

# A MARKED COX MODEL FOR THE NUMBER OF IBNR CLAIMS: ESTIMATION AND APPLICATION

BY

ANDREI L. BADESCU, TIANLE CHEN, X. SHELDON LIN  
AND DAMENG TANG

## ABSTRACT

Incurring but not reported (IBNR) loss reserving is of great importance for Property & Casualty (P&C) insurers. However, the temporal dependence exhibited in the claim arrival process is not reflected in many current loss reserving models, which might affect the accuracy of the IBNR reserve predictions. To overcome this shortcoming, we proposed a marked Cox process and showed its many desirable properties in Badescu et al. (2016).

In this paper, we consider the model estimation and applications. We first present an expectation–maximization (EM) algorithm which guarantees the efficiency of the estimators unlike the moment estimation methods widely used in estimating Cox processes. In addition, the proposed fitting algorithm can be implemented at a reasonable computational cost. We examine the performance of the proposed algorithm through simulation studies. The applicability of the proposed model is tested by fitting it to a real insurance claim data set. Through out-of-sample tests, we find that the proposed model can provide realistic predictive distributions.

## KEYWORDS

IBNR claims, loss reserving, Cox model, hidden Markov chain, temporal dependence, Pascal mixture, EM algorithm.

## 1. INTRODUCTION

For Property & Casualty (P&C) insurers, an important type of reserve is the so-called incurred but not reported (IBNR) reserve, which stems from the potential time delay between the claim occurrence date and the claim reporting date. In practice, various deterministic algorithms based on aggregated triangular data (e.g., Friedland 2010) are used to provide a best estimate for

the IBNR reserve together with the reported but not settled (RBNS) reserve. Meanwhile, there is a whole array of stochastic models which aims to interpret these so called macro-level methods and to analyze the uncertainty of their results. See Wüthrich and Merz (2008) and Wüthrich and Merz (2015) for a comprehensive overview.

Due to the limited number of data points contained in the triangular structure, these macro-level models tend to be overparameterized and thus produce unstable estimates (Verdonck et al. 2009). Furthermore, they cannot separately estimate the RBNS reserve and the IBNR reserve without some further granulation of the current data (Schinieper 1991; Liu and Verrall 2009), or the inclusion of new data such as the numbers of reported claims (Verrall et al. 2010). To address some of these issues, a class of models that are very popular nowadays and are referred to as “micro-level” models has emerged due to the seminal work of Ragnar Norberg (see Norberg 1993a; Norberg 1993b; Norberg 1999). Norberg (1993a) proposed to use a marked nonhomogeneous Poisson model and a general mathematical framework for predicting the IBNR claims and reserve calculation. Since then, quite a few authors contributed substantially to the development of the micro-level modeling framework. Antonio and Plat (2014) presented a study case based on the Poisson arrival model introduced by Norberg (1993a). An econometric comparison between the micro- and the macro-level models is presented in Charpentier and Pigeon (2016). Understanding the importance of reporting delay together with its proper calibration is essential in obtaining accurate predictions for the IBNR counts. Verrall and Wüthrich (2016) and Verbelen et al. (2018) considered reporting delays that are time dependent under a Poisson-type marked accident arrival framework. Deep learning methods for evaluating the IBNR and RBNS reserves are considered in Wüthrich (2018).

In Badescu et al. (2016), we proposed the modeling of the claim arrival process together with the reporting delays as a marked Cox process. The proposed model is mathematically tractable, statistically flexible and computationally efficient. The underlying intensity process is assumed to be a piecewise stochastic process generated by a hidden Markov model (HMM) with Erlang state-dependent distributions. As such, the model allows for the policy exposure fluctuation. Because of the potential reporting delay, one can only observe the associated reported claim process instead of the claim arrival process. Meanwhile, the associated reported claim process provides a starting point for predicting the number of IBNR claims. As a result, it is critical to understand the connections among these three processes from a modeling perspective. Under our model assumptions, we showed that both the associated reported claim process and the IBNR claim process are marked Cox processes with easily convertible intensity processes and marking distributions. Because data are usually aggregated in some discrete form before analysis in practice, we also studied the three corresponding discretely observed processes. We showed that they are all Pascal-HMMs and also preserve all the information about the claim arrivals due to an order statistics property. The flexibility of the proposed Cox

process in modeling temporal dependence is guaranteed as the joint distribution of the discretely observed claim arrival process is a multivariate Pascal mixture (Badescu et al. 2015). In addition, we derived an analytical formula for the auto-correlation function (ACF) of the discretely observed process, whose power-decaying pattern is desirable for temporal dependence interpretation. Finally, we showed that both the distributions of the number of reported claims and the number of IBNR claims come from the class of Pascal mixtures with closed-form expression for their mixing weights. Note that assuming the reporting delay distribution to be time independent can be viewed as a limitation in our model. However, as shown in the numerical analysis, our model provides a very accurate prediction of the number of the IBNR claims even if the model lacks this assumption. In Section 6, we discuss how this limitation will be addressed in our future investigations.

Given the statistical flexibility and the mathematical tractability of the model proposed in Badescu et al. (2016), in this paper we aim to provide an estimation procedure for the proposed Cox process. In the current literature, Cox processes fall into two large classes, that is, the log-Gaussian Cox process (Moller et al. 1998) and the shot-noise Cox process (Moller 2003; Avanzi et al. 2016). Usually, the maximization of the full likelihood is computationally very intensive and often requires an enormous amount of Markov chain Monte Carlo (MCMC) simulations (Moller and Waagepetersen 2003). On the other hand, closed-form expressions are available for many second-order summary statistics of Cox processes, like pair-correlation function, Palm intensity and K-function. Based on these properties, many moment estimation methods have been proposed, including composite likelihood (Guan 2006), Palm likelihood estimation (Tanaka et al. 2008) and minimum contrast method (Diggle 2003). At the expense of losing efficiency compared with maximum likelihood estimators (MLEs), the estimators from these three methods are faster to compute and are simulation-free.

Besides the balance between computational cost and efficiency of estimators, estimating a Cox process as a model for claim arrivals raises some extra challenges. Firstly, the exposure fluctuation makes the Cox process inhomogeneous. While the aforementioned three moment estimation methods can be adapted for inhomogeneous Cox processes using a two-step approach (Waagepetersen and Guan 2009), one needs to impose restrictive conditions on the Cox process, which are usually not readily suitable for insurance interpretation. Secondly, due to the reporting delay that is the key component in estimating the IBNR reserve (Verrall and Wüthrich 2016), the claim arrival process cannot be fully observed. Parameter estimation has to be carried out using the associated reported claim process, which only contains incomplete information.

In this paper, we propose an expectation–maximization (EM) algorithm for estimating the Cox process model, which can deal with the above-mentioned issues. The exposure fluctuation can be easily incorporated by varying the coefficients before the scale parameter in the proposed model and thus poses no

issues for the estimation (see Equation (2.1)). Since both the discretely observed claim arrival process and the reported claim process are Pascal-HMMs with easily convertible parameters (Proposition 2.2), one can easily cope with the incomplete data challenge by first estimating the reported claim process and then converting its parameters back to that of the claim arrival process. As MLEs, the estimators from the proposed fitting algorithm are efficient. Most importantly, the proposed algorithm can be implemented at a reasonable computational cost. The E-step reduces to calculating the forward and backward probabilities, which are recursively defined. Out of the three estimators in the M-step, two have analytical expressions and one is the unique zero root of a continuous monotone function.

The paper is organized as follows. Section 2 gives a brief review of the marked Cox model proposed in Badescu et al. (2016). In Section 3, we present an EM algorithm for fitting purposes. The effectiveness of the algorithm and the versatility of the proposed model is tested through a simulation study in Section 4. In Section 5, we examine the applicability of the proposed model by fitting it to a real insurance claim data set. Detailed out-of-sample tests are held to attest for the quality of our model. The paper concludes in Section 6 and several future research directions are given.

## 2. THE PROPOSED MARKED COX MODEL

In this section, we give a brief review of the marked Cox model proposed in Badescu et al. (2016), along with its several important properties. The relevant proofs and detailed discussions of the model can be found in that paper.

Suppose that the development of a claim until its reporting time is described as a pair of random variables  $(T, U)$ , where  $T$  is the claim arrival time and  $U$  is its reporting delay. The claims from the considered portfolio are ordered in chronology of their arrival times:  $T_1 < T_2 < \dots$ . The claim arrival process constitutes of the counting process  $\{N^a(t), t \geq 0\}$ , where  $N^a(t) = \sum_{i=1}^{\infty} I_{\{T_i \leq t\}}$  and  $I_A$  is the indicator function for event  $A$ . At a given valuation date  $\tau$ , we do not observe a claim which has occurred so far unless it has been reported, that is,  $T + U \leq \tau$ . The reported claim process comprises of the counting process  $\{N^r(t), 0 \leq t \leq \tau\}$ , where  $N^r(t) = \sum_{i=1}^{\infty} I_{\{T_i \leq t, T_i + U_i \leq \tau\}}$ . In order to predict the number of IBNR claims, we also consider the IBNR claim process which consists of the counting process  $\{N^{\text{IBNR}}(t), 0 \leq t \leq \tau\}$ , where  $N^{\text{IBNR}}(t) = \sum_{i=1}^{\infty} I_{\{T_i \leq t, T_i + U_i > \tau\}}$ . It should be emphasized that all the three processes described above are marked point processes with marks being the reporting delays, although the notation does not explicitly express them.

As in Badescu et al. (2016), the claim arrival process  $\{N^a(t), t \geq 0\}$  is modeled as a marked Cox process. The marks or the reporting delays  $\{U_1, U_2, \dots\}$  are assumed to be independent and identically distributed (i.i.d.) random variables with common density function  $p_U(u)$  and cumulative distribution function  $P_U(u)$ . The stochastic intensity process  $\Lambda(t)$  is assumed to be a

piecewise stochastic process:  $\Lambda(t) = \Lambda_l$ , for  $d_{l-1} \leq t < d_l$ ,  $l = 1, 2, \dots$  with  $d_0 = 0$ , where  $d_1, d_2, \dots$  are preset time points that may be interpreted as data collecting times.  $\{\Lambda_1, \Lambda_2, \dots\}$  is assumed to come from an Erlang hidden Markov model (Erlang-HMM) with the following structure:

- *The hidden parameter process*  $\{C_1, C_2, \dots\}$  is a time-homogeneous Markov chain with a finite state space  $\{1, 2, \dots, g\}$ . Its initial distribution and transition probability matrix are, respectively, denoted by the row vector  $\pi_1$  and the matrix  $\Gamma = (\gamma_{ij})_{g \times g}$ , where  $\gamma_{ij} = P(C_l = j | C_{l-1} = i)$ . The Markov chain is assumed to be irreducible, aperiodic and all the states are positive recurrent. As a result, there exists a unique limiting distribution which is denoted by  $\delta$ .
- *The state-dependent process*  $\{\Lambda_1, \Lambda_2, \dots\}$  is such that they are independent of each other when conditional on  $\{C_1, C_2, \dots\}$ . In addition, each  $\Lambda_l$  depends only on the current state  $C_l$ . Given that  $C_l = i$ , we assume that  $\Lambda_l$  follows an Erlang distribution with shape parameter  $m_i$  and scale parameter  $\omega_l \theta$  whose density function is given by

$$f_{\Lambda_l | C_l = i}(\lambda) = \frac{\lambda^{m_i - 1} e^{-\frac{\lambda}{\omega_l \theta}}}{(\omega_l \theta)^{m_i} (m_i - 1)!} \triangleq f(\lambda; m_i, \omega_l \theta), \tag{2.1}$$

where  $\omega_l$  represents the exposure of the considered portfolio for the  $l$ th period.

We remark that the Ammeter process (Ammeter, 1948) is a very special case of the proposed claim arrival process. In fact when  $d_1 = \infty$ , the proposed process is reduced a mixed Poisson process. When the initial distribution  $\pi_1$  is the limiting distribution  $\delta$ , each row of  $\Gamma$  is  $\delta$  and there are no exposure fluctuation  $\omega_l \equiv \omega$ , the proposed process further degenerates to an Ammeter process.

If the claim arrival process  $\{N^a(t), t \geq 0\}$  is a marked Cox process, then both the associated reported claim process  $\{N^r(t), 0 \leq t \leq \tau\}$  and the IBNR claim process  $\{N^{IBNR}(t), 0 \leq t \leq \tau\}$  are still marked Cox processes with easily convertible stochastic intensity functions and mark densities.

**Theorem 2.1.** *Assume that the claim arrival process  $\{N^a(t), t \geq 0\}$  is a marked Cox process as described above. Then for a given valuation date  $\tau$ , its associated reported claim process  $\{N^r(t), 0 \leq t \leq \tau\}$  and IBNR claim process  $\{N^{IBNR}(t), 0 \leq t \leq \tau\}$  are also marked Cox processes. Their adjusted stochastic intensity functions are  $\Lambda^r(t) = \Lambda(t)P_U(\tau - t)I_{\{0 \leq t \leq \tau\}}$  and  $\Lambda^{IBNR}(t) = \Lambda(t)(1 - P_U(\tau - t))I_{\{0 \leq t \leq \tau\}}$ , respectively, and their independent marks follow adjusted position-dependent mark density functions  $p_{U|t}^r(u) = \frac{p_U(u)}{P_U(\tau - t)}I_{\{0 \leq u \leq \tau - t\}}$  and  $p_{U|t}^{IBNR}(u) = \frac{p_U(u)}{1 - P_U(\tau - t)}I_{\{u > \tau - t\}}$ , respectively.*

For a given integer  $l$ , denote the number of claims that arrived during  $[d_{l-1}, d_l)$  by  $N_l$ . Similarly,  $N_l^r$  and  $N_l^{IBNR}$ , respectively, are the number of claims occurred during  $[d_{l-1}, d_l)$  that are reported or not by the valuation time  $\tau$ .

Without loss of generality, assume that  $\tau = d_k$ . Then  $\{N_1, N_2, \dots\}, \{N_1^r, \dots, N_k^r\}$  and  $\{N_1^{IBNR}, \dots, N_k^{IBNR}\}$  all come from the class of Pascal-HMMs as shown in the following theorem.

**Theorem 2.2.** *For the proposed claim arrival process, the discretely observed claim arrival process  $\{N_1, N_2, \dots\}$ , the discretely observed reported claim process  $\{N_1^r, \dots, N_k^r\}$ , and the discretely observed IBNR claim process  $\{N_1^{IBNR}, \dots, N_k^{IBNR}\}$  all come from the class of Pascal-HMMs. They share the same hidden parameter process  $\{C_1, C_2, \dots\}$  with  $\{N^a(t), t \geq 0\}$  and their state-dependent distributions are Pascals with the following probability functions, respectively,*

$$\begin{aligned}
 P(N_l = n | C_l = i) &= p(n; m_i, (d_l - d_{l-1})\omega_l\theta), \\
 P(N_l^r = n | C_l = i) &= p\left(n; m_i, \left(\int_{d_{l-1}}^{d_l} P_U(\tau - t) dt\right) \omega_l\theta\right), \\
 P(N_l^{IBNR} = n | C_l = i) &= p\left(n; m_i, \left(\int_{d_{l-1}}^{d_l} (1 - P_U(\tau - t)) dt\right) \omega_l\theta\right),
 \end{aligned}$$

where

$$p(n; m, \theta) = \binom{n + m - 1}{m - 1} \left(\frac{1}{1 + \theta}\right)^m \left(\frac{\theta}{1 + \theta}\right)^n. \tag{2.2}$$

### 3. PARAMETER ESTIMATION: AN EM ALGORITHM

In this section, we will present an EM algorithm for fitting our class of Pascal-HMMs, which includes the three discretely observed processes discussed above. Variations of the EM algorithm have been used to fit Erlang-based mixture distributions to data in Lee and Lin (2010), Badescu et al. (2015) and Verbelen et al. (2015). As we are dealing with time series data and fitting the proposed stochastic model to the data in this paper, the proposed EM algorithm in this section, although it has some similarity to the aforementioned EM algorithms due to the use of the Erlang distribution, is very different. The likelihood calculation now involves the use of forward and backward probabilities and hence is much more complex, which requires the employment of an approximation scheme.

The general class of Pascal-HMM has the same components as that of our proposed model in Section 2, except that the state-dependent distributions are Pascals with known time-varying scale parameters:

$$P(N_l = n | C_l = i) = p(n; m_i, a_l\theta), \quad a_l > 0, \quad l = 1, 2, \dots, \quad i = 1, \dots, g. \tag{3.1}$$

When  $a_l = (d_l - d_{l-1})\omega_l$ , the above model generates the discretely observed claim arrival process. The discretely observed reported claim process and the discretely observed IBNR claim process, respectively, come from the above model if  $a_l$  equals  $\left(\int_{d_{l-1}}^{d_l} P_U(\tau - t) dt\right) \omega_l$  or  $\left(\int_{d_{l-1}}^{d_l} (1 - P_U(\tau - t)) dt\right) \omega_l$ .

Denote the random sample from this general Pascal-HMM and our observation, respectively, by  $N^{(T)} = (N_1, \dots, N_T)$  and  $\mathbf{n}^{(T)} = (n_1, \dots, n_T)$ . Our goal is to estimate all the parameters in the model:  $g, m_1, \dots, m_g, \boldsymbol{\pi}_1, \boldsymbol{\Gamma}$  and  $\theta$  under constraints including  $\sum_{i=1}^g \pi_{1i} = 1$  and  $\sum_{j=1}^g \gamma_{ij} = 1$  for  $i = 1, \dots, g$ . Denote  $\mathbf{1}$  as the row vector with all its elements being 1 and  $\mathbf{1}^T$  as the corresponding transposed column vector. We also adopt the following notation:

$$P_t(n) = \begin{pmatrix} p(n; m_1, a_t, \theta) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p(n; m_g, a_t, \theta) \end{pmatrix}. \tag{3.2}$$

Using a similar proof to that of Corollary 4.3 in Badescu et al. (2016), the likelihood for the observed data is

$$P(N^{(T)} = \mathbf{n}^{(T)}) = \boldsymbol{\pi}_1 P_1(n_1) \boldsymbol{\Gamma} P_2(n_2) \dots \boldsymbol{\Gamma} P_T(n_T) \mathbf{1}^T \triangleq L_T. \tag{3.3}$$

The maximization of the above likelihood is a complex problem of constrained nonlinear optimization, which can be a challenging task numerically. Due to the finite mixture structure of our model together with the Erlang conditional distribution assumption for the accident intensity we will consider the model estimation in an EM algorithm context. Before presenting the details about the algorithm, we first introduce forward and back probabilities, which will play an important role in the E-step of the proposed EM algorithm.

### 3.1. Forward and backward probabilities

For  $t = 1, 2, \dots, T$  and  $i = 1, 2, \dots, g$ , define

$$\alpha_t(i) = P(N^{(t)} = \mathbf{n}^{(t)}, C_t = i), \tag{3.4}$$

where  $N^{(t)} = (N_1, \dots, N_t)$  and  $\mathbf{n}^{(t)} = (n_1, \dots, n_t)$ , and

$$\beta_t(i) = P(N_{t+1}^T = \mathbf{n}_{t+1}^T | C_t = i), \tag{3.5}$$

where  $N_{t+1}^T = (N_{t+1}, \dots, N_T)$  and  $\mathbf{n}_{t+1}^T = (n_{t+1}, \dots, n_T)$ . In the literature of HMMs, the probabilities  $\alpha_t(i)$ 's and  $\beta_t(i)$ 's are usually referred to as forward probabilities and backward probabilities, respectively. Under our model assumptions, these probabilities may be calculated recursively as shown in the

proposition below. The proof is similar to those of Propositions 2 and 3 of Section 4.1 in Zucchini and MacDonald (2009).

**Proposition 3.1.** *Denote row vector  $\alpha_t = (\alpha_t(1), \dots, \alpha_t(g))$  and column vector  $\beta_t = (\beta_t(1), \dots, \beta_t(g))^T$ . Then the following recursive relations hold:*

$$\begin{aligned} \alpha_1 &= \pi_1 P_1(n_1), \\ \alpha_t &= \alpha_{t-1} \Gamma P_t(n_t), \quad t = 2, 3, \dots, T, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} \beta_T &= \mathbf{1}, \\ \beta_t &= \Gamma P_{t+1}(n_{t+1}) \beta_{t+1}, \quad t = 1, 2, \dots, T - 1. \end{aligned} \tag{3.7}$$

We remark that there exists the computational issue of underflow when calculating the forward (backward) probabilities. The values of the probabilities can become exceedingly small when  $t$  becomes large (small). When we implement the proposed EM algorithm in the next subsection, a strategy of scaling  $\alpha_t$  and  $\beta_t$  will be used as suggested in Zucchini and MacDonald (2009). In most applications, including all of our simulation studies and real data analysis, this technique can help avoid the underflow problem.

In the next proposition, we obtain the conditional probability of the states of the HMM as well as their conditional transition probabilities.

**Proposition 3.2.** *For  $t = 1, \dots, T$ ,*

$$P(C_t = i | \mathbf{N}^{(T)} = \mathbf{n}^{(T)}) = \alpha_t(i) \beta_t(i) / L_T.$$

*For  $t = 2, \dots, T$ ,*

$$P(C_{t-1} = i, C_t = j | \mathbf{N}^{(T)} = \mathbf{n}^{(T)}) = \alpha_{t-1}(i) \gamma_{ij} p(n_t; m_j, a, \theta) \beta_t(j) / L_T.$$

*Here  $L_T = \pi_1 P_1(n_1) \Gamma P_2(n_2) \dots \Gamma P_T(n_T) \mathbf{1}$ .*

The results in the both propositions are of central importance in the E-step of the proposed EM algorithm.

### 3.2. An EM algorithm

In each run of the proposed EM algorithm, we treat  $g, m_1, \dots, m_g$  as preset and aim to estimate parameters  $\Phi = (\pi_1, \Gamma, \theta)$ . We augment each observed data  $N_t = n_t, t = 1, 2, \dots, T$ , by introducing its associated state indicators  $Z_t = (Z_{t1}, \dots, Z_{tg})$ , where for  $i = 1, 2, \dots, g$ ,

$$Z_{ti} = \begin{cases} 1, & \text{if } C_t = i \\ 0, & \text{otherwise} \end{cases} \tag{3.8}$$

Based on the structure of  $\{C_1, C_2, \dots, C_T\}$ ,  $\mathbf{Z}_t$  satisfies the following distributional laws:

$$P(\mathbf{Z}_1 = \mathbf{z}_1) = \prod_{i=1}^g \pi_{1i}^{z_{1i}},$$

$$P(\mathbf{Z}_t = \mathbf{z}_t | \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}) = \prod_{i=1}^g \prod_{j=1}^g (\gamma_{ij})^{z_{t-1,i} \times z_{tj}}.$$

The likelihood of the complete data  $\{\mathbf{n}^{(T)}, \mathbf{z}^{(T)}\}$  can be obtained as

$$\begin{aligned} & \mathcal{L}(\Phi; \mathbf{n}^{(T)}, \mathbf{z}^{(T)}) \\ & \triangleq P(N^{(T)} = \mathbf{n}^{(T)}, \mathbf{Z}^{(T)} = \mathbf{z}^{(T)}) \\ & = P(\mathbf{Z}_1 = \mathbf{z}_1) \cdot \prod_{t=2}^T P(\mathbf{Z}_t = \mathbf{z}_t | \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}) \cdot \prod_{t=1}^T P(N_t = n_t | \mathbf{Z}_t = \mathbf{z}_t) \\ & = \prod_{i=1}^g \pi_{1i}^{z_{1i}} \cdot \prod_{t=2}^T \prod_{i=1}^g \prod_{j=1}^g (\gamma_{ij})^{z_{t-1,i} \times z_{tj}} \cdot \prod_{t=1}^T \prod_{i=1}^g (p(n_t; m_i, a_t \theta))^{z_{ti}}. \end{aligned}$$

Correspondingly, the log-likelihood of the complete data is

$$\begin{aligned} l(\Phi; \mathbf{n}^{(T)}, \mathbf{z}^{(T)}) &= \sum_{i=1}^g z_{1i} \ln \pi_{1i} + \sum_{i=1}^g \sum_{j=1}^g \sum_{t=2}^T z_{t-1,i} z_{tj} \ln \gamma_{ij} \\ &+ \sum_{t=1}^T \sum_{i=1}^g z_{ti} \ln p(n_t; m_i, a_t \theta). \end{aligned} \tag{3.9}$$

3.2.1. *The E-step*

At the  $k$ th iteration of the E-step, we take the conditional expectation of (3.9) given the observed data  $\mathbf{n}^{(T)}$  and the current estimator  $\Phi^{(k-1)}$  for  $\Phi$ :

$$\begin{aligned} & Q(\Phi; \Phi^{(k-1)}) \\ & \triangleq E(l(\Phi; \mathbf{n}^{(T)}, \mathbf{Z}^{(T)}) | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\ & = \sum_{i=1}^g z_{1i}^{(k)} \ln \pi_{1i} + \sum_{i=1}^g \sum_{j=1}^g \sum_{t=2}^T z_{tj}^{(k)} \ln \gamma_{ij} + \sum_{t=1}^T \sum_{i=1}^g z_{ti}^{(k)} \ln (p(n_t; m_i, a_t \theta)), \end{aligned} \tag{3.10}$$

where, according to Proposition 3.2,

$$\begin{aligned} z_{ii}^{(k)} &= E(Z_{ii} | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\ &= P(Z_{ii} = 1 | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\ &= P(C_t = i | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\ &= \frac{\alpha_t(i)^{(k-1)} \beta_t(i)^{(k-1)}}{L_T^{(k-1)}} \end{aligned} \tag{3.11}$$

and

$$\begin{aligned}
 z_{ij}^{(k)} &= E(Z_{t-1,i} \times Z_{t,j} | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\
 &= P(Z_{t-1,i} = 1, Z_{t,j} = 1 | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\
 &= P(C_{t-1} = i, C_t = j | \mathbf{n}^{(T)}; \Phi^{(k-1)}) \\
 &= \frac{\alpha_{t-1}^{(k-1)}(i) \gamma_{ij}^{(k-1)} p(n_t; m_j, a_t \theta^{(k-1)}) \beta_t(j)^{(k-1)}}{L_T^{(k-1)}}.
 \end{aligned}
 \tag{3.12}$$

Here,  $L_T^{(k-1)}$  is the likelihood for the observed data (see (3.3)) with  $\Phi = \Phi^{(k-1)}$ .

### 3.2.2. The M-step

At the  $k$ th iteration of the M-step, we maximize (3.10) under the constraints on the parameters, that is,  $\sum_{i=1}^g \pi_{1i} = 1$  and  $\sum_{j=1}^g \gamma_{ij} = 1$  for  $i = 1, 2, \dots, g$ . Observe that each of the three terms in (3.10) only contains one part of  $\Phi = (\pi_1, \Gamma, \theta)$ , which shows that this maximization neatly splits into three separate pieces.

Using Lagrange’s multiplier, it is easy to see that the constrained optimizers for  $\pi_1$  and  $\Gamma$  are

$$\pi_{1i}^{(k)} = z_{1i}^{(k)}, \quad i = 1, 2, \dots, g,
 \tag{3.13}$$

and

$$\gamma_{ij}^{(k)} = \frac{\sum_{t=2}^T z_{tij}^{(k)}}{\sum_{j=1}^g \sum_{t=2}^T z_{tij}^{(k)}}, \quad i, j = 1, 2, \dots, g.
 \tag{3.14}$$

Both of the above optimizers have intuitive interpretations.  $\pi_{1i}^{(k)}$  is the posterior probability of  $N_1$  belonging to state  $i$ .  $\gamma_{ij}^{(k)}$  is the proportion of the expected number of one-step transition from  $i$  to  $j$  to the total expected number of one-step transition from  $i$  to all the states including state  $i$  itself.

**Remark 3.3.** *If we assume the hidden Markov chain  $\{C_1, C_2, \dots\}$  is stationary, that is,  $\pi_1 = \delta$ , then (3.10) should be maximized with an extra constraint, namely,  $\pi_1 \Gamma = \pi_1$ , which may increase the complexity of this maximization. This problem is also discussed in Zucchini and MacDonald (2009). While the initial distribution can be quite different from the stationary distribution, the convergence is usually relatively fast for HMMs. This has been documented in Zucchini and MacDonald (2009).*

Finally,  $\theta^{(k)}$  satisfies equation

$$\sum_{t=1}^T \sum_{i=1}^g z_{ti}^{(k)} \frac{n_t}{\theta} = \sum_{t=1}^T \sum_{i=1}^g z_{ti}^{(k)} \left( \frac{m_i a_t}{1 + a_t \theta} + \frac{n_t a_t}{1 + a_t \theta} \right).
 \tag{3.15}$$

Since  $\sum_{i=1}^g z_{ii}^{(k)} = 1$ , it reduces to

$$\sum_{t=1}^T \frac{n_t}{\theta} = \sum_{t=1}^T \sum_{i=1}^g z_{ii}^{(k)} \frac{m_i a_t}{1 + a_t \theta} + \sum_{t=1}^T \frac{n_t a_t}{1 + a_t \theta},$$

where it can be rewritten as the following if we denote  $\sum_{i=1}^g z_{ii}^{(k)} m_i$  as  $\tilde{z}_i^{(k)}$ :

$$\sum_{t=1}^T \left( \frac{\tilde{z}_i^{(k)} a_t \theta - n_t}{1 + a_t \theta} \right) = 0. \tag{3.16}$$

Since all of  $\tilde{z}_i^{(k)}$ ,  $n_t$  and  $a_t$  are nonnegative, (3.16) is a continuous and monotone function of  $\theta$  over the domain of  $(0, \infty)$  and takes values in the range of  $(-\sum_{t=1}^T n_t, \sum_{t=1}^T \tilde{z}_i^{(k)})$  including zero. This implies that its equivalent equation (3.15) has a unique solution, which entails little computational cost to solve for.

We measure the difference between two consecutive sets of parameters  $\Phi^{(k-1)}$  and  $\Phi^{(k)}$  using the relative distance defined as

$$d(\Phi^{(k-1)}, \Phi^{(k)}) = \sum_{i=1}^g \left| \frac{\pi_{1i}^{(k-1)} - \pi_{1i}^{(k)}}{\pi_{1i}^{(k-1)}} \right| + \left| \frac{\theta^{(k-1)} - \theta^{(k)}}{\theta^{(k-1)}} \right| + \sum_{i=1}^g \sum_{j=1}^g \left| \frac{\gamma_{ij}^{(k-1)} - \gamma_{ij}^{(k)}}{\gamma_{ij}^{(k-1)}} \right|.$$

The E-step and the M-step are iterated until  $d(\Phi^{(k-1)}, \Phi^{(k)})$  becomes sufficiently small.

### 3.2.3. Initialization

Initialization is usually one of the bottlenecks for the performance of an EM algorithm, especially for models with many parameters like ours. Based on our numerical experiments and the various real data analysis from Zucchini and MacDonald (2009), we find that the following initialization strategy produces satisfactory results in most situations:

1. For a given  $g$ , set  $m_i = si$ , where  $i = 1, \dots, g$  and  $s$  is a spread factor which serves to achieve a wider spread at the initial step. If possible, one should try using a wide range of spread factors.
2. For  $i = 1, \dots, g$ , set  $\pi_{1i}^{(0)} = 1/g$ . It is known (e.g., Zucchini and MacDonald 2009) that the estimators for  $\pi_1$  will converge to a unit vector very quickly, so using these crude uniformly distributed estimates should have little influence on the efficiency of the proposed algorithm.
3. All the off-diagonal transition probabilities are initialized at small values such as 0.01, that is,  $\gamma_{ij}^{(0)} = 0.01$  for  $i \neq j$ , and  $\gamma_{ii}^{(0)} = 1 - 0.01 \times (g - 1)$  for  $i = 1, \dots, g$ .

4. Set  $\theta^{(0)} = (g/T) (\sum_t n_t/a_t) / \sum_i m_i$ . Since  $E(N_t) = \sum_i \pi_i m_i a_t \theta$ , if we assume that  $\pi_i \equiv \pi_{1i}^{(0)} = 1/g$ , it reduces to  $E(N_t/a_t) = \sum_i m_i \theta/g$ . Thus the formula for  $\theta^{(0)}$  is simply its sample version.

### 3.3. Adjustments of the shape parameters and the number of states

In the above EM algorithm, all the shape parameters are preset, which means that the resulting estimators might be suboptimal. However, it is not practical to search for the optimal shape parameters in the strict sense, since the searching range is  $\mathbb{N}^g$ . To circumvent this issue, we follow the element-wise +1/-1 variation strategy advocated in Lee and Lin (2010), which is briefly described below.

Denote the shape parameters as  $\mathbf{m} = (m_1, \dots, m_g)$  and assume that we have fitted the model with these shape parameters.

- Increase  $m_g$  by 1 and keep all the other shape parameters unchanged. Use this new set of shape parameters and the estimated parameters from the previous fit as the new initialization and perform the fitting again. Repeat increasing the value of  $m_g$  by 1 until the likelihood of the fitted model stops increasing.
- Apply the above procedure to  $m_{g-1}, \dots, m_1$  as well. One has to make sure that  $m_1, \dots, m_g$  is always in an increasing order.
- Decrease  $m_1$  by 1 and keep all the other shape parameters unchanged. Apply the procedure in step 1. Similar for  $m_2, \dots, m_g$ . As in step 2, one should keep  $m_1, \dots, m_g$  in an increasing order.
- Repeat the above steps until the likelihood of the fitted model does not increase.

The above strategy to adjust the shape parameters is quite effective according to our numerical experiments.

At the same time, we also need to determine the appropriate number of states  $g$ . While increasing the value of  $g$  can always improve the fitting effect, it introduces the statistical issue of overfitting as well. For this purpose, we propose to use a modified backward selection strategy again from Lee and Lin (2010). Each time we delete the state with the smallest limiting probability. While the estimated value of  $\theta$  in the previous run of the EM algorithm can be directly used as the initial value for the new run, the initial values of  $\pi_1$  and  $\Gamma$  can be found by normalizing their estimated values without the deleted state. This backward selection strategy can take advantage of the estimated values from the previous run and thus significantly increase the efficiency of the proposed algorithm. We stop deleting the number of states when the chosen information criteria (IC) (which can be either AIC or BIC) of the fitted model no longer decreases. If the resulting fitted model has a lower value of the chosen IC, the new number of states  $g - 1$  as well as all the other parameters would replace their old values. This procedure of reducing the number of states

continues until the value of the chosen IC no longer decreases by deleting an additional state.

#### 4. A SIMULATION STUDY

The main purpose of this section is to investigate the algorithm’s ability to recover the original model parameters when the base model comes from the same class of distributions to the one we are considering in the paper, namely the Pascal-HMM class. To this end, we consider two scenarios, the first contains a smaller state space with relatively small values of the parameters and the second one with higher state space and very large values of the shape parameters and various types of transition probability matrices.

In the first study that we present in more details, we simulated 5,000 data points  $(n_1, n_2, \dots, n_{5000})$  from a Pascal-HMM in the form of (3.1) with the following parameters:

$$\begin{aligned}
 &g = 3, \theta = 5, \\
 &(m_1, m_2, m_3) = (12, 21, 37), \\
 &\Gamma = \begin{pmatrix} 0.90 & 0.06 & 0.04 \\ 0.03 & 0.95 & 0.02 \\ 0.06 & 0.06 & 0.88 \end{pmatrix}, \\
 &\pi_1 = (1, 0, 0), \\
 &a_l \equiv 1, l = 1, 2, \dots
 \end{aligned} \tag{4.1}$$

There exists a unique limiting distribution  $\delta = (0.273, 0.545, 0.182)$  for the given  $\Gamma$  and  $a_l \equiv 1$  corresponds to the first assumption in Theorem 4.1 in Badescu et al. (2016). However,  $\pi_1$  does not equal to  $\delta$ , which means this Pascal-HMM is not stationary. On the other hand, the underlying Markov chain quickly converges to its limiting distribution ( $\pi_{50} = (0.274, 0.544, 0.182)$  for our target model). As a result, the truncated series  $(n_{51}, n_{52}, \dots, n_{5000})$  can be considered as stationary, which allows us to calculate various quantities of interest (e.g., ACF and marginal distributions). Figure 1 shows the last 500 points of our simulated sample. The three vertical lines on the histogram correspond, respectively (from left to right) to the three mean values of the state-dependent distributions (60, 105 and 185). While it seems easy to distinguish the highest level from the other two, separation between the lower two is not obvious visually. One of our aims is to detect these three groups and determine each point’s belonging based on the fitted model.

Starting from 5 states, we followed the proposed EM algorithm and obtained a sequence of fitted models corresponding to spread factors ranging from 1 to 14. The AICs of these models are shown in Figure 2. We see that the best fitted model (in terms of AIC) is obtained from an initialization with a spread factor of 4. The influence of the choice of spread factors on the fitted model signifies the necessity of introducing it in our initialization strategy.

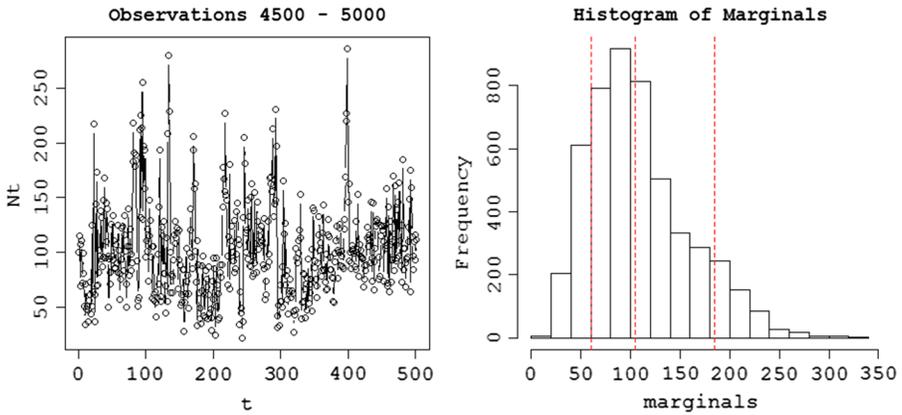


FIGURE 1: Simulated sample.

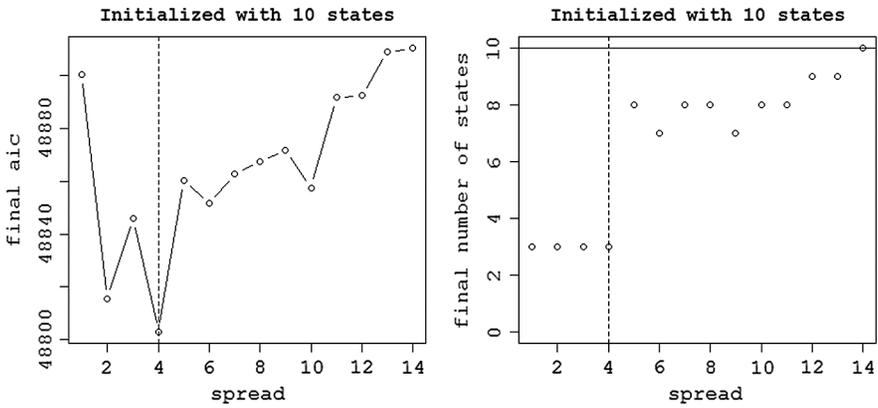


FIGURE 2: The effect of using different spread factors in the initialization.

The chosen fitted model has three states and has the following parameters:

Common Scale ( $\theta$ ): 5.005

Shapes:

(12 21 37)

Initial Probability:

(1 . .)

(4.2)

Transition Matrix:

$$\begin{pmatrix} 0.893 & 0.063 & 0.044 \\ 0.033 & 0.943 & 0.024 \\ 0.061 & 0.064 & 0.875 \end{pmatrix}$$

TABLE 1  
CLASSIFICATION OF THE SIMULATED SAMPLE.

	Accuracy
Overall	96.4%
State 1	95.7%
State 2	96.7%
State 3	96.5%

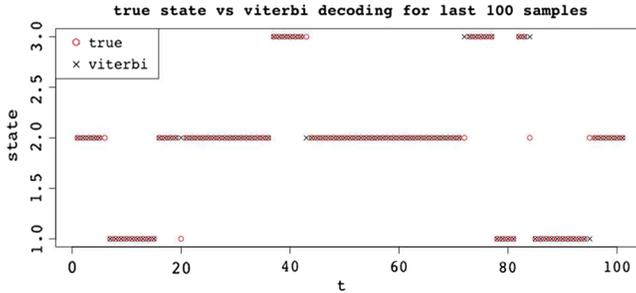


FIGURE 3: True hidden state vs. viterbi decoding.

Stationary State:

$$\begin{pmatrix} 0.2760 & 0.5253 & 0.1987 \end{pmatrix} \tag{4.3}$$

Mean Time to Return:

$$(3.623 \ 1.904 \ 5.033)$$

The fitted model successfully recognizes the three underlying states. It also fully recovers the shape parameter and the initial distribution. The estimated values for  $\theta$  and  $\Gamma$  are all very close to their target values. Closeness in parameter space indicates closeness in all distributional properties. Although not stationary, the underlying Markov chain of the fitted model quickly converges to its limiting distribution as well.

Finally, we perform data classification based on the fitted model and using the global decoding method (see Zucchini and MacDonald 2009). Our fitted model successfully recovers the states of most data points. We plotted the last 100 data points to visualize the decoding accuracy in Figure 3.

Intuitively, the state with the largest (smallest) limiting probability can be most (least) easily detected; this is confirmed by the relative errors in Table 1.

While in the first simulation study the values of the parameters are relatively nice, in the second example we purposely chose a higher dimension state space with parameter values that are very large. To this end, we simulated 1,000 time steps from a Pascal-HMM in the form of (3.1) with different number of states. The shape parameters are randomly generated and sorted in increasing

TABLE 2  
TRANSITION MATRICES WHERE  $g$  IS NUMBER OF STATES.

Transition type	$P(C_{t+1} = j   C_t = i)$
0	$1/g$
1	$((1 + g) -  i - j ) / (\sum_j (1 + g) -  i - j )$
2	$(1 +  i - j ) / (\sum_j 1 +  i - j )$

TABLE 3  
AVERAGE RELATIVE ERROR FOR PARAMETER RECOVERY.

True States	Transition Type	Fitted States	Average relative error		
			Shapes ( $m$ )	Scale ( $\theta$ )	Transition ( $\Gamma$ )
4	0	4	0.0230	0.0272	0.1063
4	1	4	0.0053	0.0041	0.1138
4	2	4	0.1478	0.1719	0.1297
8	0	9	–	0.0011	–
8	1	8	0.1150	0.1037	0.2240
8	2	8	0.3027	0.4252	0.2623

order and the common scale parameter is fixed,  $\theta = 1$ . We consider a 4- and an 8-state Pascal-HMM with the shape parameters recorded in the vector  $m = (83, 182, 271, 358, 448, 544, 633, 726)$ . Note that for the 4-state HMM we chose the first four values of the shape parameters.

Three types of transition probability behaviors are considered according to the structure presented in Table 2. For transition type 0, state transitions are uniform. Transitions of type 1 are transitions back to the current state, while transitions type 2 are transitions away from the current state.

For vectors  $v, \hat{v}$  of length  $n$ , we define the average relative error as  $E(v, \hat{v}) = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{v}_i - v_i|}{v_i}$ . For matrices, we compute the average relative error as a average over all elements. The average relative errors are presented in Table 3. As it can be observed from the table, for smaller state spaces the relative errors are small given the very large shape parameter values. However for larger state spaces (of dimension 8), the relative error increases and sometimes the algorithm is not able to recover the exact number of states. This is justified by the fact that the likelihood for such huge values of the shape parameters is insensitive to smaller changes in shape parameters, overall being almost impossible to recover the exact values of the parameters.

For 8-states, Transition Type 0, we found that the fitted model has 1 extra state:  $\hat{m} = (83, 180, 274, 359, 447, 531, 619, 661, 730)$ . This is due to the state corresponding to shape parameter 633 being misidentified as two separate states (i.e., states corresponding to shapes 619, 661), whose average shape is 640. If we combine the misidentified states as a single state with shape 640, the average relative error for shape is 0.0084.

For 8-states, Transition Type 2, the fitted common scale is 1.4252, which is far from the true scale parameter of 1. With mixture of many components, the exact parameters for each component can be hard to identify since they give similar log-likelihoods. If we compute an adjusted shape parameter for the true scale parameter by matching the component means, such that  $m^*\theta = \hat{m}\hat{\theta}$ , we find that the average relative error for the shape parameter drops to 0.0073.

To conclude, as the values of the model parameters increase it is harder and harder to recover the original parameters. However, the proposed algorithm produces very good estimates for the empirical-based model, fact that is illustrated numerically in the next section on a real insurance application.

## 5. APPLICATION: PREDICTING THE NUMBER OF IBNR CLAIMS

In this section, we aim to examine the applicability of the proposed model along with the suggested fitting algorithm through a real insurance data set. We will focus on the prediction of the number of IBNR claims.

### 5.1. Data

Our data come from an automobile liability insurance portfolio supplied by an European company. There are policies with varying policy lengths in the portfolio, with yearly policies being the most common, followed by 6-month policies, with starting dates ranging from 1/1/2005 to 6/22/2015. Among them, 321,925 claims were incurred and reported during the observation period from 1/1/2005 to 5/29/2015 with their accident and reporting dates provided. Although not directly used for the analysis in this paper, detailed records are also available for other claim information, including start and end dates of case reserves, payment amounts and times, as well as a final claim status indicating whether it is closed or not. However, information about specific claim types (bodily injury, material damage, etc.) are not provided. Since the claim development pattern can be quite different for various claim types, there might exist some degree of heterogeneity in our dataset.

We assume a valuation date  $\tau$  of 2010-01-02, based on which a training set and a validation set are extracted from the original dataset. While the training set includes all the claims incurred and reported between 2005-01-01 and 2010-01-02, the validation set shows more claims incurred during this period that were reported after 2010-01-02. A week is chosen as the unit of time. The evolution of the portfolio exposures per week over the observation period can be found in Figure 4. The exposures are calculated as earned ones, for example, if the lifetime of a policy has an overlapping of  $n$  days within a given week, then it contributes  $n/7$ th to the exposure of that week. We observe that the portfolio experienced a slowing growth from the start of valuation period until the end of valuation, on 2010-01-02. It is interesting to note that there are few sharp

TABLE 4  
CHARACTERISTICS OF THE OBSERVED REPORTING DELAYS (IN DAYS).

	Mean	Min.	1st Qu.	Median	3rd Qu.	Max.
Lower	23.14	0.00	0.00	2.00	7.00	1799.00
Upper	24.99	1.00	2.00	4.00	9.00	1801.00

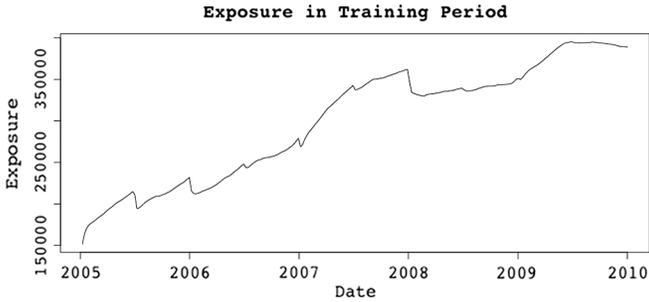


FIGURE 4: Exposure of the portfolio.

drops in the amount of covered policies during the first few years due to many policies of the same coverage length that expired simultaneously. The feature disappears slowly as policy signing dates start to vary.

Reporting delay is an important driver for the IBNR claims. Because we have some claims that are reported the same day of the accident, and as the exact times that these happen are missing from our data set, we treated this as a censored observation. We assume that a date in the table indicates that the corresponding event occurred anywhere between 00:00H and 23:59H on that day. Since an accident cannot be reported before it has occurred, we floored the lower limit of the reporting delay by 0.

$$\begin{aligned} \text{lower} &= \max\{0, \text{Reporting date} - \text{Accident date} - 1\} \\ \text{upper} &= \text{Reporting date} - \text{Accident date} + 1 \end{aligned} \quad (5.1)$$

Treating the reporting delay as censored data is not only more realistic but also more practical, since the IBNR is expected to be affected significantly by the unreported claims that occurred close to the end of the valuation period. Due to a large exposure value toward the end of the valuation day, we expect a significant number of such claims.

Some characteristics of the observed reporting delays after this modification are presented in Table 4. As it can be seen, while about three quarters of the claims were reported around 1 week, the longest reporting delay is about 4.93 years.

Since the maximum reporting delay is less than 5 years, it is reasonable to assume that most, if not all, of the claims incurred prior to 2010-01-02 have been reported by 2015-12-31.

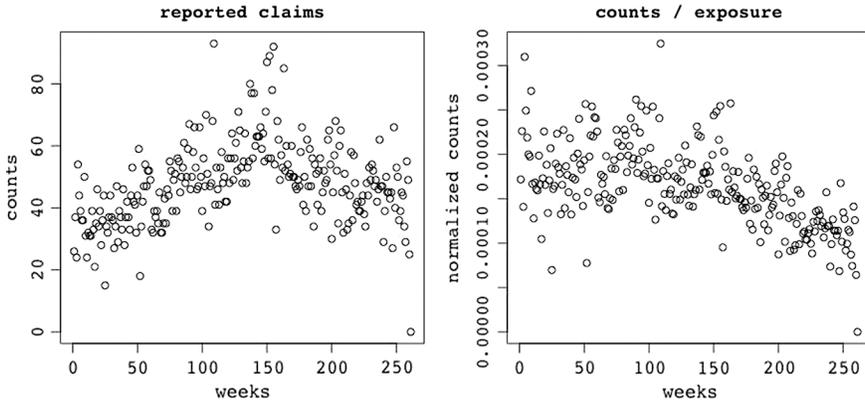


FIGURE 5: Arrivals reported by valuation.

We also plot the observed arrival counts for claims reported by the valuation date in Figure 5, as well as a normalized version, where counts are divided by exposure at that time. We note that the shape of the un-normalized version is roughly similar to the exposure during the valuation period, while the normalized version reflects a slight decay in reported claims closer to valuation time.

**5.2. Likelihood**

According to Theorem 2.1 and our model assumptions, the reported claim process  $\{N^r(t), 0 \leq t \leq \tau\}$  is a marked Cox process whose stochastic intensity function and mark density function are  $\Lambda(t)P_U(\tau - t)I_{\{0 \leq t \leq \tau\}}$  and  $\frac{p_U(u)}{P_U(\tau - t)}I_{\{0 \leq u \leq \tau - t\}}$ , respectively, where  $\Lambda(t)$  follows the Erlang-HMM described in Section 2. Without loss of generality, we assume  $\tau = d_k$ . For  $l = 1, \dots, k$ , let  $n_l^r$  stand for the number of claims incurred during period  $[d_{l-1}, d_l)$  and reported by time  $\tau$ . Ordered in terms of their arrival times, these claims together with their reporting delays are denoted as  $\{(T_i^{r(l)}, U_i^{r(l)}), i = 1, \dots, n_l^r\}$ . Using similar calculations as those in the proof of Theorem 5.1 in Badescu et al. (2016), the likelihood for observing all the arrival times and reporting delays of these claims up to time  $\tau$  (i.e., those from the training set) is

$$\begin{aligned}
 &P\left(N_l^r = n_l^r, \left(T_i^{r(l)}, U_i^{r(l)}\right) \in \left(dt_i^{r(l)}, du_i^{r(l)}\right), l = 1, \dots, k, i = 1, \dots, n_l^r\right) \\
 &= \sum_{i_1=1}^g \dots \sum_{i_k=1}^g P(C_1 = i_1, \dots, C_k = i_k) \\
 &\quad \cdot \prod_{l=1}^k \left( \int_0^\infty \lambda_l^{n_l^r} e^{-\left(\int_{d_{l-1}}^{d_l} P_U(\tau-t)dt\right)\lambda_l} f_{\Lambda_l|C_l=i_l}(\lambda_l) d\lambda_l \right) \prod_{l=1}^k \prod_{i=1}^{n_l^r} \left( dt_i^{r(l)} p_{U_i^{r(l)}}(du_i^{r(l)}) \right)
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{i_1=1}^g \dots \sum_{i_k=1}^g P(C_1 = i_1, \dots, C_k = i_k) \\
&\quad \cdot \prod_{l=1}^k \left( \int_0^\infty \frac{\left( \int_{d_{l-1}}^{d_l} P_U(\tau - t) dt \right) \lambda_l^{n_l^r} e^{-\left( \int_{d_{l-1}}^{d_l} P_U(\tau - t) dt \right) \lambda_l}}{n_l^{r_l!}} f_{\Lambda_l | C_l = i_l}(\lambda_l) d\lambda_l \right) \\
&\quad \cdot \prod_{l=1}^k \left( n_l^{r_l!} \prod_{i=1}^{n_l^r} \left( \frac{dt_i^{r(l)}}{\int_{d_{l-1}}^{d_l} P_U(\tau - t) dt} P_{U_i^{r(l)}}(du_i^{r(l)}) \right) \right) \\
&= P(N_1^r = n_1^r, \dots, N_k^r = n_k^r) \cdot \prod_{l=1}^k \left( n_l^{r_l!} \prod_{i=1}^{n_l^r} \frac{P_U(\tau - t_i^{r(l)}) dt_i^{r(l)}}{\int_{d_{l-1}}^{d_l} P_U(\tau - t) dt} \right) \cdot \prod_{l=1}^k \prod_{i=1}^{n_l^r} \frac{P_{U_i^{r(l)}}(du_i^{r(l)})}{P_U(\tau - t_i^{r(l)})}. \tag{5.2}
\end{aligned}$$

Note that (5.2) generalizes the observable likelihood under the marked Poisson process assumptions from Verrall and Wüthrich (2016) (Equation 3) breaking up into three parts. The first part is the likelihood of the discretely observed reported claim process. The second part can be interpreted as a generalization of the well-known order statistics property of the non-homogeneous Poisson process. The last part implies that the reporting delay of a claim occurred at time  $t$  is right-truncated at the threshold of  $\tau - t$  and they are independent with each other.

### 5.3. Estimation results

The maximization of the likelihood in (5.2) is a very complicated task and cannot be pursued in one piece unless certain classes of reporting delays are considered and simple numerical integration is involved (see, e.g., Weibull reporting delays in Antonio and Plat 2014). In this paper, rather than involving a global maximization of the likelihood in (5.2), we use a two step maximization. The price we pay for using this approximation for the likelihood is offset by the use of the mixture of Erlangs family for the reporting delays. This class of distributions is dense in the space of positive distributions (Tijms 1994), as well as it possesses other important properties that make it an ideal candidate for fitting purposes (see Lee and Lin 2010 for more details). More precisely, by observing that the terms involving the truncation  $P_U(\tau - t_i^{r(l)})$  from the second and third parts of (5.2) cancel each other out, in the first step we maximize the likelihood of the observed reported delay densities  $\prod_{l=1}^k \prod_{i=1}^{n_l^r} P_{U_i^{r(l)}}(du_i^{r(l)})$ . As, according to (5.1), the reporting delays are interval-censored, we use an EM algorithm for fitting a censored Erlang mixture (Verbelen et al. 2015).

The final fitted result is a mixture of 15 Erlang components as shown in Figure 6. Any additional state reduction results in a higher AIC and is thus

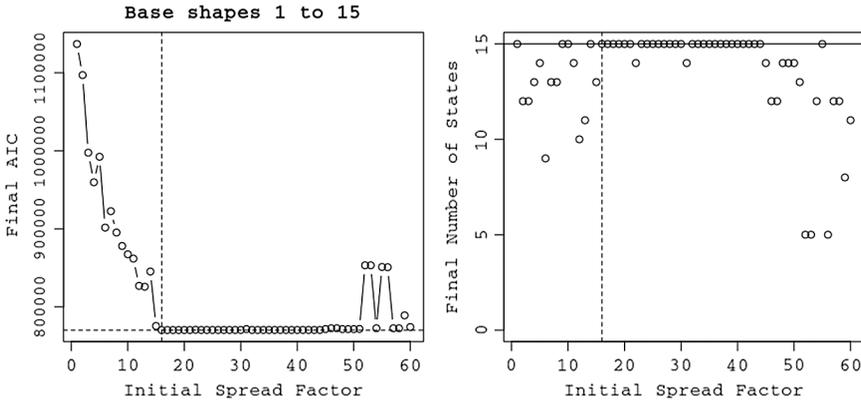


FIGURE 6: Final state reduction and AICs.

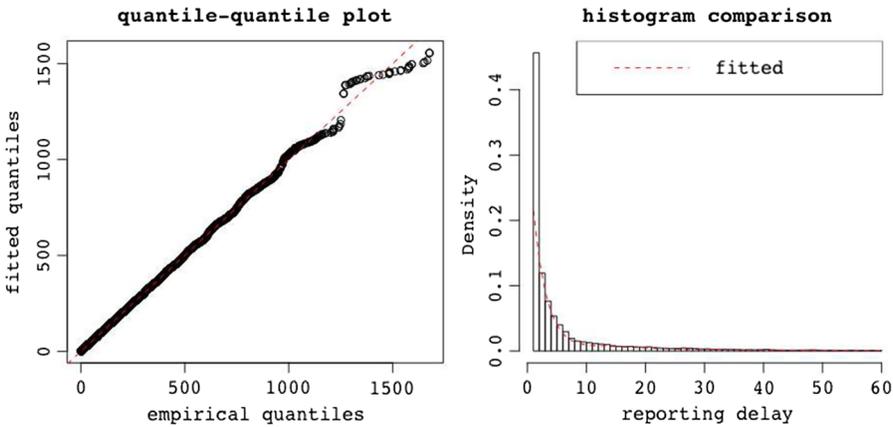


FIGURE 7: Observed reporting delays v.s. fitted Erlang mixtures.

not favoured by the fitting algorithm. Figure 7 compares the fitted result with the observed values of the censored reporting delays, via a *QQ*-plot and a histogram, where actual values are assumed to be the average of the lower and upper bounds. This is reasonable since lower and upper bounds specified above do not differ by more than 2 days. The histogram is truncated to the data of less than 60 days, as most of the mass is concentrated till 60. Visual inspection comparing the histogram to the fitted density as well as the *QQ* plot shows a satisfactory fitting both in terms of the body and tail of the data.

After the reporting delay has been fitted, the second part in (5.2) becomes a constant. Consequently, our next aim in our two-step approximation procedure is to maximize the first part therein, that is, the likelihood for the discretely observed reported claim process up to the valuation date  $\tau$ , which is shown in Figure 5. According to Theorem 2.2, this discretely observed process comes from a Pascal-HMM with the following state-dependent distribution:

$$P(N_l^r = n | C_l = i) = p \left( n; m_i, \left( \int_{d_{l-1}}^{d_l} P_U(\tau - t) dt \right) \omega_l \theta \right).$$

This is in the form of (3.1) for which we have developed an EM algorithm in Section 3, as the coefficient  $a_l$  equals  $\left( \int_{d_{l-1}}^{d_l} P_U(\tau - t) dt \right) \omega_l$ . While  $\omega_l$  can be calculated directly as shown in Figure 4,  $\int_{d_{l-1}}^{d_l} P_U(\tau - t) dt$  is unknown and we replace it with its estimator  $\int_{d_{l-1}}^{d_l} \hat{P}_U(\tau - t) dt$ , where  $\hat{P}_U(t)$  is the cumulative function of the fitted Erlang mixture for the reporting delays. Here, we assume that  $d_l = l$ , that is, the stochastic intensity function  $\Lambda(t)$  consists of piecewise random variables over daily time intervals.

**Remark 5.1.** *An important point is that due to the fact that we use only the observed reporting delays for fitting the true reporting delay distribution the likelihood in the first part of (5.2) will be underestimated. However, by using a large training period, we believe that the true reporting distribution will be accurately approximated, which is confirmed not only by our evaluation and the out of sample test, but also by the pseudo-residuals plot as a time series which shows no trend at all. For space reasoning we decided not to include the plot here. Alternatively, one can employ a similar procedure as in Verrall and Wüthrich (2016) Equation (11) and within the training period find a certain threshold  $\tau_m$  such that  $P_{U_i}(\tau - t'_i) = 1$ , for all  $t'_i > \tau_m$ . Using the reporting points until  $t_m$  only, will have the advantage that all the time points below will be fully experienced by the reserving time  $\tau$ . Obviously this approach will come at the expense of using less data points in the calibration procedure and for this reason we stick with the first approach.*

We adopt the initialization strategy recommended in Section 3.2.3 and started from 15 components with the spread factor ranging from 1 to 60. Following the proposed EM algorithm, we found that both AIC is minimized when starting with an initialization using a spread factor of 30 as seen in Figure 8.

The fitted model is a six-component HMM with the following parameters (AIC: 2833.696 loglike: -1374.848), with common scale,  $\theta = 7.19 \times 10^{-7}$ . The initial probability vector is 1 for State 2.

Shapes:

$$m = (181 \ 255 \ 305 \ 331 \ 362 \ 427)$$

Transition Matrix:

$$\Gamma = \begin{pmatrix} \cdot & \cdot & 1.0000 & \cdot & \cdot & \cdot \\ 0.0561 & 0.8586 & 0.0374 & \cdot & 0.0477 & \cdot \\ 0.0293 & 0.0226 & 0.9264 & 0.0215 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0.9366 & 0.0633 & \cdot \\ \cdot & \cdot & 0.0231 & \cdot & 0.8237 & 0.1531 \\ \cdot & 0.1557 & \cdot & 0.0614 & 0.1619 & 0.6208 \end{pmatrix} \tag{5.3}$$

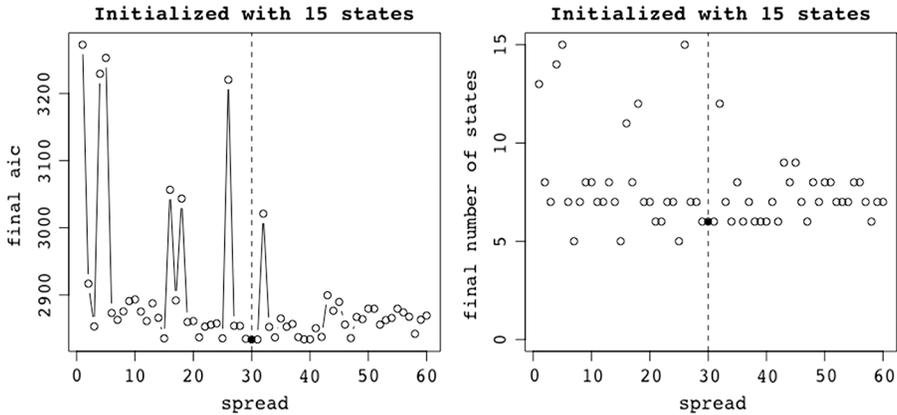


FIGURE 8: Evaluating fitting results.

Stationary State:

$$(0.0193 \ 0.1409 \ 0.3905 \ 0.2018 \ 0.0711 \ 0.1761) \tag{5.4}$$

Mean Time to Return:

$$(51.6148 \ 7.0971 \ 2.5603 \ 4.9529 \ 14.0563 \ 5.6779)$$

Most states other than the first state have a higher probability of remaining in its current state. A perhaps better image on how the transitions are made among these states can be visualized in the next section when we are decoding the states for simulation. However, an over-interpretation of the states should be avoided based on the model properties alone. For the most part, it is more appropriate to simply deem them as alternatives which can flexibly model some data characteristics. More relevant comments can be found in Zucchini and MacDonald (2009), pg 101.

**Model Diagnosis**

Due to the exposure fluctuations,  $\{N_1^r, \dots, N_\tau^r\}$  is not a stationary time series. As a result, we consider the goodness-of-fit test based on the “ordinary pseudo-residual” (Zucchini and MacDonald 2009) which computes the conditional distribution of one point given all other observations.

In our case, these ordinary pseudo-residuals are

$$z_l = \Phi^{-1} (P(N_l^r \leq n_l^r | N^{(-l)} = \mathbf{n}^{(-l)})),$$

where  $\Phi(\cdot)$  is the standard normal distribution function and  $N^{(-l)}$  and  $\mathbf{n}^{(-l)}$  stand for all the observations  $\{N_1^r, \dots, N_\tau^r\}$  minus  $N_l^r$ . The above cumulative probabilities can be calculated based on the likelihoods of modified observations:

$$P(N_l^r = n_l^r | N^{(-l)} = \mathbf{n}^{(-l)}) = \frac{\pi_1 P_1(n_1^r) B_2 \cdots B_{l-1} \Gamma P_l(n) B_{l+1} \cdots B_\tau \mathbf{1}^T}{\pi_1 P_1(n_1^r) B_2 \cdots B_{l-1} \Gamma B_{l+1} \cdots B_\tau \mathbf{1}^T},$$

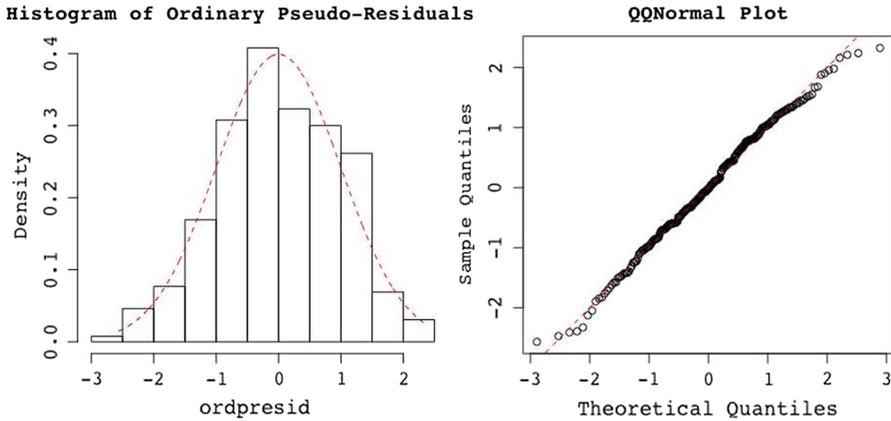


FIGURE 9: Ordinary pseudo residuals for the fitted Pascal-HMM.

and for  $l = 1$ , it is

$$P(N_1^r = n | N^{(-1)} = \mathbf{n}^{(-1)}) = \frac{\pi_1 P_1(n) B_2 \cdots B_\tau \mathbf{1}^T}{\pi_1 B_2 \cdots B_\tau \mathbf{1}^T}.$$

Here  $B_l = \Gamma P_l(n_l^r)$ , where  $P_l(n)$  is defined in (3.2).

Although  $\{N_1^r, \dots, N_\tau^r\}$  are discretely valued, they all have relatively high means (Figure 5), which means that  $z_1, z_2, \dots, z_\tau$  should be approximately  $N(0, 1)$  distributed if the fitted Pascal-HMM is valid. In Figure 9, we compare the distribution of the ordinary pseudo residuals with the standard normal distribution in terms of histogram and QQ plot. In the left panel, the red curve represents the standard normal density function. We conclude that the selected Pascal-HMM provides an adequate fit.

### 5.4. Model prediction and out-of-sample test

In this section, we aim to conduct an out-of-sample test of the fitted model. We will compare its predictive distribution of the total number of IBNR claims with the observed values from the validation set. Finally, we investigate the impact of temporal dependence on our proposed model.

The run-off triangle for our data set is shown in Table 5. The numbers in the upper triangle including the diagonal line are observed from the training set and the rest (those in bold) come from the validation set. There are 1721 claims incurred before valuation but are reported after.

#### 5.4.1. Predicting the total number of IBNR claims

We use the fitted model to predict the total number of IBNR claims at our chosen valuation date of 2010-01-02. By Proposition 2.2, the discretely observed

TABLE 5  
 RUN-OFF TRIANGLE FOR NUMBERS OF CLAIMS.

Accident year (AY)	Development year (DY)				
	1	2	3	4	5
2005	16882	1003	73	15	17
2006	21747	1241	95	23	<b>30</b>
2007	26577	1845	243	<b>50</b>	<b>48</b>
2008	24806	1831	<b>181</b>	<b>57</b>	<b>20</b>
2009	25082	<b>1180</b>	<b>110</b>	<b>38</b>	<b>7</b>

IBNR claim process comes from a Pascal-HMM with a hidden parameter process  $\{C_1, \dots, C_k\}$  described in Section 2 and the following state-dependent distributions (for  $l = 1, \dots, k; i = 1, \dots, g$ ):

$$P(N_l^{IBNR} = n | C_l = i) = p \left( n; \hat{m}_i, \left( \int_{d_{l-1}}^{d_l} (1 - \hat{P}_U(\tau - t) dt) \right) \omega_l \hat{\theta} \right),$$

where we have replaced all the model parameters by their corresponding estimators.

The predicted total number of IBNR claims given valuation date  $\tau$  is  $\sum_{l=1}^k N_l^{IBNR}$ . Although Theorem 6.1 in Badescu et al. (2016) shows that theoretically it has a closed-form expression under the model assumptions, there exist some computational issues. When one uses Proposition 6.2 in Badescu et al. (2016) to unify the different scale parameters over the time intervals, the resulting multivariate Pascal mixture will have infinite numbers of terms. There is no guarantee that keeping the first finite number of terms can give an adequate approximation. More importantly, the truncated terms might play an important role in deciding the tail shape of the predictive distribution. Due to these considerations, we choose to use the alternative approach of simulation to be described below.

As discussed in Section 5.3, we assume that  $d_l = l$ .

1. Compute the most likely sequence of hidden states  $(\{\tilde{c}_l\})$  given our fitted HMM and our observed reported arrivals process  $(\{n_l^r\})$ . For this we use the global decoding method that maximizes  $P(\mathbf{C}^{(T)} = \mathbf{c}^{(T)} | \mathbf{N}^{(T)} = \mathbf{n}^{(T)})$  over  $\mathbf{c}^{(T)}$ , where  $\mathbf{c}^{(T)} = (c_1, \dots, c_T)$ . We sample the entire sequence of hidden states by using the Viterbi algorithm (Viterbi 1967) as described in Zucchini and MacDonald (2009) and in Viterbi (1967).
2. Calculate the values of  $\left( \int_{l-1}^l (1 - \hat{P}_U(\tau - t)) dt \right) \omega_l, l = 1, \dots, \tau$ .

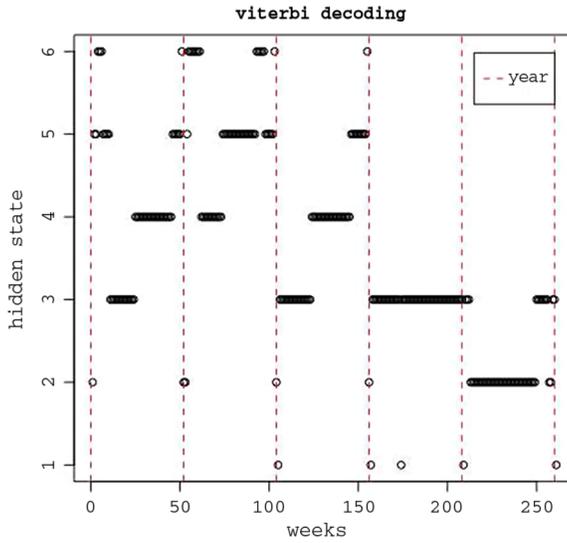


FIGURE 10: Viterbi decoding of reported claims process.

3. Conditioned on observing the underlying hidden parameter process  $(\tilde{c}_l)$ , simulate a path of the discretely observed IBNR claim process up to time  $\tau$  by  $(l = 1, \dots, \tau)$

$$P(N_l^{\text{IBNR}} = n | C_l = \tilde{c}_l) = p \left( n; \hat{m}_i, \left( \int_{l-1}^l (1 - \hat{P}_U(\tau - t) dt) \right) \omega_i \hat{\theta} \right).$$

Denote it as  $\tilde{n}_1^{\text{IBNR}}, \dots, \tilde{n}_\tau^{\text{IBNR}}$ .

4.  $\sum_{l=1}^\tau \tilde{n}_l^{\text{IBNR}}$  is one realization of the total number of IBNR claims at valuation date  $\tau$ .

Using Viterbi’s decoding procedure (Viterbi 1967), we obtained the most likely sequence of hidden states that describes the observed reported claims process. The sequence of states is shown in Figure 10. The pattern of decoded states follows approximately the normalized reported claims counts in Figure 5. As expected the model depicts some sort of seasonality effect with more claims happening in the winter periods, as the Markov chain jumps from smaller shape parameter states to larger ones. The red interrupted line in Figure 10 represents the beginning of each calendar year.

Moreover, in the initial period, we observe states which are associated with larger shape parameters (States 3,4,5 and 6) while in the period close to valuation, we observe that states associated with smaller shape parameters are decoded (States 1,2 and 3). This trend is in complete agreement with the fact that the accident arrival rate follows a decreasing trend in the last two years prior to the evaluation time in 2010 (see also Figure 5 for the accident arrival rate based on reported claims prior to the evaluation point in 2010). Our model not only can catch the potential seasonal effect in accidents, but also decreasing

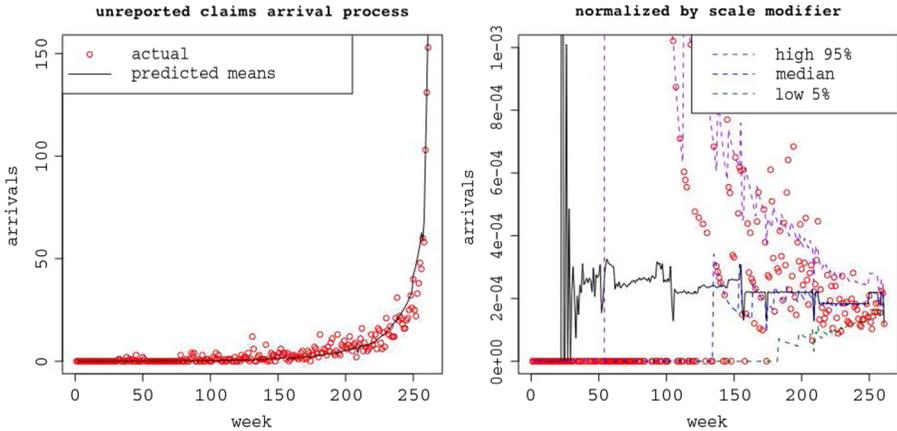


FIGURE 11: Predicted IBNR process.

**Histogram of Simulated Total IBNR**

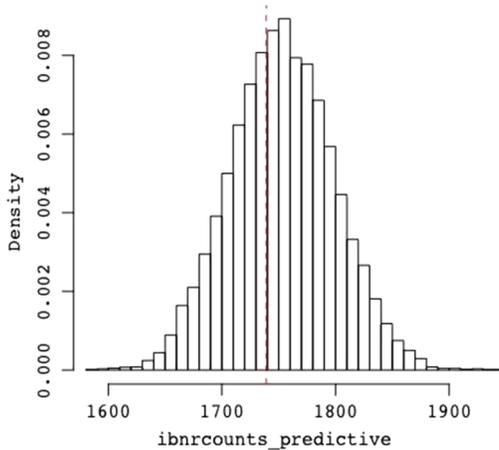


FIGURE 12: Total IBNRs from proposed model.

trends of the true claim arrival rate, and this is remarkable when one just simply assumes a stationary transition probability in the model.

We repeated the last two steps for 10,000 times to obtain 10,000 sample paths of the IBNR process as described in Figure 11. Also depicted is the IBNR predictions averaged for each day over the 10,000 samples as well as the normalized version, where we divide the counts by the scale parameter for the appropriate Pascal marginal distribution for the IBNR process. The predicted 5%, median and 95% quantiles are also plotted for each week. Actual arrivals are observed to mostly fall within the extreme quantiles. We also plotted the predictive distribution of the total number of IBNR claims at the valuation date, which is shown in Figure 12. The solid vertical line represents

the realized values observed from the validation set, which is the sum of all the bold numbers in Table 5.

## 6. CONCLUDING REMARKS

In this paper, we present an EM algorithm for calibrating the marked Cox model proposed in Badescu et al. (2016). All the model parameters, including the number of hidden states and the transition probabilities of the Markov chain, are estimated. The proposed fitting algorithm produces efficient MLEs and can be implemented at a reasonable computational cost. In particular, the E-step is equivalent to calculating the forward and backward probabilities, which are both recursively defined. The estimators in the M-step either have an explicit form or can be obtained with a minimal numerical effort. The efficiency of the fitting algorithm and the versatility of the marked Cox model are illustrated through detailed simulation studies. The usefulness of the proposed marked Cox model is demonstrated by applying it to a real insurance data set.

The work in this paper opens up few research directions that will potentially further improve the flexibility of the proposed model. One direction we intend to investigate is the improvement of the calibration procedure by globally maximizing the joint likelihood (5.2). The methodology is to treat the complexity of the data due to missing reporting delays as a missing data problem and to develop an EM type algorithm as the one proposed in Verbelen et al. (2018).

Another direction will be to focus on the time dependence of the reporting delay distribution, which is the determinant factor in estimating the number of IBNR claims (see Verrall and Wüthrich 2016). To that end, a natural extension is to assume a reporting delay distribution introduced via the same HMM that describes the latent states.

While we have only used information of policy exposure with the accident and reporting dates for our analysis in this paper, detailed records about claim amounts, such as case reserves and indicators of open/closed, are also available from our data set. As a result, one could incorporate these information as covariates of claim amounts through finite mixture regression techniques. We also plan to improve the performance of the adjustment procedures (for the number of hidden states and the shape parameters) in the EM algorithm by using some penalty functions, e.g., a modified smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li 2001).

While all the above-mentioned research directions will most likely improve the flexibility of our model, we are planning to investigate its practical utility by comparing it to various well-known macro-level models that are still used by practitioners in most instances. The macro-level models are distribution-based cell-specific models (see England and Verrall 2002) that underpin the popular Chain–Ladder method (see Wüthrich and Merz 2008 for a detailed description). Some of the advantages that such models may present are that they are easy to interpret and to implement (see, e.g., England and Verrall 2002

for a detailed study for the well known over-dispersed Poisson model). Despite this, when using such models, one needs to pay attention to the variance of the predictive distributions that can be large potentially due to the simple GLM structure assumptions, or to overparametrization that may occur due the very small data points used in model calibration. Furthermore, when using these macro-level models one cannot take full advantage of the time series claim data that are available these days. The stochastic process approach, such as the one proposed in this paper, provides an integrated modeling approach and can fully use the time series data, and as a result the predictive distributions may be more accurate. On the other hand, the implementation of such a model and the parameter estimation in particular may not be as easy as the one shown in the present paper, and a sophisticated estimation algorithm must be designed to result in a reasonable computational cost. Therefore, it will be extremely useful to investigate these comparisons at a much deeper level and this will be part of our future investigations.

## REFERENCES

- AMMETER, H. (1948) A generalization of the collective theory of risk in regard to fluctuating basic-probabilities. *Scandinavian Actuarial Journal*, (1–2), 171–198.
- ANTONIO, K. and PLAT, R. (2014) Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, (7), 649–669.
- AVANZI, B., WONG, B. and YANG, X. (2015) A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics*, **71**, 1–14.
- BADESCU, A.L., GONG, L., LIN, X.S. and TANG, D. (2015) Modeling correlated frequencies with application in operational risk management. *The Journal of Operational Risk*, **10**(1), 1–43.
- BADESCU, A.L., LIN, X.S. and TANG, D. (2016) A marked Cox model for the number of IBNR claims: theory. *Insurance: Mathematics and Economics*, **69**, 29–37.
- CAMERON, A.C. and TRIVEDI, P.K. (2013) *Regression Analysis of Count Data*, Second Edition. New York: Cambridge University Press.
- CHARPENTIER, A. and PIGEON, M. (2016) Macro vs micro methods in non-life claims reserving an econometric perspective. *Risks* **4**(2), 12.
- DIGGLE, P.J. (2003) *Statistical Analysis of Spatial Point Patterns*. New York: Oxford University Press.
- ENGLAND, P.D. and VERRALL, R.J. (2002) Stochastic claims reserving in general insurance. *British Actuarial Journal*, **8**(3), 443–518.
- FAN, J. and LI, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.
- FRIEDLAND, J. (2010) *Estimating Unpaid Claims Using Basic Techniques*. CAS Exam Study Note, Casualty Actuarial Society, Arlington, Virginia.
- GUAN, Y. (2006) A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, **101**(476), 1502–1512.
- JIN, X. and FREES, E.W.J. (2013) Comparing micro-and macro-Level loss reserving models. *Presentation at ARIA*. Washington DC.
- LEE, S.C. and LIN, X.S. (2010) Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, **14**(1), 107–130.
- LIU, H. and VERRALL, R. (2009) Predictive distributions for reserves which separate true IBNR and IBNER claims. *Astin Bulletin*, **39**(01), 35–60.
- MCKENZIE, E. (1986) Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability*, **18**(3), 679–705.

- MCLACHLAN, G. and PEEL, D. (2001) *Finite Mixture Models*. New York: John Wiley & Sons.
- MOLLER, J., SYVERSVEN, A.R. and WAAGEPETERSEN, R.P. (1998) LogGaussian Cox processes. *Scandinavian Journal of Statistics*, **25**(3), 451–482.
- MOLLER, J. (2003) Shot noise Cox processes. *Advances in Applied Probability*, **35**(3), 614–640.
- MOLLER, J. and WAAGEPETERSEN, R.P. (2003) *Statistical Inference and Simulation for Spatial Point Processes*. New York: CRC Press.
- NORBERG, R. (1993a) Prediction of outstanding liabilities in non-life insurance. *Astin Bulletin*, **23**(01), 95–115.
- NORBERG, R. (1993b) Prediction of outstanding liabilities: parameter estimation. *Proceedings of the XXIV ASTIN Coll.*, 255–266.
- NORBERG, R. (1999) Prediction of outstanding liabilities II: model variations and extensions. *Astin Bulletin*, **23**(2), 213–225.
- RENSHAW, A.E. and VERRALL, R.J. (1998) A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, **4**(04), 903–923.
- SCHNIEPER, R. (1991) Separating true IBNR and IBNER claims. *Astin Bulletin*, **21**(01), 111–127.
- TANAKA, U., OGATA, Y. and Stoyan, D. (2008) Parameter estimation and model selection for Neyman-Scott point processes. *Biometrical Journal*, **50**(1), 43–57.
- TIJMS, H.C. (1994) *Stochastic Models: An Algorithmic Approach*. New York: John Wiley & Sons Inc.
- VERBELEN, R., GONG, L., ANTONIO, K., BADESCU, A.L. and LIN, X.S. (2015) Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *Astin Bulletin*, **45**(3), 729–758.
- VERBELEN R., ANTONIO K., CLASESKENS G. and CREVECOEUR (2018) An EM algorithm to model the occurrence of events subject to a reporting delay. *Technical Report*. Leuven, Belgium: K.U. Leuven.
- VERDONCK, T., VAN WOUWE, M. and DHAENE, J. (2009) A robustification of the chain-ladder method. *North American Actuarial Journal*, **13**(2), 280–298.
- VERRALL, R., NIELSEN, J.P. and JESSEN, A.H. (2010) Prediction of RBNS and IBNR claims using claim amounts and claim counts. *Astin Bulletin*, **40**(02), 871–887.
- VERRALL R. and WÜTHRICH M.V. (2016) Understanding reporting delay in general insurance. *Risks*, **4**(3), 25.
- VITERBI, A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, **13**(2), 260–269.
- WAAGEPETERSEN, R. and GUAN, Y. (2009) Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(3), 685–702.
- WÜTHRICH, M.V. and MERZ, M. (2008) *Stochastic Claims Reserving Methods in Insurance*, Vol. 435. New York: John Wiley & Sons.
- WÜTHRICH, M.V. and MERZ, M. (2015) *Stochastic Claims Reserving Manual: Advances in Dynamic Modeling*. Available at SSRN.
- WÜTHRICH M. (2018) Neural networks applied to chain-ladder reserving. Available at SSRN.
- ZUCCHINI, W. AND MACDONALD, I.L. (2009) *Hidden Markov Models for Time Series: an Introduction Using R*. New York: CRC Press.

ANDREI L. BADESCU (Corresponding author)

*Department of Statistical Sciences*  
*University of Toronto*  
*100 St. George Street*  
*Toronto, ON M5S 3G3, Canada*  
*E-Mail: [badescu@utstat.toronto.edu](mailto:badescu@utstat.toronto.edu)*

TIANLE CHEN

*Department of Statistical Sciences*  
*University of Toronto*

*100 St. George Street  
Toronto, ON M5S 3G3, Canada  
E-Mail: [tianlechen@gmail.com](mailto:tianlechen@gmail.com)*

**X. SHELDON LIN**  
*Department of Statistical Sciences  
University of Toronto  
100 St. George Street  
Toronto, ON M5S 3G3, Canada  
E-Mail: [sheldon@utstat.toronto.edu](mailto:sheldon@utstat.toronto.edu)*

**DAMENG TANG**  
*Department of Statistical Sciences  
University of Toronto  
100 St. George Street  
Toronto, ON M5S 3G3, Canada  
E-Mail: [dameng.tang@mail.utoronto.ca](mailto:dameng.tang@mail.utoronto.ca)*

