Scandinavian Actuarial Journal

ISSN: 0346-1238 (Print) 1651-2030 (Online) Journal homepage: https://www.tandfonline.com/loi/sact20

Multivariate Cox Hidden Markov models with an application to operational risk

Tsz Chai Fung, Andrei L. Badescu & X. Sheldon Lin

To cite this article: Tsz Chai Fung, Andrei L. Badescu & X. Sheldon Lin (2019) Multivariate Cox Hidden Markov models with an application to operational risk, Scandinavian Actuarial Journal, 2019:8, 686-710, DOI: 10.1080/03461238.2019.1598482

To link to this article: https://doi.org/10.1080/03461238.2019.1598482

SCANDINAVIAN ACTUARIAL JOURNAL
2019, VOL. 2019, NO. 8, 686–710
https://doi.org/10.1080/03461238.2019.1598482

Taylor & Francis
Taylor & Francis Group

Multivariate Cox Hidden Markov models with an application to operational risk

Tsz Chai Fung, Andrei L. Badescu and X. Sheldon Lin

Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

ABSTRACT

Modeling multivariate time-series aggregate losses is an important actuarial topic that is very challenging due to the fact that losses can be serially dependent with heterogeneous dependence structures across loss types and business lines. In this paper, we investigate a flexible class of multivariate Cox Hidden Markov Models for the joint arrival process of loss events. Some of the nice properties possessed by this class of models, such as closed-form expressions, thinning properties and model versatility are discussed in details. We provide the expectation-maximization (EM) algorithm for efficient model calibration. Applying the proposed model to an operational risk dataset, we demonstrate that the model offers sufficient flexibility to capture most characteristics of the observed loss frequencies. By modeling the log-transformed loss severities through mixture of Erlang distributions, we can model the aggregate losses. Finally, out-of-sample testing shows that the proposed model is adequate to predict short-term future operational risk losses.

ARTICLE HISTORY
Received 29 November 2018
Accepted 19 March 2019

KEYWORDS
Advanced measurement approach; EM algorithm; operational losses; inter-unit and temporal dependence; multivariate risk modeling

1. Introduction

Modeling multivariate time-series aggregate losses is an important actuarial topic that applies to various actuarial research areas such as claim reserving, credibility premium calculation and operational risk management. Some of the challenges of modeling such losses are due to the heterogeneous inter-unit and temporal dependence structures that exist in reality. This paper focuses mainly on a modeling framework that is motivated by an operational risk application.

Modeling Operational Risk (OR) has been one of the major concerns for insurance and financial institutions in general. The OR constitutes a significant and growing portion of the total risks (Ames et al. 2015), and the importance of modeling and managing OR has been highly recognized by various actuarial organizations and researchers. A research paper from the Society of Actuaries (SOA) (Samad-Khan et al. 2010) emphasizes the role of OR in catastrophic losses after some serious operational failures occurred in major insurance companies during the 2008 global financial crisis. Meanwhile, the Institute and Faculty of Actuaries (IFoA) (see e.g. Kelliher et al. 2017) identifies OR as a 'core Risk Management issue' that covers all practice areas. Further, the modeling framework of the OR has been investigated by several actuarial papers (see e.g. Buch-Kromann et al. 2007, Peters et al. 2011).

The advanced measurement approach (AMA) proposed by Basel II Accord (BCBS 2004), that allows financial institutions to develop their internal models, is widely regarded as a risk sensitive

CONTACT X. Sheldon Lin sheldon@utstat.utoronto.ca Department of Statistical Sciences, University of Toronto, 100 St George Street, Toronto, ON M5S 3G3, Canada

approach to quantify and model OR, and facilitates accurate risk measurement (Lubbe & Snyman 2010, Leone et al. 2017). In particular for insurance companies, Solvency II encourages the usage of internal models similar to those proposed under the AMA, for the calculation of the OR regulatory required capital. Within the AMA framework, the most commonly used method is the loss distribution approach (LDA), also called the actuarial approach. Under LDA, financial institutions estimate the probability distributions of loss frequencies and severities separately for each business line/ event type combination (called unit of measure (UOM)), in order to obtain the aggregate loss distributions at UOM level (Frachot et al. 2001). The aggregation across all UOMs will result to the aggregate loss distribution at the top of the house (TOH) level.

Traditional LDA often implies perfectly positive correlations of losses among UOMs, which leads to the overestimation of risk (Chernobai et al. 2008). Aue & Kalkbrener (2006) and Wang et al. (2017) for e.g. adopted the copula models to capture the dependence of loss frequencies among UOMs. Badescu et al. (2015) proposed a multivariate mixed Poisson distribution to capture a wide range of dependence structures. However, the above LDA models still neglect the time-series dependence among the losses, which is less studied in the literature. Guegan & Hassani (2018) commented that autocorrelations among losses are common and intuitive for some event types and proposed a standard autoregressive and Gegenbauer process to model the serial correlations. Bardoscia & Bellotti (2011) and Gara & Belkacem (2018) modeled OR in a dynamic approach applying clustering models. For flexible dynamic modeling, Badescu et al. (2016) proposed a univariate marked Cox process to capture a wide range of temporal dependence structures, with an application to insurance claim reserving. Nonetheless, very few papers (e.g. Bardoscia & Bellotti 2011) attempted to address both time series and inter-UOM dependence behavior for operational losses.

In this paper, we propose a multivariate Cox process for the joint arrival process of loss events, which provides flexible serial and inter-UOM dependence structure among loss frequencies. Conditioned on the multivariate intensity function, the event arrival processes for each UOM are assumed to be independent heterogeneous Poisson processes. The intensity function is assumed to be a stochastic piecewise process generated from a hidden Markov model (HMM) with independent Erlang state-dependent distributions.

The proposed model possesses several important desirable properties: Firstly, the model has a natural interpretation. The underlying state in HMM can be viewed as a global environmental factor that simultaneously affects the event occurrence rates for all UOMs. The evolution of the state is governed by a transitional probability matrix, which may explain the cluster behavior of the loss arrival process. Secondly, the proposed model is highly flexible to capture the heterogeneities of various dependence structures. For the associated discretely observed process, the joint distribution of the number of events at a specific time interval is a multivariate Pascal-HMM, which can capture a broad range of inter-UOM dependence structures (Badescu et al. 2015). Further, the proposed model can capture a wide range of serial dependence structures and the changes of inter-UOM correlations over time. Thirdly, our proposed model is mathematically tractable. Closed-form expressions can be obtained for the joint distributions and the forecast distributions of loss frequencies. It is also closed under marginalization and under thinning. In many OR problems, a loss event is not observed if its severity is below a certain threshold. With the thinning property, no special treatments are needed to model the event observation processes. Also, model calibration can be easily achieved through an EM algorithm with minimal computational burden. Fourthly, the class of multivariate Pascal-HMM is identifiable under mild restrictions, making it a suitable candidate for statistical inference.

Because of its versatility, the proposed model can be applied to solve more generic actuarial problems, such as a stochastic claim reserving problem with dependent business lines. Note that Badescu et al. (2016) tackles such an insurance problem with a single business line through a special choice of our proposed model.

This paper is structured as follows. Section 2 defines and interprets the proposed model, while Section 3 discusses its desirable properties. The model calibration procedures via an EM algorithm is presented in Section 4. The application of the proposed model is first discussed in Section 5, which

explains how the data characteristics of the real OR data motivate the use of the proposed model. Section 6 provides the estimation results and validates our model through various tests. Finally, potential research directions are discussed in Section 7. The Appendix contains the proofs of some theorems and properties presented in Sections 2 and 3, as well as a simulation study that evaluates the fitting performance when a very high-dimensional dataset is involved.

2. Modeling

Suppose that there are P types of losses in a financial institution. When a risk event occurs, it is observed as a (T, Z, X) triplet, where T is the occurrence time of the event, $Z \in \{1, \ldots, P\}$ is the loss type (or the UOM, in the context of operational risk) and X is its loss severity. Hence, $\{(T_s, Z_s, X_s),\ s = 1, 2, \ldots\}$ constitute the whole event process. We model such a process through a multivariate Cox process. The Poisson process is commonly used to solve actuarial science problems (see e.g. Norberg 1993), and it is a very special case of Cox process. Marginally for the pth loss type, denote $\{(T_s^p, X_s^p),\ s = 1, 2, \ldots\}$ the corresponding event arrival process, and $\{N^p(t),\ t \geq 0\}$ the process representing the number of risk events occurred up to time t. We assume that for each $p = 1, \ldots, P$, the arrival process in the pth marginal is a Cox process with the intensity $\mathbf{\Lambda}_p(t)$. Note that Cox point process is a Poisson process with random intensity.

Mathematically, it is desirable to formulate the Cox process in the context of (marked) point process in accordance to Karr (1991) and Daley & Vere-Jones (1988). This will facilitate the analysis of several nice properties possessed by this class of models, which will be presented in the later sections. The pth marginal risk event arrival process N^p can be represented as a Cox process driven by the random measure M^p on $\mathbb{R}^+$ satisfying the following relationship

$$N^p(\mathcal{A}^p) = \sum_s \varepsilon_{T_s^p}(\mathcal{A}^p) := \sum_s 1\{T_s^p \in \mathcal{A}^p\}, \quad M^p(\mathcal{A}^p) = \int_{\mathcal{A}^p} \mathbf{\Lambda}_p(t)\, \mathrm{d}t,$$

where $\varepsilon_{T_s^p}(\mathcal{A}^p)$ and $1\{T_s^p \in \mathcal{A}^p\}$ are denoted as indicator functions such that they are both equal to 1 when $T_s^p \in \mathcal{A}^p$ and are both equal to 0 otherwise. Therefore, for any time set $\mathcal{A}^p \subseteq \mathbb{R}^+$, $N^p(\mathcal{A}^p)$ is the number of events occurred within $\mathcal{A}^p$. To define the multivariate event arrival process, we need to first extend the above pth marginal process to incorporate the loss type, such that it is a Cox process $\tilde{N}^p$ driven by the random measure $\tilde{M}^p$ on $\mathbb{R}^+ \times \{1, \ldots, P\}$, satisfies

$$\tilde{N}^p(\mathcal{A}) = \sum_s \varepsilon_{(T_s^p, Z_s^p)}(\mathcal{A}) := \sum_s 1\{(T_s^p, Z_s^p) \in \mathcal{A}\}, \quad \tilde{M}^p(\mathcal{A}) = \int_{\mathcal{A}^p} \mathbf{\Lambda}_p(t)\, \mathrm{d}t$$

for any $\mathcal{A} \subseteq \mathbb{R}^+ \times \{1, \ldots, P\}$, where $Z_s^p \equiv p$ and $\mathcal{A}^p = \{t : (t, p) \in \mathcal{A}\}$. Then, we can properly define the multivariate process as a point process N on $\mathbb{R}^+ \times \{1, \ldots, P\}$ such that

$$N = \sum_{p=1}^{P} \tilde{N}^p. \tag{1}$$

To complete the formalism of the multivariate Cox process, Proposition 2.1 states that N is still a Cox process. It is proved in Appendix 2.

Proposition 2.1: *The (multivariate) point process N defined above is a Cox process on $\mathbb{R}^+ \times \{1, \ldots, P\}$ driven by the random measure M satisfying $M(\mathcal{A}) = \sum_{p=1}^{P} \tilde{M}^p(\mathcal{A}) = \sum_{p=1}^{P} \int_{\mathcal{A}^p} \mathbf{\Lambda}_p(t)\, \mathrm{d}t$ for any $\mathcal{A} \in \mathbb{R}^+ \times \{1, \ldots, P\}$ where $\mathcal{A}^p = \{t : (t, p) \in \mathcal{A}\}$.*

To incorporate severity modeling, the Cox process N is extended to a marked Cox process on $\mathbb{R}^+ \times \{1, 2, \ldots, P\}$ with a marking space of $\mathbb{R}^+$. The marking X_s represents the loss severity of the

sth point arrival. The loss severities are assumed to be mutually independent and the p-marginal severities $\{X_s^p,\ s = 1, 2, \ldots\}$ are assumed to be identically distributed for $p = 1, \ldots, P$. Conditioned on the loss type p, the probability density of the loss severity is given by $p_{X\,|\,p}(x)$. We can define such marked point process $\bar{N}$ and determine the random measure $\bar{M}$ satisfying

$$\bar{N}(\bar{\mathcal{A}}) = \sum_s \varepsilon_{(T_s, Z_s, X_s)}(\bar{\mathcal{A}}) \quad \text{and} \quad \bar{M}(\bar{\mathcal{A}}) = \sum_{p=1}^{P} \int_{\bar{\mathcal{A}}_p} \mathbf{\Lambda}_p(t) p_{X\,|\,p}(x)\, \mathrm{d}t\, \mathrm{d}x$$

for any $\bar{\mathcal{A}} \subseteq \mathbb{R}^+ \times \{1, \ldots, P\} \times \mathbb{R}^+$, where $\bar{\mathcal{A}}_p = \{(t, x) : (t, p, x) \in \bar{\mathcal{A}}\}$. This formulation is crucial for the analysis of the thinning properties of the proposed Cox process.

The remaining task is to impose assumptions on the intensities $\mathbf{\Lambda}_p(t)$. One may simply assume that the intensities are non-random, reducing to a Poisson process. However, this implies independence of loss frequencies among loss types. Also, data are usually collected on a discrete-time basis in OR problems. To make the model feasible to implement and calibrate, we model $\mathbf{\Lambda}_p(t)$ as a piecewise stochastic process with random intensity Λ_{lp}, for $d_{l-1} \le t \le d_l$, $l = 1, 2, \ldots$ and $d_0 = 0$. Here, d_l are pre-determined time points for $l = 1, 2, \ldots$. Under this assumption, $\{N^p(t), t \ge 0\}$, $p = 1, \ldots, P$ are marginally dependent, but are independent conditioned on the intensity processes $\mathbf{\Lambda}_p(t)$, $p = 1, \ldots, P$. The multivariate intensity vector $\mathbf{\Lambda}(t) = \{\mathbf{\Lambda}_1(t), \mathbf{\Lambda}_2(t), \ldots, \mathbf{\Lambda}_P(t)\}$ is modeled by an Erlang hidden Markov model (Erlang-HMM), which is defined by the following structure:

- The hidden parameter process $\{C_1, C_2, \ldots\}$ is a time-homogeneous Markov chain with a finite state space $\{1, 2, \ldots, g\}$. Its initial distribution and transition probability matrix are respectively denoted by row vector $\boldsymbol{\delta}$ and matrix $\mathbf{\Gamma} = (\gamma_{ij})_{g \times g}$, where $\gamma_{ij} = P(C_l = j \,|\, C_{l-1} = i)$. The state C_l can be interpreted as an unobservable time-varying global environmental factor that affects the event occurrence rates for all P loss types at the same time. The change of states over time is governed by $\mathbf{\Gamma}$, where large values in its diagonal entries imply high chance of staying at the same state over time, representing the clustering behavior of operation loss arrival processes.
- The state-dependent vector process $\mathbf{\Lambda}_{l\cdot} = \{\Lambda_{l1}, \Lambda_{l2}, \ldots, \Lambda_{lP}\}$ with $l = 1, 2, \ldots$, is defined such that it depends only on the current state C_l given all the observed histories $\mathbf{\Lambda}_{\cdot}^{(l-1)} = \{\mathbf{\Lambda}_{1\cdot}, \mathbf{\Lambda}_{2\cdot}, \ldots, \mathbf{\Lambda}_{l-1,\cdot}\}$ and the hidden histories $\boldsymbol{C}^{(l)} = \{C_1, \ldots, C_l\}$, i.e. $P(\mathbf{\Lambda}_{l\cdot} | \mathbf{\Lambda}_{\cdot}^{(l-1)}, \boldsymbol{C}^{(l)}) = P(\mathbf{\Lambda}_{l\cdot} \,|\, C_l)$. Given that $C_l = i$, we further assume that $\Lambda_{l1}, \Lambda_{l2}, \ldots, \Lambda_{lP}$ are independent Erlang distributed random variables with joint density function given by

$$K_{\mathbf{\Lambda}_{l\cdot}\,|\,C_l=i}(\lambda_1, \ldots, \lambda_P) = \prod_{p=1}^{P} h(\lambda_p; m_{ip}, \theta_{ip}), \quad \text{with } h(\lambda; m, \theta) = \frac{\lambda^{m-1} e^{-\lambda/\theta}}{\theta^m (m-1)!}. \tag{2}$$

Remark 2.1: The conditional independence of random intensities seems to be a strong assumption. Hence, one may attempt to extend the model by writing the state-dependent density function as a multivariate mixture of Erlang, understanding that the class of multivariate mixture of Erlang is dense in the space of multivariate positive continuous distributions

$$K_{\mathbf{\Lambda}_{\cdot l}\,|\,C_l=i}(\lambda_1, \ldots, \lambda_P) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} h(\lambda_p; m_{kip}, \theta_{kip}). \tag{3}$$

However, it is shown in Appendix 1 that the resulting model can be easily converted to our original proposed model with the joint density function of the state-dependent intensities expressed in the form of Equation (2). Such an extension does not increase the model flexibility at all.

3. Properties of the event arrival process

The multivariate Cox process is defined using the marginal processes as a starting point. By definition, it is closed under marginalization, meaning that N^p is a univariate Cox process with random intensity $\mathbf{\Lambda}_p(t)$ for $p \in \{1, \ldots, P\}$ whenever N is a multivariate Cox process with random intensity $\mathbf{\Lambda}(t)$. Also, we can easily see that $\mathbf{\Lambda}_p(t)$ follows a univariate Erlang-HMM whereas $\mathbf{\Lambda}(t)$ follows a multivariate Erlang-HMM. This section presents other desirable properties, which are critical to the model versatility and the effectiveness of model calibrations.

3.1. Thinning properties

Often in an operational risk problem, an event is recorded only if its loss amount exceeds a certain threshold level. Immaterial events are not observed and the recorded severities are left-truncated. It is desirable to identify the relationship between the actual event arrival process (non-truncated one) and the event observation process (truncated one). This subsection shows that the model structures of the truncated and non-truncated processes are indeed the same. Therefore, no special treatments are needed for model calibrations of the event observation process. Precisely, it will be shown that if the event arrival process $\bar{N}$ is a marked Cox process with the underlying intensity vector generated by Erlang-HMM, the event observation process $\bar{N}'$ is still a marked Cox process with Erlang-HMM intensities. The proof is demonstrated in Appendix 3.

Theorem 3.1: *Assume that the risk event arrival process $\bar{N}$ is a marked Cox process with intensity vector $\mathbf{\Lambda}(t)$. Corresponding to each loss event, the loss amounts X_i are independent and are only related to its loss type $p \in \{1, \ldots, P\}$, with density function $p_{X|p}(x)$. Define the risk event observation process $\bar{N}'$, where any event arrived is recorded only if $X_i > \psi(p)$. Here, $\psi(p)$ is the recording threshold and is assumed to be constant over time. Then, $\bar{N}'$ also follows a marked Cox process with the underlying intensity vector $\mathbf{\Lambda}'(t) := \{\mathbf{\Lambda}'_1(t), \ldots, \mathbf{\Lambda}'_P(t)\}$ generated by Erlang-HMM with state-dependent intensity density function*

$$K_{\Lambda'_{l1},\ldots,\Lambda'_{lP}|C_l=i}(\lambda_1, \ldots, \lambda_P) = \prod_{p=1}^{P} h(\lambda_p; m_{ip}, \theta'_{ip}),$$

where $\theta'_{ip} = \bar{F}_{X|p}(\psi(p))\theta_{ip}$. The (independent) observed loss amounts follow the adjusted position-dependent density $p'_{X|p}(x) = (p_{X|p}(x)/\bar{F}_{X|p}(\psi(p)))1\{x > \psi(p)\}$.

Theorem 3.1 shows that the (unmarked) risk event arrival process follows exactly the same model as the (unmarked) risk event observation process, except that the parameters θ_{ip} are adjusted to θ'_{ip}. Therefore, the proposed multivariate Cox process is closed under thinning.

3.2. Distributional properties of the corresponding discrete processes

This subsection exhibits some distributional properties of the discretized event arrival process $\{\boldsymbol{N}_l;\ l = 1, 2, \ldots\}$, where $\boldsymbol{N}_l = (N_{l1}, \ldots, N_{lP})$, and N_{lp} is the number of risk events of type p occurred during the time interval $[d_{l-1}, d_l)$. This is motivated by the fact that data is only collected on a discrete-time basis. Without much loss of generality, we assume that $d_l - d_{l-1} = 1$ for $l = 1, 2, \ldots$ throughout the paper. Because of the thinning property (Theorem 3.1), the following results still hold for the discretized event observation process with adjusted scale parameters θ'_{ip}.

Proposition 3.1: *Under the proposed multivariate Cox Process, the discretized event arrival process $\{\boldsymbol{N}_l;\ l = 1, 2, \ldots\}$ follows a multivariate Pascal-HMM. Its hidden parameter process $\{C_1, C_2, \ldots\}$ is governed by the transition probability $\mathbf{\Gamma}$. Given that $C_l = i$, $\boldsymbol{N}_l$ follows a multivariate Pascal distribution*

with density function given by

$$P(\boldsymbol{N}_l = \boldsymbol{n} \mid C_l = i) = \prod_{p=1}^{P} p(n_p; m_{ip}, \theta_{ip}),$$

where $\boldsymbol{n} = (n_1, \ldots, n_P)$ *and* $p(n, m, \theta) = \binom{n+m-1}{m-1}(1/(1+\theta))^m(\theta/(1+\theta))^n$.

Using this result, the univariate distribution of N_{lp}, the k-variate joint distribution of $(N_{l_1 p}, \ldots, N_{l_k p})$ and the P-variate joint distribution of the number of event arrivals at time l $(N_{l1}, \ldots, N_{lP})$ for $p = 1, \ldots, P$ can be easily obtained:

$$P(N_{lp} = n) = \sum_{i=1}^{g} \pi_{li} p(n; m_{ip}, \theta_{ip}), \tag{4}$$

$$P(N_{l_1 p} = n_{l_1}, \ldots, N_{l_k p} = n_{l_k}) = \sum_{i_1=1}^{g} \cdots \sum_{i_k=1}^{g} \beta_{(i_1,\ldots,i_k)} \prod_{j=1}^{k} p(n_{l_j}; m_{i_j p}, \theta_{i_j p}), \tag{5}$$

$$P(N_{l1} = n_1, \ldots, N_{lP} = n_P) = \sum_{i=1}^{g} \pi_{li} \prod_{p=1}^{P} p(n_p; m_{ip}, \theta_{ip}), \tag{6}$$

where π_{li} is the ith element of the vector $\boldsymbol{\pi}_l = \boldsymbol{\delta}\boldsymbol{\Gamma}^{l-1}$, $\beta_{(i_1,\ldots,i_k)} = \pi_{l_1 i_1}\gamma_{i_1 i_2}(l_2 - l_1)\cdots\gamma_{i_{k-1} i_k}(l_k - l_{k-1})$ and $\gamma_{ij}(l)$ is the (i,j)th element of the l-step transition probability matrix $\boldsymbol{\Gamma}^l$.

Equations (4) to (6) show that the k-variate marginal of the number of events across time as well as the P-variate number of events across loss types are multivariate Pascal mixtures. The following example, which can be easily generalized to more complicated settings, demonstrates that the proposed model is versatile in capturing a wide range of dependence structures in terms of second moments. We aim to show that (almost) perfectly positive and negative correlations can be simultaneously attained by the proposed model. For simplicity, we assume $P = 2$ and $g = 2$. We set $\theta_{ip} = \theta$ as a constant and choose the following parameters:

$$\boldsymbol{\delta} = (0.5, 0.5, 0, 0), \quad \boldsymbol{\Gamma} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{m} = \begin{pmatrix} m_0 & m_0 \\ 2m_0 & 2m_0 \\ m_0 & 2m_0 \\ 2m_0 & m_0 \end{pmatrix}.$$

Directly computing the moments of multivariate Pascal mixtures, it is easy to show that $\mathrm{Corr}(N_{11}, N_{12}) \to 1$ but $\mathrm{Corr}(N_{21}, N_{22}) \to -1$ when $m_0 \to \infty$. Hence, the proposed model is flexible to capture the dependence structure across loss types, as well as the changes of dependence structures among loss types over time. Moreover, $\mathrm{Corr}(N_{11}, N_{21}) \to 1$ and $\mathrm{Corr}(N_{12}, N_{22}) \to -1$ when $m_0 \to \infty$. Therefore, the proposed model can also cater for a broad range of temporal dependence structures.

Badescu et al. (2015) demonstrate the versatility of multivariate Pascal mixture in capturing dependence structures by a simpler version of the above example. Theorem 4.1 of Badescu et al. (2016) argues (through the autocorrelation function (ACF)) that univariate Pascal-HMM, a special case of our proposed model, is flexible in capturing time-series correlations. Such an argument still holds for multivariate Pascal-HMM because of its closure under marginalization property.

It is also interesting to investigate the properties of the aggregated event arrival process $\{N^a(t),\ t \geq 0\}$ (i.e. the number of events aggregated across all loss types). Mathematically, $N^a = N(\cdot, \{1, \ldots, P\})$, where N is given by Equation (1). Therefore, N^a is a point process on $\mathbb{R}^+$. Appendix 4 shows that such process still follows univariate Cox-HMM. If $\theta_{ip} \equiv \theta_i$ does not depend on the loss type p, the proposed

model is 'closed under aggregation'. Under any choices of θ_{ip}, the distribution of the corresponding discretized process $N_l^a = \sum_{p=1}^{P} N_{lp}$ can still be written as a Pascal mixture:

$$P(N_l^a = n) = \sum_{i=1}^{g} \pi_{li} \sum_{k=1}^{\infty} \tilde{\psi}_k p(n; k, \theta_i), \tag{7}$$

where $\theta_i = \min\{\theta_{i1}, \theta_{i2}, \ldots, \theta_{iP}\}$, $\boldsymbol{k} = (k_1, k_2, \ldots, k_P)$, $\tilde{\psi}_k = \sum_{k_1+\cdots+k_P=k} \psi_{\boldsymbol{k}} \prod_{p=1}^{P} 1\{k_p \geq m_{ip}\}$ and $\psi_{\boldsymbol{k}} = \prod_{p=1}^{P} \binom{k_p-1}{m_{ip}-1} (\theta_i/\theta_{ip})^{m_{ip}} (1 - \theta_i/\theta_{ip})^{k_p - m_{ip}}$.

3.3. Forecast distributions

One important use of modeling the operational risk events is to predict the future losses, so that financial institutions can set up adequate reserves. Suppose that the discretized loss arrival processes $\boldsymbol{N}^{(L)} := (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_L)$ are observed and their realizations are $\boldsymbol{n}^{(L)} := (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L)$. Introduce

(1) The random sample and observed data up to time l are respectively given by $\boldsymbol{N}^{(l)} = (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_l)$ and $\boldsymbol{n}^{(l)} = (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_l)$.
(2) The forward probabilities row vector $\boldsymbol{\alpha}_l$, with its i^{th} element $\alpha_l(i) = P(\boldsymbol{N}^{(l)} = \boldsymbol{n}^{(l)}, C_l = i)$. Zucchini & MacDonald (2009) shows that for $l = 2, 3, \ldots, L, \boldsymbol{\alpha}_1 = \boldsymbol{\delta P}(\boldsymbol{n}_1)$, $\boldsymbol{\alpha}_l = \boldsymbol{\alpha}_{l-1} \boldsymbol{\Gamma P}(\boldsymbol{n}_l)$.
(3) The backward probabilities column vector $\boldsymbol{\beta}_l$, with its ith element $\beta_l(i) = P(\boldsymbol{N}_{l+1}^{L} = \boldsymbol{n}_{l+1}^{L} \mid C_l = i)$. It can be shown that for $l = 1, 2, \ldots, L-1$, $\boldsymbol{\beta}_L = \boldsymbol{1}$, $\boldsymbol{\beta}_l = \boldsymbol{\Gamma P}(\boldsymbol{n}_{l+1})\boldsymbol{\beta}_{l+1}$.
(4) The observed data likelihood $\mathcal{L}_L = P(\boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}) = \boldsymbol{\delta P}(\boldsymbol{n}_1)\boldsymbol{\Gamma P}(\boldsymbol{n}_2) \cdots \boldsymbol{\Gamma P}(\boldsymbol{n}_L)\boldsymbol{1}$.

Denote an arbitrary finite-dimensional vector $\boldsymbol{N}^* \subset \{N_{l,p};\ l = L+1, L+2, \ldots, p = 1, \ldots, P\}$, which represents any future information on the discretized event arrivals. Let $\boldsymbol{n}^*$ be its corresponding realizations. Applying simple probabilistic arguments, the forecast distributions are given by

$$P(\boldsymbol{N}^* = \boldsymbol{n}^* \mid \boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}) = \sum_{i=1}^{g} \delta^*(i) P(\boldsymbol{N}^* = \boldsymbol{n}^* \mid C_{L+1} = i), \tag{8}$$

where $\delta^*(i)$, the ith element of $\boldsymbol{\delta}^*$, is defined by

$$\delta^*(i) := P(C_{L+1} = i \mid \boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}) = \frac{(\boldsymbol{\alpha}_L \boldsymbol{\Gamma})_i}{\mathcal{L}_L}. \tag{9}$$

The predictive multivariate discretized process $(\boldsymbol{N}_{L+1}, \boldsymbol{N}_{L+2}, \ldots)$ is a restarted Pascal-HMM with an adjusted initial state distribution $\boldsymbol{\delta}^*$, which depends on the past observations.

3.4. Model identifiability

Model identifiability is an important issue in statistical inference. For a non-identifiable class of model, different sets of parametrization can result to the same model distribution, hence causing troubles for statistical inference. The conditions for the proposed model to be identifiable are discussed in this subsection. We first provide the definition of identifiability for the class of HMMs. This definition aligns with that by Teicher (1963) and Yakowitz & Spragins (1968) for the class of finite mixture through treating the class of models as identifiable under the existence of permutation invariance and considering only irreducible models, i.e. no components/ states always have zero weight.

Definition 3.1: Let $\mathcal{G}$ be the class of all HMMs with state-dependent distribution $\mathcal{F}_\xi$ (ξ corresponds to the parameters of the state-dependent distribution), so that each element $G_{\boldsymbol{\Phi}} \in \mathcal{G}$ has the parameter

setting $\boldsymbol{\Phi} = (\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_g)$. A subclass $\bar{\mathcal{G}} \subseteq \mathcal{G}$ is identifiable whenever $G_{\boldsymbol{\Phi}^*}, G_{\boldsymbol{\Phi}} \in \bar{\mathcal{G}}$ with $P(C_L = i; \boldsymbol{\Phi}) > 0$ for some $L = 1, 2, \ldots$ for all $i = 1, \ldots, g$ and the same for $\boldsymbol{\Phi}^*$ (irreducibility condition), if their observable likelihoods match for all $L = 1, 2, \ldots$, i.e.

$$P(\boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}; \boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\xi}^*, g^*) = P(\boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$$

it implies that $g^* = g$, $\delta^*(i) = \delta(c(i))$, $\gamma^*_{ij} = \gamma_{c(i),c(j)}$ and $\xi^*_i = \xi_{c(i)}$ for $i = 1, \ldots, g$, where $\{c(1), \ldots, c(g)\}$ is a permutation of $\{1, \ldots, g\}$.

Theorem 3.2: *The proposed model for the discretized event arrival processes (i.e.: Multivariate Pascal HMM in the form described by Proposition* 3.1*) is identifiable subject to the restriction that $\xi_1, \ldots, \xi_g$ are distinct, where $\xi_i = (m_{i1}, \ldots, m_{iP}, \theta_{i1}, \ldots, \theta_{iP})$ for $i = 1, \ldots, g$.*

The proof is detailed in Appendix 5. Consider fitting the proposed model to an OR dataset. By allowing the scale parameters θ_{ip} dependent on the state i and the loss type p, the possibilities that there exists $i \neq j$ such that $\xi_i = \xi_j$ for the fitted model are eliminated, since θ_{ip} is taking a continuous value. By Theorem 3.2, the fitted model is identifiable. On the other hand, there exists the risk of fitting a non-identifiable model if a universal scale parameter θ is assumed across all states (e.g. Badescu et al. 2016), because the shape parameter m_{ip} is taking a discrete value.

4. Model calibration: an EM algorithm

In this section, we will present an EM algorithm to estimate the parameters of the proposed model based on the observed discretized loss arrival process $\boldsymbol{N}^{(L)}$. At each run of the EM algorithm, we fix g and m_{ip}. The adjustments of such parameters will be addressed later in this section. The goal of the EM algorithm is to efficiently estimate the parameters $\boldsymbol{\Phi} = (\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\theta_{ip} : i = 1, \ldots, g;\ p = 1, \ldots, P\}$. By introducing $\boldsymbol{Z}_l = (Z_{l1}, \ldots, Z_{lg})$ and $Z_{li} = 1\{C_l = i\}$ for $i = 1, 2, \ldots, g$, the complete data likelihood and log-likelihood can respectively be written as

$$\mathcal{L}(\boldsymbol{\Phi}; \boldsymbol{n}^{(L)}, \boldsymbol{z}^{(L)}) = \prod_{i=1}^{g} \delta_i^{z_{1i}} \prod_{l=2}^{L} \prod_{i=1}^{g} \prod_{j=1}^{g} (\gamma_{ij})^{z_{l-1,i} \times z_{lj}} \prod_{l=1}^{L} \prod_{i=1}^{g} \prod_{p=1}^{P} p(n_{lp}; m_{ip}, \theta_{ip})^{z_{li}},$$

$$l(\boldsymbol{\Phi}; \boldsymbol{n}^{(L)}, \boldsymbol{z}^{(L)}) = \sum_{i=1}^{g} z_{1i} \log \delta_i + \sum_{l=2}^{L} \sum_{i=1}^{g} \sum_{j=1}^{g} z_{l-1,i} z_{lj} \log \gamma_{ij} + \sum_{l=1}^{L} \sum_{i=1}^{g} \sum_{p=1}^{P} z_{li} \log p(n_{lp}; m_{ip}, \theta_{ip}).$$

4.1. E-step

The E-step computes the expectation of the complete data log-likelihood at sth iteration given the observed data, evaluated using the parameters obtained in the previous iteration:

$$\begin{aligned} Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(s-1)}) &= E[l(\boldsymbol{\Phi}; \boldsymbol{n}^{(L)}, \boldsymbol{z}^{(L)}) \mid \boldsymbol{n}^{(L)}, \boldsymbol{\Phi}^{(s-1)}] \\ &= \sum_{i=1}^{g} z_{1i}^{(s)} \log \delta_i + \sum_{l=2}^{L} \sum_{i=1}^{g} \sum_{j=1}^{g} z_{lij}^{(s)} \log \gamma_{ij} + \sum_{l=1}^{L} \sum_{i=1}^{g} \sum_{p=1}^{P} z_{li}^{(s)} \log p(n_{lp}; m_{ip}, \theta_{ip}), \end{aligned}$$

where $z_{li}^{(s)} = E[Z_{li} \mid \boldsymbol{n}^{(L)}, \boldsymbol{\Phi}^{(s-1)}]$ and $z_{lij}^{(s)} = E[Z_{l-1,i} Z_{lj} \mid \boldsymbol{n}^{(L)}, \boldsymbol{\Phi}^{(s-1)}]$ are respectively given by

$$z_{li}^{(s)} = \frac{\alpha_l(i)^{(s-1)} \beta_l(i)^{(s-1)}}{\mathcal{L}_L^{(s-1)}}, \quad z_{lij}^{(s)} = \frac{\alpha_{l-1}^{(s-1)}(i) \gamma_{ij}^{(s-1)} \left[\prod_{p=1}^{P} p(n_{lp}; m_{jp}, \theta_{jp})\right] \beta_l^{(s-1)}(j)}{\mathcal{L}_L^{(s-1)}}, \tag{10}$$

where $\alpha_l^{(s-1)}$, $\beta_l^{(s-1)}$ and $\mathcal{L}_L^{(s-1)}$ are respectively the forward probabilities, backward probabilities and complete data log-likelihood computed using $\boldsymbol{\Phi}^{(s-1)}$ as parameters.

4.2. M-step

The M-step maximizes $Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(s-1)})$ for the sth iteration, subject to constraints of $\sum_{i=1}^{g} \delta_{1i} = 1$ and $\sum_{j=1}^{g} \gamma_{ij} = 1$ for $j = 1, 2, \ldots, g$. For $i, j = 1, 2, \ldots, g$, the following parameters can be easily updated using the method of Lagrange's multiplier and standard calculus

$$\delta_i^{(s)} = z_{1i}^{(s)}, \quad \gamma_{ij}^{(s)} = \frac{\sum_{l=2}^{L} z_{lij}^{(s)}}{\sum_{j'=1}^{g} \sum_{l=2}^{L} z_{lij'}^{(s)}}, \quad \theta_{ip} = \frac{\sum_{l=1}^{L} z_{li}^{(s)} n_{lp}}{\sum_{l=1}^{L} z_{li}^{(s)} m_{ip}}. \tag{11}$$

The E-step and M-step are repeated until the observed data log-likelihood between two consecutive iterations is smaller than a tolerance threshold of 10^{-3}.

4.3. Initialization and parameter adjustments

Proper initialization is important as it can affect the performance of the proposed EM algorithm. In this subsection, we propose a simple initialization strategy that involves randomization and first-moment matching. For $i = 1, \ldots, g$:

- For $p = 1, \ldots, P$, sample m_{ip} uniformly on $\{1, 2, \ldots, C\}$, where C is a constant. Based on our experiments, the fitting result is insensitive to C unless it is too small (say $C < 5$).
- Set $\delta_i^{(0)} = 1/g$, $\gamma_{ij}^{(0)} = 0.01$ for $i \neq j$ and $\gamma_{ii}^{(0)} = 1 - 0.01 \times (g - 1)$. See Badescu et al. (2019) for further justifications.
- For $p = 1, \ldots, P$, set $\theta_{ip}^{(0)} = \sum_{l=1}^{L} n_{lp}/(L m_{ip})$ to match the first moments.

Remark 4.1: One may attempt to initialize m_{ip} using the spread factor strategy introduced by Verbelen et al. (2015) since it is found to provide satisfactory results while fitting univariate data. However, if this strategy is applied to P-variate data, the number of initial states will become prohibitively large (g initial states in univariate case means g^P states in P-variate case).

Remark 4.2: By allowing the scale parameters θ_{ip} dependent on i and p, the fitting results are empirically much more stable, since the initialized distribution of the proposed model matches the first moment of the data (for each state/ loss type). On the other hand, assuming a universal θ, the first moment of the initialized distribution for some states/ loss types can be very far away from that implied by the data, causing numerical underflow of the initial observed likelihood.

Remark 4.3: While the proposed initialization strategy is found to be robust and stable for the real operational risk dataset in Section 5 and also the simulated dataset in Appendix 6, there can be many other initialization strategies (such as the k-means clustering strategies proposed by Gui et al. 2018) that may yield even better performances. Yet, it is not the focus of this paper and hence we leave it as a direction of potential future research.

The remaining task is to find m_{ip} and g to optimize the fitting. We adopt the element-wise $+1/-1$ variation strategy (Lee & Lin 2010) for m_{ip}. There are many ways to control the number of parameters, for example we can choose the optimal g that minimizes the Akaike Information Criterion (AIC). Since no further iterations are needed within an EM iteration, it is computationally feasible to try a wide range of g and find the optimal one. An alternative approach is to follow the backward selection strategy proposed by Lee & Lin (2010).

5. Overview of data

From this section onwards, we are to fit the operational risk data into the proposed model. The dataset comes from a North American financial institution from April 2007 to March 2012. It consists of the

occurrence date, the UOM and the loss severities for each operational risk event. While the dataset consists of a large number of UOMs, many of them are incomplete or contain very few losses, so it is not material and statistically meaningful to model the losses from these UOMs. Therefore, we choose to analyze 10 UOMs in total, covering two event types: External Fraud (EF) and Execution, Delivery and Process Management (EP) and five business lines: Card Service (CS), Commercial Banking (CB), Retail Brokerage (RB), Private Banking (PB) and Retail Banking (RE). Note that to examine the versatility of the proposed model and the efficiency of the proposed algorithm when the number of UOMs is large, a simulation study is performed in Appendix 6.

The losses were left-truncated at $15,000 for UOM2 and $30,000 for other UOMs. Any losses below the threshold will not be recorded. The loss frequencies are aggregated in monthly bins and the data is split into two parts: In sample (IS) data consists of information up to December 2011, while Out of sample (OS) data covers the losses for the remaining three months.

5.1. Data summary and challenges

This subsection aims to perform preliminary data analysis on the IS data in terms of the basic observations on loss frequency characteristics (Table 1), the serial correlation structures (Figure 1), the inter-UOM dependence structures (Table 2) and the relationships between frequencies and severities (Figure 2). The tail behavior of the loss severities will also be examined.

These figures and tables reveal great challenges to properly model the operational risk data. The number of data points for loss frequency is scarce. IS period consists of 57 months, together with 10 UOMs, there are only 570 data features available for multivariate count model fitting. To attempt increasing the number of data points, one may suggest separating the count data into weekly or daily bins. This approach, however, is undesirable because majority of operational losses are reported on the last day of each month due to monthly reporting administrative issues.

Table 1. The summary of the observed monthly number of losses by UOM in IS period.

	EF/CS	EP/CS	EF/CB	EP/CB	EF/RB	EP/RB	EF/PB	EP/PB	EF/RE	EP/RE
UOM	1	2	3	4	5	6	7	8	9	10
Mean	118.74	84.14	160.23	19.61	5.75	42.72	6.25	53.96	753.44	567.89
SD	44.16	27.54	111.95	6.44	3.29	20.50	3.01	27.53	160.26	174.20
Trend	↓	↑	No	↑	No	No	No	No	No	↑
Clustering	No	No	Yes	No	No	No	No	Yes	Yes	No

Note: The trend (number of losses increasing/ decreasing over time) and clustering (whether or not clustering of losses exists) are based on observations.

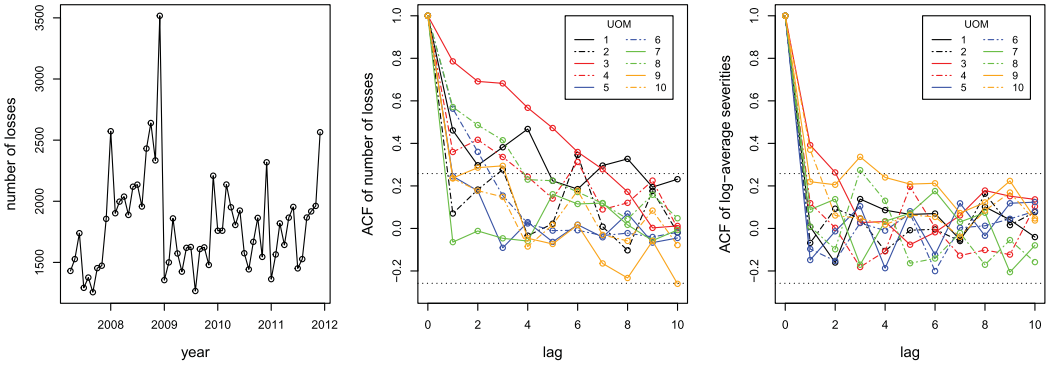

Figure 1. Monthly number of losses aggregated across all UOMs (left), the ACF of number of losses by UOM (middle) and the ACF of log monthly average loss severities (right) during IS period. The ACFs are not significantly different from zero if they fall within the two black dotted horizontal lines.

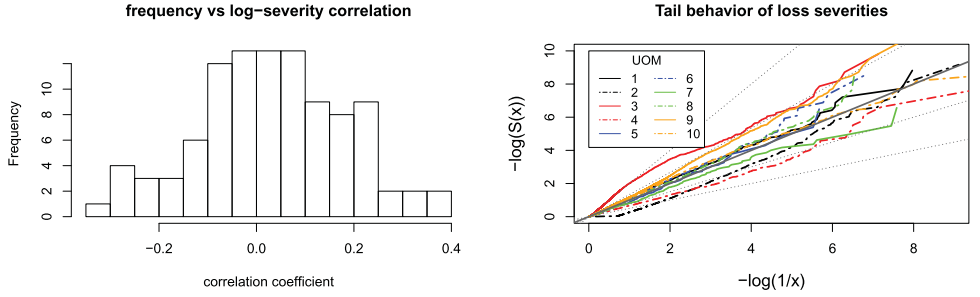

Figure 2. Left: The correlation histogram between frequencies and log-severities. For each *i* and *j* (in 1 to 10), the correlation coefficients between the number of losses of UOM *i* and the log-monthly average severities of UOM *j* is computed. Right: The empirical survival function $S(x)$ vs. $1/x$ (in log scale) for loss severities. The curves are parallel shifted such that they pass through the origin.

Table 2. Correlation characteristics of frequency data (left) and log monthly average severities (right) among UOMs for IS data.

	Correlations between monthly frequencies										Correlations between log monthly average severities									
UOM	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
1		−.10	.49	−.29	.20	.30	−.02	.35	.14	−.15		.06	.04	.01	−.19	−.03	.10	−.06	.01	−.11
2	.45		−.08	.22	−.04	.22	.03	.43	.49	.49	.66		−.03	.06	−.07	.01	.16	−.10	.09	−.01
3	**.00**	.55		−.17	.39	.48	.14	.54	.54	.12	.74	.84		−.07	.09	.19	.07	.03	.54	.27
4	**.03**	.10	.21		.02	−.17	−.02	−.25	.13	.40	.95	.65	.60		−.08	−.13	−.01	−.08	.03	−.05
5	.13	.76	**.00**	.86		.40	.19	.16	.31	.09	.16	.59	.50	.56		.35	−.06	.03	−.02	−.06
6	**.02**	.10	**.00**	.20	**.00**		.05	.73	.40	.19	.85	.95	.16	.35	**.01**		−.20	.22	−.04	.09
7	.91	.84	.31	.90	.16	.72		.04	.01	−.02	.45	.22	.59	.93	.65	.13		−.26	.22	−.19
8	**.01**	**.00**	**.00**	.06	.24	**.00**	.75		.51	.29	.67	.48	.80	.54	.82	.10	.05		−.07	.12
9	.31	**.00**	**.00**	.33	**.02**	**.00**	.91	**.00**		.68	.96	.52	**.00**	.81	.89	.77	.10	.63		.00
10	.25	**.00**	.39	**.00**	.51	.15	.86	**.03**	**.00**		.41	.93	**.04**	.71	.64	.52	.16	.39	.97	

Note: Upper triangle: correlation coefficient; Lower triangle: *p*-value to examine if the the correlation coefficient significantly deviates from zero.

Despite of the scarcity of data points, there exists several heterogeneities across UOMs. Firstly, the total number of losses over the entire IS period differ significantly across UOMs, which can be explained by different risk exposures among UOMs. Explicit measurement of OR exposures, however, are difficult (Reveiz & León 2010). Secondly, the time-series characteristics of loss frequencies varies greatly across UOMs in terms of the trends, clustering behavior (Table 1) and the ACF properties (Middle panel of Figure 1). Thirdly, the correlation structure of loss frequencies among UOMs is complex. Correlation coefficients span widely from −0.29 to 0.68 (Figure 1).

A more general way to examine the dependence structure of a multivariate time-series data is the use of cross-correlation function, which contains ACF and correlation coefficients as special cases. While the detailed results are not presented for conciseness, we find some special patterns among some UOMs. For example, while the correlation coefficient between UOM5 and 8 are insignificant, the cross-correlations are quite significant for certain lags.

The loss severities exhibit very heavy tails. Figure 2 (right) plots $-\log(S(x))$ vs. $1/x$, where $S(x)$ is the empirical survival function and x is the loss severity. The rationale to do so is the hypothesis on the tail behavior that $S(x) \sim c/x^{\alpha}$ for constant $c > 0$ and $\alpha > 0$. Then, we have $-\log(S(x)) = \alpha(-\log(1/x)) + const$. The asymptotic slope of the plots represent α. From the plot, the slopes for some (but not all) UOMs are below the 45 degree line (grey solid line), showing the possibility that the mean of loss severities can be infinite ($\alpha \leq 1$).

5.2. Model descriptions and assumption validations

Due to the complexities of data characteristics, the multivariate Pascal HMM, which is justified as a versatile model, is a suitable candidate model. This model also automatically caters for various

risk exposures among UOMs through allowing different scale parameters θ_{ip} across UOMs. The data characteristics reveal the shortfalls of modeling operational loss frequencies through traditional LDA framework, which assumes monthly loss frequencies are serially independent. The data shows that some UOMs exhibit weak trends on the number of losses, so the time series count data can be slightly non-stationary for some UOMs. Moreover, there are strong positive autocorrelations on loss frequencies for some UOMs. Therefore, predictions will be biased using traditional LDA.

Because of the extreme heavy tail structure demonstrated by Figure 2 (right panel), it is desirable to model the log-transformed severities. Motivated by the theoretical justifications by Willmot & Woo (2007) and Lee & Lin (2010), and the effectiveness of the corresponding fitting algorithm (Verbelen et al. 2015), we model the log-severities through left-truncated version of mixture Erlang distributions. The truncation marks are log(15000) for UOM2 and log(30000) for other UOMs. The density of log-severity for the ith UOM is expressed by

$$f_i(x) = \sum_{j=1}^{\infty} \alpha_{ij} \frac{x^{j-1} e^{-x/\theta_i}}{\theta_i^j (j-1)!}. \tag{12}$$

This model can cater for both finite ($\theta_i < 1$) and infinite ($\theta_i \geq 1$) mean of severities. Although mixture Erlang distributions (without transformations) are dense in the space of positive continuous distribution (Lee & Lin 2010), making it versatile to fit a wide range of distributions, we will show that the log-transformed model is more effective in modeling heavy-tailed data. The denseness property still holds for log-transformed mixture Erlang distributions.

The above modeling framework implicitly makes several assumptions, which are validated on our numerical dataset as follows. Note that all correlations below are computed based on log-transformed severities. Otherwise, the extreme tail heaviness of loss severities can make the estimations of dependence structures inconsistent.

- The loss severities are serially independent: Figure 1 (right panel) shows the ACFs of log monthly average loss severities are very weak and insignificant for most UOMs. Hence, it is reasonable to use a static distribution to model severities.
- The loss severities are independent among UOM: Table 2 (right panel) shows that the correlations of severities among UOMs are relatively small. Out of the 45 correlation values calculated for log-severities, only 3 of them are significantly different from zero, compared to 19 significant values for frequencies. Hence, most of the severities correlation values are likely caused by the randomness of data generation. This provides evidence supporting the usage of univariate distributions to model severities separately.
- Frequencies are independent of severities: Figure 2 (left panel) shows that based on 100 values computed, the correlations between loss frequencies and log-severities have approximately zero mean. Fewer than 10 correlation values exceed the 5% significant threshold (around $\pm$ 0.26), making it acceptable to assume independence between frequencies and severities.

An alternative measure of dependence structures for the assumption tests above is the kendall's tau, which can be robust in measuring the correlations among differently scaled datasets. Under this measure, the results are still aligned with our hypotheses: The correlations among severities and between frequencies and severities are much weaker than that among loss frequencies.

6. Estimation results

Using the proposed EM algorithm, we find that the optimal frequency model consists of five states. Some fitted model parameters are shown as follows:

Fitted average monthly number of losses

	UOM1	UOM2	UOM3	UOM4	UOM5	UOM6	UOM7	UOM8	UOM9	UOM10
State 1	149.78	58.56	127.66	12.89	5.44	39.44	5.67	47.33	596.91	389.76
State 2	152.5	63	335.75	18.13	6.25	40.88	6.5	59	820.5	573.75
State 3	100.05	102.84	122.23	23.09	7.04	48.9	6.55	57.87	767.93	624.76
State 4	174.5	106.5	394.75	15	9.25	97.25	6.25	124.5	1095	707.5
State 5	95.04	84.56	87.97	21.36	3.95	28.96	6.16	37.58	716.83	571.96

Figure 3. Fitted average number of losses represented by $m_{ip}\theta_{ip}, i = 1, \ldots, 5, p = 1, \ldots, 10$. For each UOM, darker color represents greater number of losses.

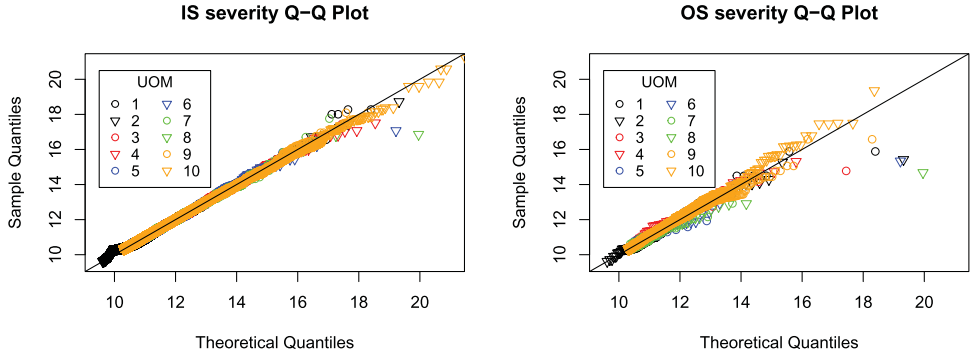

Figure 4. The IS and OS Q–Q plots for log loss severities.

$$\boldsymbol{\delta} = (1 \quad 0 \quad 0 \quad 0 \quad 0), \quad \boldsymbol{\Gamma} = \begin{pmatrix} 0.778 & 0.000 & 0.222 & 0.000 & 0.000 \\ 0.000 & 0.875 & 0.000 & 0.125 & 0.000 \\ 0.000 & 0.066 & 0.725 & 0.000 & 0.208 \\ 0.250 & 0.000 & 0.000 & 0.750 & 0.000 \\ 0.000 & 0.000 & 0.158 & 0.000 & 0.842 \end{pmatrix}$$

The fitted model has an intuitive interpretation that the global environment starts at the first state where losses are expected to occur less frequently (as shown in Figure 3, the colors on the row of 'State 1' are mostly light), and then it gradually transits to the other four states which have different loss characteristics by UOMs. State 4 may represent an 'unfavorable' environment for most UOMs while state 5 is considered as a more 'favorable' one. The large values on the diagonal entries of $\boldsymbol{\Gamma}$ also confirms the clustering behavior of OR events.

Moreover, Figure 3 gives some insights on the dependence structures of the fitted model. Considering UOM1 and UOM4, in the states where UOM1 has a greater average number of losses, fewer losses are expected to occur in UOM4, and vice versa. The loss frequencies of UOM1 and UOM4 under the fitted model look negatively correlated, matching the empirical correlation coefficient of -0.29. Similar observations and analyses among other UOMs will help us further understand how the dependence structure of the fitted model synchronizes those of empirical data.

For the fitting of the log-severity distributions using the AIC, the number of mixture components varies from 1 to 16 and the scale parameter varies from 0.03 to 1.16 among UOMs. The Q–Q plots (Figure 4) show that the fitting is adequate. On the other hand, if the severity is not log-transformed, much more mixture components (from 11 to 54) are required to fit the data.

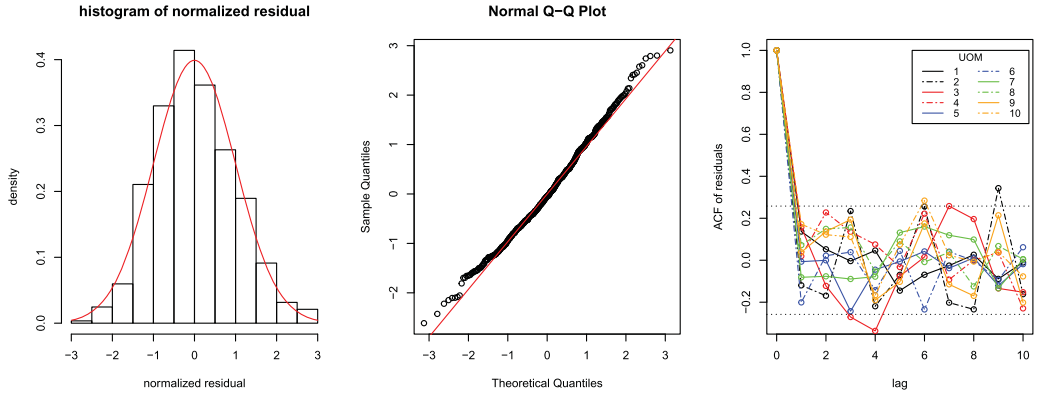

Figure 5. Ordinary pseudo residuals for the fitted model.

6.1. In-sample validation tests

This section performs in-sample tests to evaluate the adequacy of the model fitting. Generally, we are testing the null hypothesis (H_0) that the empirical data is generated from the fitted model against the alternative hypothesis that H_0 is false. The first test is 'ordinary pseudo-residual' by Zucchini & MacDonald (2009). The residuals are given by $z_{lp} = \Phi^{-1}(P(N_{lp} \leq n_{lp} \mid \boldsymbol{N}^{(-l)} = \boldsymbol{n}^{(-l)}))$ for $l = 1, \ldots, L$ and $p = 1, \ldots, P$, where $\Phi(\cdot)$ is the standard normal cdf and $\boldsymbol{N}^{(-l)}$ and $\boldsymbol{n}^{(-l)}$ contain all information other than that at time l. Under H_0, each z_{lp} should be approximately standard normal distributed and $\{z_{lp}\}_{l=1,\ldots,L}$ are serially uncorrelated for all $p = 1, \ldots, P$.

Figure 5 compares the residuals to the standard normal distribution and examines the ACFs of pseudo residuals for each UOM. Apart from a good fit based on Q–Q plot, the residual autocorrelations are generally small, with only 5 out of 100 points being slightly out of the 5% significance level. Therefore, there is evidence that model fitting is adequate.

The second test examines how well the fitted model captures the dependence structures among UOMs. The number of losses across time and UOMs are simulated from the fitted model for 10,000 times. For each simulation, the correlation coefficient matrix across UOMs is computed. Aggregating the results for all simulations, we obtain a distribution of correlations for each matrix element. The results are presented through box plots in Figure 6 and compared to the empirical correlation matrix. All empirical correlations fall within the 95% confidence intervals generated by simulations from the fitted model, except for that between UOM9 and UOM10, where the empirical correlation is relatively very high (0.682) and the fitted model seems under-capturing the empirical correlation. Note that such high correlation can also be partially caused by the sampling error of empirical data, so it is normal that the correlations are under-captured for a few UOMs.

The third test applies the similar methodology as the second test and analyzes the goodness-of-fit of our proposed model through its ability to capture data's ACFs. Out of the 100 empirical ACF values (lags 1-10 with UOM1-10), only 7 of them are slightly off the 95% confidence intervals. For conciseness, only the result for a UOM is shown in Figure 7 (left panel). It is then concluded that the fitted model decently caters for serial correlation structures of empirical data.

All the tests above suggest that the proposed multivariate Pascal-HMM fits the data very well, so the remaining problem is that such flexible model may overfit the data, because the OR frequency data are scarce. We compare the marginal distributions of the monthly fitted number of losses with the empirical data. In Figure 7 (right panel), the fitted distributions are smooth instead of matching tightly the empirical distributions, which are rather peaky due to scarcity of data points. We further evaluate the existence of overfitting of the proposed model by fitting it with a large number of states

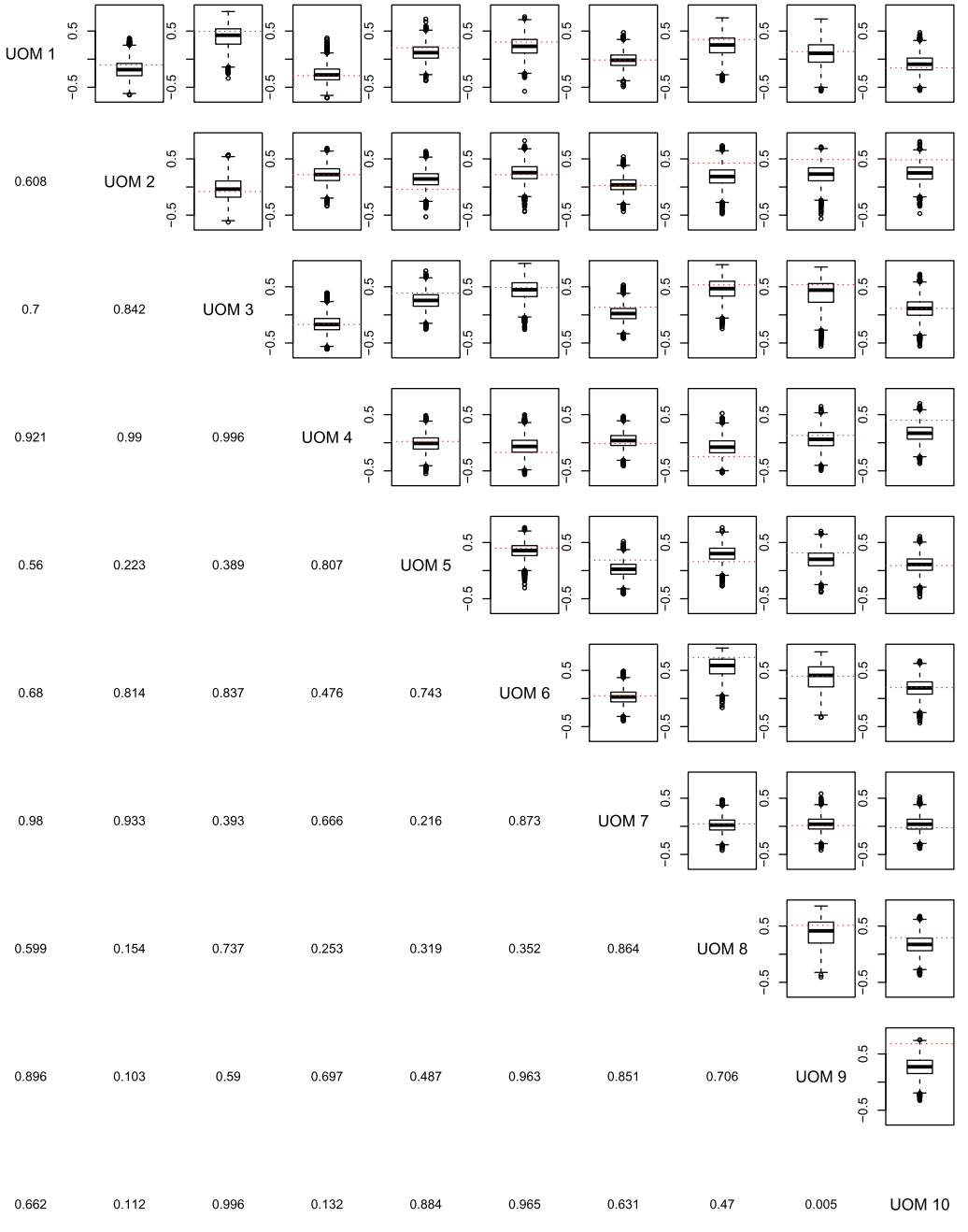

Figure 6. The simulated (from the fitted model) vs empirical correlation coefficient matrix. Upper triangle shows the box plots of simulated correlations and the dotted line represents the empirical correlations; Lower triangle shows the 2-sided *p*-values.

such that the number of parameters is greater than the number of observable features. The resulting fitted distributions, are surprisingly smooth. Therefore, overfitting problem does not exist. This phenomenon can be explained by the over-disperse nature of Pascal-HMM. Equation (6) shows that the marginal distribution of loss frequency for any month is a Pascal mixture, which guarantees a greater-than-one dispersion ratio. This makes the fitting smooth regardless of the number of mixture components and eliminates the overfitting problem.

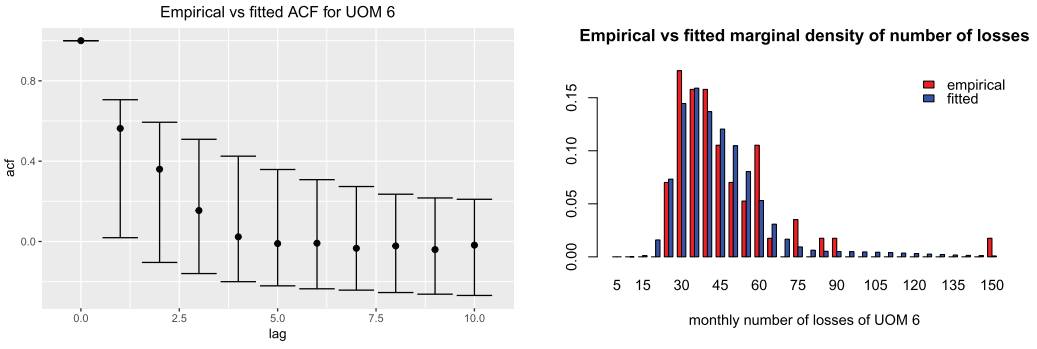

Figure 7. Left: Empirical ACF (dots) vs simulated ACF from the fitted model. The 95% CIs are displayed as bars. Right: Empirical vs. fitted marginal distributions.

6.2. Out-of-sample model prediction

This section aims to examine the predictive ability of the fitted model through an OS testing. The 1-month (for January 2012) and 3-month (for January to March 2012) predictive distribution of the number of losses aggregated in UOM and TOH levels are respectively computed and compared to observed OS data. Given the fitted model parameters, applying Equation (8) and (9), the l^*-month ($l^* = 1, 2, \ldots$) predictive distributions can be simulated by the following steps:

Step 1: Calculate the posterior probabilities for each state at the first month of the OS period (i.e. January 2012) $\boldsymbol{\delta}^*$.

Step 2: For each $l = 1, \ldots, l^*$, simulate hidden states $\hat{C}_{L+l}$ governed by the probabilities $P(C_{L+l} \mid \boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}) = \boldsymbol{\delta}^* \boldsymbol{\Gamma}^{l-1}$.

Step 3: For each $l = 1, \ldots, l^*$, simulate the number of losses $\hat{N}_{L+l,p}$ for each UOM-p ($p = 1, \ldots, 10$) governed by the probabilities $P(\hat{N}_{L+l,p} = n \mid \hat{C}_{L+l} = i) = p(n; m_{ip}, \theta_{ip})$.

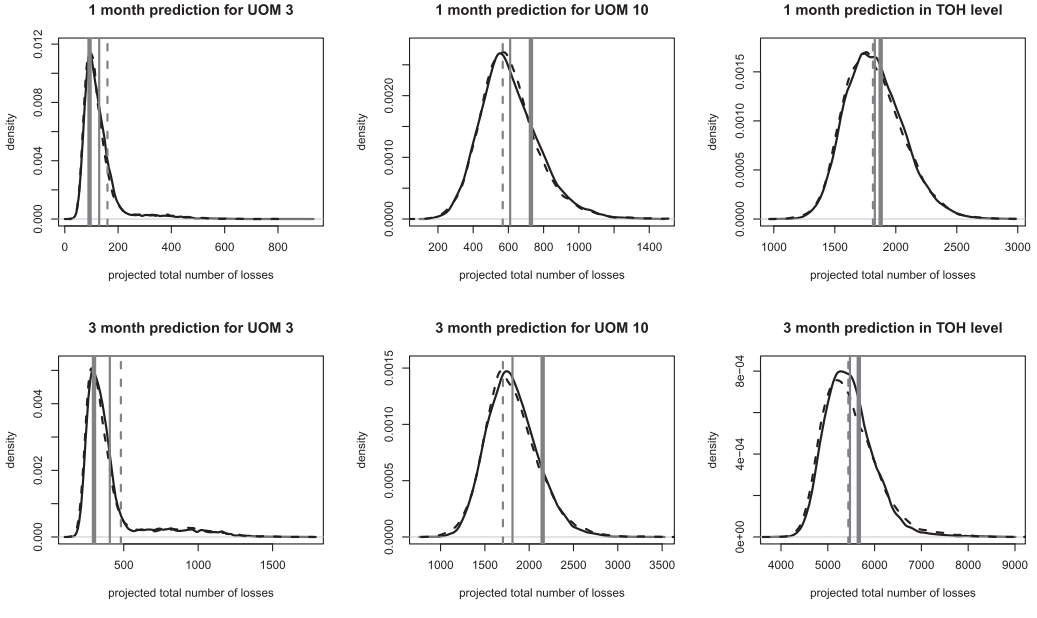

Figure 8. Predictions on loss frequencies. Solid curve and dotted curve are respectively the predictive distributions without and with parameter uncertainties. For vertical lines, thick solid, thin solid and dotted lines are respectively the true observed number, the prediction mean of the proposed model (without parameter uncertainties) and the LDA prediction.

Table 3. Prediction mean and variance at TOH level under various approaches.

	1 month prediction			3 month prediction		
	w/o PU	with PU	LDA	w/o PU	with PU	LDA
mean	1828	1822	1813	5474	5477	5438
variance	55,205	59,520	168,622	278,705	352,435	505,866

Notes: Both 'w/o PU' and 'with PU' apply our proposed model. PU stands for parameter uncertainties.

Step 4: The l^*-month prediction on the number of losses for UOM-p is given by $\sum_{l=1}^{l^*} \hat{N}_{L+l,p}$, while in TOH level it can be computed as $\sum_{p=1}^{10} \sum_{l=1}^{l^*} \hat{N}_{L+l,p}$.

After one million simulations, the predictive distributions are obtained. The above predictions, however, do not take parameter uncertainties into account. Because of the scarcity of frequency data, parameter uncertainties may be significant. To estimate its effect, we also apply parametric bootstrap for predictions. The procedures is as follows, first, simulate $B = 100$ paths of loss frequencies $\hat{\boldsymbol{N}}_l$ from the fitted model parameters $\hat{\Psi}$. Second, for each path generated, refit the model to obtain parameters $\tilde{\Psi}^b$, $b = 1, \ldots, B$. Third, for each b, simulate the predictive distribution using empirical frequencies $\hat{\boldsymbol{N}}^{(L)}$ and the refitted model parameters $\tilde{\Psi}^b$. Aggregate the results across b.

The prediction results are displayed concisely in Figure 8. We will show that our proposed stochastic model provides better predictions than any (unbiased) distributional models under traditional LDA framework, which ignore serial dependence structures. A decent distributional model should approximately match the first and second moments of empirical data in both UOM and TOH level. Under this assumption, the l-month prediction mean and variance of the number of losses (for both UOM and TOH level) are l times as the mean and variance of the in-sample empirical data, respectively. Overall, the observed values (thick solid lines in Figure 8) are more likely closer to the prediction mean under our proposed model (thin solid lines) than the in-sample mean (dotted lines, also the prediction mean for LDA). This phenomenon can be explained intuitively. The predictions by our time-series model take account for the fact that recent losses are more relevant to future losses, while this is ignored using traditional LDA. For example, under UOM3, there is a huge-loss-frequency cluster at the early stage of in-sample period, which should be less relevant to out-sample predictions. Such a cluster causes LDA to overpredict the future number of losses. In TOH level, both our proposed model and LDA provides adequate predictions, but in traditional LDA approach, it can also be caused by cancelation of biases among UOMs. To further compare the predictions performance between the proposed model and the traditional LDA, it is also useful to compare the variances of prediction distributions. Table 3 shows that the prediction variances using our proposed model (whether or not parameter uncertainties are considered) are significantly smaller than that using traditional LDA. The variance reduction effects are greater for shorter term (1-month) predictions. Moreover, this table shows that parameter uncertainties plays a significant, but not a dominant role in explaining overall variances.

To generate the 1-month and 3-month predictive distributions for aggregate loss amounts in UOM and TOH levels, we may independently simulate the loss severities for each loss based on the UOM to which the loss belongs, from the fitted left-truncated mixture Erlang distributions. Aggregating the losses by time intervals and/ or UOMs, we will obtain the desired predictive distributions. The prediction results are summarized in Table 4, which shows that the realized losses across UOMs are mostly within the 95% CI of our predictions as shown by $p > 0.05$ for most UOMs. Our prediction procedures allow easy computations of the 99.95-percentiles (also called the 99.95% value at risk (VaR)) of aggregate losses, which may represent the Economic Capital (EC) charged to a financial institution. The prediction distributions in TOH level (in log scale) are also displayed in Figure 9. It can be seen that the effect of parameter uncertainties are insignificant when aggregate losses are considered. Figure 9 (left panel) shows the one month prediction results by using Erlang mixture distribution to model severities without log-transformation. Some small peaks are observed on the

Table 4. Summary of the predictions of the aggregate losses.

	1 month prediction						3 month prediction					
		Fitted						Fitted				
	Empirical realized	5%	50%	95%	99.95%	*p*-value	Empirical realized	5%	50%	95%	99.95%	*p*-value
UOM1	0.22	0.08	0.16	0.49	2.25	0.58	0.47	0.32	0.56	1.19	3.70	0.64
UOM2	0.12	0.09	0.20	0.60	2.58	0.33	0.41	0.37	0.64	1.39	3.93	0.17
UOM3	0.07	0.04	0.09	0.28	0.80	0.53	0.24	0.17	0.29	0.84	1.62	0.60
UOM4	0.11	0.02	0.10	0.61	14.33	0.91	0.27	0.15	0.38	1.62	28.95	0.58
UOM5	0.00	0.00	0.01	0.04	0.83	0.52	0.01	0.01	0.03	0.09	1.63	0.23
UOM6	0.07	0.02	0.06	0.16	2.81	0.58	0.10	0.10	0.18	0.42	6.21	0.16
UOM7	0.01	0.00	0.01	0.10	18.95	0.64	0.02	0.01	0.04	0.35	61.33	0.48
UOM8	0.01	0.03	0.08	0.25	1.38	0.01	0.09	0.13	0.26	0.59	2.17	0.01
UOM9	1.16	0.65	0.92	1.42	3.52	0.34	2.91	2.27	2.83	3.75	6.68	0.83
UOM10	1.90	0.59	1.36	11.10	87.69	0.60	11.00	2.65	4.90	26.51	125.17	0.41
TOH	3.68	2.24	3.34	13.40	100.81	0.77	15.55	7.91	10.97	33.09	160.21	0.49
TOH*		2.22	3.31	13.31	99.63	0.75		7.83	10.94	32.99	163.79	0.49

Notes: All the amounts are in 100 million. The percentages are the percentiles. 'TOH*' considers parameter uncertainties.

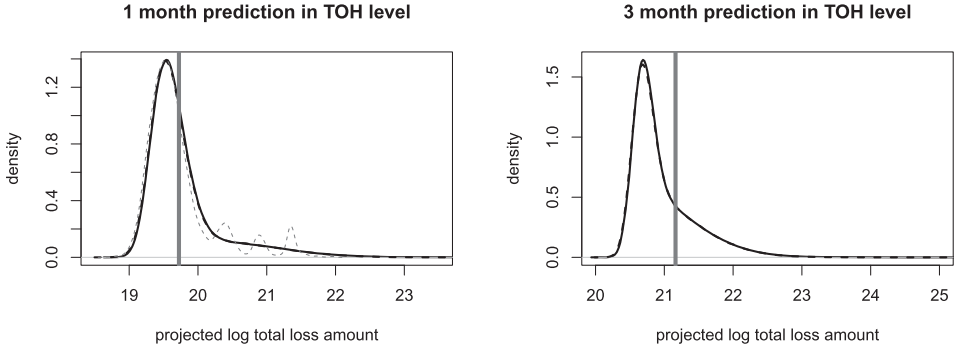

Figure 9. 1-month (left) and 3-month (right) predictive distributions on the aggregate loss amount (in log scale) in TOH level. Thick vertical line: the observed log aggregate loss; thin dotted grey curve in the left panel: the result using Erlang mixture to fit severities without log-transformation.

tail, indicating that the severities model overfits the tail. Based on the results, we conclude that our proposed model is adequate to predict short term future losses.

7. Concluding remarks

In this paper, we propose a multivariate Cox model with the underlying intensity vector following multivariate Erlang-HMM, which serves as a dynamical model for multivariate count processes. This work is motivated by the complexities and heterogeneities of our OR data. We first provide a natural interpretation of the proposed model via latent global environmental states. Various important properties, such as thinning, distributional and identifiability are investigated in the paper. The proposed frequency model is fitted to the OR dataset though the proposed EM algorithm. We have evaluated various goodness-of-fit tests, which unanimously suggest that the proposed model can cater for several heterogeneities implied by the dataset. Also, the aggregate loss can also be modeled through fitting loss severities by truncated log Erlang mixture distributions. Performing out-of-sample tests, we conclude that the proposed model can adequately predict short-term future losses, which is crucial for a financial institution to set up adequate reserve capitals.

Since the proposed model is flexible to capture a broad range of behavior of multivariate count data, it can be further applied to various actuarial areas other than OR. For example, it can potentially

be used to solve a multidimensional stochastic claim reserving problem with dependent business lines. In the insurance context, however, the claim frequencies and severities may become correlated. Therefore, in our future work, we plan to extend the current model to cater for such dependence structure, for example by introducing a state-dependent severity distribution in our HMM.

Acknowledgments

The authors would like to thank two anonymous referees for their comments that greatly improve the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada [RGPIN 284246,RGPIN-2017-06684].

References

Al-Osh M. A. & Aly E.-E. A. (1992). First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics-Theory and Methods* **21**(9), 2483–2492.

Ames M., Schuermann T. & Scott H. S. (2015). Bank capital for operational risk: a tale of fragility and instability. *Journal of Risk Management in Financial Institutions* **8**(3), 227–243.

Aue F. & Kalkbrener M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk. *Journal of Operational Risk* **1**(4), 49–93.

Badescu A. L., Chen T., Lin X. S. & Tang D. (2019). A marked Cox model for the number of IBNR claims: estimation and application. *ASTIN Bulletin, The Journal of the IAA*, in press.

Badescu A. L., Gong L., Lin X. S. & Tang D. (2015). Modeling correlated frequencies with application in operational risk management. *Journal of Operational Risk* **10**(1), 1–43.

Badescu A. L., Lin X. S. & Tang D. (2016). A marked Cox model for the number of IBNR claims: theory. *Insurance: Mathematics and Economics* **69**, 29–37.

Bardoscia M. & Bellotti R. (2011). A dynamical approach to operational risk measurement. *Journal of Operational Risk* **6**(1), 3–19.

BCBS (2004). International convergence of capital measurement and capital standards: a revised framework.

Buch-Kromann T., Englund M., Gustafsson J., Perch Nielsen J. & Thuring F. (2007). Non-parametric estimation of operational risk losses adjusted for under-reporting. *Scandinavian Actuarial Journal* **2007**(4), 293–304.

Cappé O., Moulines E. & Rydén T. (2005). *Inference in hidden Markov models*. New York: Springer-Verlag.

Chernobai A. S., Rachev S. T. & Fabozzi F. J. (2008). *Operational risk: a guide to basel II capital requirements, models and analysis*. Hoboken, NJ: John Wiley & Sons.

Daley D. J. & Vere-Jones D. (1988). *An introduction to the theory of point processes*. New York: Springer-Verlag.

Frachot A., Georges P. & Roncalli T. (2001). Loss distribution approach for operational risk. Available at SSRN.

Gara Z. & Belkacem L. (2018). Modeling catastrophic operational risk using a compound Neyman–Scott clustering model. *Journal of Operational Risk* **13**(1), 51–75.

Guegan D. & Hassani B. (2018). Using a time series approach to correct serial correlation in operational risk capital calculation. *Journal of Operational Risk* **8**(3), 31–56.

Gui W., Huang R. & Lin X. S. (2018). Fitting the Erlang mixture model to data via a GEM-CMM algorithm. *Journal of Computational and Applied Mathematics* **343**, 189–205.

Karr A. (1991). *Point processes and their statistical inference*, Vol. **7**. New York: CRC press.

Kelliher P., Acharyya M., Couper A., Grant K., Maguire E., Nicholas P., Smerald C., Stevenson D., Thirlwell J. & Cantle N. (2017). Good practice guide to setting inputs for operational risk models. *British Actuarial Journal* **22**(1), 68–108.

Lee S. C. K. & Lin X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal* **14**(1), 107–130.

Leone P., Porretta P. & Vellella M. (2017). *Measuring and managing operational risk: an integrated approach*. New York: Springer-Verlag.

Lubbe J. & Snyman F. (2010). The advanced measurement approach for banks. *IFC Bulletin* **33**, 141–149.

Norberg R. (1993). Prediction of outstanding liabilities in non-life insurance I. *ASTIN Bulletin: The Journal of the IAA* **23**(1), 95–115.

Peters G. W., Shevchenko P. V., Young M. & Yip W. (2011). Analytic loss distributional approach models for operational risk from the α-stable doubly stochastic compound processes and implications for capital allocation. *Insurance: Mathematics and Economics* **49**(3), 565–579.
Reveiz A. & León C. (2010). Operational risk management using a fuzzy logic inference system. *Journal of Financial Transformation* **30**, 141–153.
Samad-Khan A., Guharay S., Franklin B., Fischtrom B., Scanlon M. & Shimpi P. (2010). A new approach for managing operational risk: addressing the issues underlying the 2008 global financial crisis. *Society of Actuaries*.
Teicher H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34**, 1265–1269.
Teicher H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics* **38**(4), 1300–1302.
Verbelen R., Gong L., Antonio K., Badescu A. & Lin S. (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin: The Journal of the IAA* **45**(3), 729–758.
Wang Y., Li J. & Zhu X. (2017). A method of estimating operational risk: loss distribution approach with piecewise-defined frequency dependence. *Procedia Computer Science* **122**, 261–268.
Willmot G. E. & Woo J. K. (2007). On the class of Erlang mixtures with risk theoretic applications. *North American Actuarial Journal* **11**(2), 99–115.
Willmot G. E. & Woo J. K. (2015). On some properties of a class of multivariate Erlang mixtures with insurance applications. *ASTIN Bulletin: The Journal of the IAA* **45**(1), 151–173.
Yakowitz S. J. & Spragins J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* **39**(1), 209–214.
Zucchini W. & MacDonald I. L. (2009). *Hidden Markov models for time series: an introduction using R*. New York: CRC Press.

Appendices

Appendix 1. The derivations for Remark 2.1

This appendix section shows that under the conditions described by Remark 2.1, the resulting model can still be converted to our original proposed model. For demonstrative purpose, we only prove the result for the discretized event arrival process described in Section 3.2. The likelihood function (defined in Section 3.3) under this model can be written as the following, see Chapter 2.3.2 of Zucchini & MacDonald (2009) for more details:

$$\begin{aligned}\mathcal{L}_L(\boldsymbol{\Phi}; \boldsymbol{n}_1, \ldots, \boldsymbol{n}_L) = & \sum_{c_1,..,c_L=1}^{g} (\delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_3}\cdots\gamma_{c_{L-1},c_L}) \\ & \times \sum_{k_1,\ldots,k_L=1}^{K} \left(\pi_{k_1c_1}\prod_{p=1}^{P} p(n_{p1}; m_{k_1c_1p}, \theta_{kc_1p})\right)\cdots(\pi_{k_Lc_L}\prod_{p=1}^{P} p(n_{pL}; m_{k_Lc_Lp}, \theta_{kc_Lp})) \\ = & \sum_{(k_1,c_1),\ldots,(k_L,c_L)\in B} (\pi_{k_1c_1}\delta_{c_1})\prod_{p=1}^{P} p(n_{p1}; m_{k_1c_1p}, \theta_{kc_1p}) \\ & \times (\pi_{k_2c_2}\gamma_{c_1c_2})\prod_{p=1}^{P} p(n_{p2}; m_{k_2c_2p}, \theta_{kc_2p})\cdots(\pi_{k_Lc_L}\gamma_{c_{L-1}c_L})\prod_{p=1}^{P} p(n_{pL}; m_{k_Lc_Lp}, \theta_{kc_Lp}),\end{aligned}$$

where $B = \{1, \ldots K\} \times \{1, \ldots g\}$ and $\boldsymbol{\Phi}$ contains all parameters in the model. Note that the above likelihood is equivalent to that of a $g \times K$-state Pascal HMM with $\delta^*_{(k,i)} = \pi_{ki}\delta_i$, $\gamma^*_{(k,i),(k',i')} = \pi_{k'i'}\gamma_{ii'}$ and $p^*_{N_{1l},\ldots,N_{Pl}|(K_l,C_l)=(k,i)}(n_1, \ldots, n_P) = \prod_{p=1}^{P} p(n_p; m_{kip}, \theta_{kip})$, so the result for the discretized event arrival process follows. Note that the result can be easily generalized to the whole (not necessarily discretized) process because a point process can be characterized by finite-dimensional distributions (Theorem 1.12(b) of Karr 1991).

Appendix 2. The proof of Proposition 2.1

Before proving the desired result, we introduce Laplace functional transform (LFT), which is a function $L_N(f)$ defined on a point process $N = \{\boldsymbol{X}_i,\ i = 1, 2, \ldots\}$, given by

$$L_N(f) = E(e^{-\sum_i f(\boldsymbol{X}_i)}),$$

where f is a non-negative function of the point process. One nice property of LFT is that it uniquely define point processes. The proof of Proposition 2.1 is based on the computation of the LFT of the point process N:

$$L_N(f) = E[e^{-\sum_i f(T_i,Z_i)}] = E\left[\prod_{p=1}^{P} e^{-\sum_i f(T_i^p,Z_i^p)}\right] = E\left[\prod_{p=1}^{P} E\left[e^{-\sum_i f(T_i^p,Z_i^p)} \mid \tilde{M}^1, \ldots \tilde{M}^P\right]\right]$$
$$= E\left[\prod_{p=1}^{P} e^{-\int (1-e^{-f(t,p)})\, \mathrm{d}\tilde{M}^p}\right] = E[e^{-\int (1-e^{-f(t,p)})\, \mathrm{d}M}],$$

where the third equality is resulted from the conditional independence among $\tilde{N}^p, p = 1, \ldots, P$ and the forth equality follows by Example 1.15 and 1.16 in Karr (1991). Since $L_N(f)$ has the same LFT as that generated by a Cox Process driven by the random measure M, the result follows.

Appendix 3. The proofs for thinning properties

This appendix section proposes a more general result on the thinning properties of the proposed model, and use the result obtained to prove Theorem 3.1.

Theorem A.1: *Define $\bar{N}$ a marked Cox process on $\mathbb{R}^+ \times \{1, \ldots, P\}$ with a marking space of $\mathbb{R}^+$ with intensity measures $\{\mathbf{\Lambda}_1(t), \ldots, \mathbf{\Lambda}_P(t)\}$ and marks X_i, $i = 1, 2, \ldots$, i.e. $\bar{N} = \sum_i \varepsilon_{(T_i,Z_i,X_i)}$. The marks are independent but position-dependent with density function $p_{X\,|\,t,p}(x)$. Consider the thinning probabilities $\bar{\phi}(t, p, x) = 1\{(t, p, x) \in D\}$, where $D \subseteq \mathbb{R}^+ \times \{1, \ldots, P\} \times \mathbb{R}^+$. Then, the resulting thinned point process $\bar{N}' = \sum_i U_i \varepsilon_{(T_i,Z_i,X_i)}$, where $U_i \in \{0, 1\}$ is a random variable conditionally independent given $\bar{N}$ with $P(U_i = 1 \,|\, \bar{N}) = \bar{\phi}(T_i, Z_i, X_i)$, is still a marked Cox process with multivariate intensity vector $\mathbf{\Lambda}'(t) = \{\mathbf{\Lambda}_1(t)P(X \in D_{t,1} \,|\, t, 1)1\{t \in T_{D,1}\}, \ldots, \mathbf{\Lambda}_P(t)P(X \in D_{t,P} \,|\, t, P)1\{t \in T_{D,P}\}\}$ and independent yet position-dependent marks having density function $(p_{X\,|\,t,p}(x)/P(X \in D_{t,p} \,|\, t, p))1\{x \in D_{t,p}\}$, where $D_{t,p} = \{x \in \mathbb{R}^* : (t, p, x) \in D\}$ and $T_{D,p} = \{t \in \mathbb{R}^+ : \exists x \text{ s.t. } (t, p, x) \in D\}$.*

Proof: Adopting similar techniques as Theorem 3.1 of Badescu et al. (2016), which derives the thinning property under univariate setting through analyzing the LFT of the thinned point process, we compute the LFT of $\bar{N}'$ as

$$L_{\bar{N}'}(f) = L_{\bar{N}}[-\log(1 - \bar{\phi}(t, p, x) + \bar{\phi}(t, p, x)e^{-f(t,p,x)})]$$
$$= E[e^{-\sum_{p=1}^{P} \int_{T_{D,p}} (1-\int_{\mathbb{R}^+} (1-\bar{\phi}(t,p,x)+\bar{\phi}(t,p,x)e^{-f(t,p,x)}) p_{X\,|\,t,p}(x)\, \mathrm{d}x)\mathbf{\Lambda}_p(t)\, \mathrm{d}t}]$$
$$= E[e^{-\sum_{p=1}^{P} \int_{T_{D,p}} (\int_{D_{t,p}} p_{X\,|\,t,p}(x)\, \mathrm{d}x - \int_{D_{t,p}} e^{-f(t,p,x)} p_{X\,|\,t,p}(x)\, \mathrm{d}x)\mathbf{\Lambda}_p(t)\, \mathrm{d}t}]$$
$$= E[e^{-\sum_{p=1}^{P} \int_{\mathbb{R}^+} (1-\int_{\mathbb{R}^+} e^{-f(t,p,x)} (p_{X\,|\,t,p}(x)/P(X \in D_{t,p} \,|\, t,p))1\{x \in D_{t,p}\}\, \mathrm{d}x)\mathbf{\Lambda}_p(t) P(X \in D_{t,p} \,|\, t,p) 1\{t \in T_{D,p}\}\, \mathrm{d}t}].$$

From the above derivation, it can be seen from Examples 1.16 and 1.28 in Karr (1991) that $L_{\bar{N}'}(f)$ is the LFT of a marked Cox process with intensity vector $\mathbf{\Lambda}'(t) = \{\mathbf{\Lambda}_1(t)P(X \in D_{t,1} \,|\, t, 1)1\{t \in T_{D,1}\}, \ldots, \mathbf{\Lambda}_P(t)P(X \in D_{t,P} \,|\, t, P)1\{t \in T_{D,P}\}\}$ and independent yet position-dependent marks having density function $(p_{X\,|\,t,p}(x)/P(X \in D_{t,p} \,|\, t, p))1\{x \in D_{t,p}\}$, so the result follows. ■

For the proof of Theorem 3.1, choosing $D_{t,p} = \{x : x > \psi(p)\}$ and $1\{t \in T_{D,p}\} = 1$ (since $\psi(p) < \infty$), we have $P(X \in D_{t,p} \,|\, t, p) = \bar{F}_{X\,|\,p}(\psi(p))$ and $1\{x \in D_{t,p}\} = 1\{x > \psi(p)\}$. Applying Theorem A.1, $\bar{N}'$ is a marked Cox process with piecewise constant intensities $\Lambda'_{lp} = \bar{F}_{X\,|\,p}(\psi(p))\Lambda_{lp}$ for $p = 1, \ldots, P$ and the density of observed loss amounts $p'_{X\,|\,p}(x) = (p_{X\,|\,p}(x)/\bar{F}_{X\,|\,p}(\psi(p)))1\{x > \psi(p)\}$. Using the scaling properties of Erlang distributed random variables, the desired state-dependent intensity density function can be obtained.

Appendix 4. Aggregated event process across all loss types

This appendix section investigates the properties of the aggregated event arrival process $\{N^a(t),\ t \geq 0\}$ and proves Equation (7) in Section 3.2.

Proposition A.1: *Let $N^a = N(\cdot, \{1, \ldots, P\})$ be a point process on $\mathbb{R}^+$, representing the number of events aggregated across all loss types. Then, N^a is a Cox process on $\mathbb{R}^+$ driven by the random measure $M^a(\mathcal{A}^a) = \int_{\mathcal{A}^a} \sum_{p=1}^{P} \mathbf{\Lambda}_p(t)\, \mathrm{d}t$, where $\mathcal{A}^a \subseteq \mathbb{R}^+$.*

Proof: The LFT of N^a is given by

$$\begin{aligned} L_{N^a}(f) &= E[e^{-\sum_i f(T_i)}] = E[e^{-\sum_i g(T_i,Z_i)}] = L_N(g) \\ &= E[e^{-\int (1-e^{-g(t,p)})\,\mathrm{d}M}] = E[e^{-\int (1-e^{-f(t)})\sum_{p=1}^{P}\mathbf{\Lambda}_p(t)\,\mathrm{d}t}] = E[e^{-\int (1-e^{-f(t)})\,\mathrm{d}M^a}] \end{aligned}$$

where $g(t,p) = f(t)$. The result follows since $L_{N^a}(f)$ is the LFT of a Cox process driven by the random measure M^a. ■

From Proposition A.1, it is concluded that the aggregated event arrival process $\{N^a(t),\ t \geq 0\}$ is a univariate Cox process governed by the intensity $\mathbf{\Lambda}^a(t) = \sum_{p=1}^{P}\mathbf{\Lambda}_p(t)$, which is still a piecewise stochastic process with the random intensity $\Lambda_l^a = \sum_{p=1}^{P}\Lambda_{lp}$ generated by Erlang-HMM described by the previous subsection. If $\theta_{ip} \equiv \theta$ does not depend on the loss type p for any $i = 1,\ldots,g$ and $p = 1,\ldots,P$, it is obvious that $\Lambda_l^a \mid C_l = i$ still follows Erlang distribution with shape parameter $\sum_{p=1}^{P} m_{ip}$ and scale parameter θ. Therefore, the intensity $\mathbf{\Lambda}^a(t)$ is still generated by Erlang-HMM. In this case, we can say that the proposed multivariate Cox processes are 'closed under aggregation'. However, if θ_{ip} is not a constant across loss types, the following proposition will be needed to determine the state conditional distribution of the intensity Λ_l^a. It is shown that the intensity is instead generated by Erlang Mixture-HMM.

Proposition A.2: *The distribution of $\Lambda_l^a \mid C_l = i$ is given by*

$$\begin{aligned} K_{\Lambda_l^a \mid C_l=i}(\lambda) &= \sum_{k_1=m_{i1}}^{\infty}\cdots\sum_{k_P=m_{iP}}^{\infty}\psi_{\boldsymbol{k}}h(\lambda;\sum_{p=1}^{P}k_p,\theta_i) \\ &= \sum_{k=1}^{\infty}\tilde{\psi}_k h(\lambda;k,\theta_i) \end{aligned}$$

where $\theta_i = \min\{\theta_{i1},\theta_{i2},\ldots,\theta_{iP}\}$, $\boldsymbol{k} = (k_1,k_2,\ldots,k_P)$, $\tilde{\psi}_k = \sum_{k_1+\cdots+k_P=k}\psi_{\boldsymbol{k}}\prod_{p=1}^{P}1\{k_p \geq m_{ip}\}$ *and*

$$\psi_{\boldsymbol{k}} = \prod_{p=1}^{P}\binom{k_p-1}{m_{ip}-1}\left(\frac{\theta_i}{\theta_{ip}}\right)^{m_{ip}}\left(1-\frac{\theta_i}{\theta_{ip}}\right)^{k_p-m_{ip}}.$$

Proof: Applying Proposition 1 of Willmot & Woo (2015), we have

$$K_{\Lambda_{l1},\ldots,\Lambda_{lP}\mid C_l=i} = \sum_{k_1=1}^{\infty}\cdots\sum_{k_P=1}^{\infty}p_{\boldsymbol{k}}\prod_{p=1}^{P}h(\lambda_p;k_p,\theta_i)$$

where

$$\begin{aligned} p_{\boldsymbol{k}} &= \sum_{b_1=1}^{k_1}\cdots\sum_{b_P=1}^{k_P}\prod_{p=1}^{P}1\{b_p = m_{ip}\}\prod_{p=1}^{P}\binom{k_p-1}{b_p-1}\left(\frac{\theta_i}{\theta_{ip}}\right)^{b_p}\left(1-\frac{\theta_i}{\theta_{ip}}\right)^{k_p-b_p} \\ &= \left(\prod_{p=1}^{P}1\{k_p \geq m_{ip}\}\right)\psi_{\boldsymbol{k}}. \end{aligned}$$

Therefore, the joint density of $\Lambda_{l1},\ldots,\Lambda_{lP}$ conditional on $C_l = i$ can be also expressed by

$$K_{\Lambda_{l1},\ldots,\Lambda_{lP}\mid C_l=i} = \sum_{k_1=m_{i1}}^{\infty}\cdots\sum_{k_P=m_{iP}}^{\infty}\psi_{\boldsymbol{k}}\prod_{p=1}^{P}h(\lambda_p;k_p,\theta_i).$$

Since the scale parameter θ_i in the expression above does not depend on the loss type p, the result follows by the addition property of Erlang distributed random variables. ■

Then, Equation (7) is just a direct consequence from Proposition A.2 and the law of total probability.

Appendix 5. The proof of model identifiability

Theorem 3.2 can be proved in the following 3 steps, under the condition that $\xi_1,\ldots,\xi_g$ are distinct:

First, it can be proved that the class of finite mixture of univariate Pascal is identifiable. We follow Teicher (1963) as the definition of identifiability for finite mixtures. Applying Theorem 2 of Teicher (1963) and similar procedures as

Badescu et al. (2015) (but not exactly the same since it is assumed that the scale parameter θ_{ip} depends on state and loss type, instead of being a universal one), it suffices to show that:

- There exists a transform $\phi_\xi(t)$ defined for $t \in \mathcal{S}_{\phi_\xi}$ such that the mapping $M : F_\xi \to \phi_\xi$ is linear and one-to-one. In this case, $\xi = (m, \theta)$ and the PGF is taken as the transform, so that $\phi_\xi(t) = (1 + \theta - \theta t)^{-m}$ and $\mathcal{S}_{\phi_\xi} = (0, (1+\theta)/\theta)$.
- There exists a total ordering of $\mathcal{F}$ such that $F_{\xi_1} \prec F_{\xi_2}$ implies: (i) $\mathcal{S}_{\phi_{\xi_1}} \subseteq \mathcal{S}_{\phi_{\xi_2}}$; (ii) There exists $t^* \in \bar{\mathcal{S}}_{\phi_{\xi_1}}$ (t^* being independent of ϕ_{ξ_2}) such that $\lim_{t\to t^*} \phi_{\xi_2}(t)/\phi_{\xi_1}(t) = 0$. To demonstrate this, we order the Pascal distribution by $p(n; m_1, \theta_1) \prec p(n; m_2, \theta_2)$ when $(\theta_1 < \theta_2)$ or $(\theta_1 = \theta_2$ and $m_1 > m_2)$. If $\theta_1 = \theta_2 = \theta$ and $m_1 > m_2$, choose $t^* = (1+\theta)/\theta \in \bar{\mathcal{S}}_{\phi_{\xi_1}}$ and hence

$$\lim_{t\to t^*} \frac{\phi_{\xi_2}}{\phi_{\xi_1}} = \lim_{t\to t^*} (1 + \theta - \theta t)^{m_1 - m_2} = 0.$$

  If $\theta_1 < \theta_2$, choose $t^* = (1+\theta_1)/\theta_1 \in \bar{\mathcal{S}}_{\phi_{\xi_1}}$ and hence

$$\lim_{t\to t^*} \frac{\phi_{\xi_2}}{\phi_{\xi_1}} = \lim_{t\to t^*} \frac{(1 + \theta_1 - \theta_1 t)^{m_1}}{(1 + \theta_2 - \theta_2 t)^{m_2}} = 0.$$

Second, it can be proved that the class of finite mixture of P-fold-product of Pascal distribution (i.e.: independent multivariate Pascal) is identifiable. It is a direct consequence using the identifiability result of univariate case, and applying Theorem 2 of Teicher (1967).

Third, it can be proved that the class of multivariate Pascal HMM is identifiable. Some related arguments for Normal HMM are discussed in Section 12.4.4 of Cappé et al. (2005). From Equation (6), we have

$$P(\boldsymbol{N}_l = \boldsymbol{n}_l; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g) = \sum_{i=1}^{g} \pi_{li} \prod_{p=1}^{P} p(n_p; m_{ip}, \theta_{ip}),$$

which is still a multivariate Pascal mixture. Because of its identifiability, $P(\boldsymbol{N}_l = \boldsymbol{n}_l; \boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\xi}^*, g^*) = P(\boldsymbol{N}_l = \boldsymbol{n}_l; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$ implies $\xi_i^* = \xi_{c(i)}$ for all $i = 1, \ldots, g$ if $\pi_{li}^* > 0$, where $c : \{1, \ldots, g^*\} \mapsto \{1, \ldots, g\}$ and $c(1), \ldots, c(g^*)$ are distinct. Denote $B_l^* = \{i : \pi_{li}^* > 0\}$. Since the above result holds for any $l = 1, 2, \ldots$ and $\bigcup_{l=1}^{\infty} B_l^* = \{1, \ldots, g^*\}$ (irreducibility condition), we have $g^* \le g$ and $\xi_i^* = \xi_{c(i)}$ for $i = 1, \ldots, g^*$. Same arguments also yield $g \le g^*$ and $\xi_i = \xi_{c^*(i)}^*$ for $i = 1, \ldots, g$. Therefore, we have $g^* = g$ and $\xi_i^* = \xi_{c(i)}$ for $i = 1, \ldots, g$, where $c(\cdot)$ is a permutation of $\{1, \ldots, g\}$.

Consider $L = 1$, Equation (6) shows that

$$P(\boldsymbol{N}_1 = \boldsymbol{n}_1; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g) = \sum_{i=1}^{g} \delta(i) \prod_{p=1}^{P} p(n_{1p}; m_{ip}, \theta_{ip}),$$

which is a finite mixture of P-variate Pascal distribution. Using the identifiability result of finite mixture and knowing that $\xi_i^* = \xi_{c(i)}$ for $i = 1, \ldots, g$, we have $P(\boldsymbol{N}_1 = \boldsymbol{n}_1; \boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\xi}^*, g^*) = P(\boldsymbol{N}_1 = \boldsymbol{n}_1; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$ implies $\delta^*(i) = \delta(c(i))$ for $i = 1, \ldots, g$.

Consider a general $L \ge 2$, we have the equation similar to Equations (5) and (6)

$$P(\boldsymbol{N}^{(L)} = \boldsymbol{n}^{(L)}; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g) = \sum_{\boldsymbol{i} \in \{1,\ldots,g\}^L} \delta(i_1) \prod_{l=2}^{L} \gamma_{i_{l-1} i_l} \prod_{l=1}^{L} \prod_{p=1}^{P} p(n_{lp}; m_{i_l p}, \theta_{i_l p}),$$

where $\boldsymbol{i} = \{i_1, \ldots, i_L\}$. It is a finite mixture of LP-variate Pascal distribution and is still identifiable. Consider $L = 2$, we have $P(\boldsymbol{N}^{(2)} = \boldsymbol{n}^{(2)}; \boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\xi}^*, g^*) = P(\boldsymbol{N}^{(2)} = \boldsymbol{n}^{(2)}; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$ implies $\delta^*(i_1)\gamma_{i_1 i_2}^* = \delta(c(i_1))\gamma_{c(i_1)c(i_2)}$. Therefore $\gamma_{i_1 i_2}^* = \gamma_{c(i_1)c(i_2)}$ for all $i_2 = 1, \ldots, g$ if $\delta^*(i_1) \ne 0$, i.e. for $i_1 \in B_1^*$. Consider $L = 3$, $P(\boldsymbol{N}^{(3)} = \boldsymbol{n}^{(3)}; \boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\xi}^*, g^*) = P(\boldsymbol{N}^{(3)} = \boldsymbol{n}^{(3)}; \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\xi}, g)$ implies $\delta^*(i_1)\gamma_{i_1 i_2}^* \gamma_{i_2 i_3}^* = \delta(c(i_1))\gamma_{c(i_1)c(i_2)}\gamma_{c(i_2)c(i_3)}$. Therefore, $\gamma_{i_2 i_3}^* = \gamma_{c(i_2)c(i_3)}$ for all $i_2 = 1, \ldots, g$ if there exists i_1 such that $\delta^*(i_1)\gamma_{i_1 i_2}^* \ne 0$, i.e. for $i_2 \in B_2^*$. Similar arguments hold for $L > 3$ by induction. Therefore, $\gamma_{ij}^* = \gamma_{c(i)c(j)}$ for all $i \in \bigcup_{l=1}^{\infty} B_l^* = \{1, \ldots, g\}$ and $j = 1, \ldots, g$.

Appendix 6. A simulation study

While the operational risk data we studied contains 10 UOMs, it is not unusual that a financial institution consists of more than 10 UOMs (see e.g. Peters et al. 2011 mentions that an OR model should be checked with 8 business lines by 7 event types). To this end, we perform a simulation study to access the performance of the proposed frequency model in fitting high-dimensional dataset. Note that the simulation study below is performed several times. Obtaining similar results, we only present one of the replications for conciseness purpose.

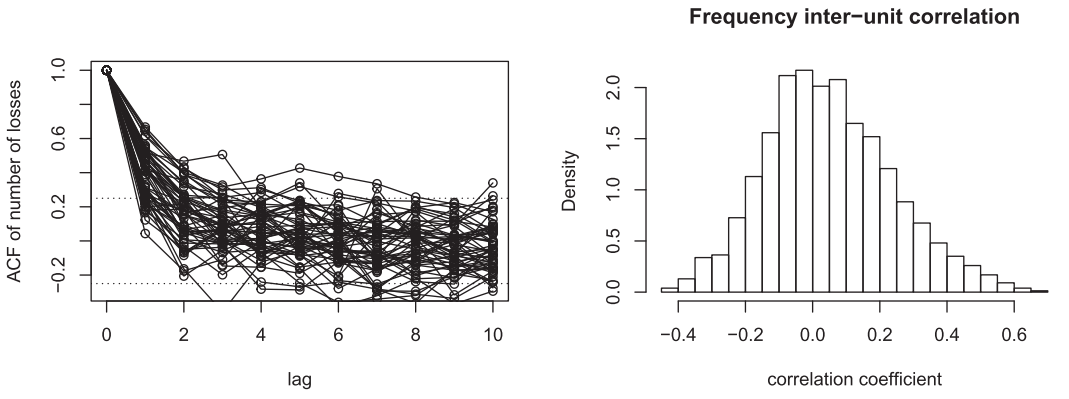

Figure A1. Left: The ACF of number of losses by UOM for the simulated data. Right: The inter-UOM correlation histogram of frequencies. For each $p \neq p'$, the correlation coefficient between the loss frequencies of UOM p and p' is computed.

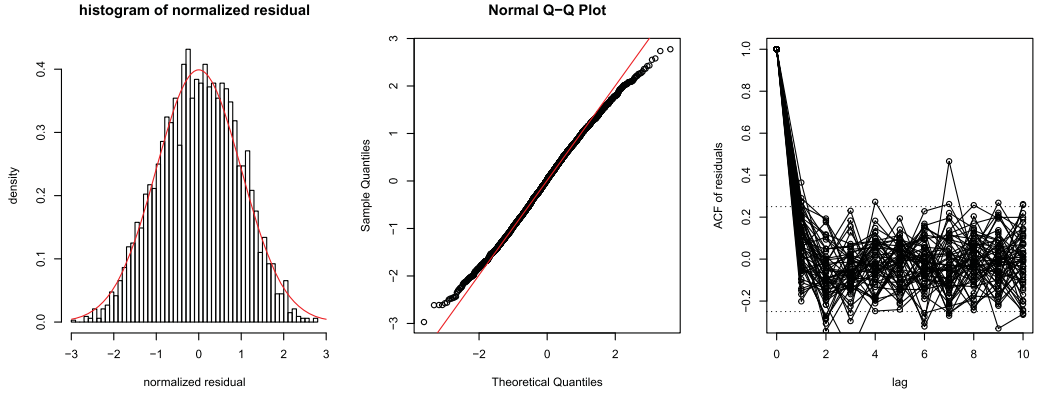

Figure A2. Ordinary pseudo residuals for the fitted model from the simulated data.

In this study, we generate the marginal frequencies $\boldsymbol{N}_p := (N_{1p}, \ldots, N_{Lp})$ with $p \in \{1, \ldots, P\}$ from a negative binomial integer-valued first order autoregressive model (NB-INAR(1)) (Al-Osh & Aly 1992). We first define the probability function of a random variable $N \sim NB(r, \beta)$:

$$P(N = n) = \binom{n + r + 1}{n} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^n, \quad n = 0, 1, \ldots.$$

Under the NB-INAR(1), $N_{1p} \sim NB(\nu_p, 1/\alpha(1-\gamma_p))$ and for $l = 2, 3, \ldots, L$,

$$N_{lp} = \alpha \circ N_{(l-1)p} + \epsilon_{lp}, \tag{A1}$$

where $\alpha \circ N_{(l-1)p} \mid B(N_{(l-1)p}) \sim NB(B(N_{(l-1)p}), 1/\alpha)$ and $B(N_{(l-1)p}) \mid N_{(l-1)p}$ independently follows binomial distribution with size $N_{(l-1)p}$ and probability $\alpha\gamma_p$. Under this setting, $N_{lp} \sim NB(\nu_p, 1/\alpha(1-\gamma_p))$ for all $l = 1, \ldots, L$ and $p = 1, \ldots, P$. Also, $\boldsymbol{N}_p$ has a geometrically decaying ACF $\rho(k) = \gamma_p^k$.

The simulation model is well motivated, not only because NB distribution is a widely adopted OR frequency model under the LDA framework, but also because it caters for serial correlations of loss frequencies that are implied by the real OR dataset (see the middle panel of Figure 1). For the parameters, we choose a fixed $\alpha = 0.5$ but different (random) γ_p and ν_p across UOMs to allow for a variety of serial correlation structures and expected number of losses among different UOMs.

The dependence of loss frequencies among UOMs is modeled by a Gaussian copula, which is a widely used dependence model for OR (Frachot et al. 2001). The copula function is

$$C(u_1, \ldots, u_P) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_P)) \tag{A2}$$

where Φ is a standard normal cdf, Φ_Σ is a joint multivariate normal cdf with mean $\mathbf{0}$, covariance matrix Σ with $\text{diag}(\Sigma) = 1$. For $p \neq p'$, we choose different (random) $\Sigma_{pp'} > 0$ when UOM p and p' belong to the same business line or event type, and $\Sigma_{pp'} = 0$ otherwise.

Overall, the dataset is simulated as follows. Firstly, simulate L independent P-variate normal vectors from the multivariate normal distribution and apply the normal cdf to transform them into Uniform$[0, 1]$ random variables $\{(\hat{u}_{l1}, \ldots, \hat{u}_{lP});\ l = 1, \ldots, L\}$. Secondly, for $l = 1, \ldots, L$ and $p = 1, \ldots, P$, simulate $\hat{n}_{lp}$ by inverting the cdf $\hat{n}_{lp} = F^{-1}(\hat{u}_{lp} \mid N_{(l-1)p} = \hat{n}_{(l-1)p})$, where F is the cdf of $N_{lp} \mid N_{(l-1)p}$ that can be computed analytically or by simulation.

Choosing $L = 60$ and $P = 56$, the simulated loss frequencies are both serially and inter-UOM correlated (Figure A1). Also, the average number of losses spans widely from 3.25 to 236.98 across UOMs. These align with the overall structure of the real OR dataset.

The fitted model contains nine states. To analyze the overall fitting quality, we again perform the ordinary pseudo-residual test similar to that performed to the real OR dataset. From Figure A2, the residuals are normal-like distributed with autocorrelations mostly within the 95% confidence interval. Hence, the fitted model is adequate in capturing the distribution and serial-correlation structures of the simulated dataset. Further, we perform an inter-unit correlation test similar to that in Figure 6. Under the $56 \times 55/2 = 1,540$ UOM combinations, only 118 (7.66%) and 68 (4.42%) of the empirical correlations fall beyond the 90% and 95% confidence intervals respectively generated by simulations from the fitted model. As a result, the model also well fits the inter-UOM correlation structures.