Nearest Neighbor Hazard Estimation with Left-Truncated Duration Data

Rafael Weißbach[1] *, Wladislaw Poniatowski[2] and Walter Krämer[3]

[1] Fakultät für Wirtschafts- und Sozialwissenschaften, Universität Rostock, Germany
[2] Deutsche Telekom, Bonn, Germany
[3] Fakultät Statistik, Technische Universität Dortmund, Germany

April 2, 2012

Abstract

Duration data often suffer from both left-truncation and right-censoring. We show how both deficiencies can be overcome at the same time when estimating the hazard rate nonparametrically by kernel smoothing with the nearest neighbor bandwidth. Smoothing Turnbull's estimator of the cumulative hazard rate, we derive strong uniform consistency of the estimate from Hoeffding's inequality, applied to a generalized empirical distribution function. We also apply our estimator to rating transitions of corporate loans in Germany.

*Keywords. kernel smoothing, hazard rate, left-truncation, right-censoring

Address for correspondence: Prof. Dr. Rafael Weißbach, Lehrstuhl Statistik, Institut für Volkswirtschaftslehre, Fakultät für Wirtschafts- und Sozialwissenschaften, Universität Rostock, 18051 Rostock, Germany, email: rafael.weissbach@uni-rostock.de, Phone: +49/381/4984428, Fax: +49/381/4984401.

JEL numbers; C14, C13, C24, C41, G24, G12

Financial support by Deutsche Forschungsgemeinschaft is gratefully acknowledged (SFB 823 and Grant WE3573/2)

1 Introduction and Summary

Recent years have witnessed a substantial increase of interest in the statistical modeling of credit risk, see e.g. Knüppel and Hermsen (2010). Usually, credit ratings are modeled as discrete-state stochastic process, the crucial parameter being - more often that not - the class-unspecific rating transitions into the adjacent classes (Weißbach and Mollenhauer 2011; Kim et al. 2012). This requires only one rating transition hazard which is usually estimated on the basis of existing selectors for the bandwidth (Weißbach et al. 2008), with lots of durations censored on the right.

However, allowing for right-censoring only reduces the data set to observations originating after the start of a study. In many applications there are lots of creditors which have already been in the portfolio for some time, the information on which is thus discarded. Weißbach et al. (2009) lose 50% of their data that way. Explicitly allowing for left-truncation on the other hand retains all observations and improves upon the efficiency of parameter estimates. In addition, for most smoothing methods, a data-adaptive bandwidth improves the bias-variance trade-off and reduces the boundary bias near the origin (Weißbach et al. 2008). This is especially important for the analysis of durations of the type which concern us here, i.e. durations which have the age origin as left boundary.

Nonparametric estimation of the distribution and the hazard rate of left-truncated duration data has a long history, see e.g. Turnbull (1976), Woodroofe (1985), Stute (1993) or Goto (1996).

In the context of kernel density estimation (of independent and identically distributed data) with data-dependent - and hence stochastic - bandwidth, strong uniform consistency has received much attention (Einmahl and Mason 2005; Wied and Weißbach 2012). The present paper studies such consistency for the case of hazard rate estimation from left-truncated durations with the frequently used nearest-neighbor bandwidth (see e.g. Grillenzoni 2008). We start with the Hoeffding inequality in order to study the local oscillation behavior of the empirical distribution, similar to Schäfer (1986).

2 Estimating the hazard rate

Let T_i, $i = 1, \dots, n^*$ be independent nonnegative event ages. For concreteness, let T_i measure the time between credit origination for debtor i and the day of default, called the 'debt age at default'. This time between credit origination and default is observed under the condition that $L_i \leq T_i \leq C_i$, where L_i denotes truncation on the left and $C_i \geq 0$ denotes censoring on the right. Let n^* be the number of debtors who received a loan since the credit portfolio was first established, till the end of the study (here the end of 2003). Assume that monitoring of the credit portfolio began at a certain date (here beginning of 1997). Then, L_i denotes the time elapsed since entry into the portfolio (which might have happened earlier than 1997) up to the start of the study. The effective sample size n will be less than n^* because some debtors will have defaulted (or being censored) prior to the start of the study. Note that L_i is negative whenever a creditor entered the portfolio after the start of the study, as illustrated by the fifth observation in Figure 1. After the start of the study, all existing debtors plus the new ones will be followed until the end of the study. Apart from the end of the study, right censoring can also happen due to loss to-follow-up, debt redemption, fusion, and so on, the corresponding age is C_i.

As has been common practice since Turnbull (1976), we condition on $L_i \leq C_i$, so we observe L_i, $X_i := \min(T_i, C_i)$ and $\delta_i = \mathbb{1}_{\{T_i < C_i\}}$, or nothing at all (which happens whenever $L_i \geq T_i$). Without loss of generality, we assume the latter to happen for observations $i = n+1, \dots, n^*$ where $n \leq n^*$. Figure 1 illustrates various possible scenarios.

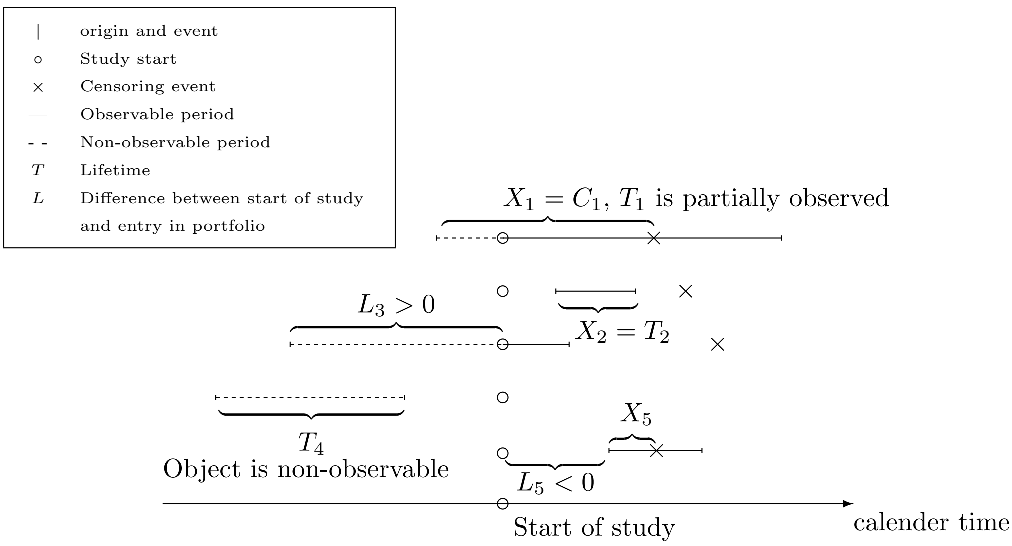

Figure 1: Scenarios of left-truncated and right-censored duration data.

For the derivation of our analytical results we impose the following assumptions:

(A1) $(T_i \in \mathbb{R}_0^+)_{i\in\mathbb{N}}$, $(C_i \in \mathbb{R}_0^+)_{i\in\mathbb{N}}$ and $(L_i \in \mathbb{R})_{i\in\mathbb{N}}$ are i.i.d. and independent from each other.

(A2) The respective distribution functions F, F^C, F^L und F^X are Lipschitz-continuous and strictly monotone.

(A3) There exist constants $0 < A < B$ such that $F(A) > 0$, $F^L(A) > 0$, $F(B) < 1$ and $F^X(B) < 1$.

Given an estimate $\Lambda_n(\cdot)$ for the cumulative hazard rate $\Lambda(\cdot)$ of the T_i, the hazard rate $\lambda(\cdot)$ can be estimated via a kernel function $K(\cdot)$ such as

$$\lambda_n(t) := \int_{\mathbb{R}_0^+} \frac{1}{R_n(s)} K\left(\frac{t-s}{R_n(s)}\right) d\Lambda_n(s). \tag{1}$$

By defining a - possibly stochastic - monotone function $\tilde{\Psi}_n(\cdot)$ and

$$R_n(t) := \inf\left\{r > 0 : \left|\tilde{\Psi}_n\left(t - r/2\right) - \tilde{\Psi}_n\left(t + r/2\right)\right| \geq p_n\right\} \tag{2}$$

we allow here for a fixed bandwidth $R_n(t) \equiv b$, but also for a variable deterministic bandwidth $R_n(t) = R(t)$. The k-nearest neighbor bandwidth of Gefeller and Dette (1992) is a special case of $R_n(t)$ when $\tilde{\Psi}_n(\cdot)$ estimates $F(\cdot)$ and p_n is equal to k/n. Li and Li (2010) suggest the k-nearest neighbor bandwidth in various econometric contexts. Throughout we require that the bandwidth parameter p_n obeys the usual restrictions $0 < p_n < 1$, $p_n \longrightarrow 0$ and $\log(n)/(np_n) \longrightarrow 0$, see e.g. Schäfer (1986).

Given assumptions (A1)-(A3), and a regularity condition on $\Lambda_n(\cdot)$ to be specified below, $\lambda_n(\cdot)$ is strongly uniformly consistent on the open interval (A,B).

Lemma 1. *There exists a constant* $0 \leq D < \infty$ *such that*

$$P\left\{\limsup_{n\longrightarrow\infty} \frac{\sup_{t\in[a,b]} |\lambda_n(t) - \lambda(t)|}{\sqrt{\log(n)/(np_n)} + p_n} = D\right\} = 1 \qquad \forall\ [a,b] \subset (A,B).$$

The proof is an application of Theorem 3.1 in Weißbach (2006) which uses empirical process technique from the 1980s. It is based on integration by parts: One decomposes the error into the total variation of the kernel and the local proximity of the stochastic processes $\Lambda_n(\cdot)$ to its limit $\Lambda(\cdot)$. The total variation is calculated in an elementary fashion. The contribution of the variability of the bandwidth to the error can be taken into account by adapting the proof in Schäfer (1986).

Weißbach (2006) requires the following local asymptotic behavior of the right-continuous and monotonous cumulative hazard rate estimator $\Lambda_n(\cdot)$: For $I := [a,b]$, $\Lambda(I) := \Lambda(b) - \Lambda(a)$ and some finite $0 \leq D < \infty$,

$$P\left\{\limsup_{n\longrightarrow\infty} \frac{\sup_{I\subset[A,B],\Lambda(I)\leq p_n} |\Lambda_n(I) - \Lambda(I)|}{\sqrt{\log(n)p_n/n}} = D\right\} = 1. \tag{3}$$

We now show that estimating the cumulative hazard rate under left-truncation as in Turnbull (1976) obeys equation (3). This is done in two steps. First, we construct a general estimator and show that it converges with the rate specified in (3), and then we establish the estimator of Turnbull (1976) as a special case.

In the classical case of right-censored durations, the observable data is a bivariate sample of X_i and δ_i. For additional left-truncation a third dimension is needed. Let $(\mathbf{S}_i)_{i=1,\dots,n}$ be a sample of independent identically distributed observable random vectors $\mathbf{S}_i : \Omega \to \mathbb{R}^3$. The first coordinate of the vector, S_i^1, is again not T_i as (A1) but X_i. The second will be the truncation age L_i and the third δ_i. The hazard rate $\lambda(\cdot)$ of T_i can be represented by the ratio of density and survival function. The numerator can be estimated by dirac measures on the observations, accounting for the respective data design. The monotonously decreasing survival function can be estimated by (one minus) the empirical distribution function. The estimator of the cumulative hazard rate is then a Stieltjes integral, a sum. For left-truncated data, the denominator cannot be the empirical distribution function anymore: When estimating the hazard rate at an observation, the population at risk in the denominator can have increased (as compared to the estimate at the preceding observation). Left-truncated observations may have entered the population at risk prior to another event having occurred. The population at risk is not monotonously decreasing anymore. Hence, we define a function $G : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ to be (only) continuous, accompanied by an estimator $G_n : \mathbb{R}_0^+ \times (\mathbb{R}^3)^n \to \mathbb{R}_0^+$, $(t, s_1, \dots, s_n) \mapsto G_n(t)(s_1, \dots, s_n)$ being symmetric for each fixed $t \in \mathbb{R}_0^+$ and $s_1, \dots, s_n \in \mathbb{R}^3$. In addition, we use the simplified notation $G_n(t, \omega)$ - or even $G_n(t)$ - for $G_n(t)(\mathbf{S}_1(\omega), \dots, \mathbf{S}_n(\omega))$.

With respect to censoring, it will in addition prove useful to define a weight function by the mapping $\Delta : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, (t, s) \mapsto \Delta^t(s)$ with simplified notation Δ_i^t for $\Delta^t(\mathbf{S}_i(\omega))$. All Δ_i^t are assumed to be bounded (from above) by a Δ_{max} on the interval $[A, B]$.

Let the function $\Psi : \mathbb{R} \to \mathbb{R}_0^+$ be continuous, positive and strictly monotonically increasing on the interval $[A, B]$. We now propose to estimate $\Psi(\cdot)$ by

$$\Psi_n(t) := \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}_{\{S_i^1 \leq t\}} \cdot \Delta_i^t}{G_n(S_i^1)}. \tag{4}$$

The idea is that Δ_i^t is supposed to account for censoring, e.g. it is δ_i for right-censored data. The denominator $G_n(\cdot)$ must account for conditions, e.g. in hazard rate estimation for the condition of being at risk. For left-truncation, the condition of an event to be observed at all can be modeled by G_n as will soon be seen. The indicator function $\mathbb{1}_{\{S_i^1 \leq t\}}$ results from the dirac measures, similar to the empirical distribution function for iid data. Note that $\Psi_n(t)$ does not need to be observable.

The local consistency of the estimate (4) requires some assumptions on the target function $\Psi(\cdot)$, on the observed random variables and on the rate at which $G_n(\cdot)$ converges to $G(\cdot)$.

(B1) There exists a finite constant $M := \sup_{t \in [A,B]} [G(t)]^{-1}$.

(B2) $[\mathbb{1}_{\{t \leq a\}} \Delta_i^a - \mathbb{1}_{\{t \leq b\}} \Delta_i^b][G(t) - G_n(t)] = 0$ for all $[a, b] \subseteq [A, B]$ and $t \notin [A, B]$.

(B3) For each fixed $t \in [A, B]$, $\mathbb{1}_{\{S_i^1 \leq t\}} \Delta_i^t / G(S_i^1)$ is an unbiased estimator for $\Psi(t)$. In case $(\mathbf{S}_i)_{i=1,\ldots,n}$ are only observable under a condition, the estimator is conditionally unbiased.

(B4) For $G(t)$ and $G_n(t)$, there exists a constant $0 \leq D_G < \infty$, such that

$$P\left\{ \limsup_{n\to\infty} \frac{\sup_{t \in [A,B]} |G_n(t) - G(t)|}{\sqrt{\log(n)/n}} = D_G \right\} = 1.$$

Theorem 2. *Under regularity conditions (A1)-(A3) and (B1)-(B4), there exists a constant*

$$0 \leq D \leq 2(\sqrt{2 \cdot (2\Delta_{max}M + \Psi(B))} + D_G M)$$

such that

$$P\left\{ \limsup_{n\to\infty} \frac{\sup_{I\subseteq[A,B],\Psi(I)\leq p_n} |\Psi_n(I) - \Psi(I)|}{\sqrt{\log(n)p_n/n}} = D \right\} = 1.$$

The proof of Theorem 2 relies on the following preliminary estimator with known $G(\cdot)$:

$$\Psi_n^*(t) := \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{S_i^1\leq t\}} \cdot \Delta_i^t}{G(S_i^1)}.$$

The aim is to split the difference $|\Psi_n(I) - \Psi(I)|$ into two parts using the measure $\Psi_n^*(I) := \Psi_n^*(b) - \Psi_n^*(a)$ for $I = [a, b]$ and to prove the almost sure convergence for each term separately. The term without $G_n(\cdot)$ represents a sum of bounded random variables so that convergence results from the inequality of Hoeffding (1963). The complete proof of Theorem 2 is in Appendix A.

Using Theorem 2 we show next that the cumulative hazard rate estimator of Turnbull (1976) has a representation (4). Recall that $F(\cdot)$ denotes the distribution function and $f(\cdot)$ the density function of T_i. It is easily seen that

$$\Lambda(t) := \int_0^t \lambda(s)ds = \int_0^t \frac{dF^{X*}(s)}{G(s)}, \tag{5}$$

where $F^{X*}(t) := P(X_i \leq t, \delta_i = 1|L_i \leq X_i)$ and $G(t) := P(L_i \leq t \leq X_i|L_i \leq X_i)$. Turnbull (1976) proposes to estimate the cumulative hazard rate by

$$\Lambda_n(t) := \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i\leq t,\delta_i=1\}}}{nG_n(X_i)} = \sum_{i:X_{(i)}\leq t} \frac{\delta_i}{\sharp\{j : L_j \leq X_{(i)} \leq X_j\}}, \tag{6}$$

where summation occurs only over cases where $L_i \leq X_i$, and where $G_n(t) := n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{L_i \leq t \leq X_i\}}$ is a consistent estimate of $G(\cdot)$. This is the Nelson-Aalen estimator for right-censored observations, additionally allowing for late entry into the population under risk.

Corollary 3. *Given (A1)-(A3) there exists for $\Lambda_n(\cdot)$ defined in (6) a constant $D \leq 2(\sqrt{2 \cdot (2M + \Lambda(B))} + 2M)$, such that*

$$P\left\{\limsup_{n\to\infty} \frac{\sup_{I\subseteq[A,B],\Lambda(I)\leq p_n} |\Lambda_n(I) - \Lambda(I)|}{\sqrt{\log{(n)} p_n/n}} = D\right\} = 1$$

with finite $M := \sup_{t\in[A,B]}[P(L_i \leq t \leq X_i | L_i \leq X_i)]^{-1}$.

To prove local convergence we check the conditions (B1)-(B4) and define its components as $\Delta_i^t := \delta_i \leq 1 =: \Delta_{max}$ for $i = 1, \dots, n$ and for each fixed $t \in [A, B]$, $\Psi(t) := \Lambda(t)$ and $D_G := 2$. The remainder of the proof is in Appendix B.

To estimate the hazard rate $\lambda(t)$, we apply (1) to obtain from $\Lambda_n(\cdot)$

$$\lambda_n(t) = \sum_{i=1}^{n} \frac{1}{R_n(X_{(i)})} K\left(\frac{t - X_{(i)}}{R_n(X_{(i)})}\right) \frac{\delta_i}{\sharp\{j : L_j \leq X_{(i)} \leq X_j\}}, \qquad (7)$$

where $R_n(\cdot)$ is the nearest-neighbor bandwidth.

One of the main conditions for the kernel estimation of $\lambda_n(\cdot)$ is the Lipschitz-continuity of $\lambda(\cdot)$ and $\Lambda(\cdot)$, which follows from the Lipschitz-continuity of $G(\cdot)$ and $F^{X*}(\cdot)$. By assumption (A2) for $F^X(\cdot)$, is $F^{X*}(\cdot)$ likewise Lipschitz-continuous. Next we rewrite $G(\cdot)$ as follows to prove its Lipschitz-continuity:

$$\begin{aligned} G(t) &= \alpha^{-1} F^L(t)(1 - F(t))(1 - F^C(t)) \\ &= \int_{-\infty}^{t} \alpha^{-1}(1 - F(t))(1 - F^C(t)) f^L(s) ds, \end{aligned}$$

where $\alpha := P(L_i \leq X_i)$. Obviously, the Lipschitz-continuity of $F^L(\cdot)$ implies the Lipschitz-continuity of $G(\cdot)$.

3 An empirical application

Next we apply the techniques described above to corporate credit risk. Since corporations default rarely, credit rating transition has become the surrogate event in empirical research. We analyze rating duration data from WestLB, Düsseldorf, which provided us with rating histories from an internal rating system with 8 non-default classes and 1 default class observed over seven years from January 1, 1997 to December 31, 2003. The data set comprises about 600 transitions for 359 debtors overall.

In this context, a constant hazard is a common assumption (see Bluhm et al., 2002). This has been questioned by Kiefer and Larson (2007) and Weißbach et al. (2009), so Weißbach and Walter (2010) propose a parametric piecewise constant model as an alternative. The asset value model of Merton (1974) allows only transitions to adjacent classes. Weißbach and Mollenhauer (2011) show that changing rating classes is not class-specific, i.e. does not depend on the class h from where the rating change starts, nor does it depend on the target class j of a transition. Hence our model for the rating transition hazard is

$$\lambda_{hj}(t) \equiv \lambda(t) \quad \text{for} \quad h = 1, \ldots, 8,\ j = 1, \ldots, 9,\ |h - j| = 1, \tag{8}$$

and $\lambda_{hj}(t) = 0$ for $|h - j| > 1$.

Next we estimate $\lambda(t)$ using (1). We apply bandwidth (2) in its specification as nearest-neighbor bandwidth, i.e. by using an estimate of F in the role of $\tilde{\Psi}_n$. The kernel function is known to have little impact; we use the bi-square kernel $K(t) = 15/16(1 - t^2)^2 \mathbb{1}_{\{|t| \leq 1\}}$.

We start by considering first transitions only, the times T_i. If, for instance, a debtor remains in its rating class for the entire observation period, T_i is not observed and C_i at the end of the observation period is recorded instead (right-censoring). About 60% of our 359 debtors produce durations which are censored on the right for several reasons. We estimate the cumulative

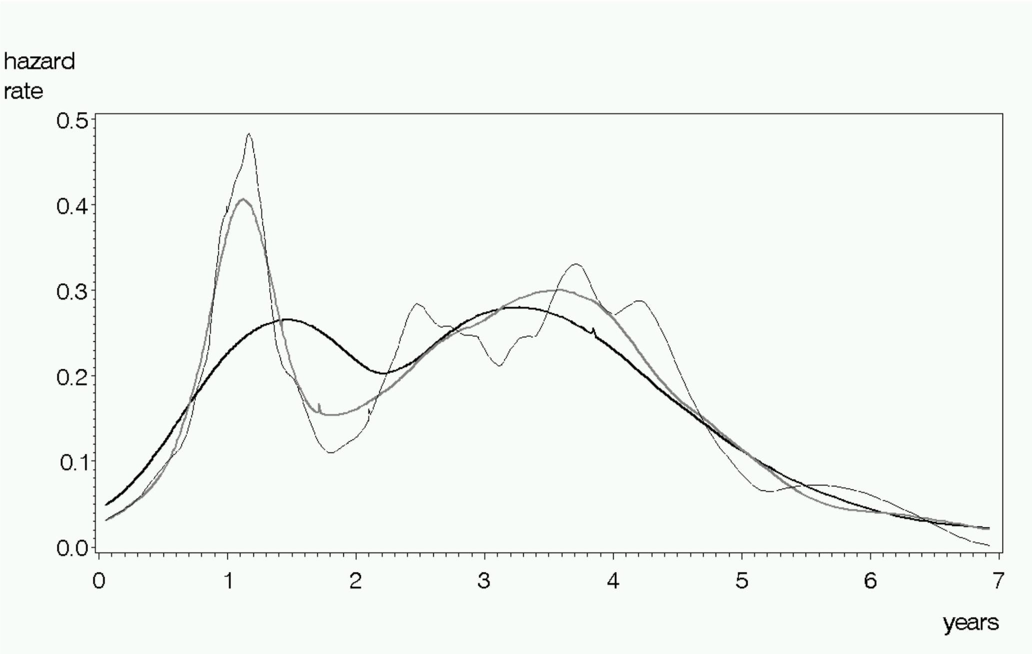

Figure 2: Estimated transition hazard $\lambda_n(t)$ from 359 right-censored rating transitions: Bandwidth selection by rule-of-thumb (thick-black), cross-validation (thick-grey), and plug-in (thin-black)

hazard by the Nelson-Aalen estimator

$$\Lambda_n(t) = \sum_{i:X_{(i)} \leq t} \frac{\delta_i}{\sharp\{j : X_{(i)} \leq X_j\}}.$$

For the specification of $R_n(t)$, F is estimated by $1 - \exp(-\Lambda_n(t))$. Additionally, the number of nearest neighbors k is crucial. We use three selectors. As fast solution, Weißbach et al. (2008) adapt the rule of thumb of Silverman (1986) which results in $k = 123$ nearest neighbors. Second, a cross-validatory selector for k as described in Gefeller et al. (1996) yields $k = 78$ nearest neighbors. Third, a plug-in rule from Weißbach (2006) yields $k = 38$. Results are displayed in Figure 2. All bandwidth selectors produce similarly shaped hazard rates, only plug-in seems to be under-smoothing. At the left edge, near the origin and up to one year, the hazard rate is small for all bandwidth selectors. It is unlikely that the well-known boundary effect is the only reason because such bias is reduced by the nearest-neighbor bandwidth, see Weißbach et al. (2008). Note also that Weißbach and Walter (2010) reject constant rating transition hazards, presumably due to the small transition hazard of the first year: Credit analysts tend not to revaluate their debtors within the first year. The mode at one year then results from an increased evaluation activity one year after grading the credit. Note that transitions to rating classes beyond the adjacent ones are censored and do not enter this analysis. As of now we cannot explain the second mode at three-and-a-half years.

Considering only the first transition for each rating history implies a loss of 40% of all transitions. This loss is avoided by incorporating second transitions by means of left-truncation. Such potential second transitions can be viewed as an additional T_i, subject to left-truncation L_i, where L_i is the first transition. The second transition is again potentially right-censored by a C_i. Very rarely, there are third and further transitions which could be treated similarly. We apply estimator (7) for the transition hazard model (8) to a sample of now 542 univariate durations. We use the three bandwidths calculated above for the right-censored data set. This is because we are interested in the possible improvement of the estimates due to the additional observations. In the nearest neighbor bandwidth specification (2), the estimator of F now involves the exponential of the Turnbull estimator (6). However, there are now two possibilities for choosing $p_n = k/n$. First we can use the additional 183 observations and define $p_n = k/542$. This will keep the number of nearest neighbors used for the bandwidth (2) equal to the number in the previous (right-censored) setting and hence decreases the window width, so the bias is reduced. Second, we keep the window width equal, i.e. continue with $p_n = k/359$, so the number of nearest neighbors

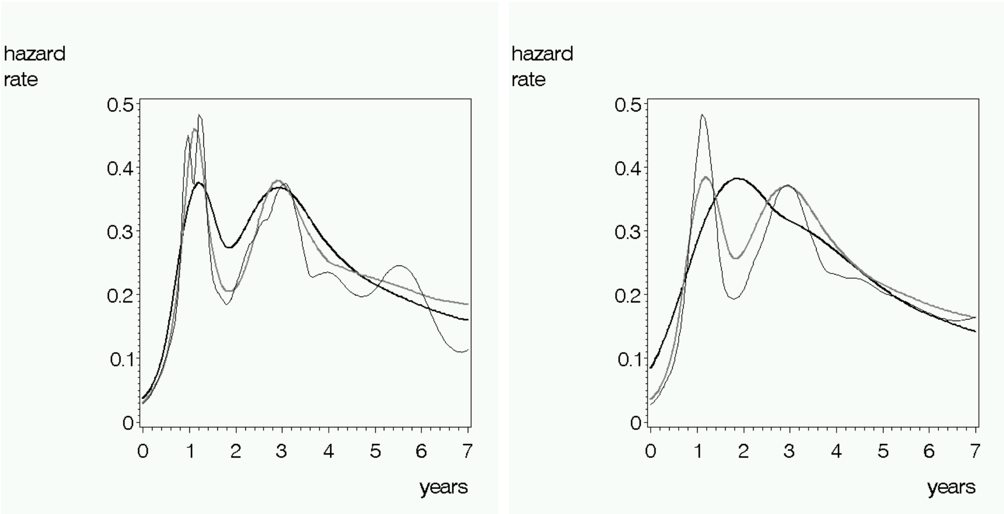

Figure 3: Estimating rating transition hazard for right-censored and left-truncated data: reduced bias (left) versus reduced variance (right) - legend as in Figure 2

in the window is increased and hence the variance reduced. Figure 3 shows both results.

Allowing for left-truncated rating transitions favors second (and third) transitions, which naturally occur later than the first. Therefore the additional 183 observations result in a more stable estimate of the hazard rate, especially from year 5 onwards. Allowing for left-truncation enables risk quantification of older debt. The steep increase near the origin is confirmed. However, only few observations are added for estimation in that region. The second mode is not pronounced anymore in the rule-of-thumb smoothing with reduced variance.

References

Bluhm, C., Overbeck, L., and Wagner, C. (2002). *An Introduction to Credit Risk Modeling.* Chapman & Hall.

Einmahl, U. and Mason, D. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33:1380–1403.

Gefeller, O. and Dette, H. (1992). Nearest neighbour kernel estimation of the hazard function from censored data. *Journal of Statistical Computation and Simulation*, 43:93–101.

Gefeller, O., Weißbach, R., and Bregenzer, T. (1996). The implementation of a data-driven selection procedure for the smoothing parameter in non-parametric hazard rate estimation using SAS/IML software. In Friedl, H., Berghold, A., and Kauermann, G., editors, *Proceedings of the 13th SAS European Users Group International Conference, Stockholm*, pages 1288–1300. SAS Institute Inc. Carry.

Goto, F. (1996). Achieving semiparametric efficiency bounds in left-censored duration models. *Econometrica*, 64(2):439–442.

Grillenzoni, C. (2008). Robust nonparametric estimation of the intensity function of point data. *Advances in Statistical Analysis*, 92:117–134.

Hewitt, E. and Savage, L. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80:470–501.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Kiefer, N. and Larson, C. (2007). A simulation estimator for testing the time homogeneity of credit rating transitions. *Journal of Empirical Finance*, 14:818–835.

Kim, Y.-D., James, L., and Weißbach, R. (2012). Bayesian analysis of multi-state event history data: Beta-dirichlet process prior. *Biometrika*, 99:127–140.

Knüppel, L. and Hermsen, O. (2010). Median split, k-group split, and optimality in continuous populations. *Advances in Statistical Analysis*, 94:53–74.

Li, D. and Li, Q. (2010). Nonparametric/semiparametric estimation and testing of econometric models with data dependent smoothing parameters. *Journal of Econometrics*, 157:179 – 190.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470.

Schäfer, H. (1986). Local convergence of empirical measures in the random censorship situation with application to density and rate estimators. *Annals of Statistics*, 14:1240–1245.

Shorack, G. and Wellner, J. (1986). *Empirical Processes with Application to Statistics*. Wiley & Sons.

Silverman, B. (1986). *Density Estimation*. Chapman & Hall, London.

Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *Annals of Statistics*, 21:146–156.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38:290–295.

Weißbach, R. and Mollenhauer, T. (2011). Modelling rating transitions. *Journal of the Korean Statistical Society*, 4:469–485.

Weißbach, R., Pfahlberg, A., and Gefeller, O. (2008). Double-smoothing in kernel hazard rate estimation. *Methods of Information in Medicine*, 47:167–173.

Weißbach, R., Tschiersch, P., and Lawrenz, C. (2009). Testing time-homogeneity of rating transitions after origination of debt. *Empirical Economics*, 36:575–596.

Weißbach, R. and Walter, R. (2010). A likelihood ratio test for stationarity of rating transitions. *Journal of Econometrics*, 155:188–194.

Weißbach, R. (2006). A general kernel functional estimator with general bandwidth - strong consistency and applications. *Journal of Nonparametric Statistics*, 18:1–12.

Wied, D. and Weißbach, R. (2012). Consistency of the kernel density estimator - a survey. *Statistical Papers*, 53:1–21.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics*, 13:163–177.

A Proof of Theorem 2

The proof of Theorem 2 is in four steps. First, for an interval $I := [a, b] \subseteq [A, B]$ we establish an exponential bound for the distribution of the difference $|\Psi_n^*(I) - \Psi(I)|$:

$$P(|\Psi_n^*(I) - \Psi(I)| > \varepsilon) < 2 \exp\left(\frac{-n\varepsilon^2}{2(2\Delta_{max}M + \Psi(B))(p + \varepsilon)}\right) \qquad (9)$$

for all $p > 0$, $\varepsilon > 0$, $n \in \mathbb{N}_{>0}$ and for each fixed $I \subseteq [A, B]$ with $\Psi(I) \leq p$.

Because of definition (4) and the boundedness of $0 \leq \Delta_i^x \leq \Delta_{max} < \infty$

$$\Psi_n^*(I) - \Psi(I) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}_{\{S_i^1 \leq b\}} \cdot \Delta_i^b}{G(S_i^1)} - \frac{\mathbb{1}_{\{S_i^1 \leq a\}} \cdot \Delta_i^a}{G(S_i^1)} - \Psi(I)\right) \qquad (10)$$

is the arithmetic mean of the n independent and bounded random variables for each fixed $I \subseteq [A, B]$, distributed as

$$T_I := \frac{\mathbb{1}_{\{S_1^1 \leq b\}} \cdot \Delta_1^b}{G(S_1^1)} - \frac{\mathbb{1}_{\{S_1^1 \leq a\}} \cdot \Delta_1^a}{G(S_1^1)} - \Psi(I).$$

The expectation, the variance and the bound of T_I can then be calculated for fixed $I \subseteq [A, B]$ with $\Psi(I) \leq p$.

The expectation of T_I follows from assumption (B3):

$$E(T_I) = E\left(\frac{\mathbb{1}_{\{S_1^1 \leq b\}} \cdot \Delta_1^b}{G(S_1^1)}\right) - E\left(\frac{\mathbb{1}_{\{S_1^1 \leq a\}} \cdot \Delta_1^a}{G(S_1^1)}\right) - \Psi(b) + \Psi(a) = 0. \quad (11)$$

From assumption (B1), we get the following bound of $|T_I|$ on $[A, B]$:

$$\begin{aligned} |T_I| &= \left|\frac{\mathbb{1}_{\{S_1^1 \leq b\}} \cdot \Delta_1^b}{G(S_1^1)} - \frac{\mathbb{1}_{\{S_1^1 \leq a\}} \cdot \Delta_1^a}{G(S_1^1)} - \Psi(I)\right| \\ &< 2\Delta_{max} M + \Psi(B) - \Psi(A) < 2\Delta_{max} M + \Psi(B) =: g. \end{aligned} \quad (12)$$

The variance of T_I can be obtained from the expectation (11) and the bound (12) as follows:

$$\begin{aligned} \sigma_I^2 :=& Var(T_I) = E\left[\left(\frac{\mathbb{1}_{\{S_1^1 \leq b\}} \cdot \Delta_1^b}{G(S_1^1)} - \frac{\mathbb{1}_{\{S_1^1 \leq a\}} \cdot \Delta_1^a}{G(S_1^1)} - \Psi(I)\right)^2\right] \\ &< 2\Delta_{max} M \cdot E\left(\frac{\mathbb{1}_{\{S_1^1 \leq b\}} \cdot \Delta_1^b}{G(S_1^1)} - \frac{\mathbb{1}_{\{S_1^1 \leq a\}} \cdot \Delta_1^a}{G(S_1^1)}\right) \\ &= 2\Delta_{max} M \cdot \Psi(I) < g \cdot p. \end{aligned} \quad (13)$$

From equations (10), (11), (12), (13) and the inequality from Hoeffding (1963) results the following right bound:

$$P(|\Psi_n^*(I) - \Psi(I)| > \varepsilon) < 2\exp\left(\frac{-n\varepsilon^2}{2(\sigma_I^2 + g\varepsilon/3)}\right) < 2\exp\left(\frac{-n\varepsilon^2}{2g(p+\varepsilon)}\right)$$

for each fixed interval $I \subseteq [A, B]$ with $\Psi(I) \leq p$.

In the second step we derive the inequality

$$\sup_{I \subseteq [A,B], \Psi(I) \leq p_n} |\Psi_n^*(I) - \Psi(I)| \leq C\sqrt{\log(n)p_n/n} \tag{14}$$

almost surely for a constant $C > \sqrt{2(2\Delta_{max}M + \Psi(B))}$ and large n.

On the right hand side of the inequality (9), p and ε can be substituted with p_n and $\varepsilon_n := C\sqrt{\log{(n)}p_n/n}$ for $C > 0$ and $n > 1$ altering the upper bound to

$$< 2 \cdot \exp\left(-\log(n)\frac{C^2}{2g}\frac{p_n}{(p_n + \varepsilon_n)}\right) = 2n^{-\frac{C^2}{2g}\frac{p_n}{(p_n+\varepsilon_n)}} =: A_n.$$

The series (A_n) is then summable starting from some large $n < \infty$ only if the exponent $\beta_n := (C^2 p_n)/(2g(p_n + \varepsilon_n)) > 1$. From $\varepsilon_n/p_n = C\sqrt{\log(n)/(np_n)}$ and the assumptions for p_n follow $\varepsilon_n/p_n \to 0$ and $p_n/(p_n + \varepsilon_n) \to 1$ for large n. The condition $\beta_n > 1$ can be then achieved with $C^2/2g > 1$ or $C > \sqrt{2g}$.

As a consequence, the series (A_n) is summable from some large $n < \infty$ and only for $C > \sqrt{2g}$. For each $I \subseteq [A, B]$ with $\Psi(I) \leq p_n$ we get then $\exists C > \sqrt{2g}\ \exists m < \infty, m \in \mathbb{N} : \sum_{n=m}^{\infty} P(|\Psi_n^*(I) - \Psi(I)| > \varepsilon_n) < \sum_{n=m}^{\infty} A_n < \infty$ and $\forall m < \infty, m \in \mathbb{N} : \sum_{n=1}^{m} P(|\Psi_n^*(I) - \Psi(I)| > \varepsilon_n) \leq m < \infty$.

Because of the summability of $P(|\Psi_n^*(I) - \Psi(I)| > \varepsilon_n)$,

$$P\left(\limsup_{n\to\infty} |\Psi_n^*(I) - \Psi(I)| > \varepsilon_n\right) = 0$$

results from the Borel-Cantelli lemma for $C > \sqrt{2g}$, i.e. $|\Psi_n^*(I) - \Psi(I)|$ does not exceed ε_n for most of the n. For large n and for all $I \subseteq [A, B]$ with $\Psi(I) \leq p_n$, we derive almost surely that $|\Psi_n^*(I) - \Psi(I))| \leq C\sqrt{\log{(n)}p_n/n}$.

The same inequality holds for the supremum of $|\Psi_n^*(I) - \Psi(I)|$ on $[A, B]$: $\sup_{I\subseteq[A,B],\Psi(I)\leq p_n} |\Psi_n^*(I) - \Psi(I)| \leq C\sqrt{\log{(n)}p_n/n}$ for $C > \sqrt{2g}$ and large n almost surely.

Using the results above we prove the following inequality in a third step:

$$\sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n^*(I)-\Psi(I)| \le C\cdot p_n\sqrt{\log(n)/n}$$

almost surely for some $C > D_G \cdot M$ and large n.

From assumption (B4) and the limes superior formulation of Hewitt and Savage (1955) we obtain the right bound $G_n(x)-G(x) \le |G_n(x)-G(x)| \le \sup_{x\in[A,B]}|G_n(x)-G(x)| \le C_1'\cdot\sqrt{\log(n)/n}$ almost surely for $C_1' > D_G$, large n and all $x\in[A,B]$. These bounds can be rewritten for $G_n(x)$ as follows: $G_n(x) \ge G(x) - C_1'\sqrt{\log(n)/n} \ge \inf_{t\in[A,B]} G(t) - C_1'\cdot\sqrt{\log(n)/n}$.

From assumption (B4) we have $\inf_{t\in[A,B]} G(t) > 0$. Because of $\sqrt{\log(n)/n}\to 0$, the following inequalities hold for $x\in[A,B]$ and large n:

$$\inf_{t\in[A,B]} G(t) - C_1'\cdot\sqrt{\log(n)/n} > 0,$$

$$\frac{1}{G_n(x)} \le \frac{1}{\inf\limits_{t\in[A,B]} G(t) - C_1'\cdot\sqrt{\log(n)/n}}$$

and

$$\frac{|G_n(x)-G(x)|}{G_n(x)} \le \frac{C_1'\cdot\sqrt{\log(n)/n}}{\inf\limits_{t\in[A,B]} G(t) - C_1'\cdot\sqrt{\log(n)/n}}.$$

The following bounds for $\Psi_n^*(I)-\Psi(I)$ and $\Psi_n^*(I)$ result from the equation (14) almost surely for $I\subseteq[A,B]$ with $\Psi(I)\le p_n$, large n and $C_2' > \sqrt{2\cdot(2\Delta_{max}M+\Psi(B))}$:

$$\begin{aligned}\Psi_n^*(I)-\Psi(I) &\le |\Psi_n^*(I)-\Psi(I)| \\ &\le \sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n^*(I)-\Psi(I)| \le C_2'\sqrt{\log(n)p_n/n}\end{aligned}$$

and consequently $\Psi_n^*(I) \le \Psi(I) + C_2'\sqrt{\log(n)p_n/n} \le p_n + C_2'\sqrt{\log(n)p_n/n}$.

We then obtain the following equation from assumption (B2) almost surely for each $I \subseteq [A, B]$ with $\Psi(I) \le p_n$ and large n:

$$
\begin{aligned}
&|\Psi_n^*(I) - \Psi(I)| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{G_n(S_i^1)} - \frac{1}{G(S_i^1)}\right|(\mathbb{1}_{\{S_i^1\le b\}}\cdot\Delta_i^b - \mathbb{1}_{\{S_i^1\le a\}}\cdot\Delta_i^a)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{G_n(S_i^1) - G(S_i^1)}{G_n(S_i^1)}\right|\frac{\mathbb{1}_{\{S_i^1\le b\}}\cdot\Delta_i^b - \mathbb{1}_{\{S_i^1\le a\}}\cdot\Delta_i^a}{G(S_i^1)}\\
&\le \frac{C_1'\sqrt{\log(n)/n}\cdot\Psi_n^*(I)}{\inf\limits_{t\in[A,B]} G(t) - C_1'\sqrt{\log(n)/n}} \le \frac{C_1'\sqrt{\log(n)/n}\cdot\left(p_n + C_2'\sqrt{\log(n)p_n/n}\right)}{\inf\limits_{t\in[A,B]} G(t) - C_1'\sqrt{\log(n)/n}}.
\end{aligned}
$$

By $p_n + C_2'\sqrt{\log(n)p_n/n} = p_n[1 + C_2'\sqrt{\log(n)/(p_n n)}]$ it is evident, that $C_2'\sqrt{\log(n)/(p_n n)}$ can be neglected for large n because of the assumptions for p_n. For large n, we can also neglect the term $\sqrt{\log(n)/n}$ in the numerator. For all $I \subseteq [A, B]$ with $\Psi(I) \le p_n$ and for large n, we derive the inequality $|\Psi_n^*(I) - \Psi(I)| \le \frac{C_1'}{\inf_{t\in[A,B]} G(t)} p_n\sqrt{\log(n)/n} = C_1' \cdot M \cdot p_n\sqrt{\log(n)/n}$ almost surely.

The requested bound $\sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n^*(I) - \Psi(I)| \le C \cdot p_n\sqrt{\log(n)/n}$ results for some $C > D_G \cdot M$ and large n almost surely.

In a final step we examine the expression $\sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n(I) - \Psi(I)|$. This overall difference can be represented by the sum of the deviations of the empirical and theoretical measures $\Psi_n(I)$ and $\Psi(I)$ from the preliminary measure $\Psi_n^*(I)$ as follows: $\sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n(I) - \Psi(I)|$
$\le \sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n(I) - \Psi_n^*(I)| + \sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n^*(I) - \Psi(I)|$.

Because $p_n\sqrt{\log(n)/n}/\sqrt{\log(n)p_n/n} = \sqrt{p_n}$ approaches zero, i.e. $p_n\sqrt{\log(n)/n} \le \sqrt{\log(n)p_n/n}$ holds for large n.

The previously mentioned upper bounds of $|\Psi_n(I) - \Psi_n^*(I)|$ and $|\Psi_n^*(I) - \Psi(I)|$ imply the existence of a constant $C > \sqrt{2\cdot(2\Delta_{max}M + \Psi(B))} + D_G \cdot M$, such that almost surely for large n

$$
\begin{aligned}
\sup_{I\subseteq[A,B],\Psi(I)\le p_n} |\Psi_n(I) - \Psi(I)| &\le C(\sqrt{\log(n)p_n/n} + p_n\sqrt{\log(n)/n})\\
&\le 2C\sqrt{\log(n)p_n/n}.
\end{aligned}
$$

Due to the symmetry of $\Psi_n(I)$ the limes superior formulation of the convergence follows from Hewitt and Savage (1955). □

B Proof of Corollary 3

The boundedness of the Δ_i^x for each $x \in [A, B]$ and conditions (B1) and (B2) follow from the definition of Δ_i^x. This is so because the variables Δ_i^x do not depend on the x.

The consistency of the estimator $G_n(\cdot)$ (B4) can be easily shown and is a slight modification of the law of the iterated logarithm (see Shorack and Wellner, 1986, p. 504).

The assumption (A2) for $F(\cdot)$ implies that the cumulative hazard rate $\Lambda(\cdot)$ is strictly increasing and the hazard rate $\lambda(\cdot)$ is obviously strictly positive on $[A, B]$.

Now only the condition (B3) needs to be verified. We note that the vectors $\mathbf{S}_i = (X_i, L_i, \delta_i)_{i=1,\dots,n}$ are observable under $L_i \le X_i$. Hence, we derive the following conditional expectation:

$$\begin{aligned} &E\left(\frac{\mathbb{1}_{\{X_i \le x\}} \cdot \Delta_i^x}{G(X_i)} \mid L_i \le X_i\right) = E\left(\frac{\mathbb{1}_{\{X_i \le x, \delta_i = 1\}}}{G(X_i)} \mid L_i \le X_i\right) \\ =&\sum_{\delta_1=0}^{1} \int_{-\infty}^{\infty} \frac{\mathbb{1}_{\{x_1 \le x, \delta_1 = 1\}}}{G(x_1)} dF^{X,\delta}(x_1, \delta_1) = \int_{-\infty}^{x} \frac{dF^{X,\delta}(x_1, 1)}{G(x_1)}, \end{aligned} \tag{15}$$

where $F^{X,\delta}(x, y) = P(X \le x, \delta \le y \mid L \le X)$ is the conditional distribution function of (X, δ).

The integral $\int_{x_1 \in I} dF^{X,\delta}(x_1, 1)$ for the intervals $I := [a, b] \subseteq [A, B]$ can now be calculated. First we express the probability $P(X_i \in I, \delta_i = 1 \mid L_i \le X_i)$ in the terms of the non-observable vector (T_i, L_i, C_i) as follows:

$$\begin{aligned} &P(X_i \in I, \delta_i = 1 \mid L_i \le X_i) = \alpha^{-1} P(X_i \in I, \delta_i = 1, L_i \le X_i) \\ =&\alpha^{-1}[P(T_i \in I, T_i \le C_i, L_i \le T_i, T_i \le C_i) \\ +&P(C_i \in I, T_i \le C_i, L_i \le C_i, C_i < T_i)] = \alpha^{-1} P(T_i \in I, L_i \le T_i \le C_i), \end{aligned} \tag{16}$$

where $\alpha = P(L_i \leq X_i)$. Hence, we express the probabilities $P(X_i \in I, \delta_i = 1 \mid L_i \leq X_i)$ and $P(T_i \in I, L_i \leq T_i \leq C_i)$ as the following expectations of the Bernoulli-variables:

$$\begin{aligned} &P(X_i \in I, \delta_i = 1 \mid L_i \leq X_i) = E(\mathbb{1}_{\{X_i \in I, \delta_i = 1\}} \mid L_i \leq X_i) \\ &= \sum_{\delta_1=0}^{1} \int_{-\infty}^{\infty} \mathbb{1}_{\{x_1 \in I, \delta_1 = 1\}} dF^{X,\delta}(x_1, \delta_1) = \int_{x_1 \in I} dF^{X,\delta}(x_1, 1) \end{aligned} \tag{17}$$

and

$$\begin{aligned} &\alpha^{-1} P(T_i \in I, L_i \leq T_i \leq C_i) = \alpha^{-1} E(\mathbb{1}_{\{T_i \in I, L_i \leq T_i \leq C_i\}}) \\ &= \int_{t \in \mathbb{R}} \int_{c \in \mathbb{R}} \int_{l \in \mathbb{R}} \alpha^{-1} \mathbb{1}_{\{t \in I\}} \mathbb{1}_{\{l \leq t\}} \mathbb{1}_{\{t \leq c\}} dF(t) dF^C(c) dF^L(l) \\ &= \int_{t \in I} \alpha^{-1} F^L(t)(1 - F^C(t)) dF(t). \end{aligned} \tag{18}$$

One can see that $dF^{X,\delta}(x, 1) = \alpha^{-1} F^L(j)(1 - F^C(j)) dF(x)$ follows from the expressions (16), (17) and (18). Consequently the expectation (15) can be written as follows:

$$\begin{aligned} &E\left(\frac{\mathbb{1}_{\{X_i \leq x\}} \cdot \Delta_i^x}{G(X_i)} \mid L_i \leq X_i\right) = \int_{-\infty}^{x} \frac{dF^{X,\delta}(x_1, 1)}{G(x_1)} \\ &= \int_{-\infty}^{x} \frac{\alpha^{-1} F^L(x_1)(1 - F^C(x_1)) dF(x_1)}{G(x_1)} \\ &= \int_{-\infty}^{x} \frac{\alpha^{-1} F^L(x_1)(1 - F^C(x_1)) dF(x_1)}{\alpha^{-1} F^L(x_1)(1 - F^C(x_1))(1 - F(x_1))} = \int_{-\infty}^{x} \frac{dF(x_1)}{1 - F(x_1)} = \Lambda(x) = \Psi(x). \end{aligned}$$

Obviously, conditions (B1)-(B4) hold and the local convergence

$$P\left\{\limsup_{n \to \infty} \frac{\sup_{I \subseteq [A,B], \Lambda(I) \leq p_n} |\Lambda_n(I) - \Lambda(I)|}{\sqrt{\log(n) p_n / n}} = D\right\} = 1$$

follows for a constant $D \leq 2(\sqrt{2 \cdot (2M + \Lambda(B))} + 2M)$. □