Estimating the time-to-event distribution for loan-level data within an asset-backed security

Jackson P. Lautier*† Vladimir Pozdnyakov‡ Jun Yan‡

November 6, 2024

Abstract

The random cash flows of asset-backed securities (ABS) depend critically on the time-to-event distribution of the individual, securitized assets. Estimating this distribution has historically been challenged by limited data. Recent regulatory changes reversed this, however, and asset-level ABS data is now publicly available to investors for the first time. The idiosyncrasies of this ABS data present new difficulties in estimating the loan-level lifetime distribution due to its discrete-time structure, finite support, and exposure to left-truncation. We propose a parametric framework for estimating the loan-level lifetime distribution while leaving the left-truncation time distribution unspecified. Through theorems developed to identify the stationary points of the likelihood, we significantly simplify a complex multiparameter constrained optimization problem. These stationary points, shown to be the roots of an estimating equation, enable asymptotic normality and large-sample inference to follow. Assuming a geometric distribution combined with an actuarial policy limit, we derive closed-form maximum likelihood estimates. These theoretical results are further generalized to accommodate right-censoring and validated through numerical and simulation studies. We then estimate the loan-level lifetime distributions for two Ally Auto Receivables Trust ABS bonds. The efficient and accurate estimates we find for these bonds offer valuable insights to structured finance investors.

Keywords: Asset-level disclosures, Incomplete data, Reg AB II, Risk management, Survival analysis

1 Introduction

If we treat financial engineering as a subclass of engineering more broadly, it is not unreasonable to argue that securitization is a marvel that would rival any bridge, railway, or

*Department of Mathematical Sciences, Bentley University, Waltham, Massachusetts, USA

†Corresponding to Jackson P. Lautier, Morison Hall, 175 Forest Street Waltham, MA 02452, USA; email: jlautier@bentley.edu.

‡Department of Statistics, University of Connecticut, Storrs, Connecticut, USA.

expressway. Consider the financial dilemma of an automobile manufacturer. The manufacturer has the main objective of selling cars. A hurdle to most potential consumers, however, is the car's prohibitive cost. Thus, most consumers will require financing in the form of an auto loan, and auto *financiers* or lenders (e.g., Ally Bank) step in to provide it. Eventually, after writing many auto loans, the lender will run into a duration cash flow mismatch on their balance sheet. The lender desires cash immediately to write more loans, but the majority of its assets are previously written, long-dated (e.g., 72-month) auto loans. Investors, conversely, have cash to invest and seek to earn returns over a longer horizon. Insert the asset-backed security (ABS) to make this connection, just as a bridge, railway, or expressway connects allied geographic regions with a shared economic interest.

Legally, the lender will collect paying consumer auto loans into a separate entity or trust (e.g., AART, 2017). This trust removes the direct financial interest of the lender from the performance of the individual auto loans, and it simultaneously removes any financial dependence of prospective investors on the lender's financial health. Both legal separations benefit the auto financier; the former creates a tradeable asset that can be marketed to investors, and the latter achieves a lower financing cost than a traditional corporate bond. For investors, this collection of loans into an ABS trust (henceforth, ABS), is now capable of generating a large, long duration monthly cash flow that may be bought and sold on the open market as an investment security (hence the name, *securitization*).

It is informative to visualize the investor level cash flows of an ABS. Each ABS will contain n consumer auto loans of various ages, x_i, $1 \leq i \leq n$. Each loan at age x_i, $L_i(x_i)$, $1 \leq i \leq n$, will then produce a monthly cash flow, either a loan payment, a prepayment, or a zero in the event of default or scheduled termination. For any month j the securitization is active, $1 \leq j \leq s$, a column-wise sum of these *loan-level* cash flows, $\mathrm{CF}_{i(x_i+j)}$, $1 \leq i \leq n$, $1 \leq j \leq s$, supports the cash flow that is ultimately returned to investors. This total investor cash flow, $\sum_i \mathrm{CF}_{i(x_i+1)}$ feeds a payment *waterfall*, denoted by f_w. The waterfall allows the ABS to be marketed to investors with differing risk appetites via its *tranches* (AART, 2017).

Loan (Age)	Month 1	Month 2	...	Month s
$\mathrm{L}_1(x_1)$	$\mathrm{CF}_{1(x_1+1)}$	$\mathrm{CF}_{1(x_1+2)}$	...	$\mathrm{CF}_{1(x_1+s)}$
$\mathrm{L}_2(x_2)$	$\mathrm{CF}_{2(x_2+1)}$	$\mathrm{CF}_{2(x_2+2)}$	...	$\mathrm{CF}_{2(x_2+s)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathrm{L}_n(x_n)$	$\mathrm{CF}_{n(x_n+1)}$	$\mathrm{CF}_{n(x_n+2)}$	...	$\mathrm{CF}_{n(x_n+s)}$
ABS CF	$f_w\left(\sum_{i=1}^{n}\mathrm{CF}_{i(x_i+1)}\right)$	$f_w\left(\sum_{i=1}^{n}\mathrm{CF}_{i(x_i+2)}\right)$	...	$f_w\left(\sum_{i=1}^{n}\mathrm{CF}_{i(x_i+s)}\right)$

Table 1: **Visualizing ABS Investor Cash Flows**. The monthly investor level cash flow, ABS CF, is generated by a sum of the loan-level cash flows, denoted $\mathrm{CF}_{i(x_i+j)}$ for consumer loan i, at initial age x_i, $L(x_i)$, $1 \leq i \leq n$, for month j, $1 \leq j \leq s$. The column-wise sum total loan-level cash flow, $\sum_i \mathrm{CF}_{i(x_i+j)}$ at month j is then fed into a payment waterfall, denoted f_w. Given this, investors often desire to model the time-to-termination random variable of the individual loans.

This monthly process repeats until the ABS is wound down, which we assume is s total months. This ABS investor cash flow generation process is summarized in Table 1.

Because the time-to-termination of each individual loan within an ABS is a random variable, the monthly investor ABS cash flow is a random variable. Hence, estimating the loan-level time-to-termination distribution is of significant interest to ABS investors (e.g., Lautier et al., 2023b; Agarwal et al., 2024). Critical to this task is access to loan-level data, which has historically been limited. This lack of transparency in ABS was deemed unacceptable after investors failed to anticipate higher than expected defaults for residential subprime mortgages packaged into ABS, triggering the 2007-2009 financial crisis (Mishkin, 2011). In response, regulators implemented several measures to improve transparency in financial markets. These include the Securities and Exchange Commission's (SEC) significant revisions to Regulation AB and new rules governing ABS disclosures (Securities and Exchange Commission, 2014). Notably, ABS issuers must now provide detailed loan-level data, including payment performance, on a monthly basis via the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (Securities and Exchange Commission, 2016).

This new, unprecedented public data access can inform the vital time-to-termination distribution estimation underpinning ABS cash flow modeling. Using this data is not with-

out its complexity, however. One major challenge is incomplete data, particularly in the form of left-truncation (Lautier et al., 2023a; Katcher et al., 2024). Beyond left-truncation, ABS loan data, like most financial loan data, also present challenges related to its discrete-time structure. While left-truncation in continuous time has been extensively studied (e.g., Woodroofe, 1985; Tsai et al., 1987; Wang, 1989; Huang and Wang, 1995), the combination of left-truncation, discrete-time, and a known, finite support for the lifetime random variable remains relatively unexplored (Lautier et al., 2023a). Because this combination is necessary to accurately estimate and model the time-to-termination random variable for the ABS setting (see Section 2), it is important to contribute to this nascent line of work.

Recent approaches have leveraged the finite, known support of the loan lifetime random variable to estimate each recoverable probability point mass in the distribution as a parameter (Lautier et al., 2023a). This allows direct proofs that the classical nonparametric estimators of Woodroofe (1985) and Huang and Wang (1995) are simultaneously parametric maximum likelihood estimates (MLE) in this setting (Lautier et al., 2023a, 2024). It also yields completely specified asymptotic multivariate normal distributions with a diagonal covariance structure for the vector of hazard rate estimators (Lautier et al., 2022, 2023a,b). A further benefit of this line of work is the assumption of a sample of fixed size n from a distribution formed after the impact of left-truncation, which more closely reflects the data generation process of loan-level data within an ABS pool (e.g., Lautier et al., 2023a). In contrast, classical treatments (e.g., Woodroofe, 1985) assumes a bivariate sample of size N, from which a left truncated random sample of size $n \leq N$ is drawn. This is akin to investors unrealistically having access to all loans, including those not in the ABS. Additionally, Lautier et al. (2023a) impose no assumptions about the shape of the left-truncation random variable, making it less restrictive than methods that require a uniformity assumption (e.g., *length-biased sampling* Asgharian et al., 2002; De Uña-Álvarez, 2004).

Despite the aforementioned advances obtained by treating each probability point mass of the lifetime distribution as a parameter (Lautier et al., 2022, 2023a,b, 2024), alternatively

using a traditional parametric distribution with fewer parameters would offer distinctive advantages. First, parametric distributions provide a natural smoothing, whereas the approach of Lautier et al. (2023a) can yield zero estimates at some months in the recoverable space. Second, parametric forms allow for the incorporation of economic variables, facilitating an analysis of their impact on loan survival times. Finally, given the need for rapid decision-making in investment contexts, a parametric approach enables quick and concise estimates via a single parameter and its interval estimate.

We thus present a novel parametric framework for analyzing left-truncated, discrete time-to-event data, with a particular focus on ABS data. We simplify a complex multidimensional constrained optimization problem into a single-parameter optimization (Theorem 3.1), significantly reducing computational complexity. This result is generalized to a vector of parameters (Corollary 3.1.1) and provides asymptotically consistent M-estimators (e.g., van der Vaart, 1998), ensuring robust inference for lifetime distribution parameters. We further derive closed-form maximum likelihood estimates (MLE) for a policy limit geometric distribution (Klugman et al., 2012, §8.4, pg. 125). All results also accommodate right-censoring (Section 4). The proposed methods are validated numerically through simulation studies and applied to loan-level data from the ABS bonds AART (2017) and AART (2019). While the focus is ABS, the framework is also applicable where incomplete, discrete time-to-event data is prevalent, such as healthcare, finance, engineering, telecommunication, and insurance.

The paper proceeds as follows. Section 2 establishes preliminaries within the context of the ABS application. Section 3 states the results under left-truncation. It is the theoretical foundation of this work and ABS setting. Section 4 then generalizes all of the results of Section 3 to the incomplete data case of both left-truncation and right-censoring. Section 5 includes a numerical verification of all results and a set of robustness simulation studies. Section 6 is a complete application of our methods to estimating the time-to-termination loan-level random variable for the consumer automobile ABS bonds AART (2017) and AART (2019), including how such estimates may be utilized by ABS investors. The paper concludes

in Section 7. The Supplemental Material provides complete proofs of all major results and references to support likelihood construction, implementation, and simulation. For reference, all data and replication code is publicly available at [...(anonymized)].

2 Asset-Backed Security Data

We now introduce notation and assumptions within the context of estimating the parameters of the loan-level time-to-termination random variable via ABS data. We denote this time-to-event random variable by X. Because the lifetime of interest is a financial instrument with monthly payments, most commonly a loan, it is a discrete random variable over $\mathbb{N}$. Further, the amortization schedule is known at contract signing, so X is known to have a finite upper bound, denoted by the nonrandom $\omega \in \mathbb{N}$. For example, for a 72-month auto loan, $\omega = 72$.

Left-truncation manifests in the data generating process of ABS loan lifetimes through the lengthy legal machinations of its formation. It can take several years for this process to complete (known colloquially as a *warehousing period*) because it takes time for ABS pools to amass the tens of thousands of loans they commonly contain (e.g., AART, 2017). It is thus possible loans will terminate during the warehousing period and never be observed by ABS investors. Hence, loan-level time-to-event data present in an ABS is conditional on survival beyond a second time-to-event random variable, denoted Y. The random variable Y represents the random number of months from loan origination to the first month an ABS trust begins paying investors. In other words, we observe X if and only if $X \geq Y$. That is, Y is a left-truncation random variable for the loan lifetime random variable, X.

We now formalize the support of X and Y. As with X, Y is a discrete random variable with a known, finite support. Loans that will eventually be included in the ABS will be originated for an observable total of nonrandom $m \in \mathbb{N}$ months (Lautier et al., 2023b). Once the trust closes to new loans, there will be a second observable period of nonrandom $\Delta \in \mathbb{N} \cup \{0\}$ months that the ABS is marketed to investors. Formally, then, the left-

truncation random variable, Y, is finite, discrete with support $v \in \mathcal{V} \equiv \{\Delta+1, \ldots, \Delta+m\}$. Because we observe X if and only if $X \geq Y$, X is a finite, discrete random variable with support $u \in \mathcal{U} \equiv \{\Delta+1, \ldots, \omega\}$ (it is assumed $\omega \geq \Delta+m$). As a minor technical point, X may not be completely *recoverable* (Woodroofe, 1985; Lautier et al., 2023a) (i.e., if $\Delta \geq 1$).

We further assume X and Y are independent, which is vital to the subsequent analysis and thus warrants discussion. The length of the warehousing period of ABS (i.e., Y) is generally dictated by the financing needs of the lender and prevailing market rates. In other words, a lender relies on ABS to exchange long-dated assets for capital now to write more loans. This need and its current cost will supersede any changes to the loan lifetime random variable, X. Furthermore, it is the legal purpose of ABS to separate a financier's vested interest in individual loan performance (i.e., X) from current market operations. Hence, we find the assumed independence of X and Y justifiable within an application to ABS data. For greater discussion, see Lautier et al. (2023b). More broadly, the independence between X and Y is sometimes supplemented by assuming Y is a uniform random variable (i.e., *length-biased sampling*) (e.g., Asgharian et al., 2002; De Uña-Álvarez, 2004). Loan originations are often cyclical (e.g., auto markets (Lautier et al., 2023a)) and not uniform throughout the year, however, and so the flexibility of Y we allow for herein is further motivated by the ABS application. This left-truncation setting is our theoretical foundation (see Section 3).

Beyond left-truncation, estimating the parameters of X from ABS data in practice will likely also require accounting for random right-censoring. ABS investors interested in modeling cash flows from an active ABS bond, for example, will need to construct estimates for X from many observations known to still be making ongoing monthly payments with yet unknown termination times. This is the incomplete data setting of right-censoring, and it will complicate the left-truncation setting further (Lautier et al., 2023b). To formalize, denote $\varepsilon \in \mathbb{N}$, $m+\Delta \leq \varepsilon \leq m+\omega$ to be the present time of the data generation process of an active ABS. If $\varepsilon < m+\omega$, then random right-censoring is present in the data (Lautier et al., 2023b). Specifically, the exact termination time is observed if and only if $X \leq Y+\varepsilon-(m+\Delta+1)$.

In other words, if we define the random variable $C = Y + \varepsilon - (m + \Delta + 1) \equiv Y + \tau$, then the random right-censoring time is a linear shift of Y. This is convenient because C is a linear function of Y and $C \geq Y$ almost surely (Lautier et al., 2023b), and so C is trivially independent of X. The right-censoring and left-truncation results reside in Section 4.

The final important preliminary point is that we assume throughout that X is dependent on the parameter, $p \in \mathcal{P}$, where $\mathcal{P}$ is a convex interval of $\mathbb{R}$. (The parameter $p \in \mathcal{P}$ need not be a scalar, though we assume so at present for ease of exposition.) Therefore, we focus all remaining effort on accurately estimating p from ABS data (i.e., subject to left-truncation, discrete-time, a known, finite support, and potentially right-censoring). We desire to estimate p because it will completely specify X, and it is from X that the random ABS cash flows may be modeled (recall Table 1 or see Lautier et al. (2023b)). Furthermore, efficient and accurate estimation of p will help ABS investors gain rapid insights into the performance of the individual loans. These insights will help with investment allocation decisions and risk assessment (see Section 6). Our methods begin with left-truncation in Section 3.

3 Left-Truncation

We consider first the left-truncation ABS setting under the notation introduced in Section 2. Denote the probability mass function (pmf) of X by $f(u \mid p)$, $u \in \mathcal{U}$, $p \in \mathcal{P}$ and denote the pmf of Y by $g(v)$, $v \in \mathcal{V}$. The distribution Y is itself a parametric distribution, where each point of the probability point mass function, $g(v)$, $v \in \mathcal{V}$, may be represented by a parameter, denoted g_v, $v \in \mathcal{V}$, $0 \leq g_v \leq 1$, under the constraint $\sum_{\mathcal{V}} g_v = 1$. For convenience of notation, therefore, we will henceforth drop the $g(v)$ representation and use only g_v, $v \in \mathcal{V}$.

By the assumed independence of X and Y, we obtain the conditional bivariate pmf,

$$h_*(u, v \mid p) = \Pr(X = u, Y = v \mid Y \leq X, p) = \frac{f(u \mid p) g_v}{\alpha}, \quad p \in \mathcal{P}, (u, v) \in \mathcal{A}, \tag{1}$$

where

$$\alpha \equiv \Pr(Y \leq X) = \sum_{u=\Delta+1}^{\omega} f(u \mid p) \left(\sum_{v=\Delta+1}^{\min(u,\Delta+m)} g_v \right) = \sum_{v=\Delta+1}^{\Delta+m} g_v \left(\sum_{u=v}^{\omega} f(u \mid p) \right), \tag{2}$$

and $\mathcal{A} = \{\mathcal{U} \times \mathcal{V} : v \leq u\}$. The distribution h_* is a parametric distribution with complete parameter vector $\boldsymbol{\Theta} = (p, \boldsymbol{g})^\top$, where $p \in \mathcal{P}$, $\boldsymbol{g} = (g_{\Delta+1}, \ldots, g_{\Delta+m})^\top \subset \boldsymbol{\mathcal{G}}$, and $\boldsymbol{\mathcal{G}}$ is an m-dimensional hypercube over the unit interval, $\mathcal{I} = (0, 1)$.

Given an independent and identically distributed (i.i.d.) sample of pairs of left-truncated observations, $\mathcal{S}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$, it is of interest to estimate the parameters of h_*. From (1) and (2), the likelihood is

$$\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{S}_n) = \prod_{v=\Delta+1}^{\Delta+m} \prod_{u=v}^{\omega} \left[\frac{f(u \mid p) g_v}{\alpha} \right]^{\sum_{i=1}^{n} \mathbf{1}_{(X_i, Y_i)=(u,v)}}. \tag{3}$$

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

$$\sup_{p, \boldsymbol{g} \in \mathcal{C}} \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{S}_n). \tag{4}$$

An approach to solve (4) without the assistance of computational programming is not immediate. Further, as the parameter space grows in dimension, performing the multidimensional constrained optimization numerically can become computationally demanding, complex in its implementation, and even potentially unfeasible (see, for example, the field of *large-scale linearly constrained optimization* problems, (e.g., Murtagh and Saunders, 1978)). We now show (4) may be reduced to a single-parameter optimization problem.

Theorem 3.1 (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 $h_*(u, v \mid p)$ *defined in* (1) *such that*

$$\hat{h}_{\bullet v} \coloneqq \sum_{u=v}^{\omega} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i,Y_i)=(u,v)} \right) > 0, \quad \textit{and} \quad \hat{h}_{u\bullet} \coloneqq \sum_{v=\Delta+1}^{\min(u,\Delta+m)} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i,Y_i)=(u,v)} \right) > 0.$$

Further assume $\partial f(u \mid p)/\partial p$ *exists and is finite for all* $p \in \mathcal{P}$, $u \in \mathcal{U}$. *Then the stationary points of* $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{S}_n)$ *are*

$$\hat{g}_v = \frac{\hat{h}_{\bullet v}}{S(v \mid \hat{p})} \left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{h}_{\bullet k}}{S(k \mid \hat{p})} \right]^{-1}, \quad v \in \mathcal{V}, \tag{5}$$

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

$$S(x \mid p) \coloneqq \Pr(X \geq x \mid p) = \sum_{u=x}^{\omega} f(u \mid p), \tag{6}$$

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

$$\hat{\mathcal{P}} = \left\{ \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{h}_{\bullet v}}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) = \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} \frac{\hat{h}_{uv}}{f(u \mid p)} \frac{\partial}{\partial p} f(u \mid p) \right\}, \tag{7}$$

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* $v \in \mathcal{V}$.

Proof. See the Supplemental Material, Section A.1. □

Remark. *The conditions* $\hat{h}_{\bullet v} > 0$ *(i.e., each empirical "row sum" of* $\mathcal{A} > 0$*) and* $\hat{h}_{u\bullet} > 0$ *(i.e., each empirical "column sum" of* $\mathcal{A} > 0$*) in Theorem 3.1 are necessary to avoid vacuous identifiability concerns for the boundaries of* $\mathcal{A}$.

The solution space (7) is a general form of an estimator for the parameter $p \in \mathcal{P}$ under the setting of Theorem 3.1 not yet derived to our knowledge. Conveniently, it, along with the closed-form solutions (5), reduce the multidimensional problem of (4) to a single dimension

problem. Aside from reducing the problem's computational complexity, it can also reduce computational demands, especially as the dimension of $\boldsymbol{g}$ grows. A large parametric space for $\boldsymbol{g}$ is common when the lifetime of interest is consumer automobile monthly loan payments from ABS data. For example, a subset of data from the ABS bond AART (2017) that we consider in Section 6 generates $\mathcal{V} = \{4, \ldots, 24\}$, which requires us to estimate 21 total parameters (including p). Without the dimension reduction of Theorem 3.1, solving (4) for this data set would not be a trivial computational task.

It may be desirable to allow f to depend on a finite, r-dimensional parameter vector $\boldsymbol{p} \equiv (p_1, \ldots, p_r)^\top \subset \boldsymbol{\mathcal{P}}$, where $r < (\omega - \Delta)$ and $\boldsymbol{\mathcal{P}}$ is an r-dimensional convex set of $\mathbb{R}^r$. This allows for greater flexibility in modeling the lifetime distribution. In this setting, an equivalent yet more general form of the estimator (7) in Theorem 3.1 may also be found, and we now provide discussion for completeness. Before doing so, we establish the following preliminaries. Denote the pmf of X by $f(u \mid \boldsymbol{p})$, $u \in \mathcal{A}$. The equivalent notation for the conditional bivariate pmf in (1) then becomes $h_*(u, v \mid \boldsymbol{p})$, $(u, v) \in \mathcal{A}$. Let $\mathcal{S}'_n = \{(X_i, Y_i)\}_{1 \le i \le n}$ be an i.i.d. sample of left-truncated observations from the distribution $h_*(u, v \mid \boldsymbol{p})$. The multidimensional, constrained optimization problem is then to find $\sup \mathcal{L}(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}'_n)$ such that $(\boldsymbol{p}, \boldsymbol{g}) \in \boldsymbol{\mathcal{C}}$, where

$$\boldsymbol{\mathcal{C}} = \left\{ \boldsymbol{\mathcal{P}} \times \boldsymbol{\mathcal{G}} : \sum_{v \in \mathcal{V}} g_v = 1 \right\} \subset \boldsymbol{\mathcal{P}} \times \boldsymbol{\mathcal{G}}, \tag{8}$$

and

$$\mathcal{L}(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}'_n) = \prod_{v=\Delta+1}^{\Delta+m} \prod_{u=v}^{\omega} \left[\frac{f(u \mid \boldsymbol{p}) g_v}{\alpha} \right]^{\sum_{i=1}^n \mathbf{1}_{(X_i, Y_i) = (u, v)}}. \tag{9}$$

We have dropped the parametric notation, $\boldsymbol{\Theta}$, of (3) in the lead up to (9) to emphasize the replacement of p with the more general $\boldsymbol{p}$ in the parametric space. For completeness, α in (9) takes the same form as (2) but with $f(\cdot \mid \boldsymbol{p})$ replacing $f(\cdot \mid p)$. The formal result is stated in Corollary 3.1.1. For reference, Lautier et al. (2023a) study the limiting case, where each probability mass, $f(u)$, is itself represented by a parameter, f_u, $\Delta + 1 \le u \le \omega$.

Corollary 3.1.1 (Stationary points of $\mathcal{L}$ over $\boldsymbol{\mathcal{C}}$)**.** *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 3.1. Further assume* $\partial f(u \mid \boldsymbol{p})/\partial p_j$ *exists and is finite for all* $j = 1, \ldots, r$. *Then the stationary points of* $\mathcal{L}(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}_n')$ *are*

$$\hat{g}_v = \frac{\hat{h}_{\bullet v}}{S(v \mid \hat{\boldsymbol{p}})} \left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{h}_{\bullet k}}{S(k \mid \hat{\boldsymbol{p}})} \right]^{-1}, \quad v \in \mathcal{V}, \tag{10}$$

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

$$\hat{\boldsymbol{\mathcal{P}}} = \{\boldsymbol{p} \in \boldsymbol{\mathcal{P}} : \xi_1(j) = \xi_2(j), \quad \textit{for all} \quad j = 1, \ldots, r\}, \tag{11}$$

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

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* $v \in \mathcal{V}$.

Proof. See the Supplemental Material, Section A.2. □

Beyond the point estimation results of Theorem 3.1 and Corollary 3.1.1 that simplify the multidimensional constrained optimization problem created by the maximum likelihood process in our financial setting, it is of interest to examine the asymptotic properties of (7) to assess estimation precision. (Recall, it is the lifetime distribution X that is of most practical importance to ABS investors.) To this end, a further advantage of (7) is that it takes the form of an asymptotically consistent M-estimator (van der Vaart, 1998, §5.3, pg. 51). This observation, which we prove in Theorem 3.2, allows us to derive the exact form of its asymptotically normal or large sample distribution.

Theorem 3.2. *Let* $\hat{p}_n \in \mathcal{P}$ *satisfy* (7) *from Theorem 3.1 and denote* p_0 *as the true parameter*

value. Define

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

Then $\Psi_n(p, \mathcal{S}_n) \equiv \Psi_n(p)$ is an asymptotically consistent M-estimator of $\mathbf{E}\psi(X_i, Y_i, p)$ for all $p \in \mathcal{P}$ (van der Vaart, 1998, §5.3, pg. 51), where

$$\psi(X_i, Y_i, p) = \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\sum_{u=v}^{\omega} W_i}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) - \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} \frac{W_i}{f(u \mid p)} \frac{\partial}{\partial p} f(u \mid p),$$

and $W_i(u, v) = \mathbf{1}((X_i, Y_i) = (u, v))$ for $1 \leq i \leq n$. Further, $\Psi_n(\hat{p}_n) = 0$. If we also assume (i) $\hat{p}_n \xrightarrow{\mathbf{P}} p_0$, (ii) $\mathbf{E}[\psi(X_i, Y_i, p_0)]^2 < \infty$, (iii) $\mathbf{E}[\partial\psi(X_i, Y_i, p_0)/\partial p]$ exists, and (iv) $\partial^2\Psi_n(\tilde{p})/\partial p^2$ is $O_{\mathbf{E}\psi}(1)$, where $\tilde{p}$ is a point between $\hat{p}_n$ and p_0, then,

$$\sqrt{n}(\hat{p}_n - p_0) \xrightarrow{\mathcal{L}} N\left(0, \frac{\mathbf{E}[\psi(X_i, Y_i, p_0)^2]}{(\mathbf{E}[\partial\psi(X_i, Y_i, p_0)/\partial p])^2}\right),$$

where

$$\begin{aligned} \frac{\partial}{\partial p}\psi(X_i, Y_i, p_0) = &\sum_{v=\Delta+1}^{\Delta+m} \left(\sum_{u=v}^{\omega} W_i \right) \left[\frac{(\sum_{u=v}^{\omega} f''(u))(\sum_{u=v}^{\omega} f(u)) - (\sum_{u=v}^{\omega} f'(u))^2}{(\sum_{u=v}^{\omega} f(u))^2} \right] \\ &- \sum_{v=\Delta+1}^{\Delta+m} \sum_{u=v}^{\omega} W_i \left[\frac{f''(u)f(u) - f'(u)^2}{f(u)^2} \right], \end{aligned}$$

and f' and f'' denote $\partial f/\partial p$ and $\partial^2 f/\partial p^2$, respectively.

Proof. See the Supplemental Material, Section A.3. □

In practical settings, the true parameter, $p \in \mathcal{P}$, will not be known. Thus, techniques to estimate the asymptotic variance we derive in Theorem 3.2 are necessary. The results of Corollary 3.2.1 provide one such approach.

Corollary 3.2.1. *Assume the conditions of Theorem 3.2 and define*

$$U = \mathbf{E}\frac{\partial}{\partial p}\psi(X_i, Y_i, p_0), \quad U_n = \frac{\partial}{\partial p}\Psi_n(\hat{p}_n).$$

Similarly define

$$V = \text{Var}[\psi(X_i, Y_i, p_0)], \quad V_n = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i, Y_i, \hat{p}_n)^2.$$

If $U_n \xrightarrow{\mathbf{P}} U$ *and* $V_n \xrightarrow{\mathbf{P}} V$*, then*

$$[V_n/U_n^2]^{-1/2}\sqrt{n}(\hat{p}_n - p_0) \xrightarrow{\mathcal{L}} N(0,1). \tag{12}$$

Additionally, if the second Bartlett identity (Ferguson, 1996, pg. 120) is also satisfied, then $U = V$ *and*

$$[V_n]^{1/2}\sqrt{n}(\hat{p}_n - p_0) \xrightarrow{\mathcal{L}} N(0,1).$$

Proof. See the Supplemental Material, Section A.4. □

Remark. *The conditions (i) through (iv) in Theorem 3.2 may be relaxed. See van der Vaart (1998, Theorems 5.21 and 5.23 pg. 51-53) for details. Further, these results may be extended to higher dimensions of parameters, such as those assumed in Corollary 3.1.1. See the discussion van der Vaart (1998, Equation (5.20) pg. 51-52) for details.*

The approaches thus far have not considered specific choices of the lifetime distribution, f. One motivation of examining a specific f is to determine if we can increase the claim in Theorem 3.1 from identifying stationary points of the constrained optimization of (4) to finding the MLEs of $p \in \mathcal{P}$ and $\boldsymbol{g} \in \boldsymbol{\mathcal{G}}$. In doing so, it is beneficial to derive an alternative form of (7), which we do in Theorem 3.3.

Theorem 3.3 (Equivalence of $\hat{\mathcal{P}}$)**.** *Assume the conditions of Theorem 3.1. Then* $p \in \hat{\mathcal{P}}$ *if*

and only if

$$\frac{\partial}{\partial p}\frac{\prod_{v=\Delta+1}^{\Delta+m} S(v \mid p)^{\hat{h}_{\bullet v}}}{\prod_{u=\Delta+1}^{\omega} f(u \mid p)^{\hat{h}_{u\bullet}}} = 0. \tag{13}$$

Proof. See the Supplemental Material, Section A.5. □

A close study of (13) suggests candidates for f to yield a direct solution and thus an MLE. See, for example, Theorem 3.4, which yields an f with a closed-form MLE of $\boldsymbol{\Theta}$.

Theorem 3.4 (MLE of $\boldsymbol{\Theta}$, PL geometric)**.** *Define the policy limit (PL) geometric distribution with parameter,* $0 < p < 1$*, as*

$$f_T(u \mid p) = \begin{cases} p(1-p)^{u-(\Delta+1)} & \Delta+1 \le u \le \omega - 1, \\ (1-p)^{u-(\Delta+1)} & u = \omega. \end{cases} \tag{14}$$

Then, for the conditional bivariate probability mass function, h_**, defined in* (1) *under the sampling conditions of Theorem 3.1, the MLE of the parameter* p *is*

$$\hat{p}_{\text{MLE}} = \frac{b}{b-a}, \tag{15}$$

where

$$a = \sum_{v=\Delta+1}^{\Delta+m} \{v - (\Delta+1)\}\hat{h}_{\bullet v} - \sum_{u=\Delta+1}^{\omega} \{u - (\Delta+1)\}\hat{h}_{u\bullet}, \tag{16}$$

and

$$b = \sum_{u=\Delta+1}^{\omega-1} \hat{h}_{u\bullet}. \tag{17}$$

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

$$\{\hat{g}_{v,\text{MLE}}\}_{v\in\mathcal{V}} = \hat{h}_{\bullet v}\left(1-\frac{b}{a}\right)^{v-(\Delta+1)}\left[\sum_{k=\Delta+1}^{\Delta+m}\hat{h}_{\bullet k}\left(1-\frac{b}{a}\right)^{k-(\Delta+1)}\right]^{-1}. \tag{18}$$

Proof. See the Supplemental Material, Section A.6. □

The density function f_T defined in Theorem 3.4 is motivated by actuarial applications of statistical analysis to insurance policy limits. Following Klugman et al. (2012, §8.4, pg. 125), a policy limit of ζ entitles an insured to the full repayment of losses for any amount below ζ with a maximum repayment of ζ for any loss greater than or equal to ζ. Hence, if L_1 denotes the loss random variable before the limit and L_2 denotes the loss random variable after the limit, the cdf of losses to the insurer becomes

$$F_{L_2}(\ell) = \begin{cases} F_{L_1}(\ell), & \ell < \zeta, \\ 1, & \ell \geq \zeta. \end{cases}$$

This setting motivates (14), and it has a natural application to the loan-level ABS analysis that motivates our study. Our reasoning stems from the observation that any probability for $u \geq \omega$ is loaded onto the final point, ω. Because many loans will stop making payments at the termination time dictated by the amortization schedule, such a weighting can be reasonable within an application to financial loan analysis. We demonstrate this in Section 6 for the AART (2017) and AART (2019) ABS bonds. Before concluding the section, we note that the Supplemental Material, Section A.7 contains a restatement of Theorem 3.4 with an alternative parameterization via a discretized, PL exponential distribution (i.e., $p > 0$). An alternative $\mathcal{P}$ space may have utility in generalized linear model (GLM) regression analysis build from the model of (1). For a further discussion, see Section 7.

4 Right-Censoring

We now generalize the theoretical results of Section 3 to also allow for random right-censoring. As with Section 3, we continue with the notation introduced in Section 2. Because of right-censoring, the observed data in this setting takes the triple $\mathcal{S}_{\tau,n} = \{(Y_i, \min(X_i, C_i), D_i)\}_{1 \leq i \leq n}$,

where $D_i = 1$ if $X_i \leq C_i$ and 0 otherwise, for $1 \leq i \leq n$. For convenience of notation, we define $Z_i = \min(X_i, C_i)$, for $1 \leq i \leq n$, and so $\mathcal{S}_{\tau,n} \equiv \{(Y_i, Z_i, D_i)\}_{1 \leq i \leq n}$. If there is no right-censoring present in the data, $D_i = 1$, for all $1 \leq i \leq n$, and $\mathcal{S}_{\tau,n}$ reduces to the left-truncation sample of Section 3, $\mathcal{S}_n$. The subscript of τ will be used liberally in the present section to emphasize right-censoring is assumed present in the data.

Consider now the likelihood under the additional incomplete data setting of right-censoring. Assume first $D_i = 0$, which implies an observation is censored; that is, $X_i > C_i$, for some $1 \leq i \leq n$. The contribution to the likelihood is the probability

$$\begin{aligned}
\Pr(Y_i = v, Z_i = u, D_i = 0) &= \Pr(Y_i = v, \min(X_i, C_i) = u, X_i > C_i) \\
&= \Pr(Y_i = v, \min(X_i, Y_i + \tau) = u, X_i > Y_i + \tau) \\
&= \Pr(Y_i = v, Y_i + \tau = u, X_i > Y_i + \tau) \\
&= \Pr(Y_i = v, v + \tau = u, X_i > v + \tau) \\
&= \Pr(Y_i = v, X_i > u)\mathbf{1}(v + \tau = u) \\
&= \Pr(Y = v, X \geq u + 1 \mid X \geq Y, p)\mathbf{1}(v + \tau = u) \\
&= \bar{h}_*(u, v \mid p)\mathbf{1}(v + \tau = u) \\
&= \bar{h}_*(u, v \mid p),
\end{aligned}$$

where

$$\bar{h}_*(u, v \mid p) = \Pr(Y = v, X \geq u + 1 \mid X \geq Y, p) = \frac{S(u + 1 \mid p) g_v}{\alpha},$$

for $p \in \mathcal{P}, (u, v) \in \mathcal{A}$. We may drop the indicator $\mathbf{1}(v + \tau = u)$ because $D_i = 0$ if and only if $v + \tau = u$ (for any i, $1 \leq i \leq n$, $D_i = 0 \implies X_i > C_i \implies Z_i = C_i = Y_i + \tau$). By the same reasoning, the contribution to the likelihood for $D_i = 1$, for some $1 \leq i \leq n$, is (1). (The Supplemental Material, Section B provides an illustrative example that h_* and $\bar{h}_*$ together form a valid density for all possible outcomes of a single sample, (Y_i, Z_i, D_i), for

any, i, $1 \le i \le n$.) Thus, the likelihood for $\mathcal{S}_{\tau,n}$ becomes

$$\begin{aligned}\mathcal{L}_\tau(\boldsymbol{\Theta} \mid \mathcal{S}_{\tau,n}) &= \prod_{\{\mathcal{S}_{\tau,n}:D_i=1\}} h_*(Z_i, Y_i \mid p) \prod_{\{\mathcal{S}_{\tau,n}:D_i=0\}} \bar{h}_*(Z_i, Y_i \mid p) \\ &= \prod_{\{\mathcal{S}_{\tau,n}:D_i=1\}} \frac{g(Y_i) f(Z_i \mid p)}{\alpha} \prod_{\{\mathcal{S}_{\tau,n}:D_i=0\}} \frac{g(Y_i) S(Z_i+1 \mid p)}{\alpha} \\ &= \alpha^{-n} \prod_{v=\Delta+1}^{m+\Delta} g_v^{n\hat{\gamma}_n(v)} \prod_{i=1}^{n} f(Z_i \mid p)^{D_i} S(Z_i+1 \mid p)^{1-D_i}, \end{aligned} \tag{19}$$

where

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

If $D_i = 1$ for all i, $1 \le i \le n$, then $\mathcal{L}_\tau(\boldsymbol{\Theta} \mid \mathcal{S}_{\tau,n})$ reduces to $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{S}_n)$ of Section 3. As with Theorem 3.1, we seek

$$\sup_{p,\boldsymbol{g}\in\mathcal{C}} \mathcal{L}_\tau(\boldsymbol{\Theta} \mid \mathcal{S}_{\tau,n}). \tag{20}$$

Remark. *It is assumed in* (19) *and all following statments that terms involving* $f(\cdot \mid p)$ *only appear when* $D_i = 1$ *and, conversely, terms involving* $S(\cdot \mid p)$ *only appear when* $D_i = 0$. *This avoids any complications when* $D_i = 1$ *and* $S(\cdot \mid p) = 0$. *It is understood this convention may be coded easily, and we assume the form of* (19) *for ease of exposition.*

Theorem 4.1 (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* $h_*(u, v \mid p)$ *defined in* (1) *under the additional incomplete data setting of right-censoring. Assume the identifiability and differentiability conditions of Theorem 3.1. Then the stationary points of* $\mathcal{L}_\tau(\boldsymbol{\Theta} \mid \mathcal{S}_{\tau,n})$ *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* (6)*, and* $\hat{p}_\tau$ *is any* $p \in \hat{\mathcal{P}}_\tau \subset \mathcal{P}$ *where*

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{f(Z_i \mid p)}\frac{\partial}{\partial p}f(Z_i \mid p) + \frac{1-D_i}{S(Z_i+1 \mid p)}\frac{\partial}{\partial p}S(Z_i+1 \mid p)\right)\Bigg\}. \quad (21)$$

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

Proof. See the Supplemental Material, Section A.8. □

The solution space (23) is a general form of an estimator for the parameter $p \in \mathcal{P}$ under the left-truncation and right-censoring incomplete data setting of Theorem 4.1 not yet derived to our knowledge. As with Theorem 3.1, Theorem 4.1 reduces a potentially complex and computationally demanding multidimensional constrained optimization problem, (20), into a single parameter optimization problem, (21). These comments echo those following Theorem 3.1, and we omit further discussion to avoid unnecessary repetition.

As in Section 3, it may be desirable to allow f to depend on a finite, r-dimensional parameter vector, $\boldsymbol{p}$, as defined just before (8). We now prepare to state the equivalent of Corollary 3.1.1 under the additional data constraint of right-censoring. The equivalent notation for the conditional bivariate pmf of a right-censored observation becomes $\bar{h}_*(u, v \mid \boldsymbol{p})$. The multidimensional, constrained optimization problem is then to find $\sup \mathcal{L}_\tau(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}_{\tau,n})$ such that $(\boldsymbol{p}, \boldsymbol{g}) \in \boldsymbol{\mathcal{C}}$, where $\boldsymbol{\mathcal{C}}$ is defined in (8), and

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

As with the discussion prior to Corollary 3.1.1, we have dropped the parametric notation, $\boldsymbol{\Theta}$, of $\mathcal{L}_\tau$ to emphasize the replacement of p with the more general $\boldsymbol{p}$ in the parametric space. The formal result is stated in Corollary 4.1.1. For reference, Lautier et al. (2023b) study the limiting case, where each probability mass, $f(u)$, is itself represented by a parameter, f_u, $\Delta + 1 \leq u \leq \min(\omega, \varepsilon - 1)$.

Corollary 4.1.1 (Stationary points of $\mathcal{L}_\tau$ over $\boldsymbol{\mathcal{C}}$)**.** *Let* $\mathcal{S}_{\tau,n}$ *be an i.i.d. sample of left-truncated observations from the distribution* $h_*(u, v \mid \boldsymbol{p})$ *under the additional incomplete*

data setting of right-censoring. Assume the identifiability and differentiability conditions of Theorem 3.1. Then the stationary points of $\mathcal{L}_\tau(\boldsymbol{g}, \boldsymbol{p} \mid \mathcal{S}_{\tau,n})$ *are*

$$\hat{g}_{\tau,v} = \frac{\hat{\gamma}_n(v)}{S(v \mid \hat{\boldsymbol{p}}_\tau)} \left[\sum_{k=\Delta+1}^{\Delta+m} \frac{\hat{\gamma}_n(k)}{S(k \mid \hat{\boldsymbol{p}}_\tau)} \right]^{-1}, \quad v \in \mathcal{V}, \tag{22}$$

where $\hat{\boldsymbol{p}}_\tau$ *is any* $\boldsymbol{p}_\tau \in \hat{\boldsymbol{\mathcal{P}}}_\tau \subset \boldsymbol{\mathcal{P}}$*, where*

$$\hat{\boldsymbol{\mathcal{P}}}_\tau = \{\boldsymbol{p}_\tau \in \boldsymbol{\mathcal{P}} : \varphi_1(j) = \varphi_2(j), \quad \textit{for all} \quad j = 1, \ldots, r'\}, \tag{23}$$

$$\varphi_1(j) = \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{\gamma}_n(v)}{\sum_{u=v}^{\omega} f(u \mid \boldsymbol{p})} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p_j} f(u \mid \boldsymbol{p}) \right),$$

and

$$\varphi_2(j) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{f(Z_i \mid \boldsymbol{p})} \frac{\partial}{\partial p_j} f(Z_i \mid \boldsymbol{p}) + \frac{1 - D_i}{S(Z_i + 1 \mid \boldsymbol{p})} \frac{\partial}{\partial p_j} S(Z_i + 1 \mid \boldsymbol{p}) \right) \Bigg\}.$$

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

Proof. See the Supplemental Material, Section A.9. □

As in Section 3, it is of interest to examine the asymptotic properties of (21) to assess estimation precision. As with (7), the estimator (21) takes the form of an asymptotically consistent M-estimator (van der Vaart, 1998, §5.3, pg. 51). We thus prove the equivalent of Theorem 3.2 in Theorem 4.2, from which the exact form of the asymptotically normal distribution of the estimator, (21), follows.

Theorem 4.2. *Let* $\hat{p}_{\tau,n} \in \mathcal{P}$ *satisfy* (21) *from Theorem 4.1 and denote* p_0 *as the true parameter value. Define*

$$\begin{aligned} \Psi_{\tau,n}(p, \mathcal{S}_{\tau,n}) = & \sum_{v=\Delta+1}^{\Delta+m} \left(\frac{\hat{\gamma}_n(v)}{\sum_{u=v}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) \\ & - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{f(Z_i \mid p)} \frac{\partial}{\partial p} f(Z_i \mid p) + \frac{1 - D_i}{S(Z_i + 1 \mid p)} \frac{\partial}{\partial p} S(Z_i + 1 \mid p) \right). \end{aligned}$$

Then $\Psi_{\tau,n}(p,\mathcal{S}_{\tau,n}) \equiv \Psi_{\tau,n}(p)$ *is an asymptotically consistent M-estimator of* $\mathbf{E}\psi_\tau(Y_i, Z_i, D_i, p)$ *for all* $p \in \mathcal{P}$ *(van der Vaart, 1998, §5.3, pg. 51), where*

$$\psi_\tau(Y_i, Z_i, D_i, p) = \sum_{v_*=\Delta+1}^{\Delta+m} \left(\frac{\mathbf{1}(Y_i = v_*)}{\sum_{u=v_*}^{\omega} f(u \mid p)} \right) \left(\sum_{u=v_*}^{\omega} \frac{\partial}{\partial p} f(u \mid p) \right) - \left(\frac{D_i}{f(Z_i \mid p)} \frac{\partial}{\partial p} f(Z_i \mid p) + \frac{1 - D_i}{S(Z_i + 1 \mid p)} \frac{\partial}{\partial p} S(Z_i + 1 \mid p) \right),$$

for $1 \leq i \leq n$*. Further,* $\Psi_{\tau,n}(\hat{p}_{\tau,n}) = 0$*. If we also assume* (i) $\hat{p}_{\tau,n} \xrightarrow{\mathbf{P}} p_0$*,* (ii) $\mathbf{E}[\psi_\tau(Y_i, Z_i, D_i, p_0)]^2 < \infty$*,* (iii) $\mathbf{E}[\partial\psi_\tau(Y_i, Z_i, D_i, p_0)/\partial p]$ *exists, and* (iv) $\partial^2\Psi_{\tau,n}(\tilde{p}_\tau)/\partial p^2$ *is* $O_{\mathbf{E}\psi_\tau}(1)$*, where* $\tilde{p}_\tau$ *is a point between* $\hat{p}_{\tau,n}$ *and* p_0*, then*

$$\sqrt{n}(\hat{p}_{\tau,n} - p_0) \xrightarrow{\mathcal{L}} N\left(0, \frac{\mathbf{E}[\psi_\tau(Y_i, Z_i, D_i, p_0)^2]}{(\mathbf{E}[\partial\psi_\tau(Y_i, Z_i, D_i, p_0)/\partial p])^2}\right),$$

where

$$\frac{\partial}{\partial p}\psi_\tau(Y_i, Z_i, D_i, p_0) = \sum_{v_*=\Delta+1}^{\Delta+m} \mathbf{1}(Y_i = v_*) \left[\frac{(\sum_{u=v_*}^{\omega} f''(u))(\sum_{u=v_*}^{\omega} f(u)) - (\sum_{u=v_*}^{\omega} f'(u))^2}{(\sum_{u=v_*}^{\omega} f(u))^2} \right] - \left(D_i \left[\frac{f''(Z_i)f(Z_i) - f'(Z_i)^2}{f(Z_i)^2} \right] + (1 - D_i) \left[\frac{S''(Z_i + 1)S(Z_i + 1) - S'(Z_i + 1)^2}{S(Z_i + 1)^2} \right] \right),$$

and f'*,* f''*,* S'*, and* S'' *denote* $\partial f/\partial p$*,* $\partial^2 f/\partial p^2$*,* $\partial S/\partial p$*, and* $\partial^2 S/\partial p^2$*, respectively.*

Proof. See the Supplemental Material, Section A.10. □

In practical settings, the true parameter, $p \in \mathcal{P}$, will not be known. Hence, we state the equivalent of Corollary 3.2.1 in the additional incomplete data setting of right-censoring in Corollary 4.2.1 for completeness.

Corollary 4.2.1. *Assume the conditions of Theorem 4.2 and define*

$$U_\tau = \mathbf{E}\frac{\partial}{\partial p}\psi_\tau(Y_i, Z_i, D_i, p_0), \quad U_{\tau,n} = \frac{\partial}{\partial p}\Psi_{\tau,n}(\hat{p}_{\tau,n}).$$

Similarly define

$$V_\tau = \mathrm{Var}[\psi_\tau(Y_i, Z_i, D_i, p_0)], \quad V_{\tau,n} = \frac{1}{n}\sum_{i=1}^{n} \psi_\tau(Y_i, Z_i, D_i, \hat{p}_{\tau,n})^2.$$

If $U_{\tau,n} \xrightarrow{\mathbf{P}} U_\tau$ *and* $V_{\tau,n} \xrightarrow{\mathbf{P}} V_\tau$*, then*

$$[V_{\tau,n}/U_{\tau,n}^2]^{-1/2}\sqrt{n}(\hat{p}_{\tau,n} - p_0) \xrightarrow{\mathcal{L}} N(0,1). \tag{24}$$

Additionally, if the second Bartlett identity (Ferguson, 1996, pg. 120) is also satisfied, then $U_\tau = V_\tau$ *and*

$$[V_{\tau,n}]^{1/2}\sqrt{n}(\hat{p}_{\tau,n} - p_0) \xrightarrow{\mathcal{L}} N(0,1).$$

Proof. See the Supplemental Material, Section A.11. □

Remark. *The conditions (i) through (iv) in Theorem 4.2 may be relaxed. See van der Vaart (1998, Theorems 5.21 and 5.23 pg. 51-53) for details. Further, these results may be extended to higher dimensions of parameters, such as those assumed in Corollary 4.1.1. See the discussion van der Vaart (1998, Equation (5.20) pg. 51-52) for details.*

We close this section with the equivalent of Theorem 3.4 under the additional data constraints of right-censoring. The formal statement comprises Corollary 4.2.2.

Corollary 4.2.2 (MLE of $\boldsymbol{g}$, p, PL geometric, right-censoring)**.** *Recall the PL geometric distribution with parameter,* $0 < p < 1$*, defined in* (14)*. Then, for the conditional bivariate probability mass function,* h_**, defined in* (1) *under the sampling conditions of Theorem 4.1, the MLE of the parameter* p *is*

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

where

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

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*

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

Proof. See the Supplemental Material, Section A.12. □

5 Simulation Studies

We first provide numerical verification of Theorems 3.1, 3.2, 3.4, 4.1, 4.2, and Corollary 4.2.2. We assume $m = 3$, $\Delta = 0$, and $\omega = 4$. This results in a 4×3 trapezoid, $\mathcal{A}$. For the lifetime random variable, we consider two parametric distributions. The first assumes X takes the form of (14) with $p = 0.3$. This results in the pmf of X as $\Pr(X = 1) = 0.3$, $\Pr(X = 2) = 0.21$, $\Pr(X = 3) = 0.147$, and $\Pr(X = 4) = 0.343$. We then assume the pmf of Y is $g_1 \equiv \Pr(Y = 1) = 0.5$, $g_2 \equiv \Pr(Y = 2) = 0.3$, and $g_3 \equiv \Pr(Y = 3) = 0.2$. This results in $\alpha = 0.808$. Observe that Y is non-uniform, which demonstrates our results may be applied outside the domain of length-biased sampling. The second assumes X takes the form of a shifted binomial distribution with the number of successes equal to $\omega-(\Delta+1) = 3$ and a probability of success equal to $0 < \theta = 0.75 < 1$. This results in the pmf of X as $\Pr(X = 1) = 0.016$, $\Pr(X = 2) = 0.141$, $\Pr(X = 3) = 0.422$, and $\Pr(X = 4) = 0.422$. Because $\boldsymbol{g}$ is unchanged, we obtain $\alpha = 0.964$. For reference, a simulation tutorial to generate

a random sample from h_* may be found in the Supplemental Material, Section D.

To perform the numeric validation for Theorem 3.1 and Theorem 3.4, we simulate a single sample size of $n = 1{,}000$ under the setting of Section 3 (i.e., $\mathcal{S}_n$). We then estimate the parameters in three ways. First, we solve (4) with a direct, constrained numeric optimization using `constroOptim` via R Core Team (2023). Next, we perform a single parameter optimization through Theorem 3.1 using `optimize` via R Core Team (2023) in combination with the closed-form solutions for (5). Finally, we report the parameter estimates with all closed-form solutions using Theorem 3.4. The results are summarized in the top-half of Table 2. All three approaches yield nearly identical solutions, all of which are quite close to the true parameter values. For completeness, we also estimate computational performance statistics for all three parameter estimation approaches with the `microbenchmark` package (Mersmann, 2023). It is immediate that both Theorems 3.1 and 3.4 provide substantial improvements in computational demands. We then repeat this procedure under the sampling assumptions of Section 4 (i.e., $\mathcal{S}_{\tau,n}$) with $\varepsilon = 6$ to verify Theorem 4.1 and Corollary 4.2.2. The comparison is again quite close. The full results may be found in Table 2.

We next validate Theorems 3.2. First, we calculate the true asymptotically normal distribution assuming a sample size of $n = 1{,}000$. Next, we perform 1,000 replicates of simulating a sample size of $n = 1{,}000$ and estimating $\hat{p}_n$ with Theorem 3.4 (for the shifted-binomial distribution, Theorem 3.1). The true asymptotic density and the empirical density of the 1,000 replicates of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\sqrt{n}(\hat{p}_n - p_0)$ may be found in Figure 1. The true density (solid line) and empirical density (dashed line) closely agree. This process is repeated under the sampling assumptions of left-truncation and right-censoring to validate Theorem 4.2. Once again, the true density (solid line) and empirical density (dashed line) closely agree. These results may also be found in Figure 1. A reference of the necessary derivative calculations may be found in the Supplemental Material, Section C.

Our second simulation study is a traditional robustness analysis of the methods we propose in Sections 3 and 4 to sample size and level of right-censoring. We also increase the

		Left-Trunc. (§3)			Left-Trunc. & Ri.-Cens. (§4)		
		PL Geometric			*PL Geometric*		
Param.	True	`constrOptim`	Thm 3.1	Thm 3.4	`constrOptim`	Thm 4.1	Cor 4.2.2
p	0.30	0.3070043	0.3069670	0.3069670	0.3033170	0.3033175	0.3033175
$g(1)$	0.50	0.4917720	0.4913635	0.4913635	0.5184552	0.5188642	0.5188642
$g(2)$	0.30	0.3039001	0.3036941	0.3036941	0.2818413	0.2820547	0.2820547
$g(3)$	0.20	0.2050589	0.2049424	0.2049424	0.1989716	0.1990811	0.1990811
Speed (μS)		562,996	6,797	861	9,027,715	169,510	12,706
		Shifted-Binomial			*Shifted-Binomial*		
Param.	True	`constrOptim`	Thm 3.1		`constrOptim`	Thm 4.1	
θ	0.75	0.7462976	0.7462779		0.7445825	0.7446186	
$g(1)$	0.50	0.4870210	0.4867887		0.5251005	0.5246253	
$g(2)$	0.30	0.2949194	0.2947756		0.2733774	0.2731187	
$g(3)$	0.20	0.2185810	0.2184357		0.2024117	0.2022560	
Speed (μS)		543,432	8,633		5,100,241	233,909	

Table 2: **Numeric Validation and Performance Summary**. Numerical verification of Theorems 3.1, 3.4, and 4.1, and Corollary 4.2.2. Trapezoid parameters are $m = 3$, $\Delta = 0$, and $\omega = 4$ (with $\varepsilon = 6$ for right-censoring). Single sample of size $n = 1{,}000$. Performance calculations per the `microbenchmark` package (Mersmann, 2023) with 100 evaluations (mean speed reported in microseconds (μS)). The computer was a `Surface Pro 9` with a `12th Gen Intel(R) Core(TM) i5-1235U 2.50 GHz` processor, 8.00GB installed RAM, 10 cores, and 12 logical processors.

size of the trapezoid $\mathcal{A}$ to better reflect what may be encountered in an application to ABS. Hence, we assume $m = 20$, $\Delta = 0$, and $\omega = 24$. The lifetime distribution, X, in this setting follows a PL geometric distribution, (14), with parameter $p = 0.05$. The left-truncation random variable is a weighted-mixture of shifted binomial distributions with 9 trials and success probability of 0.35, denoted $\mathcal{B}$. That is, $\Pr(Y = v) = 0.4 \times \mathbf{1}(1 \leq v \leq 10)\mathcal{B}(v-1) + 0.6 \times \mathbf{1}(11 \leq v \leq 20)\mathcal{B}(v-11)$. Because an application to ABS data is primarily focused on the lifetime distribution of the loans, X, we focus on the parameter p.

We consider sample sizes $n \in \{50, 100, 250, 500\}$. We may control for right-censoring by varying the current time, $m + \Delta + 1 \leq \varepsilon \leq m + \omega$. There is more right-censoring present in the data for values of ε closer to $m + \Delta + 1$. We consider $\varepsilon \in \{26, 32, 38\}$ and a fourth setting of no right-censoring present in the data. For each combination of sample size and right-censoring, we simulate a random sample from h_*. We then estimate p using either

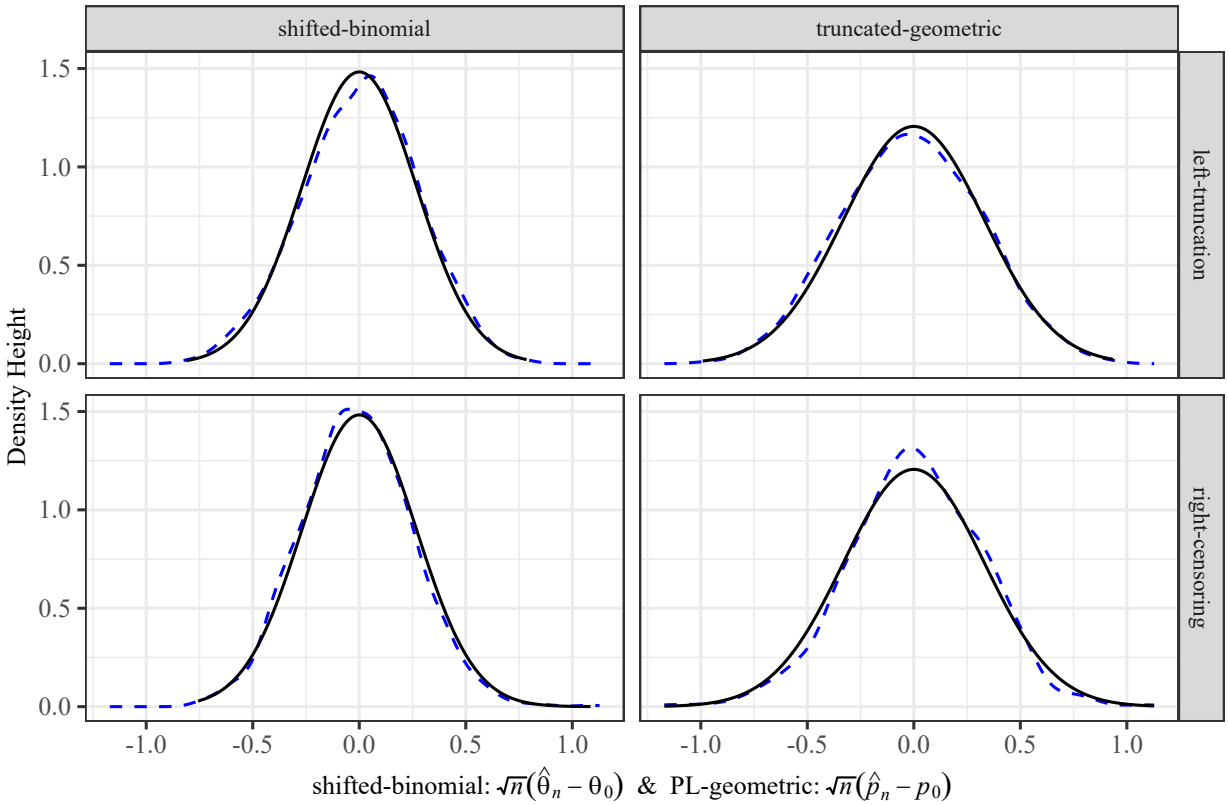

Figure 1: **Asymptotic Normality Verification**. Numeric verification of Theorem 3.2 and 4.2. Trapezoid parameters are $m = 3$, $\Delta = 0$, and $\omega = 4$ (with $\varepsilon = 6$ for right-censoring). Results for 1,000 replicates each with a sample size of $n = 1{,}000$. Lifetime random variable and left-truncation random variable parameters identical to Table 2.

Corollary 4.2.2 or Theorem 3.4, as appropriate. This process is repeated for 1,000 replicates. We then calculate the empirical mean and standard deviation of the 1,000 estimates of p. We compare the empirical mean with the true parameter value, $p = 0.05$, and we compare the empirical standard deviation with the theoretical value we calculate using Theorems 3.2 or 4.2, as appropriate. Finally, we evaluate Corollaries 3.2.1 and 4.2.1 by performing a coverage probability study of the 95% asymptotic confidence intervals. The coverage probabilities represent the percentage of 1,000 replicates such that,

$$\hat{p}_n - Z_{0.975}\sqrt{\frac{1/V_n}{n}} \leq p_0 \leq \hat{p}_n + Z_{0.975}\sqrt{\frac{1/V_n}{n}}, \tag{25}$$

where $Z_{0.975}$ represents the 97.5th percentile of a standard normal random variable. The

		$\varepsilon = 26$				$\varepsilon = 32$			
n	p_0	eMean	eSD	Thm 4.2	CP	eMean	eSD	Thm 4.2	CP
50	0.05	0.0499	0.0096	0.0097	0.942	0.0503	0.0097	0.0097	0.951
100	0.05	0.0503	0.0069	0.0069	0.946	0.0502	0.0069	0.0069	0.949
250	0.05	0.0502	0.0044	0.0043	0.943	0.0499	0.0043	0.0043	0.947
500	0.05	0.0501	0.0032	0.0031	0.950	0.0499	0.0030	0.0031	0.962
		$\varepsilon = 38$				No Censoring			
n	p_0	eMean	eSD	Thm 4.2	CP	eMean	eSD	Thm 3.2	CP
50	0.05	0.0510	0.0100	0.0097	0.933	0.0504	0.0094	0.0095	0.954
100	0.05	0.0502	0.0072	0.0069	0.931	0.0504	0.0066	0.0067	0.963
250	0.05	0.0502	0.0043	0.0043	0.956	0.0499	0.0042	0.0043	0.952
500	0.05	0.0500	0.0031	0.0031	0.944	0.0501	0.0031	0.0030	0.952

Table 3: **Robustness Simulation Study**. A robustness analysis of the methods proposed in Sections 3 and 4 for the right-truncated geometric distribution by sample size (n) and level of right-censoring (ε). Trapezoid parameters are $m = 20$, $\Delta = 0$, and $\omega = 24$. We report the empirical mean (eMean), empirical standard deviation (eSD), theoretical standard deviation (Thm 3.2, Thm 4.2), and coverage probability (CP), i.e., (25), for a 95% asymptotic confidence interval estimated using Corollaries 3.2.1 and 4.2.1. The left-truncation distribution, Y, follows a weighted binomial mixture.

complete results may be found in Table 3. The robustness of the methods we propose are quite strong to small samples and the level of right-censoring, which is an attractive property for an application to ABS data (and more generally).

6 Estimating Loan-Level Performance

We now assume the role of an ABS investment analyst to employ the statistical methods of Sections 3 and 4 to estimate the loan-level lifetime distribution parameters from data contained in the ABS bonds Ally Auto Receivables Trust 2017-3 (AART, 2017) and Ally Auto Receivables Trust 2019-3 (AART, 2019). Per Section 1, this data was obtained via the SEC's EDGAR system (Securities and Exchange Commission, 2014, 2016).

Consider first the AART-2017 ABS bond, which contains 67,797 individual consumer auto loans with origination dates ranging from February 2011 to April 2017. The AART-

2017 ABS bond was issued in May of 2017 and was active for 43 months. The mean (median) credit score of these loans is 725 (719), which corresponds to a *super-prime* credit risk tier (Consumer Financial Protection Bureau, 2019). We first demonstrate the plausibility of the PL geometric distribution as defined in (14) to describe this ABS loan-level lifetime data. In this context, the random variable X describes the time-until-loan-payments stop, either due to prepayment or default (a second statistical model may be used to simultaneously model the cause (e.g., Lautier et al., 2022) but is outside the scope of this analysis). Because the loan terms in AART-2017 range from 12 to 78 months, and each loan term must be treated separately, we focus on a subset of $n = 151$ loans with a loan term of 25 months.

For these 25-month loans, we have $m = 21$, $\Delta = 3$, and $\omega = 26$. There are thus 21 parameters to be estimated, which would make obtaining a solution to (4) or (20) numerically quite difficult without the methods we propose. We will use the full 43 months of performance data, which sets $\varepsilon = 67 > m + \omega$. Hence, there is no right-censoring present, which allows us to use methods from Section 3. As a minor data adjustment, we treat any observations with a loan termination time of 27, 28, or 29 months as a loan that terminated at full-term, 26 months. Such small extensions may occur due to delays in accounting or financial reporting, and they generally have a minimal impact on estimated profitability.

We apply Theorem 3.4, Theorem 3.2, and Corollary 3.2.1 to estimate $\hat{p}_{\text{MLE}}^{\text{17-25}} = 0.0313$ with a 95% asymptotic confidence interval of $(0.0226, 0.0399)$. But is (14) with a parameter of $\hat{p}_{\text{MLE}}^{\text{17-25}} = 0.0313$ appropriate to model this data? We investigate by comparing it to Lautier et al. (2023b), which makes no structural assumptions about the shape of the loan-level lifetime distribution. In other words, we estimate the hazard rate plus 95% asymptotic confidence intervals for each loan age via Lautier et al. (2023b). Because (14) assumes a constant hazard rate, we will first assess if $\hat{p}_{\text{MLE}}^{\text{17-25}} = 0.0313 \pm 0.0086$ falls between the 95% asymptotic confidence intervals of Lautier et al. (2023b). A visualization of this comparison may be found in the top left panel of Figure 2. The point estimates of the hazard rate via Lautier et al. (2023b) are noted by the solid blue line, and the blue ribbon represents 95%

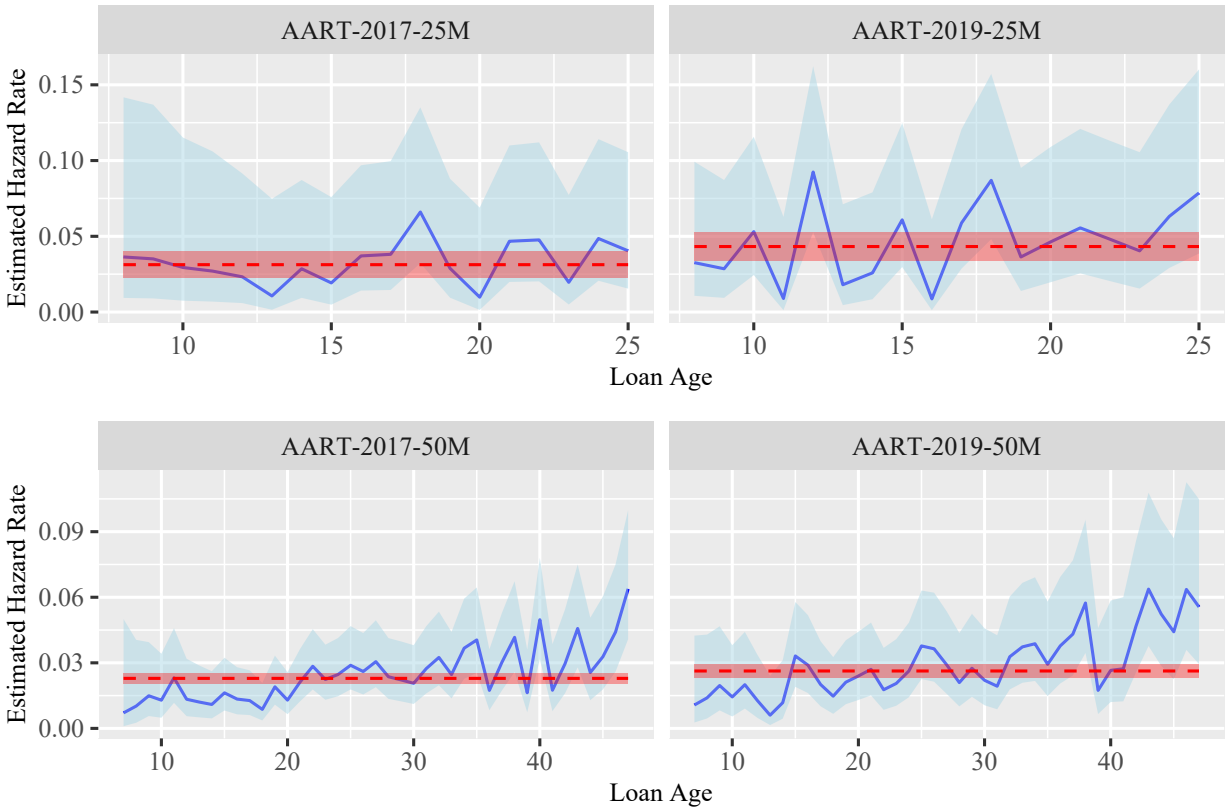

Figure 2: **PL Geometric Distribution, Empirical Validation**. A comparison of fitted hazard rates using the flexible Lautier et al. (2023b) (solid blue line and ribbon) against the policy limit geometric distribution proposed in (14) (dashed red line and ribbon) by bond AART (2017, 2019) and original loan term (25, 50 months). The consistent overlap of the asymptotic 95% confidence intervals for each model indicates the plausibility of (14) to model loan lifetime ABS data.

asymptotic confidence intervals for each hazard rate per Lautier et al. (2023b). The horizontal red dashed line plus red ribbon represents $\hat{p}_{\text{MLE}}^{\text{17-25}} = 0.0313 \pm 0.0086$, as estimated with the methods of Section 3. Because the 95% asymptotic confidence intervals for each method consistently overlap, it is evidence that the PL geometric distribution of (14) with a parameter of $\hat{p}_{\text{MLE}}^{\text{17-25}} = 0.0313 \pm 0.0086$ is reasonable to model the loan-level lifetime distribution of these 151 25-month consumer automobile loans from AART-2017.

With the viability of (14) established for this data, we now suppose we are an ABS investment analyst assessing the loan-level payment performance of the AART bonds over time. Such analysis helps identify any potential trends that may inform trading decisions. For example, an elevated parameter, p, may indicate higher defaults or earlier prepayments, both

of which may negatively impact the profitability of the AART-2019 ABS bond in comparison to AART-2017 (see Table 1). As such, we consider the AART-2019 bond and compare it to AART-2017. The AART-2019 bond consists of 67,198 individual consumer auto loans with origination dates ranging from October 2012 to July 2019. The AART-2019 bond was issued in August of 2019 and was actively paying for 46 months. The mean (median) credit score of these loans is 722 (718), which corresponds to a super-prime credit risk tier (Consumer Financial Protection Bureau, 2019). The AART-2019 bond has consumer auto loans with loan terms ranging from 13 to 79 months. Hence, the AART-2019 consumer loan pool compares quite closely to AART-2017, which would indicate to investors that both bonds will perform comparably. Let us now determine if this is indeed the case.

As with AART-2017, we consider the subset of 25-month loans, which results in $n_{2019} = 178$. For these 25-month loans, we have $m_{2019} = 19$, $\Delta_{2019} = 2$, and $\omega_{2019} = 26$. This results in 19 parameters to be estimated. We will use the full 46 months of performance data, which again sets $\varepsilon_{2019} = 67 > m_{2019} + \Delta_{2019}$. Thus, no right-censoring is present in the 2019 data for the 25-month loans. As with AART-2017, we set any consumer loans with termination times beyond month 26 to $\omega_{2019} = 26$. We next apply Theorem 3.4, Theorem 3.2, and Corollary 3.2.1 to estimate $\hat{p}_{\text{MLE}}^{\text{19-25}} = 0.0432$ with a 95% asymptotic confidence interval of $(0.0339, 0.0526)$. Because the 95% asymptotic confidence intervals of $\hat{p}_{\text{MLE}}^{\text{19-25}}$ overlap with those of $\hat{p}_{\text{MLE}}^{\text{17-25}}$, we cannot claim that the loan-level lifetime distribution has materially changed.

This may be surprising because the AART-2017 bond stopped paying in November 2020, shortly after the start of the Coronavirus pandemic. The Coronavirus was a significant economic event that had a clear negative impact on the performance of consumer auto loans (Lautier et al., 2022). The latter issuance, AART-2019, conversely, started paying in August 2019 and therefore exposed the individual consumer loans to the Coronavirus for a longer portion of the observation window. The economic interpretation of this result is that both AART bonds we analyze are generally super-prime credits. Such strong credits had a minimal default impact from COVID-19 (Lautier et al., 2022). Pleasingly, the methods

herein support this conclusion despite closed-form solutions of a single parameter estimate. Such quick estimates offer investors a single number that is easy to interpret. This ease of use may aid buy and sell decisions for consumer auto ABS holdings. A visual comparison may be found in the top row of Figure 2. For completeness, we note that (14) may be less applicable to lower credit quality borrowers, a problem we at present leave open.

We also consider loans with original termination schedules of 50-months from AART (2017, 2019). We choose to also model the loan-level lifetime distribution for 50-month loans for two reasons. First, it will result in data that is subject right-censoring in addition to left-truncation, which requires the methods of Section 4. Second, to ultimately build to the ABS cash flow, it will be necessary to fit each original loan term found in the ABS (see Table 1). Hence, we desire to fit a model to longer term loans to discuss potential modeling considerations to achieve a complete application. To begin, we find an improved fit when limiting the individual loan data to *super-prime* credits (i.e., an interest rate below 5% (Lautier et al., 2022)). As with 25-month loans, (14) may be less applicable to lower credit quality borrowers, a problem we at present leave open. We thus obtain sample sizes of 692 and 495 super-prime loans with original terms of 50 months. We also apply a minor data adjustment in that any loan with a termination time beyond 48 months was given a termination time of 48 months. This is likely an artifact of financial reporting in that consumer auto loans are commonly written in increments of 6 and 12 months (i.e., 24-, 36-, 48-, 60-, 72-months). The visual analysis may be found in the bottom row of Figure 2.

Upon inspection, we see the PL geometric distribution potentially overfits in earlier loan ages and underfits in later loan ages for these 50 month loans (second row, Figure 2). This suggests that an increasing linear hazard rate suitable for discrete data may be more appropriate. One example of a discrete parametric distribution capable of modeling an increasing hazard rate is the discrete Weibull of Nakagawa and Osaki (1975). It is a two parameter distribution, and so Corollaries 3.1.1 and 4.1.1 may be applied. Furthermore, given the theoretical plausibility of (14), a policy limit version of the discrete Weibull (i.e.,

Klugman et al., 2012, §8.4, pg. 125) may better model the longer term loan-level lifetime distribution of ABS data. As such, we suggest this as an area of further research in Section 7.

7 Discussion

The cash flows of an ABS are a random variable, which stems from the randomness of the time-to-termination of the individual loans within an ABS pool (i.e., Table 1). Hence, properly modeling ABS cash flows begins by estimating the loan-level time-to-termination lifetime distribution. Historically, this estimation was obfuscated by a lack of loan-level data. In the aftermath of the 2007-2009 financial crisis, the SEC adopted significant revisions to Regulation AB, which governs the offering, disclosure, and required reporting of ABS (Securities and Exchange Commission, 2014). One significant component of these regulatory changes is that public issuers of ABS are now required to make pertinent loan-level and payment performance data freely available on a monthly basis (Securities and Exchange Commission, 2016). It is this new glut of financial data, its immediate widespread application to estimating the loan-level lifetime distribution (and thus broadly ABS investment management), and its resulting statistical curiosity that attracts our study.

The data generating process of loan-level data within an ABS is known to be subject to the major incomplete data challenge of left-truncation (Section 2). Though many classical models exist to model continuous left-truncated data (e.g., Woodroofe, 1985; Tsai et al., 1987; Wang, 1989), the ABS setting also requires assuming discrete-time over a known, finite support. This combination has begun to attract study (Lautier et al., 2022, 2023a,b, 2024), but the problem of assuming a more traditional parametric form for the lifetime random variable in this ABS setting remains open. Successfully addressing this problem offers three advantages for estimating the loan-level time-to-event distribution. First, parametric distributions provides a natural smoothing in comparison to current approaches (e.g., Figure 2). Second, parametric forms open the door to allowing for the incorporation of

economic variables, which is desirable to assess their impact on loan payment performance. Third and finally, the fast pace of active investment management for ABS requires rapid decision-making, which benefits from a focus on a single or few key parameters (Section 6).

Our results are as follows. Under random left-truncation (Section 3), we first reduce a complex multidimensional constrained optimization problem into a single parameter optimization problem (Theorem 3.1). It is a general statement for a lifetime random variable X that depends on a single parameter. We then generalize this to a vector parameter for X (Corollary 3.1.1). Next, we show the solution set for the parameter estimate in Theorem 3.1 is an estimating equation (Theorem 3.2). This produces an asymptotically normal distribution of the estimator. For practical settings, we then provide large sample estimates for the variance structure (Corollary 3.2.1). We next assume a PL geometric distribution (Klugman et al., 2012, §8.4, pg. 125), which is justifiable to model ABS data (Figure 2). In this setting, we derive a closed-form MLE of the parameter (Theorem 3.4). In Section 4, we derive all of the results of Section 3 but generalized for the incomplete data setting of random left-truncation and right-censoring. This requires first deriving an updated likelihood equation. As with Lautier et al. (2023a), the form of the left-truncation random variable remains flexible throughout. Complete proofs are in the Supplemental Material, Section A.

The theoretical results of Sections 3 and 4 are then made practical (Section 5). All results are first verified numerically to support the theoretical proofs and to illustrate the performance efficiency gains (Table 2). We next verify Theorems 3.2 and 4.2 with simulation (i.e., Figure 1), and we provide a coverage probability robustness analysis to sample size and amount of censoring. In Section 6, we apply our methods to data from the ABS bonds AART (2017) and AART (2019). We provide evidence that the PL geometric distribution can plausibly model the lifetimes of loan-level ABS data (Figure 2). The application then uses our methods to draw rapid inference on the stability of loan-level performance over issuance years. Drawing inference quickly, aided by the closed-form MLE solutions we derive, may be valuable to ABS investors often required to make expeditious buy-sell decisions.

We close by suggesting avenues of further research. First, prepayments are an important component of the analysis of loans, and the present model requires treating all loan termination causes (i.e., default and prepayment) as equivalent. This is viable in that an ABS investor can use our approach to model the time-until-loan-payments stop, and then overlay a second model for the cause. Nonetheless, it is of interest to explore if an extension to a competing risks framework (e.g., Lautier et al., 2024) is possible. Second, the bottom row of Figure 2 evinces that a policy limit version of the discrete Weibull (Nakagawa and Osaki, 1975), constructed like (14), may better capture the apparent linearly increasing hazard rate we observe for longer term loans. Similarly, it is of interest to examine the PL geometric distribution with an application to lower credit quality borrowers. Third, it is desirable to connect the lifetime random variable to economic variables. One natural next step is to attempt to link the parameter of X to a set of covariates in the form of a GLM. We believe our efforts have moved closer towards this goal. Fourth, we suggest to revisit the pricing model of Lautier et al. (2023b) with the PL geometric distribution of (14). It is possible attractive closed-form solutions would materialize upon close study. Finally, we postulate our methods will be applicable to other fields where incomplete, discrete time-to-event data frequently occur, such as healthcare, finance, engineering, telecommunication, and insurance.

References

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

— (2019), "Ally Auto Receivables Trust," Prospectus 2019-3, Ally Auto Assets LLC.

Agarwal, N., Marion, S., and Wu, J. (2024), "Moody's Global Approach to Rating Auto Loan and Lease-Backed ABS," Rating Methodology 426427, Moody's Ratings.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002), "Length-Biased Sampling With Right Censoring," *Journal of the American Statistical Association*, 97, 201–209.

Consumer Financial Protection Bureau (2019), "Borrower Risk Profiles," Url: https://www.consumerfinance.gov/data-research/consumer-credit-trends/auto-loans/borrower-risk-profiles/ (Accessed: 2022-06-15).

De Uña-Álvarez, J. (2004), "Nonparametric Estimation Under Length-Biased Sampling and Type I Censoring: A Moment Based Approach," *Annals of the Institute of Statistical Mathematics*, 56, 667–681.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall.

Huang, Y. and Wang, M.-C. (1995), "Estimating the Occurrence Rate for Prevalent Survival Data in Competing Risks Models," *Journal of the American Statistical Association*, 90, 1406–1415.

Katcher, B., Li, G., Mezza, A., and Ramos, S. (2024), "One Month Longer, One Month Later? Prepayments in the Auto Loan Market," Finance and Economics Discussion Series (FEDS) 2024.056, Federal Reserve Board, Washington, D.C.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012), *Loss Models: From Data to Decisions, Fourth Edition*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Lautier, J. P., Pozdnyakov, V., and Yan, J. (2022), "On the Convergence of Credit Risk in Current Consumer Automobile Loans," ArXiv preprint, https://arxiv.org/abs/2211.09176.

— (2023a), "Estimating a Discrete Distribution Subject to Random Left-Truncation with an Application to Structured Finance," *Econometrics and Statistics*, forthcoming.

— (2023b), "Pricing Time-to-Event Contingent Cash Flows: A Discrete-Time Survival Analysis Approach," *Insurance: Mathematics and Economics*, 110, 53–71.

— (2024), "On the Maximum Likelihood Estimation of a Discrete, Finite Support Distribution under Left-Truncation and Competing Risks," *Statistics & Probability Letters*, 207, 109973.

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

Mishkin, F. S. (2011), "Over the Cliff: From the Subprime to the Global Financial Crisis," *Journal of Economic Perspectives*, 25, 49–70.

Murtagh, B. A. and Saunders, M. A. (1978), "Large-Scale Linearly Constrained Optimization," *Mathematical Programming*, 14, 41–72.

Nakagawa, T. and Osaki, S. (1975), "The Discrete Weibull Distribution," *IEEE Transactions on Reliability*, R-24, 300–301.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Securities and Exchange Commission (2014), "17 CFR Parts 229, 230, 232, 239, 240, 243, and 249 Asset-Backed Securities Disclosure and Registration," .

— (2016), "17 CFR S 229.1125 (Item 1125) Schedule AL - Asset-Level Information," .

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987), "A Note on the Product-Limit Estimator under Right Censoring and Left Truncation," *Biometrika*, 74, 883–886.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wang, M.-C. (1989), "A Semiparametric Model for Randomly Truncated Data," *Journal of the American Statistical Association*, 84, 742–748.

Woodroofe, M. (1985), "Estimating a Distribution Function with Truncated Data," *The Annals of Statistics*, 13, 163–177.