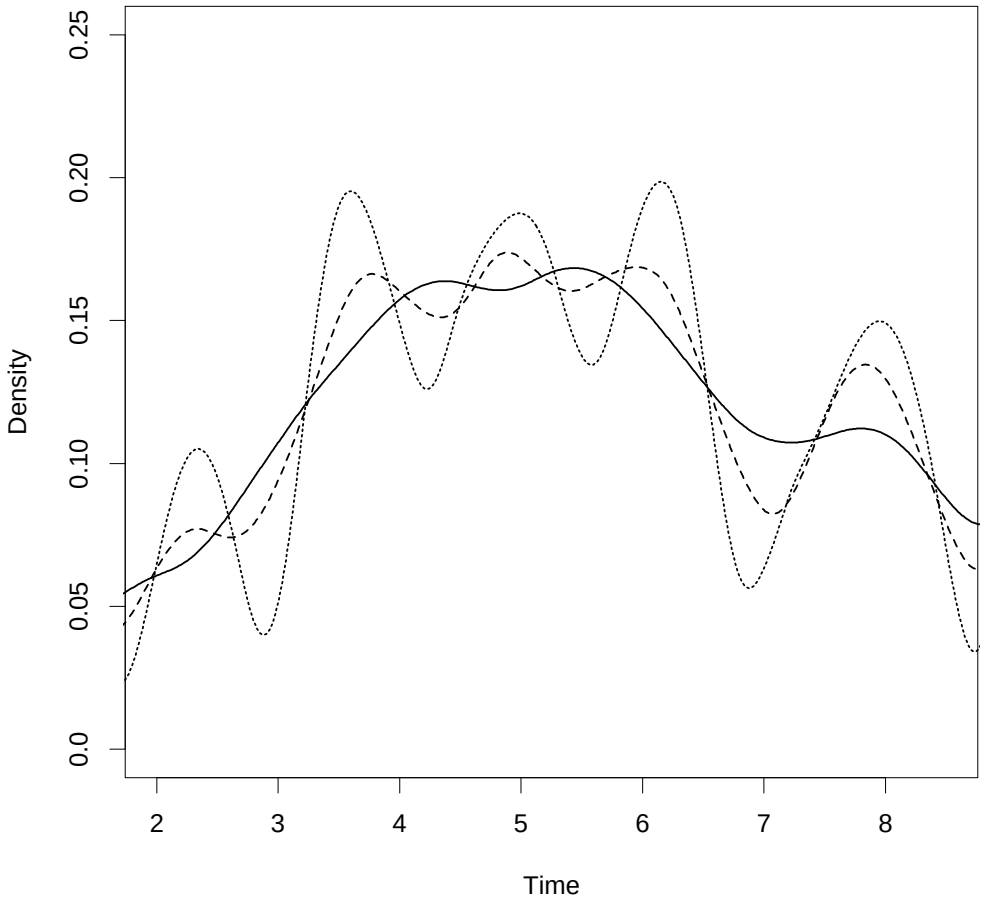


Figure A.2: Density estimation with three different bandwidths

Appendix B

Expected versus randomly drawn seroconversion dates

In section 4.2.2 we used two methods to impute a date of seroconversion. First, for case i we calculated the expected seroconversion date based on the individual seroconversion distribution and the seroconversion interval of case i. Second, for case i we randomly drew a seroconversion date from his seroconversion distribution. To see the difference of the two methods on the estimate of the incubation time distribution we perform a simulation study with 1000 cases.

- Simulate 1000 seroconversion dates $x_1, \ldots, x_{1000}$ according to the following distribution $X = \sqrt{5U}$, where U is uniformly distributed on the interval $(0, 5]$. The expected seroconversion date $\mathbb{E}(X) = 10/3$. The empirical distribution function of X is plotted in figure B.1.

- Simulate 1000 incubation times $v_1, \ldots, v_{1000}$ according to a Weibull(2,6) distribution. Let the unobserved moments of diagnoses be defined by $y_i = x_i + v_i$.

- Construct the observed incubation times by $t_i^1 = y_i - \mathbb{E}(X)$ (expected seroconversion dates) and $t_i^2 = y_i - x_i'$, where x_i' is a randomly drawn seroconversion date from the seroconversion distribution X. To deal with the fluctuation induced by randomly drawn seroconversion dates we constructed 5 sets of t_i^2.

- Figure B.2 shows the empirical distribution functions of $v_1, \ldots, v_{1000}$ the 'true' incubation times, $t_1^1, \ldots, t_{1000}^1$ the incubation times based on expected seroconversion dates and $t_1^2, \ldots, t_{1000}^2$ (5 times) the incubation times based on randomly drawn seroconversion dates.

We see that the two estimates based expected and randomly drawn seroconversion dates resemble the estimate of the 'true' incubation times. However, the estimates based on randomly drawn seroconversion times systematically overestimate the incubation time distribution at the begin and systematically underestimate the incubation time distribution at the end.

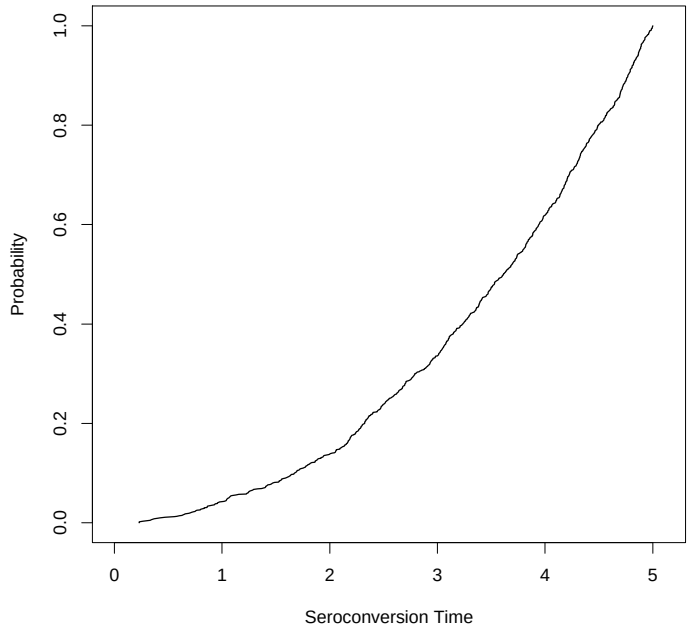


Figure B.1: Empirical distribution function of the seroconversion dates $x_1, \ldots, x_{1000}$

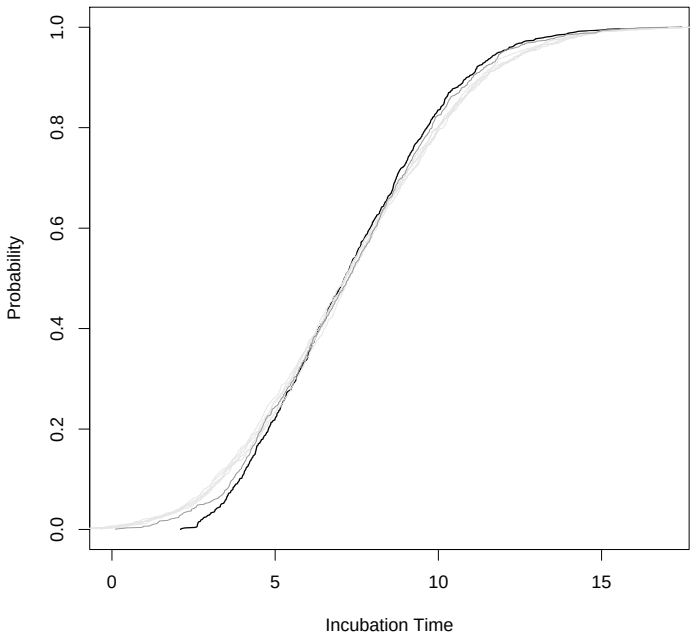


Figure B.2: Empirical distribution functions: v_i (black), t_i^1 (grey), t_i^2 (light grey)

Appendix C

Software for Double Censoring

Most standard software packages (SPSS, Splus, Stata, SAS) are well equipped with routines for ordinary survival analyses, *i.e.* Kaplan Meier and Cox proportional hazards model. However, none of these packages have routines that can deal with nonparametric estimation in the case of interval or doubly censorerd survival data. The methods described in this report have been implemented in C and put together in one software package. This package is still under construction, but beta releases can be obtained by sending an e-mail to Ronald Geskus at the Municipal Health Service in Amsterdam (rgeskus@gggd.amsterdam.nl). Once a data file has been read the incubation time distribution can be estimated nonparametrically via interval transformation or double censoring.

Parametric estimates of the incubation time distribution with double censoring could be obtained in a general way. Every software package equiped with a non-constrained optimization routine should be able to maximize the likelihood. However, the likelihood function in the double censoring case contains integrals, so the optimization routine must have a subroutine to deal with these integrals. Splus has such routines. However, evaluating many integrals in the optimization routine in Splus makes the estimation procedure very slow. The best thing to do is to program the estimation procedure in a language such as C, using integration and optimization procedures from Numerical Recipes [7]. This has also been done in the package.

Summary

Doubly censored data arise when both the time origin and event time can be censored. Due to periodic screening and limited follow-up, double censoring is likely to occur in cohort studies on HIV infection. Several methods have been discussed to analyse doubly censored data. All of these methods have their own advantages and disadvantages, depending on the specific structure of a data set.

The Kaplan Meier estimator described in chapter 2 is the simplest method. It just ignores the interval censored nature of the seroconversion dates by imputing a date of seroconversion. For data sets with small seroconversion intervals (seroconverter groups) this is probably the best nonparametric method, since it is easy to calculate, it has easy to calculate confidence intervals and the induced bias by imputing a date of seroconversion is small. For data sets with wider seroconversion intervals the interval censored nature of the data can be taken into account by using a nonparametric maximum likelihood estimator for interval censored data, as described in chapter 2.

With the double censoring method information from both the seroconversion time scale and the incubation time scale are used in the analysis. We looked at the piecewise uniform approach to analyse doubly censored data. Two fixed grids are chosen in both the seroconversion and the incubation time scale. On these grids we let the incubation time and seroconversion time distribution be piecewise uniform, *i.e.* the density between two gridpoints is constant. It turns out that the number of gridpoints has little effect on the estimate of the incubation time. However, one should not take the number of gridpoints too small to avoid trivialities and not too large to avoid non-uniqueness of the estimators and long calculation times. For cohorts with not too many wide intervals, there is not much difference in the location of the survival curves between the Kaplan Meier approach and double censoring approach (see table 3.1 and figures 3.6 and 3.8). However, the double censoring method reflects the uncertainty in the date of seroconversion better. This uncertainty is expressed by the fewer jumps in the survival curve.

Altough the double censoring method uses information from both time scales and can handle wide intervals, the seroprevalent group of the HOM-study is still problematic for two reasons. First, all the wide intervals are located at one position (1980-1985). Second, the structure of the incubation time distribution is very unfavourable, it is almost uniformly distributed from 0 to 15 years. This means that the information we can get from the incubation time scale is too little to say anything about the seroconversion pattern of the seroprevalent

group.

To include the seroprevalent cases we looked at two approaches in chapter 4. First we chose a seroconversion structure for the prevalent group. For the seroprevalent group an exponential form of the seroconversion distribution is assumed to model a lower probability of seroconversion in the early period of the epidemic. See figure 4.1 and table 4.1. Second, we used additional CD4 data to reveal the seroconversion structure of the seroprevalent group. In this approach every individual seroprevalent case will receive its own seroconversion distribution based on its CD4 concentration at study entry. Given these individual seroconversion patterns we can calculate the expected dates of seroconversion for the prevalent cases and use the Kaplan Meier estimator to estimate the incubation time. An alternative approach is to use double censoring based on the individual seroconversion curves. There is not much difference in the survival curves of the two approaches, see figure 4.10. If we looked at the survival curves of the prevalent cases and the seroconverters separately, we saw some difference between the two groups. The seroprevalent cases have a better survival, see figure 4.11. However, it turnrd out that this difference is not significant.

Chapter 5 discusses the use of parametric models. It appeared that in the HOM-study there was some evidence of a log normal parametric form for the incubation time distribution, but no evidence of a Weibull or lognormal form for the seroconversion time distribution. However if we only looked at the seroconverter group then a Weibull model for the seroconversion time and incubation time distribution resulted in a reasonable good fit, see figures 5.3 and 5.4. It is not unusual that the choice of a parametric form depends on the specific data set. So despite the easy to calculate estimators and confidence intervals one has to be careful in using parametric models for double censoring.