
The Kaplan-Meier Integral in the Presence of Covariates: A Review

2

Thomas A. Gerds, Jan Beyersmann, Liis Starkopf, Sandra Frank, Mark J. van der Laan and Martin Schumacher

2.1 Introduction

In survival analysis with covariates, many parameters of interest are special cases of the integral:

$$\theta(\varphi) = \int_{\mathbb{R}^p} \int_0^\infty \varphi(t, z) F(dt | z) H(dz). \quad (2.1)$$

Here, T is the time of an event and Z a p -dimensional vector of covariates, φ a square integrable function, and $F(t | z) = P(T \leq t | Z = z)$ and $H(dz) = P(Z \in dz)$ denote the conditional survival distribution and the marginal law of Z , respectively. For example, $\theta(I\{t > t^*\})$ is the marginal survival probability at time t^* , $\theta(I\{t > t^*, z_1 > z_1^*\})$ the bivariate distribution at (t^*, z_1^*) (Akritas 1994), and $\theta([I\{t > t^*\} - m(t^*|z)]^2)$ the expected Brier score of a regression model m which predicts survival at time t^* conditional on the covariates (Graf et al. 1999). In the absence of covariates, using the integrand $\varphi_s(t) = \exp(st)$ in (2.1) has been used for expressing the moment generating function of multi-state survival times (Hudson et al. 2014).

T.A. Gerds (✉) · L. Starkopf

Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark
e-mail: tag@biostat.ku.dk

J. Beyersmann · S. Frank

Institute of Statistics, University of Ulm, Ulm, Germany

M.J. van der Laan

Division of Biostatistics, School of Public Health, University of California, Berkeley, USA

M. Schumacher

Institute for Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany

© Springer International Publishing AG 2017

D. Feger et al. (eds.), *From Statistics to Mathematical Finance*,
DOI 10.1007/978-3-319-50986-0_2

In a remarkable series, Stute (1993, 1996, 1999) analyzed an estimator of (2.1) for right censored observations of the survival time. The estimator is called the Kaplan-Meier integral. In this paper we first show that Stute's estimator can be written as an inverse of the probability of censoring weighted (IPCW) estimator (Van der Laan and Robins 2003) and then review the structural assumptions of the estimation problem and the asymptotic properties of the estimator.

In biostatistics, Stute's method has recently been put to prominent use for estimating transition probabilities in non-Markov illness-death models (e.g., Meira-Machado et al. 2006; Andersen and Perme 2008; Allignol et al. 2014; de Uña-Álvarez and Meira-Machado 2015). For instance in oncology, illness-death models are used to jointly model progression-free survival and overall survival, and Kaplan-Meier integrals apply interpreting progression-free survival as the covariate and overall survival as time-to-event. We illustrate the general program of the present paper in this example. Using IPCW representations, we obtain simplified estimators that even allow for left-truncated data. Left-truncation is another common phenomenon in survival analysis describing a situation of delayed study entry where individuals are included in prospective cohorts after time origin, conditional on still being alive (Keiding 1992).

2.2 The Kaplan-Meier Integral

Let C be a positive random variable (the censoring time) and suppose that instead of (T, Z) one observes $X = (\tilde{T}, \Delta, Z)$ where $\tilde{T} = \min(T, C)$ and $\Delta = I\{T \leq C\}$. Stute's estimate of (2.1) is defined on a set of n *iid* right censored observations X_1, \dots, X_n . Let $\tilde{T}_{1:n} \leq \dots \leq \tilde{T}_{n:n}$ denote the ordered values of $\tilde{T}_1, \dots, \tilde{T}_n$, and $(\delta_{i:n}, Z_{i:n})$ the concomitant status and covariate values. Stute (1993) introduced the estimate

$$\hat{\theta}(\varphi) = \sum_{i=1}^n W_{in} \varphi(\tilde{T}_{i:n}, Z_{i:n}) \quad (2.2)$$

where

$$W_{in} = \frac{\delta_{i:n}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{j:n}}.$$

The weights W_{in} do not only match the initials of their inventor's first name, they are also equal to the jump sizes of the Kaplan-Meier estimator for the marginal survival function of T_i and thereby justify the name "Kaplan-Meier integral".

Lemma 1 Assume that there are no tied event times, i.e., $\tilde{T}_{i:n} < \tilde{T}_{(i+1):n}$, $i = 1, \dots, n-1$. The product limit forms of the Kaplan-Meier estimators of the marginal survival time distribution $S(t) = P(T > t)$ and the marginal censoring time distribution $G(t) = P(C > t)$ are given by

$$\hat{S}_0(t) = \prod_{i: \tilde{T}_{i:n} \leq t} \left\{ 1 - \frac{\delta_{i:n}}{n-i+1} \right\} \quad \hat{G}_0(t) = \prod_{i: \tilde{T}_{i:n} \leq t} \left\{ 1 - \frac{(1-\delta_{i:n})}{n-i+1} \right\}.$$

The corresponding IPCW sum forms are:

$$\begin{aligned} \hat{S}_0(t) \hat{G}_0(t) &= \frac{1}{n} \sum_{i=1}^n I\{\tilde{T}_{i:n} > t\} \\ \hat{S}_0(t) &= 1 - \frac{1}{n} \sum_{i=1}^n \frac{I\{\tilde{T}_{i:n} \leq t\} \delta_{i:n}}{\hat{G}_0(T_{i:n})}, \end{aligned}$$

and

$$\hat{G}_0(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{I\{\tilde{T}_{i:n} \leq t\} (1 - \delta_{i:n})}{\hat{S}_0(T_{i:n})}.$$

Proof These relations were readily noted by Gill (1980, page 36) in slightly more general form, that is allowing for tied times.

Lemma 2 Under the assumption of Lemma 1 the weights of the Kaplan-Meier integral equal the jump size of the Kaplan-Meier estimator:

$$W_{i:n} = \hat{S}_0(T_{(i-1):n}) - \hat{S}_0(T_{i:n})$$

The Kaplan-Meier integral has the following IPCW representation:

$$\hat{\theta}(\varphi) = \frac{1}{n} \sum_{i=1}^n \frac{\varphi(T_{i:n}, Z_{i:n}) \delta_{i:n}}{\hat{G}_0(T_{i:n})}.$$

Proof It follows from Lemma 1 that

$$\hat{S}_0(T_{(i-1):n}) - \hat{S}_0(T_{i:n}) = -\frac{1}{n} \sum_{j=1}^{i-1} \frac{\delta_{j:n}}{\hat{G}_0(T_{j:n})} + \frac{1}{n} \sum_{j=1}^i \frac{\delta_{j:n}}{\hat{G}_0(T_{j:n})} = \frac{1}{n} \frac{\delta_{i:n}}{\hat{G}_0(\tilde{T}_{i:n})}.$$

The claim follows since

$$\begin{aligned} nW_{i:n} &= \frac{n \delta_{i:n}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{j:n}} = \frac{n \delta_{i:n}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{j:n}}{n-j+1} \right) \\ &= \delta_{i:n} \frac{n}{n-i+1} \hat{S}_0(\tilde{T}_{(i-1):n}) = \frac{\delta_{i:n}}{\hat{G}_0(\tilde{T}_{i:n})}. \end{aligned}$$

Interestingly, Lemma 2 shows that the IPCW sum form of the Kaplan-Meier estimator (Satten and Datta 2001) is the special case of the Kaplan-Meier integral where $\varphi(t, z) = \hat{\theta}(I\{t > t^*\})$ Akritas (2000).

2.3 Identifiability and Structural Assumptions

2.3.1 Support

In biomedical applications of survival analysis, due to limited follow up times, the support of the censoring times is usually strictly smaller than the support of the survival times. This means that inference on the tail of the survival distribution is not feasible and to identify the parameter in (2.1) based on the right censored observations we have to truncate the parameter at some point in time. To formalize all this let $\tau_0 = \inf_s P(C > s) = 0$ and $\tau_1 = \inf_s P(T > s) = 0$ denote the limits of the supports of C and T , respectively. To meet the setting of typical biomedical applications of survival analysis, we assume $\tau_0 < \tau_1$, and to achieve identifiability we assume that φ satisfies the following condition for some function φ^* of the covariates only and $\epsilon > 0$:

$$\varphi(t, z) = \varphi(t, z)I\{t \leq \tau_0 - \epsilon\} + \varphi^*(z)I\{t > \tau_0 - \epsilon\}. \quad (\text{A1})$$

For example, the mean restricted lifetime is defined as $\theta(tI\{t > t^*\})$ for a suitably chosen truncation time t^* . We refer to Stute (1993, 1996) for a rigorous discussion of the borderline cases where $\epsilon \rightarrow 0$.

2.3.2 Independence

Assumption (A1) is not sufficient to achieve identifiability and a further assumption is needed regarding the independence of the censoring mechanism (Tsiatis 1975; Grüger et al. 1991; Gill et al. 1995). To discuss the different assumptions that lead to identifiability we introduce the function G whose values are the conditional probabilities that an observation is uncensored given the event time and the covariates:

$$P(\Delta = 1 \mid Z = z, T = t) = P(C > t \mid Z = z, T = t) = G(t, z). \quad (2.3)$$

Even without further independence assumptions, the density of a right censored observation X (with respect to an appropriately chosen dominating measure) can be decomposed as

$$\begin{aligned} P(\tilde{T} \in dt, \Delta = \delta, Z \in dz) &= \{P(\Delta = 1 \mid Z = z, T = t) P(T \in dt, Z \in dz)\}^\delta \\ &\quad + \{P(\Delta = 0 \mid Z = z, C = t) P(C \in dt, Z \in dz)\}^{(1-\delta)}. \end{aligned}$$

The first term can be expressed as

$$\begin{aligned} P(\tilde{T} \in dt, \Delta = 1, Z \in dz) &= P(\Delta = 1 \mid Z = z, T = t) P(T \in dt, Z \in dz) \\ &= G(t, z) F(dt \mid z) H(dz) = P^{(1)}(dt, dz) \end{aligned}$$

and this relation motivates the general form IPCW estimation equations for θ :

$$\theta_\varphi(F, H) = \int \varphi(t, z) F(dt \mid z) H(dz) = \int \varphi(t, z) \frac{P^{(1)}(dt, dz)}{G(t, z)} = \nu_\varphi(P^{(1)}, G). \quad (2.4)$$

Since $P^{(1)}$ only depends on the right censored observations it can be estimated non-parametrically, i.e., by the empirical law of the uncensored observations $\hat{P}_n^{(1)}(A, B) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}\{\tilde{T}_i \in A, \Delta_i = 1, Z_i \in B\}$. The general form of the IPCW estimate of θ is then obtained by also substituting an estimate \hat{G} for G :

$$\hat{\theta}_n(\varphi) = \hat{\nu}_n(\varphi; \hat{P}_n^{(1)}, \hat{G}) = \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \varphi(T_i, Z_i)}{\hat{G}(T_i, Z_i)}.$$

To justify the IPCW estimate defined in Lemma 2 above, Stute (1993, 1996) restricted the model for G by assuming

$$T \text{ and } C \text{ are independent}, \quad (\text{A2})$$

$$P(T \leq C \mid T, Z) = P(T \leq C \mid T). \quad (\text{A3})$$

These two conditions together imply

$$G(t, z) = P(C > t \mid T = t, Z = z) \stackrel{\text{A3}}{=} P(C > t \mid T = t) \stackrel{\text{A2}}{=} P(C > t). \quad (2.5)$$

Alternatively, we may assume

$$T \text{ and } C \text{ are conditionally independent given } Z \quad (\text{A4})$$

which is familiar from the Cox regression model (compare Begun et al. 1983, page 448). Under (A4) we have

$$G(t, z) = P(C > t \mid Z = z). \quad (2.6)$$

Comparing (2.5) and (2.6) shows that under (A2) and (A3) the function G is a simpler parameter, because it does not depend on the covariates. Note also that neither (A2) implies (A4) nor (A4) implies (A2), and that in generality both assumptions permit that the censoring times depend on the covariates. However, we emphasize that under (A2) and (A3) the function G does not depend on the covariates and hence the conditional censoring distribution may depend on the covariates only in regions of the underlying probability space that are irrelevant for estimation of θ_φ .

Under (A2) and (A3) the function $G(t) = P(C > t)$ equals the marginal survival function of the censoring times and can be estimated consistently by the marginal reverse Kaplan-Meier estimator for the survival function of the censoring times as defined in Lemma 1. Under (A4) we need to estimate the conditional censoring distribution. Only when all covariates are discrete variables this can be done without further modelling assumptions.

2.4 Large Sample Properties of the Kaplan-Meier Integral

Lemma 2 shows that the plug-in IPCW estimator $\hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_0)$ equals Stute's Kaplan-Meier integral (2.2). Stute (1993, 1996) proves strong consistency and weak convergence of $\hat{\theta}(\varphi) = \hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_0)$ and obtains the following *iid* representation (translated to our notation)

$$\sqrt{n}(\hat{\theta}(\varphi) - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{IC}_{\hat{\theta}(\varphi)}(\tilde{T}_i, \Delta_i, Z_i) + o_P(1)$$

where the influence function $\text{IC}_{\hat{\theta}(\varphi)}$ of the Kaplan-Meier integral is given in the following theorem.

Theorem 1 *Under (A1), (A2) and (A3) the Kaplan-Meier integral*

$$\hat{\theta}(\varphi) = \hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_0)$$

is consistent and regular, asymptotically Gaussian linear with influence function

$$\begin{aligned} \text{IC}_{\hat{\theta}(\varphi)}(\tilde{T}_i, \Delta_i, Z_i) &= \Delta_i \frac{\varphi(\tilde{T}_i, Z_i)}{G(\tilde{T}_i)} + \frac{(1 - \Delta_i)}{W(\tilde{T}_i)} \int_{\tilde{T}_i}^{\infty} \varphi(s, z) F(ds | z) H(dz) \\ &\quad - \int \left\{ \int_0^{\tilde{T}_i \wedge s} \frac{G(du)}{W(u)G(u-)} \right\} \varphi(s, z) F(ds | z) H(dz) - \theta(\varphi) \end{aligned} \quad (2.7)$$

where $W(t) = P(\tilde{T}_i > t)$.

Proof See Stute (1993, 1996). An alternative proof can be obtained by applying the functional delta method (e.g. Van der Vaart 1998, Theorem 20.8) to the Hadamard differentiable functional $\nu(G, P^{(1)})$ using that both the Kaplan-Meier estimator for the censored times \hat{G}_0 and the empirical distribution function $\hat{P}_n^{(1)}$ are \sqrt{n} -consistent in appropriately normed spaces of distributions (Reeds 1976; Van der Vaart 1991).

2.5 Bias and Efficiency

The Kaplan-Meier integral can have a large sample bias and it is not efficient even not when assumptions (A2) and (A3) are satisfied. The bias can be seen when the conditional survival distribution of the censoring times depend on the covariates $P(C > t | Z = z) \neq P(C > t)$. In this case the marginal Kaplan-Meier estimator for the censored times $\widehat{G}_0(t)$ converges in probability to $\widetilde{G}(t) = \int_{\mathbb{R}^p} G(t | z) H(dz)$ and the large sample bias of $\hat{\theta}(\varphi)$ is given by the following limit as $n \rightarrow \infty$:

$$\left| \hat{\theta}(\varphi) - \theta(\varphi) \right| \rightarrow \left| \int \varphi(t, z) \left\{ \frac{1}{\widetilde{G}(t)} - \frac{1}{G(t, z)} \right\} P^{(1)}(dt, dz) \right|.$$

Rotnitzky and Robins (1995) were the first to observe that the Kaplan-Meier integral is not efficient even not when it is consistent and the survival distribution of the censored times does not depend on the covariates.

The following is a special case of Van der Laan and Robins (2003, Theorem 1.1 and Example 1.12), see also Gerds (2002).

Proposition 1 *The efficient influence function for estimation of θ based on the right censored data $(\tilde{T}_i, \Delta_i, Z_i)$ is given by*

$$\begin{aligned} \text{IC}^{\text{eff}}(\tilde{T}_i, \Delta_i, Z_i) = & \Delta_i \frac{\varphi(\tilde{T}_i, Z_i)}{G(\tilde{T}_i | Z_i)} + \frac{(1 - \Delta_i)}{\widetilde{W}(\tilde{T}_i | Z_i)} \int_{\tilde{T}_i}^{\infty} \varphi(s, Z_i) F(ds | Z_i) \\ & - \int_0^{\tilde{T}_i \wedge s} \frac{G(ds | z)}{\widetilde{W}(s | z) G(s - | z)} \varphi(s, Z_i) F(ds | Z_i) - \theta(\varphi) \quad (2.8) \end{aligned}$$

where $\widetilde{W}(t | z) = P(\tilde{T} > t | Z = z)$.

A regular, asymptotically linear estimator is asymptotically efficient if and only if the influence function of the estimator equals the efficient influence function for the estimation problem. Hence, comparing (2.8) with (2.7) shows that $\hat{\theta}(\varphi)$ is inefficient except for the case where $G(t, z) = G(t, z')$ and $F(t, z) = F(t, z')$ for all z, z' , i.e. where the covariates are independent of both survival and censoring times (Malani 1995). At first glance, the inefficiency of the Kaplan-Meier integral may appear counter-intuitive as it is not so obvious where the information is lost. A closer look however reveals that the covariate values corresponding to the right censored observations do not enter the statistic (2.2). But, there is information in the fact that no event happened until the end of followup (right censored). This information can be recovered by a model for the conditional survival function of the censored times given the covariates. For example, a standard Cox regression model fitted to the censored times yields

$$\widehat{G}_1(t, z) = \exp \left\{ - \int_0^t \exp(\widehat{\beta} z) \widehat{\Lambda}_0(ds) \right\}$$

where $\widehat{\beta}$ and $\widehat{\Lambda}_0$ are the partial likelihood estimates of the regression coefficients and the Breslow estimate of the cumulative baseline hazard function, respectively. The corresponding plug-in IPWC estimator $\widehat{\nu}_n(\widehat{\mathbf{P}}_n^{(1)}, \widehat{G}_1)$ is more efficient than the IPCW estimator using Kaplan-Meier for the censoring, but it is still inefficient. The influence curve for this estimator equals $(\Delta_i \varphi)/G - \theta(\varphi)$ minus its projection on the tangent space of the scores of the censoring model, as shown in Van der Laan and Robins (2003, Sect. 2.3.7). The principle of adaptive estimation (Bickel et al. 1993) in this situation can be expressed as follows: The bigger the censoring model the more efficient the IPCW estimator. In particular, if one has available a consistent estimator in a saturated model for G , then the correspondingly defined IPCW estimator is fully efficient. Similarly, it is known that in general the traditional survival rank test needs the whole nonparametric model for its efficiency (Neuhaus 2000). But if the covariates are continuous or high dimensional such estimators perform not very nicely in small samples due to the curse of dimensionality. A practical solution is given by doubly robust estimators which rely on models for both G and F and are locally efficient if both models are correctly specified. If either the model for G or the model for F is correctly specified then the estimator is consistent and asymptotically linear.

2.6 Empirical Results

This section illustrates the magnitude of the potential bias and efficiency loss in the special case $\theta(I\{t > t^*\})$, i.e., where the parameter is the marginal survival function at t^* . Note that in this case the Kaplan-Meier integral (with \widehat{G}_0) equals the ordinary Kaplan-Meier estimate. See (e.g. Gerds and Schumacher 2006) for a similar simulation study of IPCW estimators of a more complex parameter. We consider two simulation scenarios. For both settings, a binary covariate is drawn from the binomial distribution with $P(X = 1) = 0.5$. The survival and censoring times were generated using parametric Cox proportional hazard models $\lambda_0^T \exp(1.5Z)$ and $\lambda_0^C \exp(\gamma Z)$, respectively, as described in Bender et al. (2005). In the first setting we set $\gamma = 1.2$ so that the censoring time distribution depends on the covariate. In the second setting we set $\gamma = 0$ so that only the survival times depend on the covariate. In both settings the baseline hazards λ_0^T and λ_0^C were chosen so that $S(t = 70) = 62\%$ and $P(C \leq 70, T > C) = 60\%$. We contrast estimates of the parameter $\theta(I\{t > 70\})$ obtained with the Kaplan-Meier estimate $\widehat{\nu}_n(\widehat{\mathbf{P}}_n^{(1)}, \widehat{G}_0)$ and with the IPCW estimate $\widehat{\nu}_n(\widehat{\mathbf{P}}_n^{(1)}, \widehat{G}_2)$ where

$$\widehat{G}_2(t, z) = \prod_{i: Z_i=z, \tilde{T}_{i:n} \leq t} \left\{ 1 - \frac{(1 - \delta_{i:n})}{n - i + 1} \right\}$$

Table 2.1 Summary of simulation study for estimating $P(T > 70) = 62\%$ based on right censored data where $P(C \leq 70, T > C) = 60\%$. In setting 1 both the survival time distribution and the censoring time distribution depend on a binary covariate. In setting 2 only the survival time distribution depends on a binary covariate

Setting	Estimate	Bias (%)	Variance (%)	MSE (%)
Censoring dependent on covariate	$\hat{\nu}_n(\hat{\mathbf{P}}_n^{(1)}, \hat{G}_0)$	5.0118	0.309	0.560
	$\hat{\nu}_n(\hat{\mathbf{P}}_n^{(1)}, \hat{G}_2)$	-0.0366	0.266	0.266
Censoring independent of covariate	$\hat{\nu}_n(\hat{\mathbf{P}}_n^{(1)}, \hat{G}_0)$	-0.0228	0.286	0.286
	$\hat{\nu}_n(\hat{\mathbf{P}}_n^{(1)}, \hat{G}_2)$	-0.0369	0.261	0.261

is the stratified Kaplan-Meier estimate for the censored times conditional on the strata defined by $Z = z$. We report averaged small sample bias and mean squared errors across 2000 simulated data sets.

Table 2.1 shows the results for sample size 200. In the first setting there is a large bias in the marginal Kaplan-Meier estimate whereas $\hat{\nu}_n(\hat{\mathbf{P}}_n^{(1)}, \hat{G}_2)$ is less biased. In addition, the variance of the marginal Kaplan-Meier estimate is bigger. In the second setting the marginal Kaplan-Meier IPCW estimate is no longer biased. The same holds for the stratified Kaplan-Meier IPCW estimate. However, the marginal Kaplan-Meier IPCW estimate still has a larger variance than the stratified Kaplan-Meier IPCW estimate (see Table 2.1).

Figure 2.1 illustrates the difference between the estimators as a function of the sample size. We see that the MSE can be large and this is due to a large bias as can be seen from the data in Table 2.1. The left panel of the figure indicates that the difference in MSE decreases with increasing sample size. However, Fig. 2.2 reveals that the relative advantage of the stratified Kaplan-Meier IPCW estimate does not decrease with increasing sample size. The figure also shows that the magnitude of the relative gain in MSE depends on the predictiveness of the covariate and on the amount of censoring.

2.7 Non-Markov Illness-Death Model Without Recovery

The illness-death model without recovery has important biostatistical applications, for example in oncology. In this section we make the connection with the Kaplan-Meier integral. We therefore consider a stochastic process $(X_t)_{t \in [0, \infty)}$ which has state space $\{0, 1, 2\}$, right-continuous sample paths, initial state 0, $P(X_0 = 0) = 1$, intermediate state 1 and absorbing state 2. This process describes an illness-death model without recovery when also the probability of a recovery event is zero, i.e.,

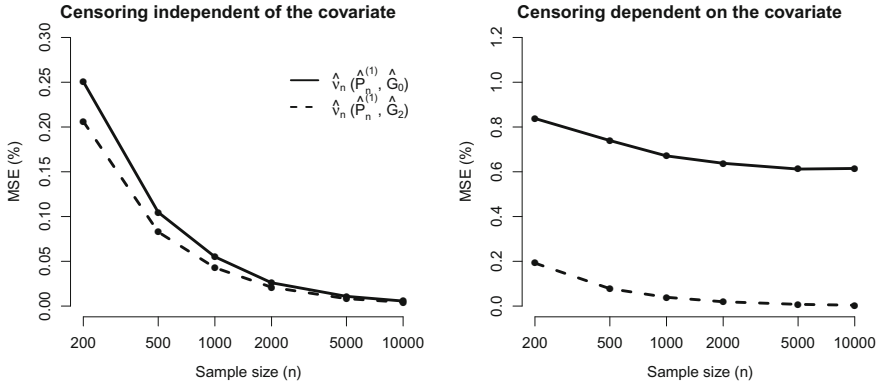


Fig. 2.1 Estimation of $S(70)$. Mean squared error as a function of sample size for Kaplan-Meier integral ($\hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_0)$, solid line) and IPCW estimator based on stratified Kaplan-Meier for censoring time distribution ($\hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_2)$, dashed line). In the left panel the log-hazard ratio of a binary covariate on survival is 3 and on censoring is 0. In the right panel the log-hazard ratio of a binary covariate on survival is 3 and on censoring is 1.2. Data show averages across 2000 simulation runs for each sample size

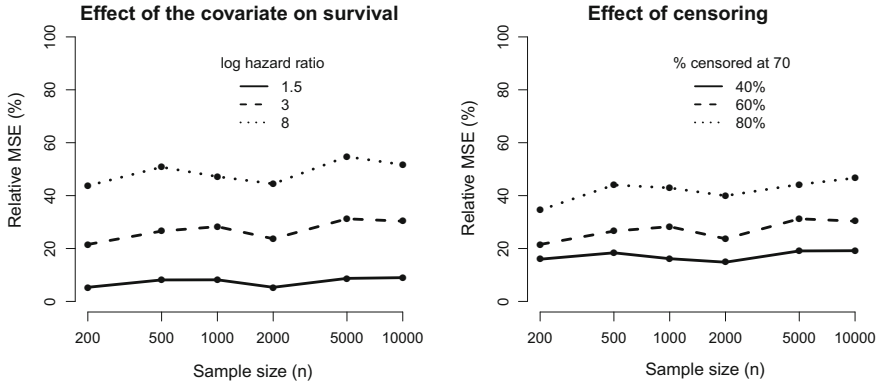


Fig. 2.2 Estimation of $S(70)$. Ratio of mean squared error for Kaplan-Meier integral ($\hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_0)$) and IPCW estimator based on stratified Kaplan-Meier for censoring time distribution ($\hat{\nu}_n(\hat{P}_n^{(1)}, \hat{G}_2)$). In both panels the log-hazard ratio of a binary covariate on censoring hazard is 0. In the left panel the effect of the binary covariate on survival hazard is varied and in the right panel the percentage censored is varied. Data show averages across 2000 simulation runs for each sample size

when $P(X(t) = 0 | X(s) = 1) = 0$ for all $s \leq t$. The process can equivalently be described by a pair of random variables

$$T_0 = \inf\{t : X_t \neq 0\} \text{ and } T = \inf\{t : X_t = 2\}$$

so that T_0 is the waiting time in the initial state, $X_{T_0} \in \{1, 2\}$, and T the time until the absorbing state is reached. The process passes through the intermediate state 1, if and only if $T_0 < T$, and $T_0 = T$ if the process does not pass through the intermediate state. Our aim is to estimate the transition probabilities between state $l \in \{0, 1\}$ and state $j \in \{1, 2\}$

$$P_{lj}(s, t) = P(X_t = j | X_s = l) \quad (2.9)$$

for pairs of time points (s, t) that satisfy $s \leq t$.

Based on right censored data of the illness-death process Meira-Machado et al. (2006) derive an estimator for (2.9) starting with the following representations:

$$\begin{aligned} P_{01}(s, t) &= \frac{P(s < T_0 \leq t, t < T)}{P(T_0 > s)}, \\ P_{11}(s, t) &= \frac{P(T_0 \leq s, t < T)}{P(T > s) - P(T_0 > s)}. \end{aligned} \quad (2.10)$$

The challenge in estimating the right hand sides in (2.10) stems from the numerators, while straightforward Kaplan-Meier estimation applies to estimating $P(T_0 > s)$ and $P(T > s)$. For the numerators, Meira-Machado et al. (2006) apply Stute's Kaplan-Meier integral with 'covariate' $Z = T_0$. Allignol et al. (2014) showed that the estimator of Meira-Machado et al. (2006) can alternatively be derived from a suitably defined competing risks process and they also obtain an IPCW representation of the estimator of Meira-Machado et al. (2006) for $P_{01}(s, t)$ in a similar fashion as we have for the Kaplan-Meier integral in Sect. 2.2. In bivariate (T_0, T) -time several IPCW estimators are available, and Allignol et al. (2014) also discuss an IPCW estimator which uses the estimate of the survival function of the censored times suggested by Tsai and Crowley (1998). This results in a simplified estimator which could easily be extended to left-truncated data. Unfortunately, the Tsai and Crowley (1998) approach is not applicable for estimating $P_{11}(s, t)$.

In what follows we discuss the Kaplan-Meier-integral based estimator of $P_{11}(s, t)$ from the IPCW-perspective. For this we express $P(T_0 \leq s, t < T)$ as a special case of (2.1):

$$P(T_0 \leq s, t < T) = \int I(z \leq s, y > s) P^{T_0, T}(dz, dy).$$

For estimation we assume i.i.d. replications $(\tilde{T}_{0i}, \tilde{T}_i, \Delta_i)$, $i = 1, \dots, n$, where $\tilde{T}_{0i} = \min(T_{0i}, C_i)$, $\tilde{T}_i = \min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$. It is convenient to introduce counting processes

$$\begin{aligned} N(u) &= \sum_{i=1}^n I\{i : \tilde{T}_i \leq u, \Delta_i = 1\}, \\ N^*(u) &= \sum_{i=1}^n I\{i : \tilde{T}_i \leq u, \Delta_i = 1, T_{0i} \leq s, T_i > t\}, \\ Y(u) &= \sum_{i=1}^n I\{i : \tilde{T}_i \geq u\}. \end{aligned}$$

Straightforward algebra shows that the estimator of Meira-Machado et al. (2006) for $P(T_0 \leq s, t < T)$ equals

$$\sum_u \prod_v \left(1 - \frac{\Delta N(v)}{Y(v)}\right) \frac{\Delta N^*(u)}{Y(u)}, \quad (2.11)$$

where both the sum and the product in (2.11) are over all observed unique times to the absorbing state and ΔN and ΔN^* denote the increments of the counting processes. Since $\prod_v (1 - \frac{\Delta N(v)}{Y(v)})$ is a standard Kaplan-Meier estimator, the IPCW-representations discussed earlier give rise to different possible IPCW-variants of (2.11),

$$\frac{1}{n} \sum_u \left(\hat{P}_a(C \geq u) \right)^{-1} \Delta N^*(u),$$

where $\hat{P}_a(C \geq \cdot)$ is some consistent estimator of the censoring survival function. Recall that in bivariate time there are several possible Kaplan-Meier-type estimators of $P(C \geq \cdot)$, simple choices only using either $\{\tilde{T}_{0i} : T_{0i} > C_i, i = 1, \dots, n\}$ or $\{\tilde{T}_i : T_i > C_i, i = 1, \dots, n\}$. Using representation (2.10), we may estimate $P_{11}(s, t)$ by

$$\frac{\frac{1}{n} \sum_u \left(\hat{P}_a(C \geq u) \right)^{-1} \Delta N^*(u)}{\frac{|\{i : \tilde{T}_i > s\}|}{n \hat{P}_b(C > s)} - \frac{|\{i : \tilde{T}_{0i} > s\}|}{n \hat{P}_c(C > s)}}, \quad (2.12)$$

where $\hat{P}_b(C \geq \cdot)$ and $\hat{P}_c(C \geq \cdot)$ are some consistent estimators of the censoring survival function. Because $P_{11}(s, t)$ conditions on being in state 1 at time s , the idea is now to estimate the censoring survival function using the censoring times of the

subjects that are uncensored by time s and are in the intermediate state at the end of followup. In order to formalize this, introduce

$$\begin{aligned} Y(u; s) &= \sum_{i=1}^n I\{i : T_{0i} < s, \tilde{T}_i \geq u\}, \quad u > s, \\ N(u; s) &= \sum_{i=1}^n I\{i : T_{0i} < s, \tilde{T}_i \geq s, \tilde{T}_i \leq u, \Delta_i = 1\}, \quad u > s, \\ N^C(u; s) &= \sum_{i=1}^n I\{i : T_{0i} < s, \tilde{T}_i \geq s, \tilde{T}_i \leq u, \Delta_i = 0\}, \quad u > s. \end{aligned}$$

In words, $Y(u; s)$ is the number of individuals at risk of absorption at u —in the subset of the data of subjects who are in the intermediate state and uncensored at time s with associated counting process of observed absorption event $N(u; s)$. N^C is the censoring counting process in this data subset. Note that N^* only counts events in the data subset at hand.

Now, define the following estimator of $P(C \geq u)$, $u > s$,

$$\begin{aligned} \tilde{P}(C \geq u; s) &= \tilde{P}(C \geq u \mid C > s) \tilde{P}(C > s) \\ &= \prod_{v \in (s, u)} \left(1 - \frac{\Delta N^C(v; s)}{Y(v; s) - \Delta N(v; s)} \right) \tilde{P}(C > s), \end{aligned} \quad (2.13)$$

where the product in the last display is over all unique jump times of $N^C(\cdot; s)$ and $\tilde{P}(C > s)$ is some consistent estimator of $P(C > s)$.

Using (2.13) in (2.12) (and the same $\tilde{P}(C > s)$ also for \hat{P}_b and \hat{P}_c) leads to the estimator

$$\hat{P}_{11}(s, t) = \sum_u \prod_v \left(1 - \frac{\Delta N(v; s)}{Y(v; s)} \right) \frac{\Delta N^*(u; s)}{Y(u; s)}. \quad (2.14)$$

We note four important facts about $\hat{P}_{11}(s, t)$. Firstly, the estimator is similar to (2.11) but evaluated in the data subset ‘in the intermediate state 1 at time s and under observation at s ’. Secondly, this data subsetting accounts for the conditioning on $X_s = 1$, and such data subsetting is, in biostatistics, known as *landmarking* (e.g., Anderson et al. 2008; van Houwelingen and Putter 2012). Thirdly, the new estimator (2.14) is just the right-hand limit of the standard Aalen-Johansen estimator of a cumulative incidence function (irrespective of $X(t)$ being Markov or not) and inherits its asymptotic properties (e.g., Andersen et al. 1993, Sect. 4.4). And finally, data subsetting (or landmarking) can easily be extended to random left-truncation (delayed study entry). We illustrate this last aspect with a brief simulation study comparing the Aalen-Johansen estimator of $P_{11}(s, t)$ with the new $\hat{P}_{11}(s, t)$ in a left-truncated non-Markov illness-death model. Recall that the original estimator of Meira-Machado et al. (2006) has only been developed for right-censored data, but an

IPCW-perspective on Kaplan-Meier-integrals has led to an estimator that naturally accounts for left-truncation via landmarking.

To this end, consider n i.i.d. units under study with data $(L_i, \tilde{T}_{0i}, \tilde{T}_i, \Delta_i)$ as before but with the addition of left-truncation times L_i . We assume that (T_{0i}, T_i) is independent of (L_i, C_i) with $P(L_i < C_i) = 1$. We also assume that these n units are under study in the sense that $L_i < \tilde{T}_i$ for all i . In order to account for delayed study entry at time L_i , we re-define

$$Y(u; s) = \sum_{i=1}^n I\{i : T_{0i} < s, \tilde{T}_i \geq u, L_i < s\}, \quad u > s,$$

and analogously for $N(u; s)$ and $N^C(u; s)$. Then $Y(u; s)$ still denotes the number of individuals at risk of absorption at u — in the subset of subjects who are in the intermediate state and under observation at time s , but now in the presence of left-truncation.

Our simulation design is similar to the one of Meira-Machado et al. (2006). We simulate waiting times T_0 in the initial state from an exponential distribution with parameter $0.039 + 0.026$ and entries into the intermediate state, $X_{T_0} = 1$, with binomial probability $0.039/(0.039 + 0.026)$. For individuals moving through the intermediate state, we set $T = 4.0 \cdot T_0$, making the model non-Markov. Random right-censoring times were drawn from an exponential distribution with parameters 0.013 and 0.035, respectively, and random left-truncation was simulated from a skew normal distribution with location parameter -5 , scale 10 and shape 10. We report averages of 1000 simulation runs per scenario, each with a simulated sample size of 200 units.

Table 2.2 shows bias (negative values indicate underestimation) and empirical variance of our new estimator (2.13) and the standard Aalen-Johansen estimator for $P_{11}(s, t)$ (Table 2.3),

$$\prod_{u \in (s, t]} \left(1 - \frac{|\{i : L_i < u = T_i \leq C_i, T_{0i} < T_i\}|}{|\{i : L_i < u \leq \tilde{T}_i, T_{0i} < u\}|} \right)$$

for $s = 25$. In the scenarios considered, the new estimator underestimates and the Aalen-Johansen estimator over-estimates the true probability. The absolute bias in general favours the new estimator, save for early time points and in particular with more pronounced censoring. The empirical variance of the Aalen-Johansen estimator tends to be smaller save for later time points, one possible explanation being that the new estimator uses less data.

2.8 Discussion

The Kaplan-Meier integral can be written as an inverse of probability of censoring weighted estimator for which the weights are estimated with the usual Kaplan-Meier

Table 2.2 Simulation results for estimating $P_{11}(25, t)$ from left-truncated and right-censored non-Markovian data

t	censoring hazard 0.013				censoring hazard 0.035			
	$\hat{P}_{11}(25, t)$		Aalen-Johansen		$\hat{P}_{11}(25, t)$		Aalen-Johansen	
	Bias	Variance	Bias	Variance	Bias	Variance	Bias	Variance
30	-0.0063	0.0029	0.0052	0.0023	-0.1090	0.0311	0.0066	0.0041
40	-0.0089	0.0065	0.0376	0.0047	-0.1101	0.0386	0.0404	0.0094
50	-0.0093	0.0081	0.0818	0.0056	-0.1049	0.0419	0.0877	0.0136
60	-0.0072	0.0082	0.1266	0.0060	-0.0999	0.0415	0.1341	0.0172
70	-0.0100	0.0077	0.1650	0.0061	-0.0974	0.0344	0.1691	0.0221
80	-0.0077	0.0064	0.2019	0.0058	-0.0728	0.0235	0.2132	0.0270
90	-0.0044	0.0037	0.2350	0.0055	-0.0378	0.0115	0.2530	0.0328

Table 2.3 True values of $P_{11}(25, t)$ to be estimated in the simulation study

t	30	40	50	60	70	80	90
$P_{11}(25, t)$	0.8890	0.6930	0.5256	0.3843	0.2649	0.1623	0.0744

method for the censoring times. With this representation the large sample properties of the Kaplan-Meier integral and various modifications can be directly derived with the functional delta method. We further showed in Sect. 2.3 that the conditions imposed by Stute (1993, 1996, 1999) and followers (e.g. Orbe et al. 2003; De Uña Álvarez and Rodríguez-Campos 2004) are practically equivalent to assuming that the censoring is independent of the survival time and of the covariates. Then we showed that it can be advantageous to derive estimators under the conditional independence assumption allowing that the censoring distribution depends on covariates. This improves efficiency and simultaneously reduces the risk of a large sample bias (Robins and Rotnitzky 1992). Our empirical results illustrate the potential bias and the inefficiency of the Kaplan-Meier integral in a specific setting (Table 2.1).

However, in real data applications there is a tradeoff between the simplicity of weighting all the uncensored observations with the Kaplan-Meier for the censoring times and the potential advantages obtained with a working regression model for the conditional censoring distribution. For example in a multi-state framework it is possible to define consistent IPCW estimators for transition probabilities by using the marginal Kaplan-Meier for the censoring (see e.g. Meira-Machado et al. 2006). But, this approach implies that every censored process is weighted unconditional on the state which is occupied at the censoring time. On the other hand, the methods comprised in van der Laan et al. (2002); Van der Laan and Robins (2003) show how to derive more efficient estimators based on an estimate of the survival function of the censored times conditional on the history of the multi-state process and other covariates. In Sect. 2.7, we have exploited this to derive a new estimator of a transition

probability in a non-Markovian illness-death model. Starting with an estimator based on Kaplan-Meier integrals and using the IPCW principle, we also extended the estimator for the case of right-censored and left-truncated data.

Stute's theory of Kaplan-Meier integrals has arguably not entered the mainstream literature on survival analysis, at least not the more biostatistically oriented one, notable exceptions also including Orbe et al. (2002). On the other hand, Kaplan-Meier integrals may form the basis for attacking complex survival models and finding efficient estimators, which we have illustrated for the important illness-death model. We believe that the theory deserves more attention, another possible field of application being competing risks models with a continuous mark (e.g. Gilbert et al. 2008).

References

- Akritis, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Statist.*, 22:1299–1327.
- Akritis, M. G. (2000). The central limit theorem under censoring. *Bernoulli*, pages 1109–1120.
- Allignol, A., Beyersmann, J., Gerds, T., and Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 20:495–513.
- Andersen, P. and Perme, M. (2008). Inference for outcome probabilities in multi-state models. *Lifetime Data Analysis*, 14(4):405–431.
- Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York.
- Anderson, J., Cain, K., and Gelber, R. (2008). Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. *Journal of Clinical Oncology*, 26(24):3913–3915.
- Begun, J. M., Hall, W. J., Huang, W.-M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11:432–452.
- Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 24:1713–1723.
- Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins.
- De Uña Álvarez, J. and Rodríguez-Campos, M. (2004). Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present. *Statistics*, 38:483–496.
- de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study. *Biometrics*, 71(2):364–375.
- Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.
- Gerds, T. A. (2002). *Nonparametric efficient estimation of prediction error for incomplete data models*. PhD thesis, Albert-Ludwig Universität Freiburg.
- Gilbert, P. B., McKeague, I. W., and Sun, Y. (2008). The 2-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. *Biostatistics*, 9(2):263–276.
- Gill, R. D. (1980). Censoring and stochastic integrals. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

- Gill, R. D., Van der Laan, M. J., and Robins, J. M. (1995). Coarsening at random: Characterizations, conjectures and counter-examples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer Lecture Notes in Statistics.
- Graf, E., Schmoor, C., Sauerbrei, W. F., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statist. Med.*, 18:2529–2545.
- Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inference based on incomplete observations in disease state models. *Biometrics*, 47:595–605.
- Hudson, H. M., Lô, S. N., John Simes, R., Tonkin, A. M., and Heritier, S. (2014). Semiparametric methods for multistate survival models in randomised trials. *Statistics in medicine*, 33(10):1621–1645.
- Keiding, N. (1992). Independent delayed entry. In Klein, J. and Goel, P., editors. *Survival analysis: state of the art*, Kluwer, Dordrecht, pages 309–326.
- Malani, H. M. (1995). A modification of the redistribution to the right algorithm using disease markers. *Biometrika*, 82:515–526.
- Meira-Machado, L., de Uña Álvarez, J., and Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-markov illness-death model. *Lifetime Data Analysis*, 12(3):325–344.
- Neuhaus, G. (2000). A method of constructing rank tests in survival analysis. *Journal of Statistical Planning and inference*, 91(2):481–497.
- Orbe, J., Ferreira, E., and Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, 21(22):3493–3510.
- Orbe, J., Ferreira, E., and Nunez-Anton, V. (2003). Censored partial regression. *Biostatistics*, 4:109–121.
- Reeds, J. A. (1976). *On the definition of von Mises Functionals*. PhD thesis, Havard University, Cambridge, Massachusetts.
- Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology, Methodological Issues*, pages 297–331. Birkhäuser, Boston.
- Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82:805–820.
- Satten, G. and Datta, S. (2001). The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.*, 45:89–103.
- Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23:461–71.
- Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 9:1089–1102.
- Tsai, W. and Crowley, J. (1998). A note on nonparametric estimators of the bivariate survival function under univariate censoring. *Biometrika*, 85(3):573–580.
- Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci.*, 72:20–22.
- Van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.
- van der Laan, M. J., Hubbard, A. E., and Robins, J. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association*, 97:494–507.
- Van der Vaart, A. W. (1991). On differentiable functionals. *Ann. Statist.*, 19:178–204.
- Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- van Houwelingen, J. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

From Statistics to Mathematical Finance

Festschrift in Honour of Winfried Stute

Ferger, D.; González Manteiga, W.; Schmidt, T.; Wang,
J.-L. (Eds.)

2017, XIII, 440 p. 43 illus., 20 illus. in color., Hardcover

ISBN: 978-3-319-50985-3