Annals of Actuarial Science (2024), **18**, pp. 126–151
doi:10.1017/S1748499523000180

ORIGINAL RESEARCH PAPER

Lapse risk modeling in insurance: a Bayesian mixture approach

Viviana G. R. Lobo[1,2], Thaís C. O. Fonseca[1,2] and Mariane B. Alves[1,2]

[1]Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil; and [2]Laboratório de Matemática Aplicada, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
Corresponding author: Viviana G. R. Lobo; Email: viviana@dme.ufrj.br

(Received 31 October 2022; revised 27 July 2023; accepted 02 August 2023; first published online 1 September 2023)

Abstract

This paper focuses on modeling surrender time for policyholders in the context of life insurance. In this setup, a large lapse rate at the first months of a contract is often observed, with a decrease in this rate after some months. The modeling of the time to cancelation must account for this specific behavior. Another stylized fact is that policies which are not canceled in the study period are considered censored. To account for both censoring and heterogeneous lapse rates, this work assumes a Bayesian survival model with a mixture of regressions. The inference is based on data augmentation allowing for fast computations even for datasets of over millions of clients. Moreover, frequentist point estimation based on Expectation–Maximization algorithm is also presented. An illustrative example emulates a typical behavior for life insurance contracts, and a simulated study investigates the properties of the proposed model. A case study is considered and illustrates the flexibility of our proposed model allowing different specifications of mixture components. In particular, the observed censoring in the insurance context might be up to 50% of the data, which is very unusual for survival models in other fields such as epidemiology. This aspect is exploited in our simulated study.

Keywords: Churn; lapse rate; persistency; Bayesian mixture survival model

1. Introduction

1.1 Background

Lapse rate risk modeling is an important issue that is getting attention from insurance markets. In the context of life insurance, this is an even more important issue, since contracts usually have longer policy term and higher surrender rates. Originally, the term lapse means termination of an insurance policy and loss of coverage because the policyholder has failed to pay premiums (Gatzert *et al*., 2009; Kuo *et al*., 2003; Eling & Kochanski, 2013). In this paper, lapse risk refers to the life policies surrendered before their maturity or canceled contracts when the policyholder fails to comply with their obligations (e.g., premium payment). In other words, when a customer cancels their policy or surrenders this policy, either to switch insurance companies or because someone is no longer interested, we consider that the customer has churned.

Due to the large impact lapses may produce on an insurer's portfolio, particularly in the first periods of the contracts, it is important to understand the factors that drive its risk. Large changes in lapse rates can potentially lead to financial losses which can prevent insurers from complying with their contractual obligations. Furthermore, lapse rates can be difficult to model due to the fact that while doing so, one needs not only to take into account the policyholder's behavioral features but also the characteristics of the life insurance products being acquired. Once factors associated with cancelation or surrender are identified, customer retention programs can be developed and actions can be taken (Günther *et al.*, 2014). Moreover, good persistence is of vital importance to the financial performance of life insurance companies.

We aim to address some issues related to churning, such as the existence of trends in the persistence of specific products or groups of products. Those factors may enhance the persistence curve in the insurance company. In addition, churning/lapsing impacts many actuarial tasks, such as product design, pricing, hedging, and risk management.

Cancelation rates vary through product policyholders' profiles. Statistical models can be used to identify risk factors associated with persistency (or lapse) rate over time as well as with pricing, taking the cancelation risk into account. In this context, the adoption of hierarchical regression models, survival regression models, and time series models can be useful.

Several works address lapse risk through binary (lapse vs. persistency) classification models along a single period of interest. Logistic regression models and machine learning classification techniques have been proposed to predict lapse. In this context, Brockett *et al.* (2008) pay particular attention to household customer behavior considering households for which at least one policy has lapsed and investigating the effects of the lapse rates of other policies owned by the same household. They use logistic regression and survival analysis techniques to assess the probability of total customer withdrawal. Günther *et al.* (2014) consider a logistic longitudinal regression model that incorporates time-dynamic explanatory variables and their interactions. Other examples of churn modeling via binary classification models, including binary regressions, decision trees, and neural networks, can be seen in Guillen *et al.* (2003), Ishwaran *et al.* (2008), Bolancé *et al.* (2016), and Hu *et al.* (2020, 2021).

In the present work, instead of binary classification of individuals, as to their lapse or persistence, our focus is on characterization of the temporal evolution of the instantaneous potential for lapse – equivalently, on the survival curve $S(t)$, which describes, for each time t, the probability that a policyholder will persist in a contract. We adopt survival analysis techniques aiming at jointly capturing, in a unique survival curve for each policy profile, the aggregate behavior of policyholders with a high or low potential for early cancelation. The survivor curves are stratified by policyholders' profiles defined through structured regression which can help identify modifiable features, aiming at customers' retention.

The use of survival regression analysis in the context of lapse risk is not a novelty. Milhaud & Dutang (2018) consider a competing risk approach, Eling & Kiesenbauer (2014) consider the proportional hazard models and generalized linear models to show that contracts' features such as product type or contract age, as well as policyholders' characteristics, are important drivers for lapse rates, illustrated by a dataset provided by a German life insurer. In this same line of work, Brockett *et al.* (2008) use survival analysis techniques to evaluate the time between the first cancelation and subsequent customer withdrawal. Guillen *et al.* (2012) considers survival analysis to study how long policyholders maintain their policy after the first policy lapsing. Dong *et al.* (2022) propose multistate modeling using multinomial logistic regression that allows specific behavior over a large number of different combinations of insurance coverage and across multiple periods. Our work is based on the experience gained from data of a specific company, in which higher cancelation rates are observed at the beginning of the contracts, decreasing after some months. In the database that motivated this study, the frequency of surrender is much higher than that of other events that can lead to the termination of a contract (e.g., the death of the policyholder). Thus, unlike competing or multinomial risk approaches, we implicitly assume independence between the lapse behavior of a policyholder and policy termination due to reasons other than a lapse. The interest lies in investigating, through regression structures, issues such as if there are specific persistency patterns associated with different products or groups of products, and if there are factors that can enhance the persistency curve in the company. Possibly, these factors can be controlled by the insurer, leading to increased persistency.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

Traditional parametric survival models, such as Weibull, Gamma, and Log-Normal, may fail to properly capture the shape of the instantaneous potential for lapse (known as hazard function), through time, for instance, in situations where the risk of cancelation is quite high in the first months of a contract, assuming more stable patterns in the following months. Some works seek to produce more flexible survival models by adopting semi-parametric approaches in which the modeling focus on the hazard function, indirectly inducing the descriptor model for the time to the event under analysis. In this context, constructions based on the piece-wise exponential model (PEM) are quite frequent (Friedman, 1982; Gamerman, 1994; Hemming & Shaw, 2002; Kim *et al.*, 2007; Demarqui *et al.*, 2014; Mayrink *et al.*, 2021). PEM assumes a piece-wise constant hazard function (thus an Exponential model for the time to event), per time interval. Some of the literature on PEM is dedicated to the specification of the number and cut points of the time interval partition, as well as on the transition of the constant hazard rate between consecutive time intervals. According to Demarqui *et al.* (2014) although parametric in the strict sense, the PEM can be considered as a nonparametric approach due to the large and unspecified number of parameters involved in modeling the hazard function. We propose a more parsimonious approach, directly focusing on the mixture of single-parametric components for the modeling of times to event, making the behavior of the mixed survival functions (and the implied hazard functions) more flexible than the ones associated with usual single-component parametric choices. A detailed description of Bayesian mixture models is found in Frühwirth-Schnatter (2006). We propose a mixture based on Log-Normal components, to describe the time to lapse. As will be shown further, mixing a finite and small number of Log-Normal components is sufficient to ensure flexibility to lapse risk curves that escape the rigid standards dictated by single-parametric component specifications. Although the proposed formulation depends on regressors, we assume that the heterogeneity in the persistence curves may not be completely captured by the features which are available in the analysis. Mixing components can naturally accommodate extra variability due to latent or unobserved factors, such as unobserved regressors or different temporal regimes. For instance, different economic scenarios could affect the persistence behavior, and some works explicitly take time-varying features into account (see Kuo *et al.*, 2003; Knoller *et al.*, 2015). In our formulation, the temporal indexation of observations is not exploited. Instead, we assume that a finite number of regime changes associated with temporal dynamics could be captured by latent clusters of persistence patterns. The particular choice of Log-Normal components enables the use of an efficient computational scheme for a fully Bayesian approach.

Specifically, the focus of this work is on modeling the time to lapse of a contract (or surrender). This includes the contracts which are terminated by the policyholder or terminated by the insurer due to lack of premium payment and do not include external events such as death. Policies that are not canceled are defined as censored, as the actual time to cancelation has not yet been observed in the study period.

Our contribution lies in the combination of the following points: (i) specification of a simple model, based on the mixture of Log-Normal survival regression models, which extends usual survival approaches to censored data (Ibrahim *et al.*, 2001; Kalbfleisch & Prentice, 2002) in the sense of making survival and hazard curves more flexible and fitting specific policyholders' profiles, especially in the presence of complex censoring schemes; (ii) adoption of a fully Bayesian approach, based on data augmentation and Gibbs sampler as well as other Markov chain Monte Carlo (MCMC) techniques, allowing for fast and feasible computations (see Tanner & Wong, 1987) which results in efficiency to deal with datasets of thousands of policyholders; (iii) frequentist estimation based on Expectation–Maximization (EM) algorithms, which provides point estimates for the mixture model parameters.

In short, our proposed method for lapse risk analysis allows accommodating heterogeneous behavior in the lapse hazard/survivor function via a Bayesian mixture model, taking into account a usually large volume of censored data, with computational efficiency.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

1.2 Outline of the paper

The remaining of the paper is organized as follows. Section 2 presents the Bayesian mixture model for lapse risk, beginning with a brief review on single-component parametric survival models (section 2.1), followed by the proposed mixture models and their properties (section 2.2). The mixture formulation based on Log-Normal components is presented in section 2.2.1. Sections 2.2.2 and 2.2.3 provide details on the fully Bayesian inferential and computational procedures for mixture survival modeling via data augmentation, and section 2.2.4 describes the adoption of a frequentist EM algorithm which produces point estimates. Section 3 presents two applications of the proposed method considering different simulated datasets. For the first one, the analysis focuses on the effectiveness of our proposal in modeling survival curves that have different behaviors over time with low computational costs. In the second application, a mixture of survival Log-Normal regressions is adopted to emulate a life insurance company lapse modeling, through an artificial dataset, aiming to obtain feasible results via our proposed method. As the applications show, the proposed formulation enables to compute churning probabilities, in specified time intervals, for groups of policyholders sharing similar features. Also, we consider a case study based on the Telco customer churn data which are available in IBM Business Analytics Community Connect, containing information about home phone and internet services (IBM, 2019). The idea is to illustrate the flexibility of our proposed model when compared to well-known models used in the literature of survival analysis. Section 4 concludes with a final discussion and remarks. Some aspects of the posterior distribution and simulated datasets are presented in Appendices A and B, respectively.

2. Modeling Lapse Risk Via Bayesian Mixture of Parametric Survival Models

This section presents a detailed description of the proposed mixture model, as well as the adopted inferential and computational schemes. It begins with a brief review of fundamentals on parametric single-component survival models, and an illustration is presented, based on a simulated dataset, aiming at showing that traditional parametric models can result in more rigid patterns for the hazard and survivor functions than demanded by the data.

2.1 Single-component parametric survival models

For the purpose of survival analysis, interest lies on the nonnegative random variable T_i, which denotes the duration of a policy i before termination (cancelation). The adoption of a probability density function $f(t)$ for T_i implies a survival function given by:

$$S(t) = P(T_i > t) = \int_t^{\infty} f(u)du, \tag{1}$$

which describes persistency probability evolution, through time. $S(t)$ is monotonic and decreasing function, starting at $S(0) = 1$ and converging numerically to zero, since $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

The hazard function $\lambda(t)$ describes the instantaneous potential for cancelation and is given by:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T_i \le t + \Delta t \mid T_i > t)}{\Delta t} \approx \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}, \tag{2}$$

where Δt is a small-time increment. In particular, $\lambda(t)\Delta t$ is the approximate probability of a cancelation occurring in the interval $(t, t + \Delta t)$, that is, the lapse rate, given that the policy has survived until time t (for more details, see Ibrahim *et al.*, 2001; Kalbfleisch & Prentice, 2002).

Survival analysis data typically contain censored observations, that is, data on sample units for which the event of interest was not observed during the study time. In the context of the present

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

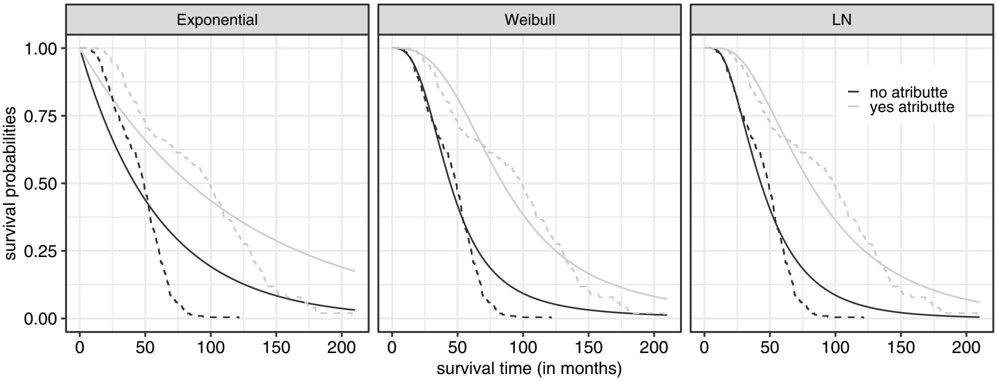


Figure 1. Simulated dataset: empirical Kaplan–Meier survival curves (dashed line) and Exponential, Weibull and Log-Normal (solid line) fitted models.

work, we define censored data as current (not canceled) policies, claim occurrences (death of the insured), and terminations of contracts.

To illustrate the larger surrender rates at the first months of a contract and a smaller rate later in time, we consider an example and assume an usual single-component parametric survival model, which proves to be inadequate for this kind of insurer portfolio behavior. An artificial database was simulated in order to emulate the cancelation behavior in insurance products with 1,000 policies, 40% censored data, and considering a dichotomous covariate x (0/no attribute, 1/yes attribute). We let the observed failure-time data be denoted by:

$$d_i = (t_i, \delta_i, x_i), \quad i = 1, \ldots, 1,000, \tag{3}$$

where the t_i is the time recorded for the ith policy, x_i is a covariate associated with the i-th policy, and δ_i is an indicator of the censoring status, given by:

$$\delta_i = \begin{cases} 1 & \text{if} \quad t_i \quad \text{is an observed lapse time} \\ 0 & \text{if} \quad t_i \quad \text{is a censored time.} \end{cases}$$

Thus, $\{\delta_i = 1\}$ is the event representing the occurrence of lapse or cancelation (with $f(t_i)$ probability density), whereas $\{\delta_i = 0\}$ denotes censored lapse time or, equivalently, nonoccurrence of lapse during the study period (with survival probability $S(t_i)$), that is, lapse/cancelation time is only known to be greater than t_i. The observations on different policies are assumed to be independent. The data were simulated considering variable $Y_i = \log(T_i)$ as the logarithmic the duration of a policy i before termination from a mixture of Gaussian distributions. For a more detailed discussion about data generation, see Appendix B.

Fig. 1 presents the empirical Kaplan–Meier survival curve (dashed line) adjusted to the simulated data, for both attributes, and it is clear that there is a heterogeneous behavior between the levels of the dichotomous covariate. The absence of the attribute described by the covariate is associated with an increased premature risk of cancelation. In addition, for both levels of the covariate, there is a difference in the survival behavior in the initial times when compared to the following ones.

In the context of survival analysis, parametric models play a key role in modeling the phenomenon of interest. If T_i follows an Exponential model, then $\lambda(t) = \lambda$ is constant over time. If a Weibull model with parameters λ and κ is considered for T_i, then $\lambda(t) = \lambda\kappa t^{\kappa-1}$, resulting in constant ($\kappa = 1$), growing ($\kappa > 1$), or decreasing ($\kappa < 1$) hazard through time. The Log-Normal model with parameters μ (mean of the logarithmic failure time) and σ (standard deviation of the logarithmic failure time) is an usual choice when the hazard function is not monotonous, reaching

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

a maximum point and then decreasing. Structured regression models based on covariates are usually considered, relating the parameters in the sampling model with covariates responsible for the description of each contract profile. The simulated data were fit by the Exponential, the Weibull, and the Log-Normal survival regression models.

Panels in Fig. 1 (see solid line) make it clear that usual survival models are not flexible enough to accommodate the behavior of the empirical survival function, per stratum, over time. Even though the Log-Normal model produces a good performance when compared with the competing models in the initial instants, it fails to adapt in later times as in the other competing models. That is, the rates tend to decrease faster over time, and usual parametric survival models are not able to accommodate this behavior. Our working premise is that simple parametric models can serve as block builders of more flexible models, via mixtures.

2.2 Bayesian mixture of parametric survival models

In this section, we present our proposal of a mixture of parametric models for censored survival data. Finite mixture models are described in detail in Frühwirth-Schnatter (2006). As seen in the illustration with artificial data presented in section 2.1, the competing fitted models apparently do not reflect the empirical distribution of the data. A possible alternative is to use more flexible structures such as mixture models, which allow the incorporation of behavioral change in the probability distribution of the data. Our proposal is based on the classical finite mixture model, where observations are assumed to arise from the mixture distribution given by:

$$f(t_i) = \sum_{j=1}^{K} \eta_j f_j(t_i), \tag{4}$$

where $f(t_i)$ is the probability density function of T_i and $f_j(t_i)$ denotes K component densities occurring with unknown proportions η_j, with $0 \leq \eta_j \leq 1$ and $\sum_{j=1}^{K} \eta_j = 1$. It follows that the survival function $S(t)$ has the mixed form:

$$S(t) = \sum_{j=1}^{K} \eta_j S_j(t) = \sum_{j=1}^{K} \eta_j \int_t^{\infty} f_j(u)du, \tag{5}$$

and the hazard function has the mixed form:

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{\sum_{j=1}^{K} \eta_j S'{}_j(t)}{S(t)} = \frac{\sum_{j=1}^{K} \eta_j \lambda_j(t) S_j(t)}{S(t)} = \sum_{j=1}^{K} \eta_j h_j(t), \tag{6}$$

where $h_j(t) = \lambda_j(t)S_j(t)/S(t)$.

For each policy i, we define a latent group indicator $I_i \mid \boldsymbol{\eta} \sim Categorical(K, \boldsymbol{\eta})$, with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$, following a categorical distribution given by $P(I_i \mid \boldsymbol{\eta}) = \prod_{j=1}^{K} [\eta_j]^{I_{ij}}$, where $I_{ij} = 1$ if observation i is allocated to group j and is null, otherwise. These auxiliary non-observable variables aim to identify which mixture component each observation has been generated from and their introduction in the formulation makes it simple to express the likelihood function for observation i:

$$f_{T_i}(t_i \mid I_i) = \prod_{j=1}^{K} [f_j(t_i)]^{I_{ij}}, \quad i = 1, \ldots, n. \tag{7}$$

Notice that in the example shown in section 2.1, $K = 2$ components are assumed in addition to a regressor with two levels, for the generation of the artificial lapse risk data. In this case, the lapse risk can be decomposed into two overlapping processes in time. The first process is associated with the period that immediately follows the contracting of the policy. In general, it covers the first

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

three months when the lapse risk is relatively high. The second process refers to the subsequent period when persistence decreases smoothly over time. Thus, the mixture distribution is written as $f_{T_i}(t_i \mid I_i) = [f_1(t_i)]^{I_{i1}} [f_2(t_i)]^{(1-I_{i1})}$. It is worth noting that single-component models are obtained as a particular case of the mixture structure when $\eta_1 = 1$, $\eta_j = 0, j > 1$.

For the estimation process in this work, the number K of mixture components is assumed to be known and subjective to a sensitivity analysis. As usual in applied contexts, K is subject to uncertainty. Frühwirth-Schnatter (2006, Chap 4) provides a good discussion of informal methods seeking to identify the number of components in a mixture, including the evaluation of the predictive behavior of a statistic $T(Y)$ for future realizations of the response, conditional on its past values $y_1, \ldots, y_n$ and on a model with fixed K components, for different specifications of K. The choice of K is then indicated by the specification that leads to the best value of the predictive performance evaluation statistic, among the proposed mixture structures. We performed sensitivity analysis to choose the number of components, observing the quality of the fit of the survival and hazard curves, compared to their respective empirical versions, which can be summarized by statistics indicating the performance of the fit. If K is not fixed, its estimation would lead to models with variable-dimensional parametric space. Details on Bayesian methods in this context can be seen in Frühwirth-Schnatter (2006, Chap 5). As shown in the following subsections, we offer a computationally efficient method for estimating a model with fixed K, and we consider that the sensitivity analysis to different specifications of K, which can be performed in reduced time, is sufficient for our purposes.

2.2.1 Mixture of Log-Normal components

In this paper, interest lies in modeling $Y_i = \log(T_i)$, the logarithmic duration of a policy i before termination (cancelation), such that $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, implying that in the original scale $T_i \sim \mathcal{LN}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. Survival times associated with a different outcome than the lapse, such as survival times past the end of our study and deaths, are assumed to be censored for policies, $i = h+1, \ldots, n.$, then

$$f(t_i \mid \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} (t_i \sigma)^{-1} \exp\left\{ -\frac{1}{2\sigma^2} (\log(t_i) - \mu_i,)^2 \right\}. \tag{8}$$

The resulting survival function is given by:

$$S(t_i \mid \mu_i, \sigma^2) = 1 - \Phi\left(\frac{\log(t_i) - \mu_i}{\sigma} \right). \tag{9}$$

We can thus write the survival likelihood function of (μ, σ^2) implied by a Log-Normal model and based on data D as:

$$L(\mu_i, \sigma^2 \mid D) = \prod_{i=1}^{n} f(t_i \mid \mu_i, \sigma^2)^{\delta_i} S(t_i \mid \mu_i, \sigma^2)^{(1-\delta_i)}. \tag{10}$$

Adopting the finite mixture approach and assuming $Y_i = log(T_i) \mid \mu_i, \sigma^2 \sim \mathcal{N}(\mu_i, \sigma^2)$ and latent variables I_{ij}, $i = 1, \ldots, n$, it follows that:

$$f_{Y_i}\big(y_i \mid \mathrm{I}_i, \mu_i, \sigma^2\big) = \prod_{j=1}^{K} [\mathcal{N}_j(y_i \mid \mu_{ij}, \sigma_j^2)]^{\mathrm{I}_{ij}}, \tag{11}$$

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

where $\mathcal{N}_j$ is a normal distribution for the component group $j, j = 1, \ldots, K$. Then,

$$\begin{aligned} y_i \mid \{\mathrm{I}_{i1} = 1\}, \boldsymbol{\beta}_1, \sigma_1^2 &\sim \mathcal{N}_1(\mu_{i1}(\boldsymbol{\beta}_1), \sigma_1^2) \\ y_i \mid \{\mathrm{I}_{i2} = 1\}, \boldsymbol{\beta}_2, \sigma_2^2 &\sim \mathcal{N}_2(\mu_{i2}(\boldsymbol{\beta}_2), \sigma_2^2) \\ &\vdots \quad \vdots \\ y_i \mid \{\mathrm{I}_{iK} = 1\}, \boldsymbol{\beta}_K, \sigma_K^2 &\sim \mathcal{N}_K(\mu_{iK}(\boldsymbol{\beta}_K), \sigma_K^2), \end{aligned}$$

with $\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})$ characterizing the unknown mean and σ_j^2, the variance, respectively, for $j = 1, \ldots, K$ and $i = 1, \ldots, n$.

Using the latent indicators of categorical allocation, the likelihood simplifies to

$$\begin{aligned} f(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2) &= \prod_{j=1}^{K} \prod_{i=1}^{n} \eta_j^{I_{ij}} [\mathcal{N}_j(y_i \mid \boldsymbol{\beta}_j, \sigma_j^2)]^{I_{ij}} \\ &= \prod_{j=1}^{K} \eta_j^{n_j} \left[\prod_{i:I_{ij}=1} \mathcal{N}_j(y_i \mid \boldsymbol{\beta}_j, \sigma_j^2) \right], \end{aligned} \tag{12}$$

where $n_j = \sum_i I_{ij}$ is the number of observations allocated to group j and $n = \sum_{j=1}^K n_j$, for $j = 1, \ldots, K$, $i = 1, \ldots, n$. Thus, the mixture survival likelihood function is given by:

$$f(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2) = \prod_{j=1}^{K} \eta_j^{n_j} \left[\prod_{i:I_{ij}=1} \mathcal{N}_j(y_i \mid \boldsymbol{\beta}_j, \sigma_j^2)^{\delta_i} S(y_i \mid \boldsymbol{\beta}_j, \sigma_j^2)^{1-\delta_i} \right], \tag{13}$$

where δ_i in the censorship indicator, as previously seen in section 2.1.

From a Bayesian point of view, we are interested in the posterior $p(\boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$. The posterior distribution are generally not available analytically, and numerical integration and simulation are considered, in particular, MCMC methods (Gamerman & Lopes, 2006) are used in this paper. Notice that to compute posterior distributions, we need to take into account the censored quantities, which in practice can be computationally prohibitive, depending on the percentage of censored data and the large dataset. Thus, inference is facilitated through data augmentation.

2.2.2 Inference based on data augmentation

The presence of censored data is a common feature when considering time data until the occurrence of an event and the likelihood function takes this fact into account, as seen in equation (10). Following the Bayesian approach, the estimation procedure can be based on MCMC algorithm using the data augmentation technique (see Tanner & Wong, 1987). Suppose we observe logarithmic survival times $\mathbf{y}^{obs} = (y_1^{obs}, \ldots, y_h^{obs})$. Then the idea is to define the logarithmic survival times for the $n - h$ censored policies as missing data which we denote as $\mathbf{z} = (z_{h+1}, \ldots, z_n)$.

Assume that the full data is given by:

$$\mathbf{y} = (\mathbf{y}^{obs}, \mathbf{z}). \tag{14}$$

Let $\boldsymbol{\theta}$ be the parametric vector of interest. The data augmentation approach is motivated by the following representation of the posterior density:

$$p(\boldsymbol{\theta} \mid \mathbf{y}^{obs}, \boldsymbol{\delta}) = \int_{\mathbf{Z}} p(\boldsymbol{\theta} \mid \mathbf{y}^{obs}, \mathbf{z}, \boldsymbol{\delta}) p(\mathbf{z} \mid \mathbf{y}^{obs}, \boldsymbol{\delta}) d\mathbf{z}, \tag{15}$$

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

where the vector $\boldsymbol{\delta}$ is composed by censorship indicators $\delta_i \in \{0, 1\}$; $p(\boldsymbol{\theta} \mid \mathbf{y}^{obs}, \boldsymbol{\delta})$ denotes the posterior density of the parameter $\boldsymbol{\theta}$ given the observed data $\mathbf{y}^{obs}$; $p(\mathbf{z} \mid \mathbf{y}^{obs}, \boldsymbol{\delta})$ denotes the predictive density of latent data $\mathbf{z}$ given $\mathbf{y}^{obs}$; and $p(\boldsymbol{\theta} \mid \mathbf{y}^{obs}, \mathbf{z}, \boldsymbol{\delta})$ denotes the conditional density of $\boldsymbol{\theta}$ given the augmented data $\mathbf{y}$.

In practice, we do not know *a priori* to which group a given observation belongs. Thus, in addition to the censored observations, whose outcome is unknown, the variable I_i in equation (7) is also latent and is estimated in our inferential algorithm. Assuming that policies are independent, the likelihood function for the complete data can be written as:

$$f(\mathbf{y}^{obs}, \mathbf{z}, \boldsymbol{\delta} \mid \boldsymbol{\theta}) = \prod_{j=1}^{K} \eta_j^{n_j} \left[\prod_{i:\delta_i=1, I_{ij}=1} f_j(y_i^{obs} \mid \boldsymbol{\theta}_j) \prod_{i:\delta_i=0, I_{ij}=1} f_j(z_i \mid \boldsymbol{\theta}_j) \mathcal{I}(z_i \geq y_i^{obs}) \right], \tag{16}$$

where $\boldsymbol{\theta}_j = (\eta_j, \boldsymbol{\beta}_j, \sigma_j^2)$, $n_j = \sum_i I_{ij}$, $\sum_{j=1}^{k} n_j = n$, and $\mathcal{N}_j \sim f_j(\cdot \mid \boldsymbol{\theta}_j)$. Following Bayes' theorem, the posterior distribution of the model parameters and latent variables, given the complete data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_K)'$, is proportional to

$$\begin{aligned} p\big(\boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \boldsymbol{\delta}\big) &\propto f(\mathbf{y}^{obs}, \mathbf{z}, \boldsymbol{\delta} \mid \boldsymbol{\theta}) p(I \mid \boldsymbol{\eta}) \pi(\boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2) \\ &\propto \prod_{j=1}^{K} \eta_j^{n_j} \left[\prod_{i:\delta_i=1, I_{ij}=1} f_j(y_i^{obs} \mid \boldsymbol{\theta}_j) \prod_{i:\delta_i=0, I_{ij}=1} f_j(z_i \mid \boldsymbol{\theta}_j) \mathcal{I}(z_i \geq y_i^{obs}) \right] \\ &\times \prod_{i=1}^{n} p(I_i \mid \boldsymbol{\eta}) \pi(\boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^2). \end{aligned} \tag{17}$$

The Bayesian mixture model is completed by the prior distribution specification. We assume independence in the prior distribution with $\boldsymbol{\eta} \sim Dirichlet(K, \alpha)$, where $\sum_{j=1}^{K} \eta_j = 1$ and $\alpha = (\alpha_1, \ldots, \alpha_K)$ is a vector of hyperparameters, such that $\alpha_j > 0$; $\phi_j = \frac{1}{\sigma_j^2} \sim Gamma(a_j, b_j)$, the regression coefficients, $\boldsymbol{\beta}_j \sim \mathcal{N}_j(\boldsymbol{m}_j, \tau_j^2 \boldsymbol{I}_p)$ and $P(I_{ij} = 1) = \eta_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, K$.

The resulting posterior distribution in equation (17) does not have closed form, and we appeal to MCMC methods to obtain samples from the posterior distribution. In particular, posterior samples are obtained through a Gibbs sampler algorithm, where the Markov chain is constructed by considering the complete conditional distribution of each hidden variable given the others and the observations. The scheme is presented in the following section.

2.2.3 Computational scheme for the mixture of Log-Normal components

Assuming K groups, it is possible to consider the Gibbs sampler algorithm in order to overcome the numerical integration condition of the data augmentation techniques, resulting in a computationally efficient algorithm.

We consider the following Bayesian Gaussian mixture survival model with data augmentation:

$$\begin{aligned} \mathbf{y}_j \mid I_i, \boldsymbol{\beta}_j, \phi_j &\sim \mathcal{N}_j(x_i^T \boldsymbol{\beta}_j, \phi_j^{-1}) \\ I_i \mid \boldsymbol{\eta} &\sim Categorical(K, \boldsymbol{\eta}) \\ \boldsymbol{\eta} &\sim Dirichlet(\alpha_1, \ldots, \alpha_K) \\ \boldsymbol{\beta}_j &\sim \mathcal{N}_j(\boldsymbol{m}_j, \tau_j^2 \boldsymbol{I}_p) \end{aligned}$$

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

$$\phi_j \sim Gamma(a_j, b_j)$$

$$p(z_i) \propto \mathcal{I}\{z_i \geq y_i\}$$

Algorithm 1 shows the scheme to estimate the parameters via data augmentation with censored observations. Details involved in obtaining the full conditional distributions can be seen in Appendix A.

Algorithm 1. Gibbs sampler for a finite Gaussian mixture survival model with data augmentation.

```
Input: Initialize all parameters $\boldsymbol{\eta}^{(0)}, \boldsymbol{\beta}_j^{(0)}, \phi_j^{(0)}$ for all $j = 1, \ldots, K$
step 1  Update $I_i$ sampling from $I_i^{(k+1)} \sim I_i \mid \mathbf{y}, \boldsymbol{\eta}^{(k)}, \boldsymbol{\beta}_j^{(k)}, \phi_j^{(k)}$.
step 2  Update $\boldsymbol{\eta}$ sampling from $\boldsymbol{\eta}^{(k+1)} \sim \boldsymbol{\eta} \mid \mathbf{y}, I_i^{(k+1)}$.
    if $\delta_i = 0$ for $i = 1, \ldots, n$ then consider the data augmentation and define a latent variable $\mathbf{z}$ as;
        $y_i^{obs} \mid z_i, \left\{I_{ij}^{(k+1)} = 1\right\} \sim \mathcal{NT}_{(-\infty, z_i)}\left(x'_{ij}\boldsymbol{\beta}_j^{(k)}, \phi_j^{-1(k)}\right)$
step 3  Update $\phi_j$ sampling from $\phi_j^{(k+1)} \sim \phi_j \mid \mathbf{y}, I_i^{(k+1)}$.
step 4  Update $\boldsymbol{\beta}_j$ sampling from $\boldsymbol{\beta}_j^{(k+1)} \sim \boldsymbol{\beta}_j \mid \mathbf{y}, I_i^{(k+1)}, \phi_j^{(k+1)}$.
step 5  Order $\boldsymbol{\beta}_j^{(k+1)}$ and arrange $\boldsymbol{\eta}^{(k+1)}$ and $\phi_j^{(k+1)}$ accordingly.
step 6  $k = k + 1$. Go back to step 1 until convergence.
```

2.2.4 Point estimation via EM

Optimization methods to obtain maximum likelihood estimates are less computationally expensive than Monte Carlo estimation because they depend uniquely on numerical convergence. In order to obtain point estimates efficiently, we consider the maximization of log-likelihoods for the mixture model.

Consider K mixture components. In the classical context, the mixture model without censored data is described by equations (11) and (12), respectively. Furthermore, without considering censored data and latent indicators of the mixture components, the log-likelihood function to be maximized is given by:

$$l\left(\{\eta_j\}_{j=1}^{K}, \left\{\boldsymbol{\beta}_j\right\}_{j=1}^{K}, \left\{\sigma_j^2\right\}_{j=1}^{K}\right) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \eta_j f_j(Y_i \mid \boldsymbol{\beta}_j, \sigma_j^2). \tag{18}$$

Notice that the expression depends on logarithms of sums, which cannot be simplified through logarithmic properties. The estimation, in this context, is exhaustive and without analytic or recursive forms for the maximum likelihood estimators (MLEs) of the model parameters.

In the mixture distribution context, it is very common to use the EM algorithm proposed by Dempster *et al.* (1977) which is an iterative mechanism to calculate the MLE in the presence of missing observations. Given the use of latent variables, and conditional on I, equation (11) is valid. Thus, it provides a probability distribution over the latent variables together with a point estimate for parameters. If a prior distribution is assumed for the parameters, the joint posterior mode is obtained by the method.

Besides that, when we take into account the censored observed data, the data augmentation technique, as previously seen, can be applied by including a new latent variable vector $\mathbf{z}$. According to our model, censored observations are originated from a truncated normal distribution.

Assume that observation y_i is censored. That is, there is an unobserved datum z_i such that $z_i \mid \{I_{ij}=1\} \sim f_j$ and $y_i^{obs} \mid z_i, \{I_{ij}=1\} \sim \mathcal{NT}_{(-\infty, z_i)}(x_{ij}'\boldsymbol{\beta}_j, \sigma_j^2)$. The strategy that we will adopt in the algorithm is to remove the truncation from the observed data y_i^{obs} to obtain z_i, at each iteration (k), so that $y_{1:n}^{(k)} = (y_1^{obs}, y_2^{obs}, \ldots, y_h^{obs}, z_{h+1}^{(k)}, \ldots, z_n^{(k)})$, as previously seen in section 2.2.2.

Algorithm 2 is adapted for this context. For more details, see Jedidi *et al.* (1993). Notice that $E(z_i \mid y_i^{obs}, \{I_{ij}=1\})$ and $Var(z_i \mid y_i^{obs}, \{I_{ij}=1\})$ denote the expected value and variance of a truncated Gaussian distribution, respectively. Although the computational cost (to obtain a point estimate) is smaller when compared to the proposal in sections 2.2.2 and 2.2.3, a disadvantage of this approach is that the EM algorithm is quite sensitive to the choice of initial parameters and does not take into account the uncertainty associated with parameter estimates. Besides, the EM algorithm will converge very slowly if a poor choice of initial value $\boldsymbol{\eta}^{(0)}$, $\boldsymbol{\beta}_j^{(0)}$, and $\sigma_j^{2(0)}$ is selected.

Algorithm 2. Expectation–Maximization algorithm for a finite Gaussian mixture survival model with data augmentation.

```
Input: Initialize all parameters $\boldsymbol{\eta}^{(0)}, \boldsymbol{\beta}_j^{(0)}, \sigma_j^{2(0)}$ for all $j = 1, \ldots, K$
step 1 Define the latent variable $z$ as
    $z_i^{(k)} = \begin{cases} y_i^{obs}, & \text{if} \quad \delta_i = 1 \;\; \forall i = 1, \ldots, h \\ \sum_{j=1}^{k} w_{ij}^{(k-1)} E(z_i \mid y_i^{obs}, I_i), & \text{if} \quad \delta_i = 0 \;\; \forall i = h+1, \ldots, n. \end{cases}$
step 2 Compute $w_{ij}$, the posterior probability that the observation was generated from
the $j$-th mixture component
    $w_{ij}^{(k+1)} = \frac{\hat{\eta}_j^{(k)} f_j(z_i^{(k)} \mid \boldsymbol{\beta}_j^{(k)}, \sigma_j^{(k)})}{\sum_{j=1}^{K} \hat{\eta}_j^{(k)} f_j(z_i^{(k)} \mid \boldsymbol{\beta}_j^{(k)}, \sigma_j^{(k)})},$
with $i = 1, \ldots, n$ and $j = 1, \ldots, K$.
step 3 Update $\eta_j$ as $\hat{\eta}_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^{(k+1)}$.
step 4 Update $\boldsymbol{\beta}_j$ and $\sigma_j^2$ as
    - $\hat{\boldsymbol{\beta}}_j^{(k+1)} = \left(X'W_j^{(k+1)}X\right)^{-1} X'W_j^{(k+1)}\mathbf{z}^{(k)}$, $j = 1, \ldots, K$ and
      $W_j^{(k+1)} = diag\left(\left\{w_{ij}^{(k+1)}\right\}_{i=1}^{n}\right)$
    - $\hat{\sigma}^{2}{}_j^{(k+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(k+1)}\left(z_i^{(k)} - \mu_{ij}^{(k+1)}\right)^2 + \sum_{\{i:\delta_i=0\}} w_{ij}^{(k+1)} Var\left(z_i^{(k)} | y_i^{obs}, \{I_{ij}=1\}\right)}{\sum_{i=1}^{n} w_{ij}^{(k+1)}}$, where
      $\mu_{ij}^{(k+1)} = x_{ij}^T \boldsymbol{\beta}_j^{(k+1)}$.

Run until convergence is achieved.
```

3. Applications

This section presents three applications of our proposed model: (i) a simulated dataset study to evaluate the performance and computational cost of our proposed methodology via data augmentation; (ii) a realistic simulated dataset that emulates a real portfolio and features often present in real actuarial data; and (iii) a case study based on the Telco customer churn data which are available in IBM Business Analytics Community Connect, containing information about home phone and internet services `https://community.ibm.com/community/user/businessanalytics/home`, IBM (2019). The idea of the case study is to illustrate the

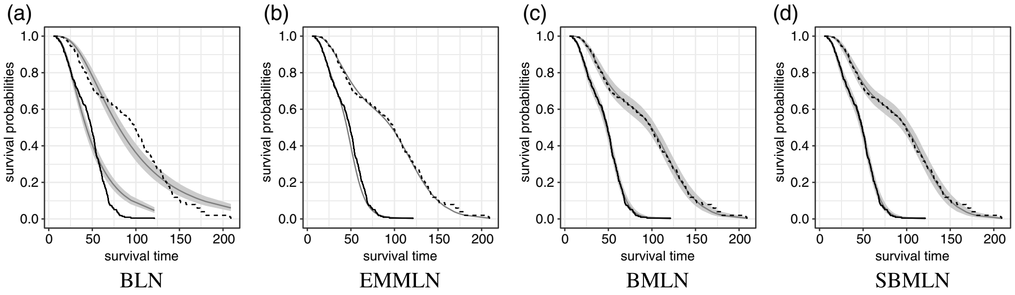


Figure 2. Simulated with 40% censored dataset: posterior survival probabilities with mean (gray line) and limits of 95% credible interval: (a) the Bayesian Log-Normal model (BLN), (b) the point estimation mixture Log-Normal model via EM algorithm (EMMLN), (c) the Bayesian mixture Log-Normal model (BMLN) with data augmentation, and (d) Bayesian mixture Log-Normal model via Stan (SBMLN), considering $n = 1,000$ policies. Categories: yes attribute (black dashed line) and no attribute (black solid line).

flexibility of our proposed model when compared to well-known models used in the literature of survival analysis. Codes with descriptive data analysis, inference procedure, and database are available at https://github.com/vivianalobo/LapseRiskAAS.

3.1 Simulated dataset

In this section, we return to the illustrative dataset seen in section 2.1. Our aim is to compare the usual and mixture survival Log-Normal models under a Bayesian approach through our proposals described in sections 2.2.1 and 2.2.2.

We simulate three scenarios: (i) a dataset with 10% of censored data; (ii) a dataset with 40% of censored data; and (iii) a dataset with 60% of censored data, considering a mixture of $K = 2$ components, with $\eta_1 = \eta = 0.6$. We would like to assess whether our proposal is efficient in sampling from the posterior distribution as well as its computational efficiency, for the model of interest. In addition, we vary the sample size ($n = 1,000;10,000;50,000;100,000$) in order to evaluate the computational cost through (a) our proposal with data augmentation for censored data. To allow fast computations, we perform part of the code implementation with the C++ programming language to build improved loops of complete conditional distributions through the RcppArmadillo package available in R software (see Eddelbuettel & Sanderson, 2014); and (b) without data augmentation via RStan package available in R (Stan Development Team, 2018; Carpenter *et al.*, 2017), that is, considering the survival likelihood given by equation (13). Stan is a C++ library for Bayesian modeling and inference that primarily uses the No-U-Turn sampler (NUTS) (see Hoffman & Gelman, 2014) to obtain posterior simulations given a user-specified model and data.

The survival time can be analyzed according to $\log(T_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$, with x the covariate that takes values ($x = 0$ or $x = 1$) and error $\varepsilon_i \sim N(0, \phi^{-1})$. We assign vague independent priors to the parameters in $\boldsymbol{\theta}$ with $\phi_j \sim Gamma(0.01, 0.01)$, $\boldsymbol{\beta}_j \sim \mathcal{N}_j(\mathbf{0}, 100\boldsymbol{I}_2)$, and $\boldsymbol{\eta} \sim Dirichilet(\alpha_1 = 2, \alpha_2 = 2)$, for $j = 1, 2$. We run an MCMC chain for 20,000 iterations and consider the first 10,000 out as burn-in. The burn-in and lag for spacing of the chain were selected so that the effective sample size was around 1,000 samples.

Fig. 2 illustrates the fit of the survival curves by the competing models considering a sample with 1,000 policies and 40% rate of censorship (as seen in Fig. 1 (a)), the usual Bayesian Log-Normal (BLN) model (without mixture, like in Fig. 1 (d)) and the Bayesian mixture Log-Normal model (BMLN) (see panels (c) and (d) in Fig. 2). As can be seen, the proposed mixture model is able to accommodate different behaviors in the survival curves when compared to the usual

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

Table 1. Posterior summaries comparison: with mean and 95% credibility for the Bayes LN, the Bayes Mixture LN via Stan, and the Bayes Mixture LN; the point estimation via EM Mixture LN, for a simulated dataset with 40% censorship rate and considering $n = 1,000$ policies.

	Bayes LN		Bayes MLN (Stan)		Bayes MLN		EM MLN
True	Mean	IC 95%	Mean	IC 95%	Mean	IC 95%	Pointwise
$\beta_{0,j=1} = 3.3$	3.76	(3.70,3.82)	3.30	(3.17,3.44)	3.30	(3.16,3.44)	3.39
$\beta_{0,j=2} = 4.0$	–	–	4.05	(4.01,4.09)	4.05	(4.02,4.08)	3.98
$\beta_{1,j=1} = 0.5$	0.62	(0.53,0.72)	0.51	(0.39,0.63)	0.51	(0.38,0.64)	0.51
$\beta_{1,j=2} = 0.8$	–	–	0.77	(0.71,0.83)	0.77	(0.72,0.82)	0.84
$\sigma^2_{j=1} = 0.3$	0.61	(0.58,0.65)	0.23	(0.17,0.31)	0.24	(0.17,0.31)	0.28
$\sigma^2_{j=2} = 0.039$	–	–	0.04	(0.03,0.06)	0.04	(0.03,0.06)	0.04
$\eta = 0.60$	–	–	0.56	(0.47,0.63)	0.56	(0.47,0.63)	0.53

Log-Normal model. In addition, the uncertainty associated with estimates is lower for our proposed mixture model. Panel (b), in Fig. 2, exhibits the point estimation via EM for the Log-Normal mixture modeling. The estimated survival curves via the EM algorithm follow the behavior of the empirical Kaplan–Meier curves. Point estimates of the parameters of interest are reasonable compared to those obtained via Gibbs sampler techniques. See more details about the simulated dataset in Appendix B.

Table 1 shows the posterior summaries for the survival Log-Normal model without mixture (Bayes LN) and considering Log-Normal mixtures via our proposal via data augmentation (BMLN) and via Stan (BMLN and SBMLN), respectively. As already mentioned, the non-mixture model is not able to capture the behavior of the survival curve. The structure of the non-mixture model does not allow the incorporation of mixture components in the coefficient estimates. On the other hand, the mixture Log-Normal model is capable of producing suitable estimates for the true parameters. Although the data augmentation proposal and the Stan method lead to similar point and interval estimates, the processing computational cost via Stan is much higher for all scenarios, as can be seen in Table 2. In this way, the use of Stan for large samples, highly censored observations, and considering more covariates in the survival model can be prohibitive. For the EM algorithm, 58 iterations were required until the parameters converged, which resulted in a computational time of 3.32 seconds. However, as already stated, the EM algorithm does not generate uncertainty measures associated with estimates.

3.2 A simulated dataset in insurance

In this section, we simulated a dataset aiming to emulate the behavior of a realistic portfolio in the private life insurance sector. We simulate a large dataset emulating 100,000 policies over 100 months, taking into account heterogeneous lapse rates and realistic censored data with an approximate 42.7% censorship rate and two mixing components based on $\eta = 0.6$, via the mixture Log-Normal model previously seen in section 2.2.1. To illustrate, this dataset contains individual policyholder information as well as information about the subscription.

The factors considered in this study are `gender` (male and female), `age group` (18–29 years, 30–49 years, and 60+ years), `policy type` (standard and gold), where the level gold represents a segmentation of insureds that have a high insured capital and the premium `payment` mode (monthly, yearly, i.e., regular premium or single premium) of the policy. Time to churn is the response variable of interest.

Panel (a) in Fig. 3 presents the simulated survival times in the log scale, indicating the mixing of two distributions. In panels (b)–(c), the empirical survival curve and hazard rate behavior show that the lapse rate is higher and sharply falls for the first periods of time after subscription

Table 2. Comparison of computational times (in seconds) involved in the Bayesian competing methods for simulated datasets with 10%, 40%, and 60% rates of censorship and considering *n* (size) policies.

% censored	Size *n*	Bayes MLN	Bayes MLN (Stan)
10%	1,000	2.20	637.81
	10,000	4.27	7,274.85
	50,000	76.45	23,913.10
	100,000	168.98	65,452.20
40%	1,000	2.33	831.70
	10,000	18.81	8,442.68
	50,000	108.01	44,036.70
	100,000	234.18	83,442.90
60%	1,000	3.33	1,005.89
	10,000	21.42	9,568.67
	50,000	119.18	49,751.10
	100,000	263.75	106,977.0

Computational system Ubuntu, Intel Core i5-10400F CPU 2.90 GHZ, 16 GB RAM.

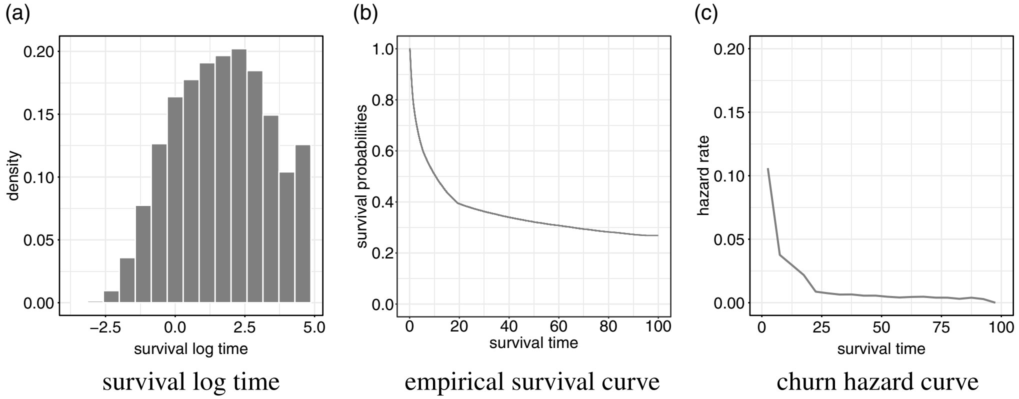


Figure 3. Simulated dataset: a summary about the survival time with (a) survival time in log scale, (b) empirical survival curve, and (c) churn hazard curve.

initiation, then it stabilizes for some time, exhibits a peak close to 20 months, and then gradually decreases.

Fig. 4 presents the marginal empirical Kaplan–Meier survival curve for the variables of this study. As we can be seen, categories for each variable present particular behavior and could be useful to understand the time to churn. There is a noticeable drop in active insured in the first periods after the policy subscription. Some reasons could be discussed, such as the policyholders who subscribe just to test a service or there may be some association with the premium payment mode. For the age group, there is no visible difference in patterns of survival probability between the 18–29 years and 30–59 years of age groups, except for a drop in during the first months of subscription and nearby the final portion of the survival curve.

Fig. 5 shows the performance of the fitted curves for the competing models, the Log-Normal model, and the Log-Normal mixture model for some scenarios. Panel (a) represents the characteristics of the policyholder in scenario 1 (male, standard, 30–59 years of age, and monthly payment),

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

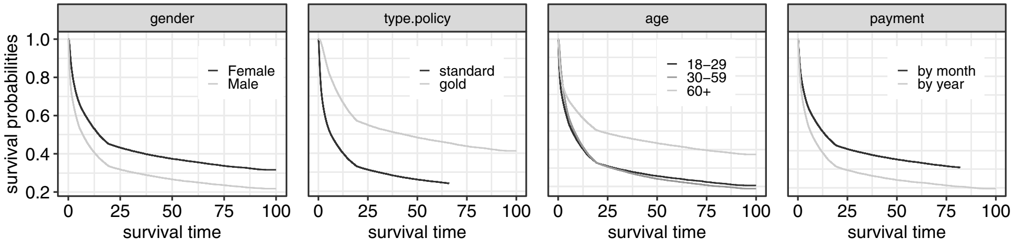


Figure 4. Simulated dataset: marginal empirical survival curves via Kaplan–Meier for the covariates: gender, type policy, age, and payment.

panel (b) exhibits the fitted curves for scenario 2 (male, gold, 60+ years of age, and year), and panel (c) exhibits the estimated survival curves for scenario 3 (female, standard, 60+ years of age, and month). As we can see, policyholders in scenario 1 have lower persistence when compared to scenarios 2 and 3, respectively. This behavior is understood due to the fact of the standard group and 30–59 years of age group present survival empirical curves with more abrupt decay than 60+ years of age and gold categories.

Our proposed Bayesian mixture survival model is able to capture the behavior of the empirical curves (see Fig. 5). Note that the BLN produces a poor fit due to the fact this model is not allowed to access distinct behavior of the curve at different times of the study. Besides that, the BMLN and EMMLN converge for the true values generated from the simulated dataset. In order to understand the lapse of the policyholders, especially in the initial periods of the contract, we consider obtaining some probabilities to profile the policyholder to customer retention via the BMLN. We hope that the longer customers are with the insurance company, the less likely they are to cancel a policy. Table 3 shows a summary of the probabilities conditional on the hypothesis that the policyholder has survived the first three months, that is, conditional on the event $L = \{T > 3\}$ for the scenarios in Fig. 5. As one can see, the risk probability of a lapse occurring in the first year is higher for scenario 1 (0.394) when compared to scenarios 2 and 3 (0.247 and 0.123), respectively. On the other hand, the conditional probability that the policyholder maintains the contract term after 36 months, having survived to the first months ($P(T \geq 36 \mid T > 3)$), is high for all considered scenarios. Other probabilities can be evaluated in order to understand the profile of the company's policyholders. To illustrate, Table 4 exhibits the results obtained conditioning on other events, such as $\{T > 12\}, \{T > 24\}$, and $\{T > 36\}$. As expected, the probability of churn decreases, having the insured persisted in the insurance company for long periods, that is, the lapses reduce substantially with increasing policy age. Furthermore, when identifying insured profiles, we are able to understand which profiles need more attention; thus, the insurance company could develop retention strategies focused on these policyholders.

3.3 Case study: Telco customer churn

The Telco customer churn data is available on IBM Business Analytics Community Connect (IBM, 2019) and contains information about a Telco company that provides home phone and internet services to 7,043 customers in California in the United States. The dataset contains 18 variables about the profile of the customers and a variable indicating who has left or stayed in the service resulting in approximately 73.4% of censoring. The response variable `tenure` represents the number of months the customer has stayed with the company with an average of 32 months. Lifetimes equal to zero were removed resulting in a dataset of 7,032 customers. Demographic characteristics are included for each customer, as well as customer account information and service information. Following previous studies, we aim to predict customer churn rate behavior and identify the most important factors that contribute to high- or low-lapse risks.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

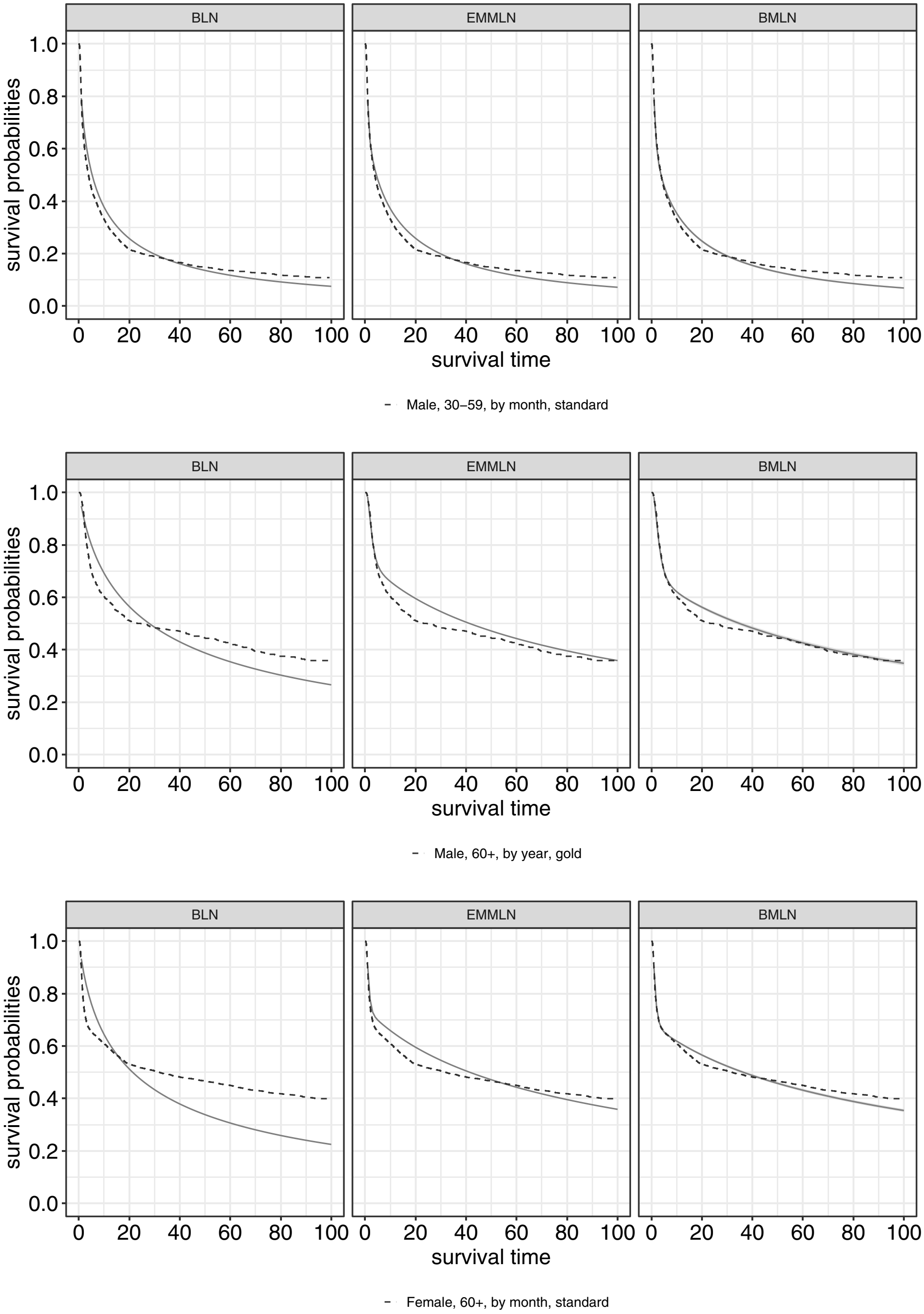


Figure 5. Simulated dataset: summary survival curves with the empirical (dashed line), and the Bayes LN, the EM mixture LN, and the Bayes mixture LN with 95% credible interval (solid gray line): scenario 1 (first row), scenario 2 (second row), and scenario 3 (third row).

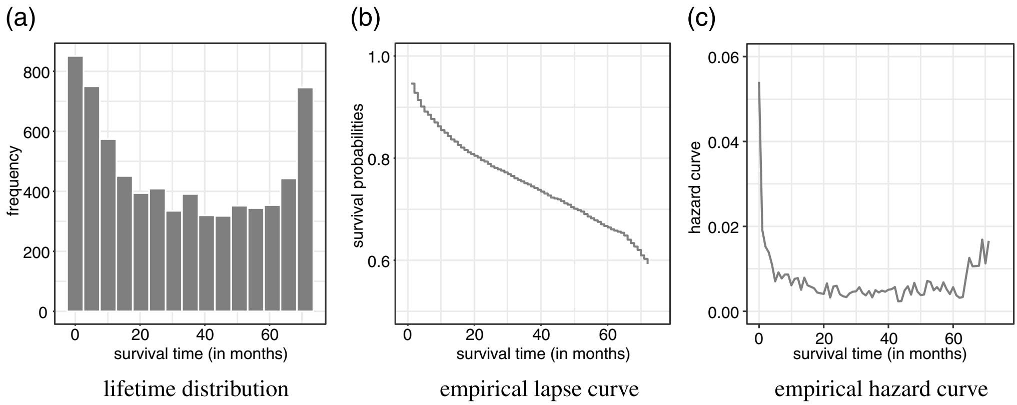
https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

Table 3. Simulated dataset: probability for the time to churn conditional on the event $L = \{T > 3\}$ for the scenarios 1, 2, and 3.

Profile	$P(T \leq 12 \mid L)$	$P(12 < T < 24 \mid L)$	$P(24 < T < 36 \mid L)$	$P(T \geq 36 \mid L)$
Scenario 1	0.394	0.194	0.100	0.312
Scenario 2	0.247	0.078	0.059	0.616
Scenario 3	0.123	0.086	0.067	0.724

Table 4. Simulated dataset: probability for the time to churn conditional on the events $\{T > 12\}$, $\{T > 24\}$ and $\{T > 36\}$ for the scenarios 1, 2, and 3.

Profile	$P(T \leq 24 \mid T > 12)$	$P(T \leq 36 \mid T > 24)$	$P(T \leq 48 \mid T > 36)$
Scenario 1	0.320	0.243	0.200
Scenario 2	0.104	0.087	0.077
Scenario 3	0.098	0.085	0.076



lifetime distribution

empirical lapse curve

empirical hazard curve

Figure 6. Case study: summary about the survival time with (a) lifetime distribution, (b) empirical survival curve, and (c) churn hazard rate.

Panels in Fig. 6 illustrate the pattern of the lifetime of the customers. Although a large admission of new customers is observed for the first months of contract, panel (c) indicates that there is a high lapse rate at the beginning, that is, there is a large portion of customers that will leave the company after just a few months of service. In the other months, the lapse risk rate remains around 0.005. Note that after 60 months of contract, there is a steep increase in the lapse rate. We expect that our proposed model is able to capture these sudden changes of pattern in the hazard function over time.

To check if usual parametric models are able to accommodate the different regimes of the survival curve over time, we consider fitting three alternative models: Exponential, Weibull, and Log-Normal and compare them with the BMLN model proposed in this paper. We present the Bayesian solution as follows instead of EM point estimates as it is desirable to access uncertainty in the curves. Figs. 7 and 8 present the general survival and hazard curves for model comparison, respectively. Indeed, the Exponential alternative seems naive, since it is a well-known fact that it assumes constant lapse rates over time. This hypothesis does not reflect reality and in this case, we will not take into account this model to model the profiles risk. Regarding our proposed mixture model, a sensitivity analysis for the number of components K was performed. Thus, the time to churn is modeled based on the mixture model with K components for different values of K. Then,

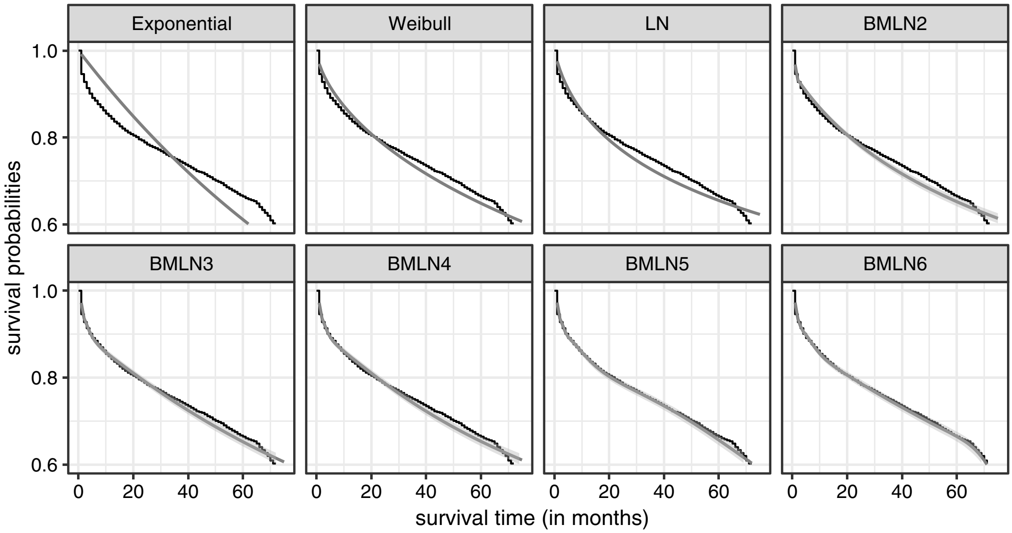


Figure 7. Case study: model comparison compared with the empirical lapse curve for the Exponential model, the Weibull model, the Log-Normal model, and the Bayesian mixture Log-Normal model considering $K = 2, 3, 4, 5, 6$ mixture components, respectively. The gray line represents the fitted model and the black line is the empirical curve.

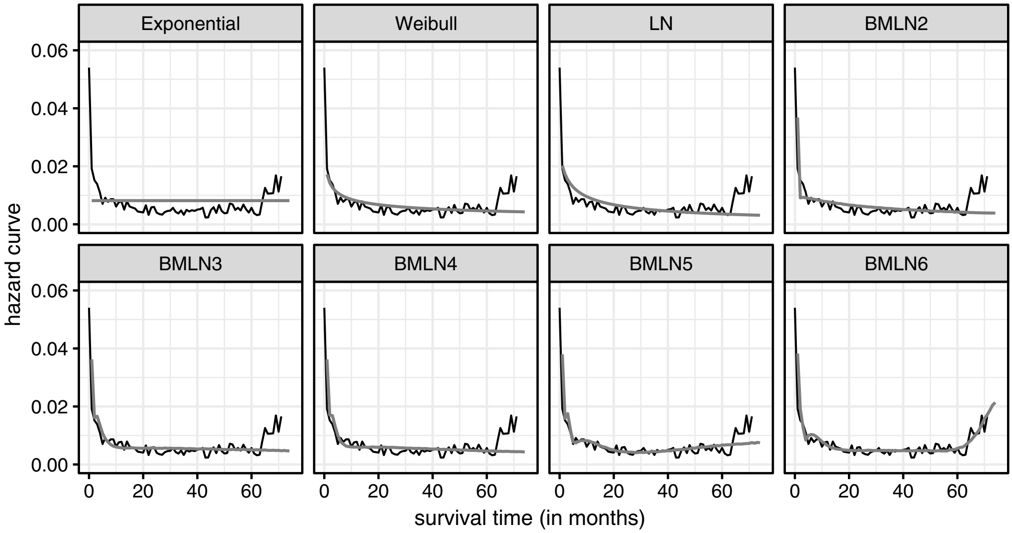


Figure 8. Case study: model comparison compared with the empirical lapse curve (first row) and the hazard curve (second row) for the Exponential model, the Weibull model, the Log-Normal model, and the Bayesian mixture Log-Normal model considering $K = 2, 3, 4, 5, 6$ mixture components, respectively. The gray line represents the fitted model and the black line is the empirical curve.

the best model can be selected based on model comparison measures computed for each value of K. As we can see, the BMLN with $K = 2$ and $K = 3$ components return fits that are similar to the Weibull model, failing to capture the curves from 30 to 70 months of contract. For $K > 4$, the mixture models provide better fits, with the general curve recovering the risk to churn in all periods of time including the first and last months of the contract. This indicates that our

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

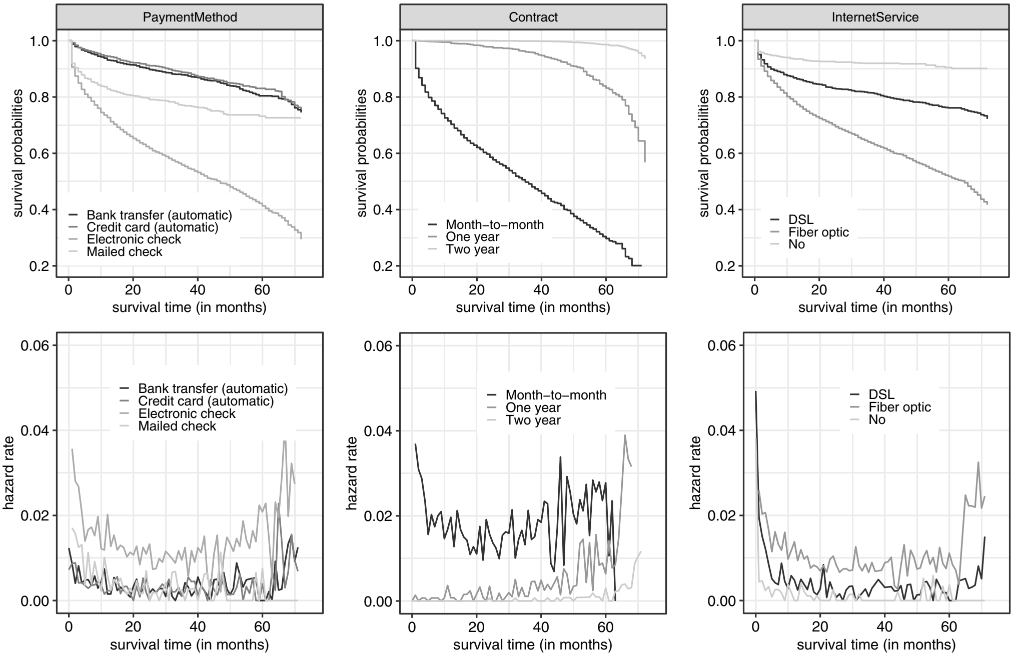


Figure 9. Case study: marginal empirical survival curve (first row) and hazard curves (second row) for covariates Payment Method, Contract, and Internet Service.

proposed model allows for more flexibility, besides being computationally efficient even for a large proportion of censorship in the data.

Although there is a large set of covariates available to explain the lapse risk, a group of them is not informative for the survival time under study. For example, the lapse rates for the two categories of gender in the data are nearly the same. Similar behavior is observed for some customer service information such as PhoneService, MutipleLines, and OnlineSecurity. Also, there is multicollinearity for groups of variables, in particular OnlineSecurity, TechSupport, OnlineBackup, and DeviceProtection are all dependent on the OnlineService variable. For instance, the covariates MonthlyCharges and TotalCharges present high collinearity. A detailed descriptive analysis for this data is presented at https://github.com/vivianalobo/LapseRiskAAS. We discuss ways to aggregate categories in each covariate considered in the analysis to achieve more interpretability of effects.

In the final models, we considered the covariates PaymentMethod, InternetService, and Contract resulting in eight risk profiles after rearranging the categories established in the descriptive analysis. Fig. 9 presents empirical survival and hazard curves for these risk profiles. For the customer account information, see that customers with Month-to-Month contracts have higher lapse rates compared to clients with One-year and Two-year contracts. Notice that the hazard curve of those with One-year payment is very similar to the one associated with more than one-year payment regime. Thus, we considered aggregating these categories in a unique group called “One-year +”. For the payment method, see that the behavior of survival curves and churn rates for Credit Card (automatic) and Bank Transfer (automatic) are quite similar. On the other hand, when we consider the instantaneous lapse curve, notice that customers that prefer an Electronic check as a payment method are the majority to leave the company, so we merged the other classes in a new category called “others.”

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

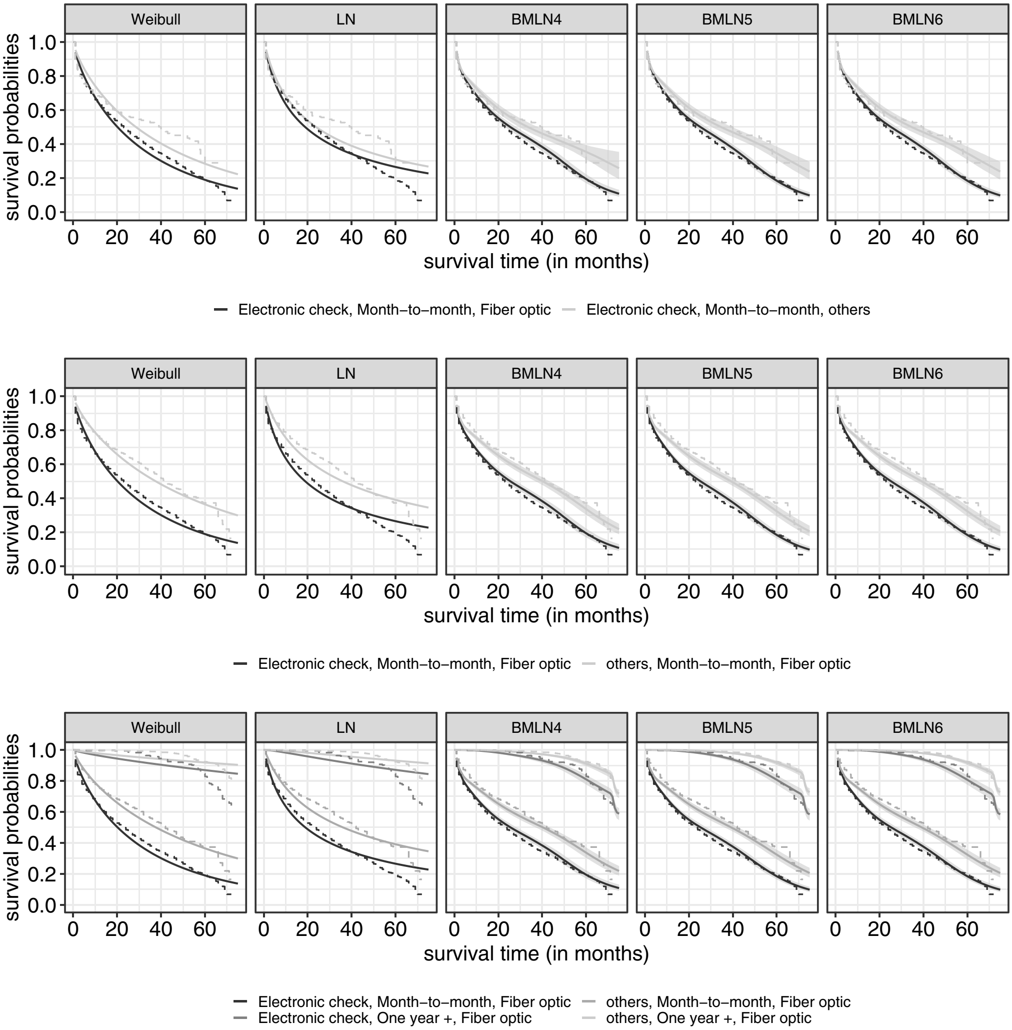


Figure 10. Case study: model comparison survival curves for the scenarios "varying Internet Service" (first row), "varying Payment Method" (second row), and "varying both Payment Method and Contract" (third row) for the Weibull model, the Log-Normal model, and the Bayesian mixture Log-Normal model with $K = 4, 5, 6$ components. The solid line represents the fitted curve and the dashed line is the empirical curve. For the BMLN, the 95% credible interval is provided in gray.

Figs. 10 and 11 present some scenarios for the profile risks. The first scenario (first row) refers to customers that have Fiber optics internet service and others (DSL, No) (for Month-to-Month and Electronic check). See that individuals that own Fiber optics internet service tend to have a higher propensity to churn when compared to others. Notice that the choice of the BMLN with $K = 4$ components is enough to capture the trend of the lapse curve and the lapse instantaneous rate. The posterior mean mixture weights estimated for the BMLN with four components (and respective 95% credible interval) are: $\eta_1 = 0.603(0.719, 0.888)$, $\eta_2 = 0.229(0.174, 0.298)$, $\eta_3 = 0.102(0.095, 0.107)$, $\eta_4 = 0.066(0.048, 0.085)$. A posterior summary of this model is detailed in Table 5 which gives some insights into the parameters of the model.

As customers who have internet services via Fiber optics have the propensity to churn faster, we would like to understand whether this trend is maintained by varying the payment method.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

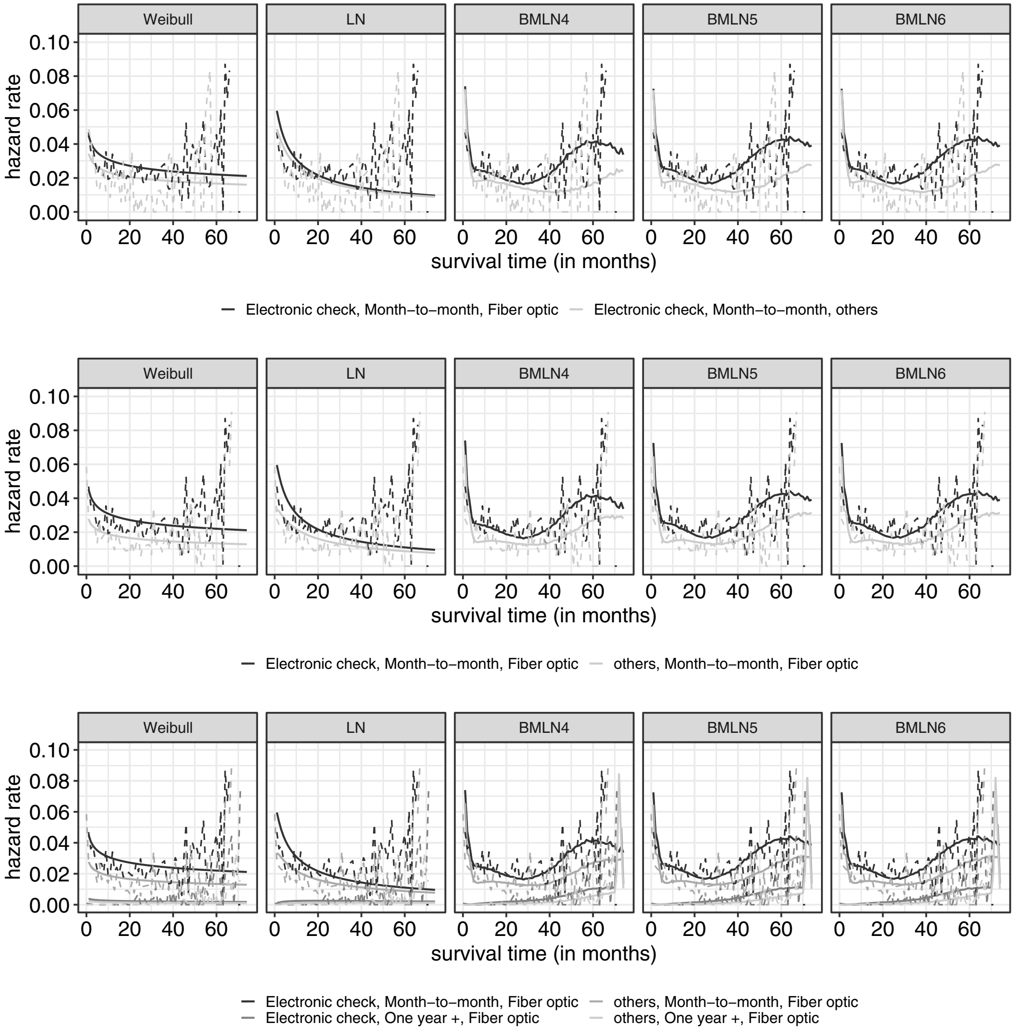


Figure 11. Case study: model comparison hazard curves for the scenarios "varying Internet Service" (first row), "varying Payment Method" (second row), and "varying both Payment Method and Contract" (third row) for the Weibull model, the Log-Normal model, and the Bayesian mixture Log-Normal model with $K = 4, 5, 6$ components. The solid line represents the fitted curve and the dashed line is the empirical curve. For the BMLN, the 95% credible interval is provided in gray.

For this purpose, scenario 2 (second row) presents customers' profiles that prefer to deal through Electronic versus customers with Bank Transfer, Credit Card, or Mailed Check (for Month-to-Month and Fiber optics). Again, the BMLN shows to be the best model, compared to the Weibull and Log-Normal. Notice that there is a significant difference between customers that prefer Electronic check and others: the churn risk for the profile Electronic check, Month-to-Month, and Fiber optics presents higher lapses over several months (see Fig. 10, second row, after Month 40).

In scenario 3, four profile risks are considered taking into account both types of contracts and payment regimes. As we can see, customers whose contract is for One year or more (One year +) have a lower chance of lapsing when compared to Month-to-Month contracts. This can be seen if one varies the payment method (for the ones who have Fiber optics). See that risk lapse rates are higher for Electronic check payment method.

Table 5. Posterior summary for regressors of the BMLN model with $K = 4$ mixture components: with mean, standard error, and 95% credible interval, for the case study.

Coefficients	Mean	Standard error	95% CI
$intercept_{j=1}$	3.087	0.105	(2.854, 3.266)
$PaymentMethodothers_{j=1}$	0.493	0.070	(0.361, 0.630)
$ContractOneyear+_{j=1}$	2.006	0.120	(1.785, 2.232)
$InternetServiceothers_{j=1}$	0.265	0.075	(0.118, 0.410)
$intercept_{j=2}$	3.998	0.052	(3.895, 4.094)
$PaymentMethodothers_{j=2}$	0.161	0.065	(0.038, 0.292)
$ContractOneyear+_{j=2}$	0.215	0.083	(0.072, 0.394)
$InternetServiceothers_{j=2}$	0.545	0.375	(0.191, 1.546)
$intercept_{j=3}$	0.000	0.001	(−0.001, 0.001)
$PaymentMethodothers_{j=3}$	0.000	0.001	(−0.001, 0.002)
$ContractOneyear+_{j=3}$	4.285	0.003	(4.281, 4.291)
$InternetServiceothers_{j=3}$	0.000	0.001	(−0.001, 0.002)
$intercept_{j=4}$	1.002	0.069	(0.866, 1.139)
$PaymentMethodothers_{j=4}$	0.108	0.062	(−0.016, 0.232)
$ContractOneyear+_{j=4}$	3.879	0.581	(3.301, 5.520)
$InternetServiceothers_{j=4}$	−0.020	0.065	(−0.148, 0.098)

Overall the proposed mixture models with $K \geq 4$ present the best fits when compared with well-known survival models such as the Weibull model. The flexibility of our proposal captures changes of regimes in the lapse rates over time, exhibited by the data.

4. Conclusions

We have proposed a flexible survival mixture model that extends the usual models usually considered for censored data and accommodates different behavior of T_i over time. This is done via mixtures of parametric models that accommodate censored survival times. The proposal combines the mixture distributions based on Frühwirth-Schnatter (2006) and data augmentation techniques proposed by Tanner & Wong (1987). Full uncertainty quantification is available through the Bayesian distributions which are obtained via MCMC methods. Furthermore, we proposed an efficient sampling algorithm for inference summarized in point estimates, via the EM algorithm. Notice that the Bayesian and frequentist solutions provide similar results in terms of point estimation for the regression coefficients. However, the Bayesian posterior distribution also allows for uncertainty quantification and estimation of nonlinear functions such as the survival curve with no asymptotic approximation required.

We performed extensive simulation studies to investigate the ability of the proposed mixture model to capture different survival curves. Our simulated examples indicate that the data-generating model with a mixture of distributions provides the best fit and indicates that the usual survival models are not able to capture the change in behavior for the churn times. Besides, our designed solution combining the Gibbs sampler and data augmentation is faster than the solution provided by the Stan package.

We have exploited three applications to illustrate the usefulness of our proposed mixture model for fitting time to churn. The first one is a simulated study that shows that our model is feasible

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

even for very large datasets (order of thousands) and that our proposed implementation is computationally efficient and recovers well all model parameters and risk behaviors studied. The second application emulates an insurance dataset with risks changing over time and high censoring. The proposed model is also able to recover well all risk behaviors. Lastly, our case study illustrates that our mixture model is more flexible than the well-known survival models usually fitted in these applications.

We conclude that allowing for a flexible survival model enables more realistic description of the behavior of the survival probability curves, for the factors affecting times to an event such as surrender. In particular, this is crucial when the lapse rates change regime over time. An alternative way to accommodate heterogeneity in the hazard and survival functions is to define a dummy variable indicating the first months of contract. However, this solution would work for this period, but often it is not easy to set *a priori* which are the periods of time when changes of regime occur. In that context, our model allows the data to inform which are the regime changes necessary for a good fit of the regression model. In the present work, calendar effects were not considered. As an extension, we could investigate if changes in regime would be observed over several years, in which case dynamic models would be necessary. It is also a research interest to investigate if a finite number of components, as proposed in our work, could capture changes in the regime.

Regarding model complexity, an excessive number of components could lead to overfitting, while too few might not be able to accommodate heterogeneities in the hazard and survival curves. In this work, different values of K were set, subject to subsequent sensitivity analysis based on the quality of fit metric. If our interest lies in the estimation of the effects to understand the profile customer retention, then a large number of parameters and components would not be a concern. However, if the purpose of the analysis is to predict whether a customer will be subject to churn or not, an alternative would be to apply regularization methods, which take into account both the quality and the complexity of the fit, penalizing the model's coefficients and shrinking them towards zero. Among such methods, least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996; Park & Casella, 2008) can be seen as a form of variable selection, as it is able to nullify coefficients and could be used to select the number of components while preserving the adequacy of the fit.

Data availability statement. Codes for reproducing the descriptive data analysis, inference procedure, and database of the paper are available at https://github.com/vivianalobo/LapseRiskAAS.

Acknowledgments. We are grateful to LabMA/UFRJ (Laboratório de Matemática Aplicada of the Universidade Federal do Rio de Janeiro) based in Brazil for financial support and to its members for the very enriching discussions. We are very thankful to Vitor Gabriel Capdeville da Silva for reading the preview of this paper and providing support in this research developing a user-friendly interface using a smart implementation with `C++` library through `RcppArmadillo` tool available in `R` allowing fast computations to take into account different values of mixture components.

References

Bolancé, C., **Guillen, M.** & **Padilla Barreto, A.** (2016). *Predicting Probability of Customer Churn in Insurance*, vol. **254**

Brockett, P.L., **Golden, L.L.**, **Guillen, M.**, **Nielsen, J. P.**, **Parner, J.** & **Perez-Marin, A.M.** (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk & Insurance*, **75**(3), 713–737.

Carpenter, B., **Gelman, A.**, **Hoffman, M.D.**, **Lee, D.**, **Goodrich, B.**, **Betancourt, M.**, **Brubaker, M.A.**, **Guo, J.**, **Li, P.** & **Riddell, A.** (2017). Stan: a probabilistic programming language. *Grantee Submission*, **76**(1), 1–32.

Demarqui, F.N., **Dey, D.K.**, **Loschi, R.H.** & **Colosimo, E.A.** (2014). Fully semiparametric Bayesian approach for modeling survival data with cure fraction. *Biometrical Journal*, **56**(22), 198–218.

Dempster, A.P., **Laird, N.M.** & **Rubin, D.B.** (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

Dong, Y., Frees, E.W., Huang, F. & Hui, F.K.C. (2022). Multi-state modelling of customer churn. *ASTIN Bulletin*, **52**(3), 735–764.

Eddelbuettel, D. & Sanderson, C. (2014). Rcpparmadillo: accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, **71**, 1054–1063.

Eling, M. & Kiesenbauer, D. (2014). What policy features determine life insurance lapse? an analysis of the german market. *Journal of Risk & Insurance*, **81**(2), 241–269.

Eling, M. & Kochanski, M. (2013). Research on lapse in life insurance: what has been done and what needs to be done? *Journal of Risk Finance*, **14**(4), 392–413.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, **10**(1), 101–113.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Berlin, Germany: Springer Science & Business Media.

Gamerman, D. (1994). Bayes estimation of the piece-wise exponential distribution. *IEEE Transactions on Reliability*, **43**(1), 128–131.

Gamerman, D. & Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. United Kingdom: Chapman & Hall/CRC.

Gatzert, N., Hoermann, G. & Schmeiser, H. (2009). The impact of the secondary market on life insurers' surrender profits. *Journal of Risk & Insurance*, **76**(4), 887–908.

Guillen, M., Nielsen, J.P., Scheike, T.H. & Pérez-Marín, A.M. (2012). Time-varying effects in the analysis of customer loyalty: a case study in insurance. *Expert Systems with Applications*, **39**(3), 3551–3558.

Guillen, M., Parner, J., Densgsoe, C. & Perez-Marin, A.M. (2003). *Using Logistic Regression Models to Predict and Understand Why Customers Leave an Insurance Company*. Intelligent and Other Computational Techniques in Insurance: Theory and Applications. Singapore: World Scientific Publishing.

Günther, C.-C., Tvete, I.F., Aas, K., Sandnes, G.I. & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, **2014**(1), 58–71.

Hemming, K. & Shaw, J.E.H. (2002). A parametric dynamic survival model applied to breast cancer survival times. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**(4), 421–435.

Hoffman, M.D. & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, **15**(1), 1593–1623.

Hu, S., O'Hagan, A., Sweeney, J. & Ghahramani, M. (2021). A spatial machine learning model for analysing customers' lapse behaviour in life insurance. *Annals of Actuarial Science*, **15**(2), 367–393.

Hu, X., Yang, Y., Chen, L. & Zhu, S. (2020). Research on a customer churn combination prediction model based on decision tree and neural network (pp. 129–132). In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA).

IBM. (2019). IBM business analytics community connect.

Ibrahim, J.G., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer Series in Statistics (SSS) Berlin, Germany: Springer.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. (2008). Random survival forests. *The Annals of Applied Statistics*, **2**(3), 841–860.

Jedidi, K., Ramaswamy, V. & DeSarbo, W.S. (1993). A maximum likelihood method for latent class regression involving a censored dependent variable. *Psychometrika*, **58**(3), 375–394.

Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. John Wiley & Sons.

Kim, S., Chen, M.-H., Dey, D. & Gamerman, D. (2007). Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis*, **13**(1), 17–35.

Knoller, C., Kraut, G. & Schoenmaekers, P. (2015). On the propensity to surrender a variable annuity contract: an empirical analysis of dynamic policyholder behavior. *The Journal of Risk and Insurance*, **83**(4), 979–1006.

Kuo, W., Tsai, C. & Chen, W.-K. (2003). An empirical study on the lapse rate: the cointegration approach. *The Journal of Risk and Insurance*, **70**(3), 489–508.

Mayrink, V.D., Duarte, J.D.N. & Demarqui, F.N. (2021). pexm: a jags module for applications involving the piecewise exponential distribution. *Journal of Statistical Software*, **100**(8), 1–28.

Milhaud, X. & Dutang, C. (2018). Lapse tables for lapse risk management in insurance: a competing risk approach. *European Actuarial Journal*, **8**(1), 97–126.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

Stan Development Team. (2018). RStan: the R interface to Stan. R package version 2.17.3.

Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82**(398), 528–540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, **58**(1), 267–288.

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

A. Posterior distribution

The prior distributions considered for the parameters, the complete conditional distributions, and proposal densities used in the MCMC algorithm are detailed as follows.

A.1 Bayesian mixture survival model

Consider the likelihood given in equation (16) and mixture components in equation (11).

The conditional distribution of each I_i, $i = 1, \ldots, n$, given all other parameters and prior distribution $I_i \mid \boldsymbol{\eta} \sim Categorical(K, \boldsymbol{\eta})$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ is given by:

$$p(I_{ij} = 1 \mid y_i, \boldsymbol{\eta}, \cdot) \propto f(y_i \mid \left\{I_{ij} = 1\right\}, \boldsymbol{\eta})\pi(I_{ij} = 1 \mid \boldsymbol{\eta})$$
$$\propto \frac{\eta_j f_j(y_i)}{\sum_{j=1}^{K} \eta_j f_j(y_i)}.$$

If $y_i \mid \left\{I_{ij} = 1\right\} \sim \mathcal{N}_j(\mu_{ij}, \phi_j^{-1})$, then

$$p(\left\{I_{ij} = 1\right\} \mid y_i, \boldsymbol{\eta}, \boldsymbol{\beta}_j, \phi_j) = \frac{\eta_j \mathcal{N}_j(y_i \mid \boldsymbol{\beta}_j, \phi_j)}{\sum_{j=1}^{K} \eta_j \mathcal{N}_j(y_i \mid \boldsymbol{\beta}_j, \phi_j)}. \tag{A.1}$$

Thus, the marginal posterior distribution is $I \mid \mathbf{y} \sim Categorical\left(K; \frac{\eta_1 \mathcal{N}_1(y_i|\boldsymbol{\beta}_1,\phi_1)}{\sum_{j=1}^{K} \eta_j \mathcal{N}_j(y_i|\boldsymbol{\beta}_j,\phi_j)}, \ldots, \frac{\eta_K \mathcal{N}_K(y_i|\boldsymbol{\beta}_K,\phi_K)}{\sum_{j=1}^{K} \eta_j \mathcal{N}_j(y_i|\boldsymbol{\beta}_j,\phi_j)}\right)$.

The conditional distribution for, considering a prior distribution as $\boldsymbol{\eta} \sim Dirichilet(K; \alpha_1, \ldots, \alpha_K)$ is given by:

$$p(\boldsymbol{\eta} \mid I, \cdot) \propto f(I \mid \boldsymbol{\eta}, \cdot)\pi(\boldsymbol{\eta}) = \prod_{i=1}^{K} p(I_i \mid \boldsymbol{\eta}, \cdot)\pi(\boldsymbol{\eta}) = \left[\prod_{i=1}^{n}\prod_{j=1}^{K} \eta_j^{[I_{ij}=1]}\right]\prod_{j=1}^{K} \eta_j^{\alpha_j - 1} \tag{A.2}$$
$$\propto \prod_{j=1}^{K} \eta_j^{\alpha_j - 1 + \sum_{i:I_{ij}=1}[I_{ij}=1]} = \prod_{j=1}^{K} \eta_j^{(\alpha_j + n_j) - 1},$$

where $n_j = \sum_i I_{ij}$. Thus, $\boldsymbol{\eta} \mid I, \cdot \sim Dirichilet(\alpha_1^*, \ldots, \alpha_K^*)$, with $\alpha_j^* = n_j + \alpha_j, j = 1, \ldots, K$.

For each group j, we can compute a sample from each of the components $\phi_j = 1/\sigma_j^2$ individually. For $\phi_j \sim Gamma(a_j, b_j)$,

$$p(\phi_j \mid \mathbf{y}_j, \cdot) \propto f(\mathbf{y}_j \mid \phi_j, I, \cdot)\pi(\phi_j)$$
$$\propto (\phi_j)^{n_j/2} exp\left\{-\phi_j\left(\frac{(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)^T(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)}{2}\right)\right\} \phi_j^{a_j - 1} \exp\{-\phi_j b_j\}$$
$$\propto \phi_j^{a_j + n_j/2 - 1} exp\left\{-\phi_j\left(b_j + \frac{(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)^T(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)}{2}\right)\right\}. \tag{A.3}$$

The conditional distribution of $\phi_j \mid \mathbf{y}_j, \cdot \sim Gamma\left(a_j + \frac{n_j}{2}; b_j + \frac{(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)^T(\mathbf{y}_j - \mathbf{x}_i^T\boldsymbol{\beta}_j)}{2}\right)$.

For $\boldsymbol{\beta}_j \sim \mathcal{N}_j(m_j, \tau_j^2)$, with $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})$. The conditional distribution is given by:

$$p(\boldsymbol{\beta}_j \mid \mathbf{y}_j, \cdot) \propto f(\mathbf{y}_j \mid \boldsymbol{\beta}_j, \cdot)\pi(\boldsymbol{\beta}_j) \tag{A.4}$$

https://doi.org/10.1017/S1748499523000180 Published online by Cambridge University Press

For latent variable Z, when $\delta_i = 0$, $i = 1, \ldots, n$, we consider the data augmentation. Thus, $p(z_i \mid \left\{I_{ij} = 1\right\}, \boldsymbol{\beta}_j, \phi_j, \cdot) \propto \mathcal{I}(y_i \geq y_i)$, we have

$$\begin{aligned} p(z_i \mid I\left\{I_{ij} = 1\right\}, y_i, \cdot) &\propto f(y_i \mid \left\{I_{ij} = 1\right\}, \cdot)\pi(z_i \mid \left\{I_{ij} = 1\right\}) \\ &\propto f(y_i^{obs}, z_i \mid \left\{I_{ij} = 1\right\}, \cdot)\pi(z_i) \qquad \text{(A.5)} \\ &\propto f(z_i \mid \left\{I_{ij} = 1\right\})\pi(z_i) = \exp\left\{-\frac{\phi_j}{2}(z_i - \mathbf{x}_i^T\boldsymbol{\beta}_j)^2\right\}\mathcal{I}(z_i \geq y_i). \end{aligned}$$

The conditional distribution of $z_i \mid \left\{I_{ij} = 1\right\}, y_i \sim \mathcal{NT}(\mathbf{x}_i^T\boldsymbol{\beta}_j, \phi_j^{-1})$.

B. Some results for simulated datasets

In this appendix, we present the generation of observations via mixture distribution considering the presence of censored data introduced in section 2.1 and explored in section 3.1. The dataset was simulated considering variable $Y_i = \log(T_i)$ as the logarithmic the duration of a policy i before termination from a mixture of Gaussian distribution with $K = 2$ components given by:

$$f(y_i \mid \boldsymbol{\theta}) = \eta\,\mathcal{N}_1(\mu_{i1}, \sigma_1^2) + (1-\eta)\mathcal{N}_2(\mu_{i2}, \sigma_2^2), \quad i = 1, \ldots, n, \qquad \text{(B.1)}$$

where $\boldsymbol{\theta} = (\eta, \boldsymbol{\beta}, \sigma^2)$ the parametric vector of interest. The mean for $j = 1, 2$ is given by $\mu_{i,j=1} = \beta_{01} + \beta_{11}x = 3.3 + 0.5x$ and $\mu_{i,j=2} = \beta_{02} + \beta_{12}x = 4.0 + 0.8x$ and variance as $\sigma_1^2 = 0.3$ and $\sigma_2^2 = 0.039$, respectively. The weight $\eta_1 = \eta$ is equal to 0.6 and x represents the covariate that takes values ($x = 0$, no attribute and $x = 1$, yes attribute). In the presence of censored observation, that is, $\delta_i = 0$, we generate the censored observation y_i^c from a truncated Gaussian distribution as:

$$y_i^c \mid \cdot \sim \mathcal{NT}_{(-\infty, y_i]}\left(\mu_{ij}, \sigma_j^2\right), \forall\ j = 1, \ldots, K. \qquad \text{(B.2)}$$

Thus, in inferential processes with simulated data, the likelihood will be based on the original observations y_i for $\delta_i = 1$ and the censored ones, y_i^c, for $\delta_i = 0$, emulating the observable information in real practical situations. Fig. B.1 shows the mixture distribution with 40% censored data considering $n = 1,000$ policies. Panels (a) show the behavior in the original scale, that is, T_i from a Log-Normal distribution, and (b) in the log scale. See that the mixture proportions are indeed based on the probabilities that were defined, that is, $\eta = 0.6$ and $1 - \eta = 0.4$, respectively.

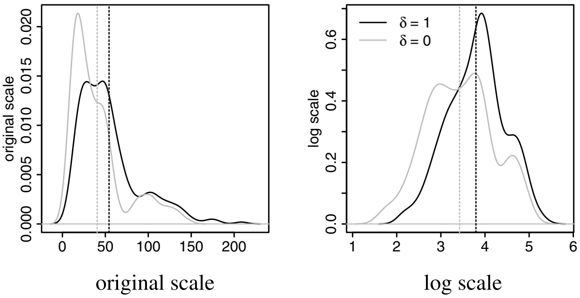


Figure B.1. Simulated with 40% censored dataset considering $n = 1,000$ policies: (a) original scale T_i and (b) log scale $\log(T_i)$. Besides, 60% come from $\mathcal{N}_1(\cdot,\cdot)$ and 40% come from $\mathcal{N}_2(\cdot,\cdot)$. Dashed lines are the mean for $\delta = 0$ and $\delta = 1$, respectively.

Cite this article: Lobo VGR, Fonseca TCO and Alves MB (2024). Lapse risk modeling in insurance: a Bayesian mixture approach, *Annals of Actuarial Science*, **18**, 126–151. https://doi.org/10.1017/S1748499523000180