A discrete-time, semi-parametric time-to-event model for left-truncated and right-censored data

Jackson P. Lautier[1] Vladimir Pozdnyakov[2] Jun Yan[2]

[1]Department of Mathematical Sciences, Bentley University
[2]Department of Statistics, University of Connecticut

New England Statistics Symposium
University of Connecticut
May 23, 2024

Table of Contents

BENTLEY UNIVERSITY

Asset-Backed Securities

Figure: A recent estimate of total issuance of asset-backed securities (ABS) in the U.S. securities market is a stunning $297,763.3 million (SIFMA, 2022).

Visualizing ABS cash-flows

Suppose an ABS of n loans is active for s months:

Loan (Age)	Month 1	Month 2	$\dots$	Month s
$\mathrm{L}_1(x_1)$	$CF_{1(x_1+1)}$	$CF_{1(x_1+2)}$	$\dots$	$CF_{1(x_1+s)}$
$\mathrm{L}_2(x_2)$	$CF_{2(x_2+1)}$	$CF_{2(x_2+2)}$	$\dots$	$CF_{2(x_2+s)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathrm{L}_n(x_n)$	$CF_{n(x_n+1)}$	$CF_{n(x_n+2)}$	$\dots$	$CF_{n(x_n+s)}$
ABS CF	$\sum_{j=1}^{n} CF_{j(x_j+1)}$	$\sum_{j=1}^{n} CF_{j(x_j+2)}$	$\dots$	$\sum_{j=1}^{n} CF_{j(x_j+s)}$

The ABS cash-flows are random variables that are heavily influenced by the time-to-termination probability distribution.

Application specific data challenges

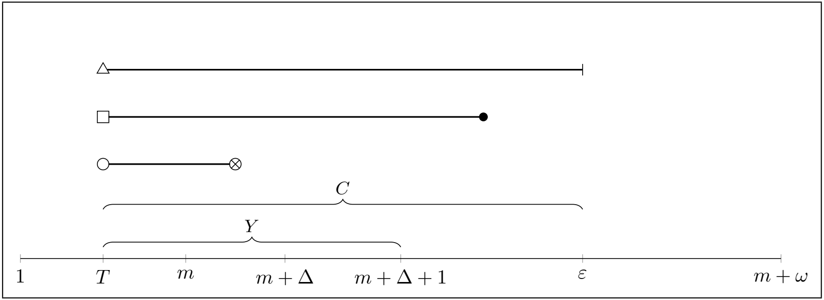

Figure: Asset-level lifetime data sampled from ABS will be subject to: left-truncation, right-censoring, and discrete-time over a known, finite support (i.e., a 72-month consumer auto loan). The triplet of left-truncation, discrete-time, and a known, finite support has received limited study (Lautier et al., 2023a).

Conditional bivariate sample space, $\mathcal{A}$

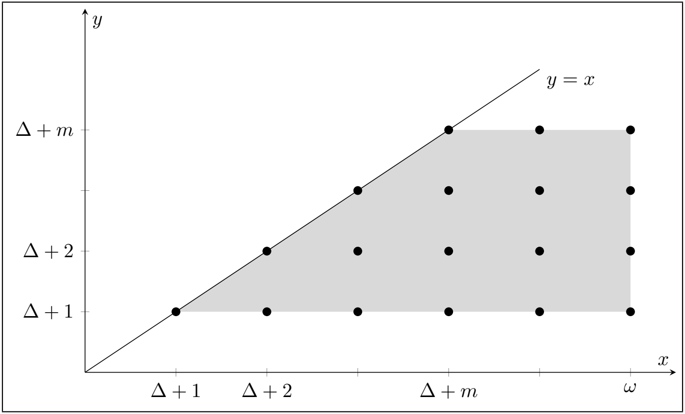

Figure: The conditional bivariate distribution between the left-truncation random variable, Y, and the lifetime random variable, X, is $h_*(u, v) = \Pr(X = u, Y = v \mid Y \leq X)$, for $(u, v) \in \mathcal{A}$. ABS loan-level data is sampled from h_*, and we seek to recover X.

A semi-parametric question

- Previous results treat h_* as a "parametric-non-parametric" distribution (Lautier et al., 2023a,b, 2024, E&S, IME, SPL).
- That is, the parameters of Y are $g(v)$, $\Delta + 1 \leq v \leq \Delta + m$, and the parameters of X are $f(u)$, $\Delta + 1 \leq u \leq \omega$. Hence, by $X \perp Y$,
$$h_*(u, v) = \frac{f(u)g(v)}{\alpha}, \quad (u, v) \in \mathcal{A},$$
where $\alpha = \Pr(Y \leq X)$.
- For economic modeling, it is desirable that X depends on economic variables. Hence, we consider the "semi-parametric"
$$h_*(u, v \mid p) = \frac{f(u \mid p)g(v)}{\alpha}, \quad (u, v) \in \mathcal{A}, p \in \mathcal{P}.$$

Table of Contents

Left-truncation: Estimating p

Given an i.i.d. sample of pairs of left-truncated observations, $\mathcal{S}_n = \{(X_i, Y_i)_{1 \le i \le n}\}$, it is of interest to estimate the parameters of h_*. From h_*, the likelihood is

$$\mathcal{L}(\boldsymbol{g}, p \mid \mathcal{S}_n) = \prod_{v=\Delta+1}^{\Delta+m} \prod_{u=v}^{\omega} \left[\frac{f(u \mid p) g_v}{\alpha}\right]^{\sum_{i=1}^{n} \mathbf{1}_{(X_i, Y_i)=(u,v)}},$$

where $\boldsymbol{g} = (g(\Delta+1), \ldots, g(\Delta+m))^\top \subset \boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{G}}$ is an m-dimensional hypercube over the unit interval, $\mathcal{I} = (0, 1)$. If we denote the convex subset,

$$\mathcal{C} = \left\{\mathcal{P} \times \boldsymbol{\mathcal{G}} : \sum_{v \in \mathcal{V}} g(v) = 1\right\} \subset \mathcal{P} \times \boldsymbol{\mathcal{G}},$$

then we seek

$$\sup_{p, \boldsymbol{g} \in \mathcal{C}} \mathcal{L}(\boldsymbol{g}, p \mid \mathcal{S}_n).$$

Theorem 1: Stationary points of $\mathcal{L}$ over $\mathcal{C}$

Let $\mathcal{S}_n$ be an i.i.d. sample of left-truncated observations from the distribution h_*. Then the stationary points of $\mathcal{L}(\boldsymbol{g}, p \mid \mathcal{S}_n)$ are

$$\hat{g}_v = \frac{\hat{h}_{\bullet v}}{S(v \mid \hat{p})} \left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{h}_{\bullet k}}{S(k \mid \hat{p})} \right]^{-1}, \quad v \in \mathcal{V},$$

where $S(\cdot)$ denotes the survival function,

$$S(x \mid p) := \Pr(X \geq x \mid p) = \sum_{u=x}^{\omega} f(u \mid p),$$

and $\hat{p}$ is any $p \in \hat{\mathcal{P}} \subset \mathcal{P}$, where

$$\hat{\mathcal{P}} = \left\{ \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{h}_{\bullet v}}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) = \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} \frac{\hat{h}_{uv}}{f(u \mid p)} \frac{\partial}{\partial p} f(u \mid p) \right\},$$

and

$$\hat{h}_{uv} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i, Y_i)=(u,v)}.$$

Further, $\hat{p} \in \mathcal{C}$ and $\hat{g}_v \in \mathcal{C}$ for all $v \in \mathcal{V}$.

Theorem 1 comments

- To our knowledge, $\hat{p}$ in Theorem 1 represents a new estimator.
- No restrictions on the distribution of Y (e.g., *length-bias sampling* requires Y to be uniform).
- Computational savings: the forms of $\hat{\mathcal{P}}$ and $\hat{g}$ reduce a multi-dimensional constrained optimization problem into a single-parametric optimization problem.
- General form of $f(\cdot \mid p)$ allows flexibility in choice of f.

Corollary 1.1: Stationary points of $\mathcal{L}$ over $\mathcal{C}$

Let $\mathcal{S}_n'$ be an i.i.d. sample of left-truncated observations from the distribution $h_*(u, v \mid \boldsymbol{p})$ subject to the same identifiability conditions of Theorem 1. Then the stationary points of $\mathcal{L}(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}_n')$ are

$$\hat{g}_v = \frac{\hat{h}_{\bullet v}}{S(v \mid \hat{\boldsymbol{p}})}\left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{h}_{\bullet k}}{S(k \mid \hat{\boldsymbol{p}})}\right]^{-1}, \quad v \in \mathcal{V},$$

where $\hat{\boldsymbol{p}}$ is any $\boldsymbol{p} \in \hat{\boldsymbol{\mathcal{P}}} \subset \boldsymbol{\mathcal{P}}$, with

$$\hat{\boldsymbol{\mathcal{P}}} = \{\boldsymbol{p} \in \boldsymbol{\mathcal{P}} : \xi_1(j) = \xi_2(j), \quad \text{for all} \quad j = 1, \dots, r\},$$

$$\xi_1(j) = \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{h}_{\bullet v}}{\sum_{u=v}^{\omega} f(u \mid \boldsymbol{p})}\right)\left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p_j} f(u \mid \boldsymbol{p})\right),$$

and

$$\xi_2(j) = \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} \frac{\hat{h}_{uv}}{f(u \mid \boldsymbol{p})} \frac{\partial}{\partial p_j} f(u \mid \boldsymbol{p}).$$

Further, $\hat{\boldsymbol{p}} \in \boldsymbol{\mathcal{C}}$ and $\hat{g}_v \in \boldsymbol{\mathcal{C}}$ for all $v \in \mathcal{V}$.

BENTLEY UNIVERSITY

Theorem 2: Equivalence of $\hat{\mathcal{P}}$

Assume the conditions of Theorem 1. Then $p \in \hat{\mathcal{P}}$,

$$\hat{\mathcal{P}} = \left\{ \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{h}_{\bullet v}}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) = \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} \frac{\hat{h}_{uv}}{f(u \mid p)} \frac{\partial}{\partial p} f(u \mid p) \right\},$$

if and only if

$$\frac{\partial}{\partial p} \frac{\prod_{v=\Delta+1}^{\Delta+m} S(v \mid p)^{\hat{h}_{\bullet v}}}{\prod_{u=\Delta+1}^{\omega} f(u \mid p)^{\hat{h}_{u\bullet}}} = 0,$$

where

$$\hat{h}_{\bullet v} := \sum_{u=v}^{\omega} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i, Y_i)=(u,v)} \right),$$

and

$$\hat{h}_{u\bullet} := \sum_{v=\Delta+1}^{\min(u,\Delta+m)} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i, Y_i)=(u,v)} \right).$$

Theorem 3: MLE of $\boldsymbol{g}$, p, RT geometric

Define the right-truncated geometric distribution with parameter, $0 < p < 1$, as

$$f_T(u \mid p) = \begin{cases} p(1-p)^{u-(\Delta+1)} & \Delta+1 \leq u \leq \omega-1, \\ (1-p)^{u-(\Delta+1)} & u = \omega. \end{cases}$$

Then, for the conditional bivariate probability mass function, h_*, under the sampling conditions of Theorem 1, the MLE of the parameter p is

$$\hat{p}_{\text{MLE}} = \frac{b}{b-a},$$

where

$$a = \sum_{v=\Delta+1}^{\Delta+m} \{v-(\Delta+1)\}\hat{h}_{\bullet v} - \sum_{u=\Delta+1}^{\omega} \{u-(\Delta+1)\}\hat{h}_{u\bullet},$$

and

$$b = \sum_{u=\Delta+1}^{\omega-1} \hat{h}_{u\bullet}.$$

Further, the MLE of $\boldsymbol{g}$ is

$$\{\hat{g}_{v,\text{MLE}}\}_{v\in\mathcal{V}} = \hat{h}_{\bullet v}\left(1-\frac{b}{a}\right)^{v-(\Delta+1)}\left[\sum_{k=\Delta+1}^{\Delta+m} \hat{h}_{\bullet k}\left(1-\frac{b}{a}\right)^{k-(\Delta+1)}\right]^{-1}.$$

Right-censoring: Estimating p

From Lautier et al. (2023b), define the right-censoring random variable, $C = Y + \varepsilon - (m + \Delta + 1) \equiv Y + \tau$ (note: $C \perp X$) The observed data takes the triple $\mathcal{S}_{\tau,n} \equiv \{Y_i, Z_i, D_i\}_{1 \le i \le n}$, where $Z_i = \min(X_i, C_i)$ and $D_i = 1$ if $X_i \le C_i$ and 0 otherwise. Thus, the likelihood for $\mathcal{S}_{\tau,n}$ becomes

$$\mathcal{L}_\tau(\boldsymbol{g}, p \mid \mathcal{S}_{\tau,n}) = \prod_{\{\mathcal{S}_{\tau,n}: D_i = 1\}} \frac{g(Y_i) f(Z_i \mid p)}{\alpha} \prod_{\{\mathcal{S}_{\tau,n}: D_i = 0\}} \frac{g(Y_i) S(Z_i + 1 \mid p)}{\alpha}$$
$$= \alpha^{-n} \prod_{v=\Delta+1}^{m+\Delta} g(v)^{n \hat{\gamma}_n(v)} \prod_{i=1}^{n} f(Z_i \mid p)^{D_i} S(Z_i + 1 \mid p)^{1-D_i},$$

where

$$\hat{\gamma}_n(v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i = v).$$

As with Theorem 1, we seek

$$\sup_{p, \boldsymbol{g} \in \mathcal{C}} \mathcal{L}_\tau(\boldsymbol{g}, p \mid \mathcal{S}_{\tau,n}).$$

BENTLEY UNIVERSITY

Theorem 4: Stationary points of $\mathcal{L}_\tau$ over $\mathcal{C}$

Let $\mathcal{S}_{\tau,n}$ be an i.i.d. sample of left-truncated observations from the distribution h_* under the additional incomplete data setting of right-censoring. Assume the identifiability conditions of Theorem 1. Then the stationary points of $\mathcal{L}_\tau(\boldsymbol{g}, p \mid \mathcal{S}_{\tau,n})$ are

$$\hat{g}_\tau(v) = \frac{\hat{\gamma}_n(v)}{S(v \mid \hat{p}_\tau)} \left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{\gamma}_n(k)}{S(k \mid \hat{p}_\tau)} \right]^{-1}, \quad v \in \mathcal{V},$$

where $S(\cdot)$ denotes the survival function defined in Theorem 1, and $\hat{p}_\tau$ is any $p \in \hat{\mathcal{P}}_\tau \subset \mathcal{P}$ where

$$\hat{\mathcal{P}}_\tau = \Bigg\{ p \in \mathcal{P} : \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{\gamma}_n(v)}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right)$$
$$= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{f(Z_i \mid p)} \frac{\partial}{\partial p} f(Z_i \mid p) + \frac{1 - D_i}{S(Z_i + 1 \mid p)} \frac{\partial}{\partial p} S(Z_i + 1 \mid p) \right) \Bigg\}.$$

Further, $\hat{p}_\tau \in \mathcal{C}$ and $\hat{g}_\tau(v) \in \mathcal{C}$, for all $v \in \mathcal{V}$.

Theorem 4 comments

- To our knowledge, $\hat{p}_\tau$ in Theorem 4 represents a new estimator.
- The ability to handle right-censoring greatly expands potential applications.
- No restrictions on the distribution of Y (e.g., *length-bias sampling* requires Y to be uniform).
- Computational savings: the forms of $\hat{\mathcal{P}}_\tau$ and $\hat{g}_\tau$ reduce a multi-dimensional constrained optimization problem into a single-parametric optimization problem.
- General form of $f(\cdot \mid p)$ allows flexibility in choice of f.
- The equivalent to Corollary 1.1 may be shown (i.e., $\boldsymbol{p}_\tau$) but is omitted from this talk for brevity.

Corollary 4.1: MLE of $\boldsymbol{g}$, p_τ, RT geometric, right-censoring

Recall the right-truncated geometric distribution with parameter, $0 < p < 1$, defined in Theorem 3. Then, for the conditional bivariate probability mass function, h_*, under the sampling conditions of Theorem 4, the MLE of the parameter p is

$$\hat{p}_{\tau,\mathrm{MLE}} = \frac{b_\tau}{b_\tau - a_\tau},$$

where

$$a_\tau = \sum_{v=\Delta+1}^{\Delta+m} \{v-(\Delta+1)\}\hat{\gamma}_n(v) - \frac{1}{n}\sum_{i=1}^{n}\left(\{Z_i-(\Delta+1)\}D_i + \{Z_i+1-(\Delta+1)\}(1-D_i)\right),$$

and

$$b_\tau = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i \neq \omega)D_i.$$

Further, the MLE of $\boldsymbol{g}$ is

$$\{\hat{g}_{\tau,\mathrm{MLE}}(v)\}_{v\in\mathcal{V}} = \hat{\gamma}_n(v)\left(1-\frac{b_\tau}{a_\tau}\right)^{v-(\Delta+1)}\left[\sum_{k=\Delta+1}^{\Delta+m}\hat{\gamma}_n(k)\left(1-\frac{b_\tau}{a_\tau}\right)^{k-(\Delta+1)}\right]^{-1}.$$

BENTLEY UNIVERSITY

Table of Contents

Two Illustrations

Let $m = 3$, $\Delta = 0$, and $\omega = 4$. Hence, the bivariate distribution h_* is a 4×3 trapezoid with nine possible combinations (see next slide). For the left-truncation random variable, Y, we assume $g(1) = 0.5$, $g(2) = 0.3$, and $g(3) = 0.2$. We consider:

(1) **Theorem 1**: Set $\varepsilon = 7 = \omega + m$ (no right-censoring) and $X \sim \text{Binom}(\omega - 1 \in \mathbb{Z}, 0 < \theta = 0.3 < 1)$. That is,

$$f(u \mid \theta) = \binom{3}{u-1}\theta^{u-1}(1-\theta)^{3-(u-1)}, \quad 1 \leq u \leq 4.$$

(2) **Corollary 4.1** Set $\varepsilon = 6 \implies \tau = \varepsilon - (m + \Delta + 1) = 2$ (right-censoring is present) and $X \sim f_T(p = 0.6)$.

Simulation study sample space

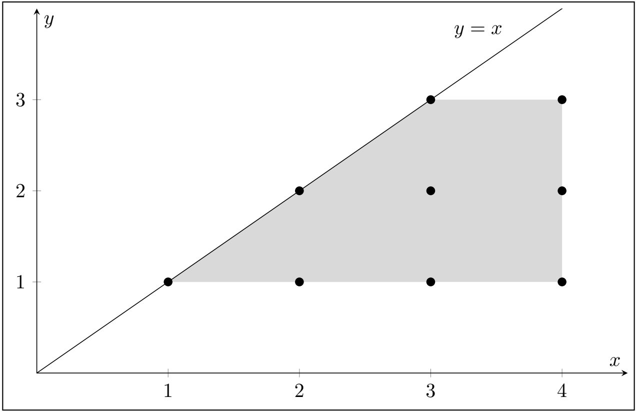

Figure: Visualization of the simulation study sample space.

Results summary

Parameter	Actual	`constrOptim`	Speed (Ns)	Theorem 1	Speed (Ns)
θ	0.30	0.3165660	1554.602	0.3165729	3.539
$g(1)$	0.50	0.5114408		0.5114206	
$g(2)$	0.30	0.2934628		0.2934616	
$g(3)$	0.20	0.1951772		0.1951178	

Parameter	Actual	`constrOptim`	Speed (Ms)	Corollary 4.1	Speed (Ms)
p	0.60	0.5992329	1331.172	0.5991903	6.596
$g(1)$	0.50	0.4652415		0.4655774	
$g(2)$	0.30	0.2972558		0.2975195	
$g(3)$	0.20	0.2367859		0.2369030	

Table: **Numeric Validation and Performance Summary**. Sample sizes $n = 982$ (top) and $n = 983$ (bottom). Direct multidimensional optimization (`constroOptim` via R Core Team (2023)). The performance calculations were measured with the `microbenchmark` package (Mersmann, 2023) (reported times approximate).

BENTLEY UNIVERSITY

Table of Contents

BENTLEY UNIVERSITY

Ally Auto Receivables Trust 2017-3

- We consider a subset of $n = 151$ 25-month consumer auto loans from the Ally Auto Receivables Trust 2017-3 securitized bond (Ally, 2017).
- The time-to-event of interest is the time-until-monthly-payments stop (either default or prepayment).
- Loans with observed termination times beyond 26 months (i.e., 27, 28, and 29 months) were treated as full-term 26 month loans. Such an adjustment has minimal practical significance.
- For this data, $\Delta = 3$, $m = 21$, $\omega = 26$, and $\varepsilon = 67 \implies \tau = 42$ (and thus no right-censoring).
- There are thus 21 parameters to estimate, which limits the effectiveness of computational approaches.

Model fitting results

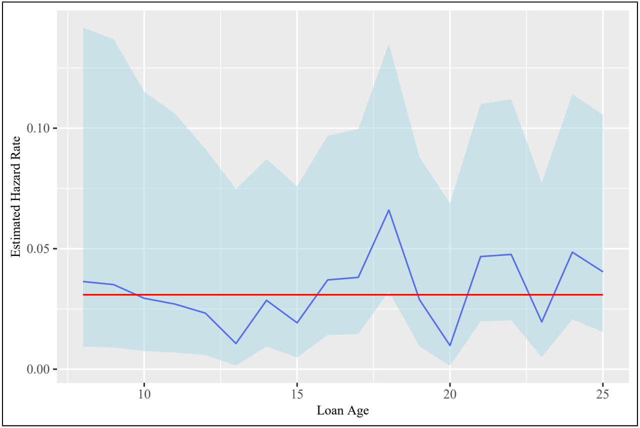

Figure: A comparison of the "non-parametric-parametric" approach of Lautier et al. (2023b) with 95% confidence intervals (blue line + ribbon) of the hazard rate to $\hat{p}_{\text{MLE}} = 0.0309$ using Corollary 4.1. A chi-square goodness of fit test results in a p-value of 0.3271.

Table of Contents

BENTLEY UNIVERSITY

Conclusion

- We propose a new estimator for discrete lifetime data with a known, finite support under incomplete data (left-truncation & right-censoring).
- It does not require any assumptions about the left-truncation random variable (i.e., *length-biased sampling*) and offers computational savings.
- For a right-truncated geometric distribution, appropriate for consumer loan analysis, we derive the MLE for the parameter, p. All results verified numerically.
- We illustrate our results with the Ally Auto Receivables Trust 2017-3 securitized bond.
- Next, we return to the original question: can we link p to a set of economic variables?

Thank you!

Jackson P. Lautier, PhD, FSA, CERA, MAAA
e: jlautier@bentley.edu
w: www.jacksonlautier.com

Ally (2017). "Ally Auto Receivables Trust." Prospectus 2017-3, Ally Auto Assets LLC.

J. P. Lautier, V. Pozdnyakov and J. Yan (2023a). "Estimating a discrete distribution subject to random left-truncation with an application to structured finance." *Econometrics and Statistics* Forthcoming.

J. P. Lautier, V. Pozdnyakov and J. Yan (2023b). "Pricing time-to-event contingent cash flows: A discrete-time survival analysis approach." *Insurance: Mathematics and Economics* **110**, 53–71.

J. P. Lautier, V. Pozdnyakov and J. Yan (2024). "On the maximum likelihood estimation of a discrete, finite support distribution under left-truncation and competing risks." *Statistics & Probability Letters* **207**, 109973.

O. Mersmann (2023). *microbenchmark: Accurate Timing Functions*. R package version 1.4.10.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

SIFMA (2022). "US ABS securities: Issuance, trading volume, outstanding." https://www.sifma.org/resources/research/us-asset-backed-securities-statistics/. Online; accessed 24 February 2022.