Joint estimation of insurance loss development factors using Bayesian hidden Markov models

Conor Goold[1]

[1]Ledger Investing, Inc.

June 28, 2024

Loss development modelling is the actuarial practice of predicting the total *ultimate* losses incurred on a set of policies once all claims are reported and settled. This poses a challenging prediction task as losses frequently take years to fully emerge from reported claims, and not all claims might yet be reported. Loss development models frequently estimate a set of *link ratios* from insurance loss triangles, which are multiplicative factors transforming losses at one time point to ultimate. However, link ratios estimated using classical methods typically underestimate ultimate losses and cannot be extrapolated outside the domains of the triangle, requiring extension by *tail factors* from another model. Although flexible, this two-step process relies on subjective decision points that might bias inference. Methods that jointly estimate 'body' link ratios and smooth tail factors offer an attractive alternative. This paper proposes a novel application of Bayesian hidden Markov models to loss development modelling, where discrete, latent states representing body and tail processes are automatically learned from the data. The hidden Markov development model is found to perform comparably to, and frequently better than, the two-step approach on numerical examples and industry datasets.

Keywords: actuarial science, loss reserving, mixture modelling, time series

1 Introduction

Loss development modelling in actuarial science is the practice of predicting the total losses incurred from all reported and settled claims on a set of insurance policies. At any moment in time, these so-called *ultimate* losses, less the losses paid on already reported claims, constitute the overall loss reserves, generally referred to as *incurred but not reported* (IBNR). The IBNR amount includes future payments on claims already reported ('incurred by not enough reported'), as well as payments for claims not yet reported ('incurred by not yet reported'). In general or non-life insurance, IBNR can be substantial as it may take years for all claims on a set of policies to be reported and settled. Therefore, the accurate estimation of IBNR is key to evaluating an insurance company's performance and solvency (Beard, 1960; Bornhuetter and Ferguson, 1972; Friedland, 2010; England and Verrall, 2002; Wüthrich and Merz, 2008). Transparent assessment of the uncertainty in IBNR is also critical to the legal responsibilities of insurers and reinsurers, such as for the Solvency II directive (England et al., 2019; Fröhlich and Weng, 2018; Munroe et al., 2018).

Estimation of loss reserves is a challenging prediction problem that has received considerable attention from actuaries, econometricians, and statisticians for decades (e.g. for some key developments, see Bornhuetter and Ferguson, 1972; Clarke and Harland, 1974; Taylor, 1977; Taylor and Ashe, 1983; Mack, 1993; Barnett and Zehnwirth, 2000; England and Verrall, 2001, 2002; Taylor et al., 2003; Wüthrich and Merz, 2008). From early deterministic and algebraic approaches (e.g. Scurfield, 1968; Bornhuetter and Ferguson, 1972; Clarke and Harland, 1974; Taylor, 1977), loss development modelling has advanced to utilise a wide, and growing, variety of statistical methods, including parametric and non-parametric regressions (Mack, 1994; England and Verrall, 2001, 2002; Lally and Hartman, 2018), Bayesian estimation (England and Verrall, 2002; De Alba, 2002; Zhang et al., 2012; Meyers, 2015), differential equations (Gesmann and Morris, 2020), and neural networks and machine learning methods (Kunce and Chatterjee, 2017; Kuo, 2019; Al-Mudafer et al., 2022). Despite their differences, the majority of these approaches are applied to aggregated insurance risks, typically displayed as a triangular matrix of experience periods and development periods or lags called a *loss triangle* (see Figure 1). Each experience period represents losses on a distinct set of policies, commonly losses from all policies with accidents occuring in a

specific time period, and each development lag records the cumulative or incremental development of those losses through time since reporting.

A key inferential quantity from any loss development model is the set of multiplicative factors transforming the losses at development period one to the ultimate losses at development period ∞, known as the *loss development factors* or *link ratios*. While many models include link ratios as an explicit parameter to be inferred, notably the family of chain ladder methods (e.g. Mack, 1993; England and Verrall, 2002), others derive link ratios as a generated quantity (England and Verrall, 2001; Meyers, 2015). Ideally, link ratios will smoothly decline over time towards unity, but in practice often include periods of volatility, particularly for early development periods, and may further encode systematic and non-systematic effects across experience periods or date of development evaluation (e.g. the influence of the Covid-19 pandemic). Depending on the degree of volatility and the amount of data available, link ratios estimated from a triangle of finite risks will be insufficient to infer the ultimate losses for each experience period, because losses may still be emerging at the latest development period available in the data. Thus, estimation of ultimate losses will require extending the link ratios to outside the domains of the focal triangle to include *tail factors*. Like general loss development modelling, tail factor estimation has had its own expansive history of deterministic and stochastic methods (CAS Tail Factor Working Party, 2013), and is of particular importance in 'long-tailed' lines of business, such as workers' compensation or general liability, where experience periods might display continued loss cost growth and volatility at relatively late development lags.

Of the various approaches to calculating tail factors, many use a second model fit to a portion of the focal triangle that conveys how the triangle may behave in the tail (CAS Tail Factor Working Party, 2013), or to the set of link ratios directly from the first model. These models typically infer a parametric, monotonically increasing growth curve of losses from the training data, such as various forms of inverse power curves (e.g. Sherman, 1984; Evans, 2015; Clark, 2017). The link ratios derived from this tail model are then appended to the link ratios estimated from the primary loss development model to produce predictions an arbitrary number of development lags into the future. For clarity, the primary link ratios will be referred to in this paper as the 'body' link ratios to distinguish them from link ratios estimated in

the tail. Crucially, this two-step process includes a number of subjective decisions. The body-to-tail switch-over development lag is frequently chosen to reflect when the development process settles to a reasonable plateau, while the training windows for both models will be chosen based on which sections of the triangle best match each model's assumptions. For instance, including periods of non-monotonic growth into tail models might bias tail factors unreasonably high. These decisions are difficult to reproduce and opens analysts to many 'researchers degrees of freedom' (Simmons et al., 2011) – selecting one approach among many possible alternatives that might have non-trivial impacts on predicting ultimate loss.

Methods that simultaneously estimate body and tail link ratios present an attractive alternative to traditional two-step processes. However, only a few solutions have been proposed. England and Verrall (2001) presented a generalised additive model for smooth estimation of loss development factors that can be extrapolated to points futher than the existing data. Although flexible and able to integrate different functional forms and covariates of loss development processes, this approach still necessitates careful selection of training data to include in the model so that regions of volatility do not bias loss development curves. Generalised additive models further require selecting the family of splines and number of knots to apply, which may lead to another set of decisions analysts must act on. Zhang et al. (2012), alternatively, implemented a hierarchical Bayesian logistic growth curve model to cumulative loss data. However, fitting a single parametric curve that assumes monotonicity in expectation might under-estimate systematic volatility in portions of the triangle that analysts do not want to label as residual noise. Finally, Verrall and Wüthrich (2012) use reversible jump Markov chain Monte Carlo to combine a Bayesian chain ladder model, applied before some tail cut-off point, with an exponential decay process after the cut-off, and allow the model to infer where the cut-off should occur. Additionally, Verrall and Wüthrich (2015) demonstrate how the same model can be estimated in a Bayesian model averaging context. Although this approach is arguably the most flexible, it requires either bespoke sampling algorithms (i.e. reversible jump Monte Carlo) or multiple model fits (for model averaging purposes), which may dissuade analysts and researchers.

This paper proposes the use of hidden Markov models to simultaneously estimate

body and tail link ratios in a single model from a variety of loss triangles. Hidden Markov models are primarily discrete mixture models postulating an unknown latent state underlying and generating patterns of observed data, and have found a multitude of applications, from speech recognition (e.g. Rabiner, 1989) to animal behaviour (e.g. Leos-Barajas et al., 2017). Indeed, Markov processes have been previously used in micro-level claim modelling (e.g. Hesselager, 1994), but have not been applied to aggregate insurance loss triangles. As described in this paper, hidden Markov development models decompose a loss triangle into a sequence of body and tail processes. Hidden Markov models are easily fit in existing, open-source software, and can cater for complex data-generating assumptions, such as understanding the impact of covariates or including non-parametric patterns of loss development (England and Verrall, 2001), Moreover, the approach presented here is flexible enough to integrate any of the parametric and non-parametric models that have been presented in the actuarial literature for modelling body and tail processes.

Below, the hidden Markov model is formulated and validated with simulated examples, and subsequently compared to a traditional two-step process using a number of data sets. The hidden Markov models are fit using Bayesian estimation in Stan (Carpenter et al., 2017) following a modern Bayesian workflow (Gelman et al., 2020). All Python and Stan code, and all datasets, to reproduce the results are accessible at the Github repository `https://github.com/LedgerInvesting/hidden-markov-development-2024`.

2 Materials and methods

2.1 Hidden Markov development model

Consider the typical loss development context, where a large set of homogeneous insurance risks (e.g. private car insurance policies) are aggregated into a loss development triangle with cumulative loss amounts denoted $\mathcal{Y}$, defined by

$$\mathcal{Y} = \{y_{ij} : i = 1, ..., N; j = 1, ..., N - i + 1\} \tag{1}$$

where $i = (1, ..., N)$ denotes experience periods, most commonly indexing all claims occurring during period i, and $j = (1, ..., M)$ denotes development periods or lags. For a particular point in time, development information for period i is only known up to lag $j = N-i+1$, and therefore $\mathcal{Y}$ represents the left upper diagonal of the loss triangle. The complementary, lower diagonal triangle, denoted $\tilde{\mathcal{Y}}$, with $(N + M)\frac{N+M-1}{2}$ data points, is unknown and the goal of prediction.

The generative process considered here (Figure 1) is that cumulative losses in $\mathcal{Y}$ develop according to i) a period that is characterised by largely, but not strictly, monotonically increasing losses (the body), followed by ii) a period of smooth growth to ultimate (the tail). Traditional methods treat estimation of body and tail as a two-step process. By contrast, the hidden Markov development model introduces a latent, discrete state $\boldsymbol{z} = (z_{11}, z_{12}, ..., z_{ij}, ..., z_{NM}) \in \{1, K\}$ with $K = 2$ that takes value $z_{ij} = 1$ if the body process generated the losses in the ith accident period and jth development period, and $z_{ij} = 2$ if the tail process generated the losses. The latent state at one time point, z_{ij}, is dependent on the state at the previous time point, z_{ij-1}, and subsequent states are connected via a state transition matrix, Θ_{ij}. Depending on the latent state and any other parameters $\boldsymbol{\phi}$, the marginal likelihood of the data, $p(y_{ij} \mid z_{ij}, \boldsymbol{\phi})$, is given by suitable *emission* distributions. Emission distributions are the observation data distributions, which in loss development modelling are typically positive-bound, continuous probability density functions, such as lognormal or Gamma. In this paper, all models use lognormal distributions as the likelihood distribution.

The hidden Markov development model can be written generally as:

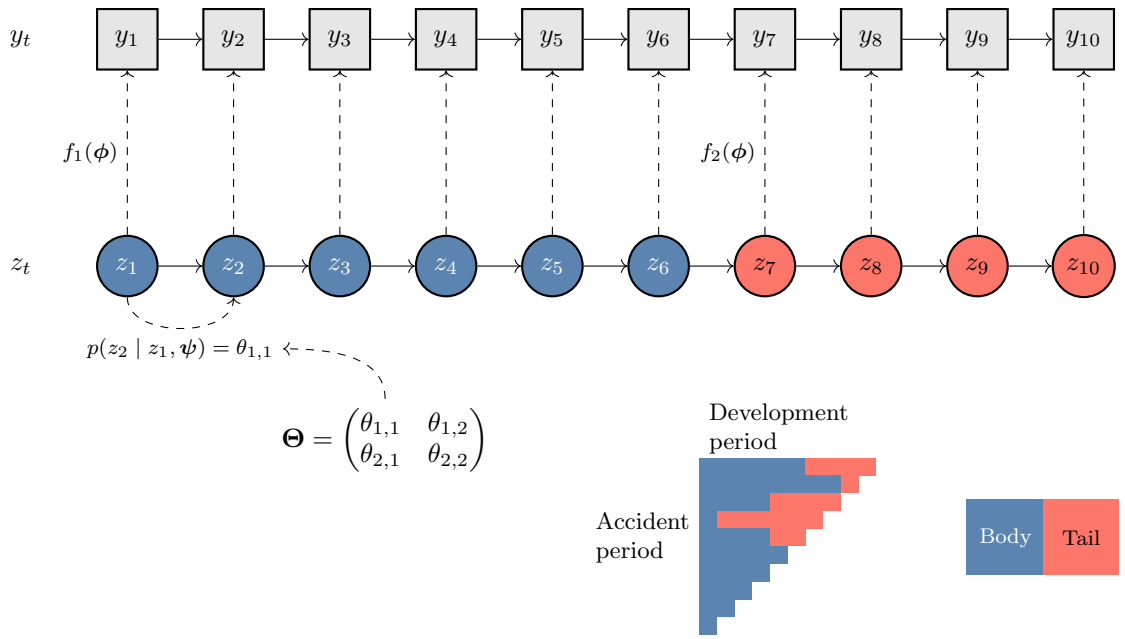

Figure 1: A schematic of the hidden Markov model process. A loss triangle of observed data is shown with 10 accident periods, and each development period generated from two processes: the body (blue) and the tail (orange). The dynamics of the body and tail states vary over accident periods. For a single accident period, observed losses at development lag t, y_t, are shown as grey squares, and are assumed generated from the latent discrete random variable z_t (circles), which transition according to the probabilities in matrix Θ.

$$
\begin{aligned}
y_{ij} &\sim \begin{cases} \text{Lognormal}(\mu_1, \sigma_{ij}) & z_{ij} = 1 \\ \text{Lognormal}(\mu_2, \sigma_{ij}) & z_{ij} = 2 \end{cases} \\
z_{ij} &\sim \text{Categorical}(\Theta_{ij}^{z_{ij-1}}) \\
\Theta_{ij} &= \begin{pmatrix} \pi & 1-\pi \\ \nu & 1-\nu \end{pmatrix} \\
\log \frac{\pi}{1-\pi} &\sim \text{Normal}(0, 1) \\
\log \frac{\nu}{1-\nu} &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{2}
$$

The cumulative losses at accident period i and development period j are assumed

positive-bound, lognormally-distributed random variates, with log-scale location μ_1 if the process is in the body or μ_2 if the process is in the tail, and with scale σ_{ij}. The latent state z_{ij} determines the body or tail state at each time point, and follows a categorical distribution with unit simplex probabilities determined by row z_{ij-1} of the state transition matrix Θ_{ij}. As discussed below, the state transition matrix may be time-homogeneous or -inhomogeneous, depending on the context. In Θ, π denotes $p(z_{ij} = 1 \mid z_{ij-1} = 1)$, the probability of staying in the body process at lag j, and $\nu = p(z_{ij} = 2 \mid z_{ij-1} = 2)$. Their complements, $1 - \pi$ and $1 - \nu$, represent the probabilities of transitioning from body to tail, and tail to body, respectively.

The functions $\boldsymbol{\mu} = (\mu_1, \mu_2)$ provide the conditional expectations of the two processes assumed to describe the data. The models here use two canonical body and tail development models: the chain ladder model for body loss development factors (Mack, 1993; England and Verrall, 2002), and an exponential decay curve following the generalised Bondy model for the tail process (CAS Tail Factor Working Party, 2013). The full model specification is completed by choosing these functional forms, along with the functional form for the variance and the remaining prior distributions.

$$
\begin{aligned}
\mu_{1_{ij}} &= \log(\alpha_{j-1} y_{ij-1}) \quad \forall j > 1 \\
\mu_{2_{ij}} &= \log(\omega^{\beta^j} y_{ij-1}) \quad \forall j > 1 \\
\sigma_{ij}^2 &= \exp(\gamma_1 + \gamma_2 j + \ln(y_{ij-1})) \\
\log \boldsymbol{\alpha}_{1:M-1} &\sim \text{Normal}(0, 1/1 : M-1) \\
\log \omega &\sim \text{Normal}(0, 1) \\
\log \frac{\beta}{1-\beta} &\sim \text{Normal}(0, 1) \\
\boldsymbol{\gamma}_{1:2} &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{3}
$$

Due to the multiplicative autoregressive nature of typical loss development models, the first data point is not modelled, and the hidden Markov process is assumed to start in the body state. The $M - 1$ body link ratios are given by the vector $\boldsymbol{\alpha}$, and the tail link ratios are given by ω^{β^j}, for any j, allowing extrapolation out to arbitrary development periods. The parameter ω is constained to be strictly greater than 1.0, such that growth is monotonically increasing, and β is constrained to lie in the

interval $(0, 1)$, to avoid tail factors growing without bound to ∞. The expression for the variance encodes the assumption of less volatility at higher development periods, and is proportional to the losses at the previous time point.

The prior distribution on $\boldsymbol{\alpha}$ is regularised towards a link ratio of 1 in inverse proportion to the development period. This assumption is imposed to ensure link ratios, when $z_{ij} = 2$ (i.e. the latent state is in the tail process), do not become unrealistically large due to the resulting non-identifiability. Even when the tail process is most likely, some samples could be generated from the body process because z_{ij} is a random variable, and large values of α_{ij-1} might have unwanted influence on the predictions. This is akin to the use of pseudo-priors in Bayesian mixture models (Carlin and Chib, 1995), which ensure parameters in the mixture not being sampled do not become implausible.

2.1.1 Model variants

Three variations of the hidden Markov model are used (Table 1). The base model (referred to as HMM) has homogeneous transition matrix Θ, and sets ν to zero. This implies that once the tail process is active, the model cannot switch back to the body process. This is the primary assumption underlying tail modelling generally: at some development point, the losses smoothly develop to ultimate. Secondly, the HMM-ν model estimates ν, allowing for tail processes to switch back to body processes. Some triangles may illustrate unexpected late-development volatility, at which point the more flexible body process is a better explanation of the data. Finally, the HMM-lag variant allows π to vary across development periods, $\boldsymbol{\pi} = (\pi_1, ..., \pi_{j-1}, ..., \pi_{M-1})$, on which an ordered assumption is imposed such that the probability of transitioning to the tail increases with development period, meaning $\pi_{j-1} < \pi_{j-2}$. Many extensions of these variants are possible, including the addition of covariates on the estimation of Θ or other parametric or non-parametric forms. This paper restricts focus on these three variants as they present the simplest use cases to compare to the two-step process. Although the HMM-lag and HMM-ν variant could be combined, early tests of the hidden Markov development model by the author indicated that this model risked unidentifiability without further data or covariates.

Name	Θ_{ij}
HMM	$\begin{pmatrix} \pi & 1-\pi \\ 0 & 1 \end{pmatrix}$
HMM-ν	$\begin{pmatrix} \pi & 1-\pi \\ \nu & 1-\nu \end{pmatrix}$
HMM-lag	$\begin{pmatrix} \pi_{ij-1} & 1-\pi_{ij-1} \\ 0 & 1 \end{pmatrix}$

Table 1: The three hidden Markov model transition matrix variants used in the examples.

2.1.2 Estimation

We fit the models in Stan (Carpenter et al., 2017) using Bayesian inference via Hamiltonion Monte Carlo, via the `cmdstanpy` (Stan Development Team, 2024b) Python package and `cmdstan` (Stan Development Team, 2024a). Stan requires specifying a statement proportional to the joint log density of the data and parameters. For the above model, the log likelihood for a single data point, grouping all transition matrix parameters as $\boldsymbol{\psi} = \{\pi, \nu\}$ and emission distribution parameters as $\boldsymbol{\phi} = \{\alpha, \omega, \beta, \gamma\}$, can be factored as:

$$p(y_{ij}, z_{ij}, z_{ij-1}, \boldsymbol{\phi}, \boldsymbol{\psi}) = p(y_{ij} \mid z_{ij}, \boldsymbol{\phi}) p(z_{ij} \mid z_{ij-1}, \boldsymbol{\psi}) p(z_{ij-1} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi}) p(\boldsymbol{\phi}) \tag{4}$$

However, it is often more computationally efficient to marginalise the latent discrete parameters z_{ij} and z_{ij-1} out of the density, leading to a double summation over possible states for z_{ij} and z_{ij-1}:

$$\begin{aligned} p(y_{ij}, \boldsymbol{\phi}, \boldsymbol{\psi}) = & \sum_{k=1}^{K} p(y_{ij} \mid z_{ij} = k, \boldsymbol{\phi}) p(\boldsymbol{\phi}) \\ & \sum_{h=1}^{K} p(z_{ij} = k \mid z_{ij-1} = h, \boldsymbol{\psi}) p(z_{ij-1} = h \mid \boldsymbol{\psi}) p(\boldsymbol{\psi}) \end{aligned} \tag{5}$$

This recursive estimation over possible states is the forward algorithm (Rabiner, 1989). After model fitting, the hidden states on the training data can be recovered using the Viterbi algorithm (Rabiner, 1989), which provides the most likely joint sequence of latent states that generated the data. For future data, new samples of z_{ij} can be taken from a categorical distribution with estimated parameters of the transition matrix Θ_{ij} for that data point.

2.2 Two-step approach

The hidden Markov development model is compared to a more traditional two-step modelling approach. Denote $\tau \in \{2, ..., M\}$ the final body training data point, and $\boldsymbol{\rho} = (\rho_1, \rho_2) \in \{2, ..., M\}$, where $\rho_1 < \rho_2$, a vector of tail start and end training window development points, respectively. While these constants could in theory vary over experience periods, there is typically insufficient data to do so. In the traditional approach, the chain ladder method is first fit to training data up to development period τ inclusive and predictions of the lower diagonal loss triangle are made up to and including τ, only. Secondly, the tail model is fit to data lying within the development period interval $[\rho_1, \rho_2]$, and predictions from τ made to some arbitrary development lag, j^*. The challenge and art of this two-step process is in finding a value for $\tau + 1$ that identifies the development lag where losses are plateauing in the tail, and finding values for $\rho_{1,2}$ that identify a suitable decaying curve of link ratios. While $\tau = \rho_1$ in some cases, more generally $\rho_1 \leq \tau$. Note, this presents a difference between the two-step and hidden Markov model approaches: the hidden Markov model identifies clear body-to-tail switch-over points, whereas the the interval of data delineated by $\rho_{1,2}$ in the two-step process might overlap the final body training point, τ.

To maintain comparability to the hidden Markov model above, the two-step approach is also implemented in Stan and both models estimated with a shared variance. In real applications, uncertainty from body to tail models is typically ignored, but this may unfairly penalise the two-step process compared to the hidden Markov models. The two-step approach differs from equation 2 in only two ways:

$$
\begin{aligned}
y_{ij} &\sim \begin{cases} \text{Lognormal}(\mu_1, \sigma_{ij}) & j \leq \tau \\ \text{Lognormal}(\mu_2, \sigma_{ij}) & \rho_1 \leq j \leq \rho_2 \end{cases} \\
\log \boldsymbol{\alpha}_{1:\tau-1} &\sim \text{Normal}(0, 1)
\end{aligned} \tag{6}
$$

where now the decision between the two models is decided by τ and $\rho_{1,2}$. The further exception in the two-step approach is the use of a standard normal prior on the log-scale link ratios, rather than the constrained prior used in the hidden Markov models. The two-step model's $\tau - 1$ link ratios are all estimated directly and do not suffer from non-identifiability.

2.3 Simulation-based calibration

To validate the hidden Markov model, we use both simulated examples (shown in the results) to build intuition, alongside simulation-based calibration (Talts et al., 2018; Modrák et al., 2023). Simulation-based calibration leverages the self-consistency of the Bayesian joint distribution of parameters and data: fitting a Bayesian model to datasets generated from its prior predictive distribution and averaging over datasets should return the prior distribution.

Focusing on the base HMM model variant, 1000 full-triangles with $N = M = 10$ were generated from the prior predictive distribution, and the HMM model fit to each upper diagonal $\mathcal{Y}$. The prior distributions were the same as in equation 2, except for the priors on (γ_1, γ_2), which were given more informative normally-distributed priors with locations and scales of $(-3, -0.25)$ and $(-1, 0.1)$, respectively. Due to the multiplicative autoregressive forms in the location and scales of the likelihood in equation 2, particularly large values for σ can cause overflow in the sampled data.

Each model was summarised by calculating the rank statistics of quantities of interest. The rank statistic is the number of times a simulated value is greater than the posterior values, and should be approximately uniformly distributed if the model has been implemented correctly and is unbiased (Talts et al., 2018). To reduce the autocorrelation in the posterior distributions, the posteriors were thinned to every

10th posterior draw. Rank statistics were calculated for each parameter in the HMM model, as well as the joint log likelihood and an the ultimate loss prediction at data point $(i = 1, j = 10)$, since Modrák et al. (2023) recommend using test quantities that average over the entire data space in evaluating SBC.

2.4 Datasets and model performance

We compared the hidden Markov development model to the two-step approach in two different ways. Firstly, we estimated both models on the 200 industry paid loss triangles from Meyers (2015), 50 triangles for the four lines of business: private passenger auto (PP), workers compensation (WC), commerical auto (CA), and other occurrence (i.e. general) liability (OO). We removed triangles with zero loss values first, leaving 49 CA triangles, 48 OO triangles, and 50 PP and WC triangles. Each triangle covers 10 years of historical accident periods and development, allowing splitting triangles into an upper diagonal of training data, $\mathcal{Y}$, and lower diagonal of test data, $\tilde{\mathcal{Y}}$. As these are yearly triangles, and have been chosen for previous model validation exercises (Meyers, 2015), it is reasonable to assume that the losses at development period 10 are close to ultimate. For the two-step approach, we inspected the mean and standard deviations of the empirical link ratios across triangles (shown in Figure 2), and selected $\tau = 6$ and $\boldsymbol{\rho} = (4, 10)$. These were chosen given that the link ratios showed smooth patterns of decay from approximately development period 4 onwards, and the triangles were close, on average, to their values at period 10 by development period $\tau + 1 = 7$.

When fitting the models to industry triangles, a small number of hidden Markov and two-step models failed due to producing very large posterior predictions on out-of-sample data, which numerically overflowed. The multiplicative autoregressive nature of both the hidden Markov and two-step models mean that large predictions at one time point can quickly compound to unrealistic and computationally-unstable values. For this reason, we capped the predictions at 100 times the maximum value across the training and test data for each triangle.

Secondly, we used five triangles presented in the relatively recent literature on papers on loss development modelling, including tail development, that provide more histori-

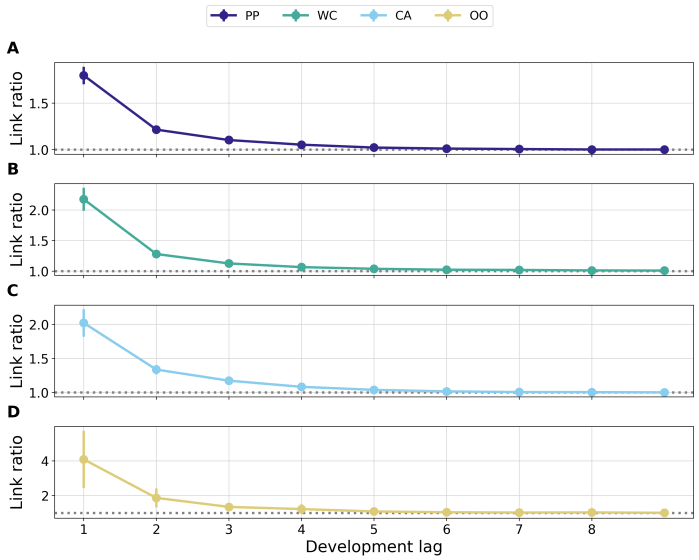

Figure 2: The empirical link ratios by line of business (panels A-D) in the Meyers (2015) dataset. Points indicate the mean across triangles, and vertical line segments show 2 standard deviations.

cal data than the industry triangles: the long-tailed liability and short-tailed property quarterly triangles from medium-size insurers in Balona and Richman (2022), the annual liability triangle in Merz and Wüthrich (2015), the Swiss annual liability triangle from Gisler (2009), and the annual liability triangle in Verrall and Wüthrich (2012). These triangles had between 17 and 22 periods of data. For each triangle, we used the latest diagonal as the test dataset to evaluate predictive performance. For the two-step modelling approach, we chose the τ and $\boldsymbol{\rho}$ constants by visual inspection of the empirical link ratios, selecting $(12, 5, 5, 5, 12)$ for τ values for each triangle, and $[(4, 16), (3, 20), (3, 21), (3, 16), (10, 21)]$ for $\boldsymbol{\rho}$ values, respectively. The link ratios are displayed in Figure 8.

2.4.1 Model performance

Model performance can be split into two facets: accuracy and calibration. We summarised model accuracy using two metrics applied purely to the out-of-sample data: the expected log predictive density (ELPD) and the root mean square error (RMSE). The ELPD (Vehtari et al., 2017) is based on the logarithmic scoring rule, and is de-

fined for a single triangle as the joint log likelihood of the out-of-sample data. In this way, it more heavily penalises models further from the true data generating process. Across each accident period i and development period j in $\tilde{\mathcal{Y}}$, we take the sum of log likelihood values marginalized across the posterior samples:

$$\begin{aligned} \text{ELPD} &= \sum_{i=1}^{N} \sum_{j=N-i+1}^{M} \log p(\tilde{y}_{ij} \mid \mathcal{Y}) \\ &= \sum_{i=1}^{N} \sum_{j=N-i+1}^{M} \log \int p(\tilde{y}_{ij} \mid \theta) p(\theta \mid \mathcal{Y}) d\theta \\ &\approx \sum_{i=1}^{N} \sum_{j=N-i+1}^{M} \frac{1}{S} \sum_{s=1}^{S} \log p(\tilde{y}_{ij}^{(s)} \mid \mathcal{Y}) \end{aligned} \tag{7}$$

where $p(\tilde{y}_{ij} \mid \mathcal{Y})$ is the posterior predictive distribution for the ith accident period at lag j, θ is used to generically refer to all model parameters, i.e. $\theta = \{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{z}\}$, and the super-script in $\tilde{y}_{ij}^{(s)}$ denotes the sth sample from the posterior distribution with S total samples. The second sum over $N - i + 1$ development periods in the ith accident period assumes a typical loss triangle with a full lower diagonal of test data. To compare models, we calculated the difference in ELPD for each triangle, t, and its standard error, where the standard error of the difference is the square root of the product of i) the sample variance of log predictive density differences between models, and ii) the number of data points (Vehtari et al., 2017; Sivula et al., 2020). For the industry data, we combined ELPD values for each triangle by taking the mean of the ELPD differences and mean of the standard errors (Sivula et al., 2020). Approximate 95% confidence intervals were then derived by using a range of 2 standard errors around the estimate. Although the test data is known with certainty, it is still just a portion of accident periods and development lags for each triangle, and therefore uncertainty in ELPD was still calculated.

The RMSE was defined per out-of-sample data point in $\tilde{\mathcal{Y}}$

$$\text{RMSE}_{\text{ij}} = \sqrt{\frac{1}{S} \sum_{s=1}^{S} (\hat{y}_{ij}^{(s)} - \tilde{y}_{ij})^2} \tag{8}$$

where $\hat{y}_{ij}^{(s)}$ is the sth sample from the posterior predictive distribution. In contrast to ELPD, RMSE penalises models that produce predictions further from the test data points using a quadratic scoring rule, and may demonstrate different results depending on the context. As with ELPD, the average differences in RMSE between models per triangle were used to compare models, and their standard errors were derived as the square root of the product of i) the sample variance of the differences in RMSE and ii) the number of data points.

Model calibration was inspected using histograms of the percentiles of the true data on the posterior predictive distributions, i.e. the empirical cumulative distribution functions. Well-calibrated models' percentiles should be uniformly distributed, as in the case of simulation-based calibration above.

3 Results

To build intuition for the two models, Figure 3 shows a numerical simulation of data from a HMM model, and the results for the HMM model, in panel A, and the two-step model, in panel B. Whereas the HMM model's posterior predictions switch from body to tail depending on the particular latent state path, the two-step models have fixed $\tau = 6$ and $\rho_{1,2} = (6, 10)$. As shown by the ELPD values above each plot, the HMM variant outperforms the two-step approach for all experience periods except the first, where our chosen value τ matches the HMM data-generating process exactly. In other experience periods, the two-step approach generalises poorly. This illustrates that, if τ is chosen correctly, and the tail model is trained on suitable development lags, the two-step approach can provide more exact predictions, because uncertainty in the latent state from the HMM model hurts predictive accuracy. However, if experience periods differ in their body-to-tail switch-over dynamics, which is expected, then overall performance suffers due to growing generalisation error.

Simulation-based calibration of the HMM model indicated no problems with model calibration (Figure 4), with all histograms matching the assumptions of uniformity. However, 10 models were removed for poor convergence, which typically occurred when the simulated link ratios from the body process were higher than values expected

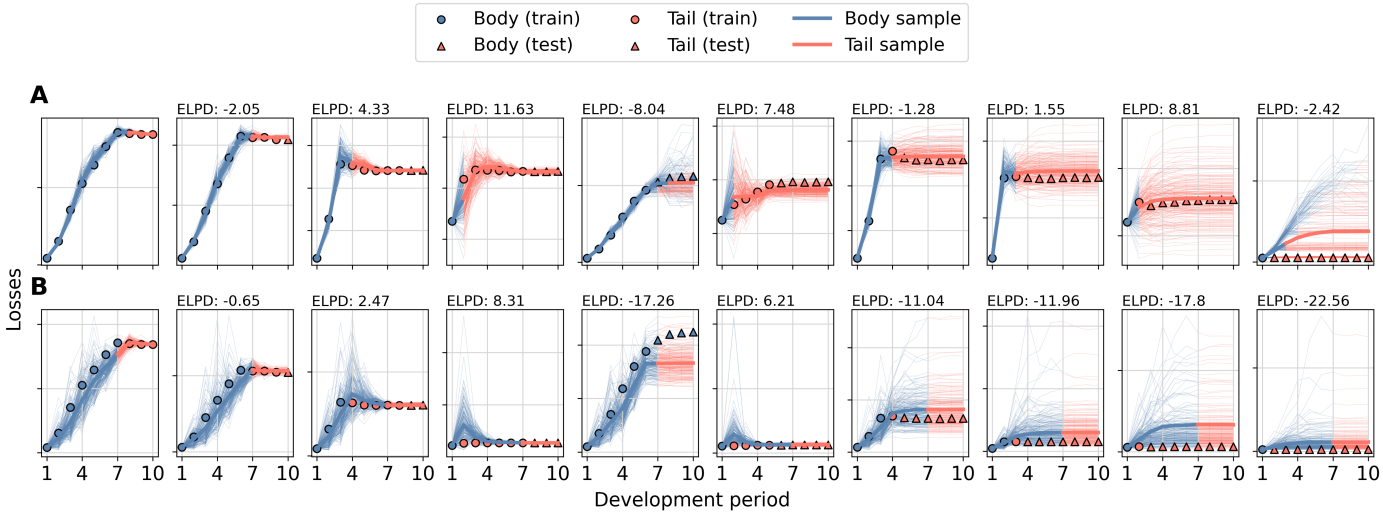

Figure 3: A simulated data example comparing the HMM to two-step process to build intuition. Panels A (HMM) and B (two-step) show 10 experience periods from a single loss triangle simulated from the HMM model. The true losses are shown as points, with colours identifying body (blue) or tail (orange) points. Circles denote training data and triangles test data. Thin lines are samples from the posterior predictive distributions, coloured by latent state, and the thicker line shows the mean paths. The ELPDs on the test data points in each experience period are shown above each plot. The two-step model was fit assuming $\tau = 6$ *and* $\rho_{1,2} = (6, 10)$.

from real loss triangles. Given this occurred rarely, the priors were left unchanged, although suitable prior distributions for Bayesian chain ladder models is an area with a dearth of literature. For the 990 models, the average classification accuracy of the recovered latent state values $\boldsymbol{z}$, across both training and test data, was 97% with a 95% highest density interval (i.e. the 95% most likely values) of [91, 100].

The ELPD and RMSE differences between models (compared to the HMM base model) are shown in Figure 5. When calculating ELPD, a small number of pointwise log densities showed very small, negative values, indicating poor predictions on the out-of-sample data leading to numerical instability. We decided to remove any log predictive densities for all models that had values < -100, which for a single out-of-sample data point was particularly low, given that most ELPD values for a single data point lie within [-5, 5]. This retained 99.91% of values for the PP results, 99.42% of values for WC, 99.64% for CA, and 98.38% for OO. The full log densities are given in the supplementary materials, as well as results for different levels of filtering.

For ELPD and RMSE, at least one of the hidden Markov model variants out-performed the two-step approach, apart from ELPD for PP and WC lines of business, where a small proportion of the two standard errors included zero (Figure 8). Overall, the HMM-ν model attained 75% of the best average ELPD scores, and 50% of the best average RMSE scores, meaning that allowing for tail processes to revert to the body is important to making future predictions. Evaluating the predictions at the lag-10 out-of-sample values only, to mirror the estimation of ultimate loss, mirrored the results in Figure 5 apart from some minor changes in ranks of the HMM models (full results supplied in the supplementary materials).

Model calibration histograms (Figure 6) indicated that both the hidden Markov models and two-step approaches typically have predictions that are too uncertain, indicated by the inverted-U shaped histograms. In WC, the hidden Markov models had the most uniform percentiles, whereas the two-step approach showed signs of both under- and over-estimation.

The hidden Markov model variants had different implications for body-to-tail switch-over points depending on the particular line of business (Figure 7). The PP and CA lines showed, in general, the quickest development to the tail state, at development

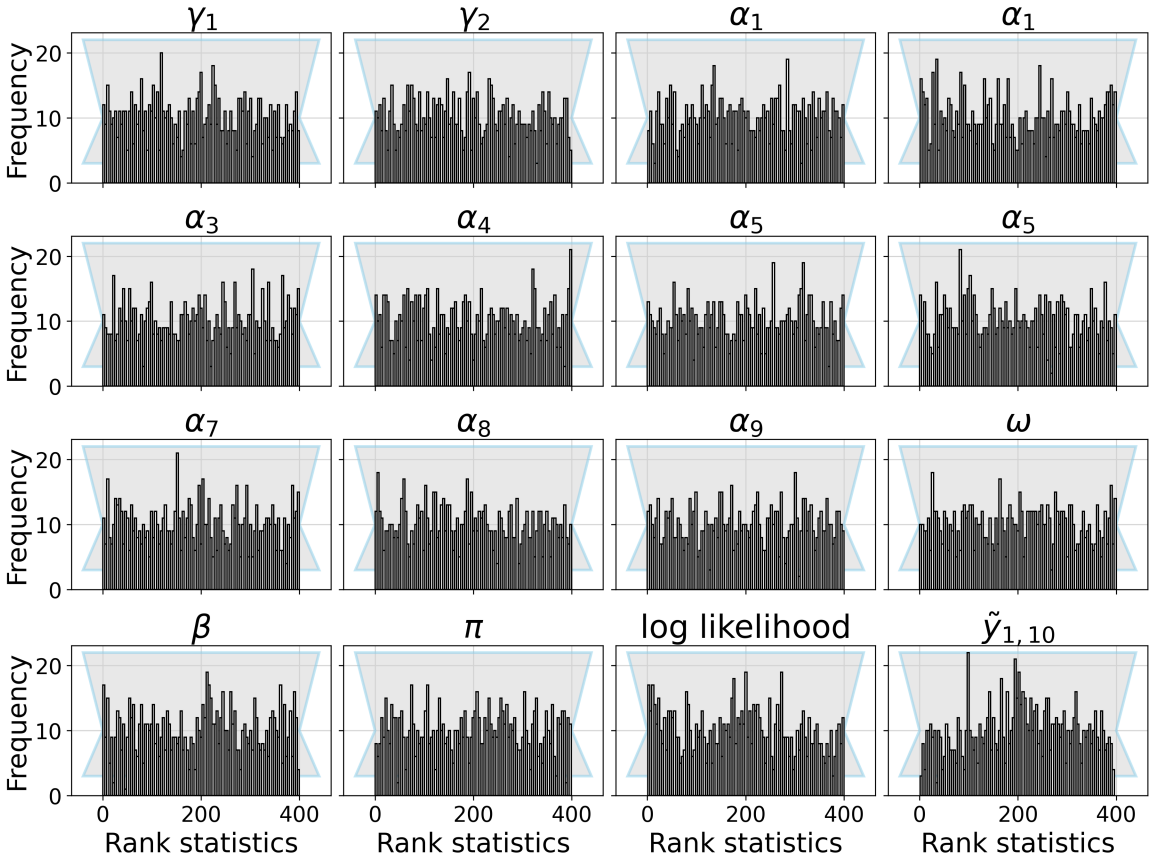

Figure 4: Histograms of simulation-based calibration rank statistics for the HMM model, with the 99% percentile interval from a discrete uniform distribution shown in the grey shaded band. Each histogram shows a key model parameter, and the final two panels show the ranks for the joint log likelihood and the first ultimate loss distribution. For each model, we sampled 4000 draws from the posterior distribution, and thinned the samples by 10 to remove any autocorrelation, meaning a maxmimum rank statistic of 400. Of the 1000 models, 10 were removed due to poor convergence.

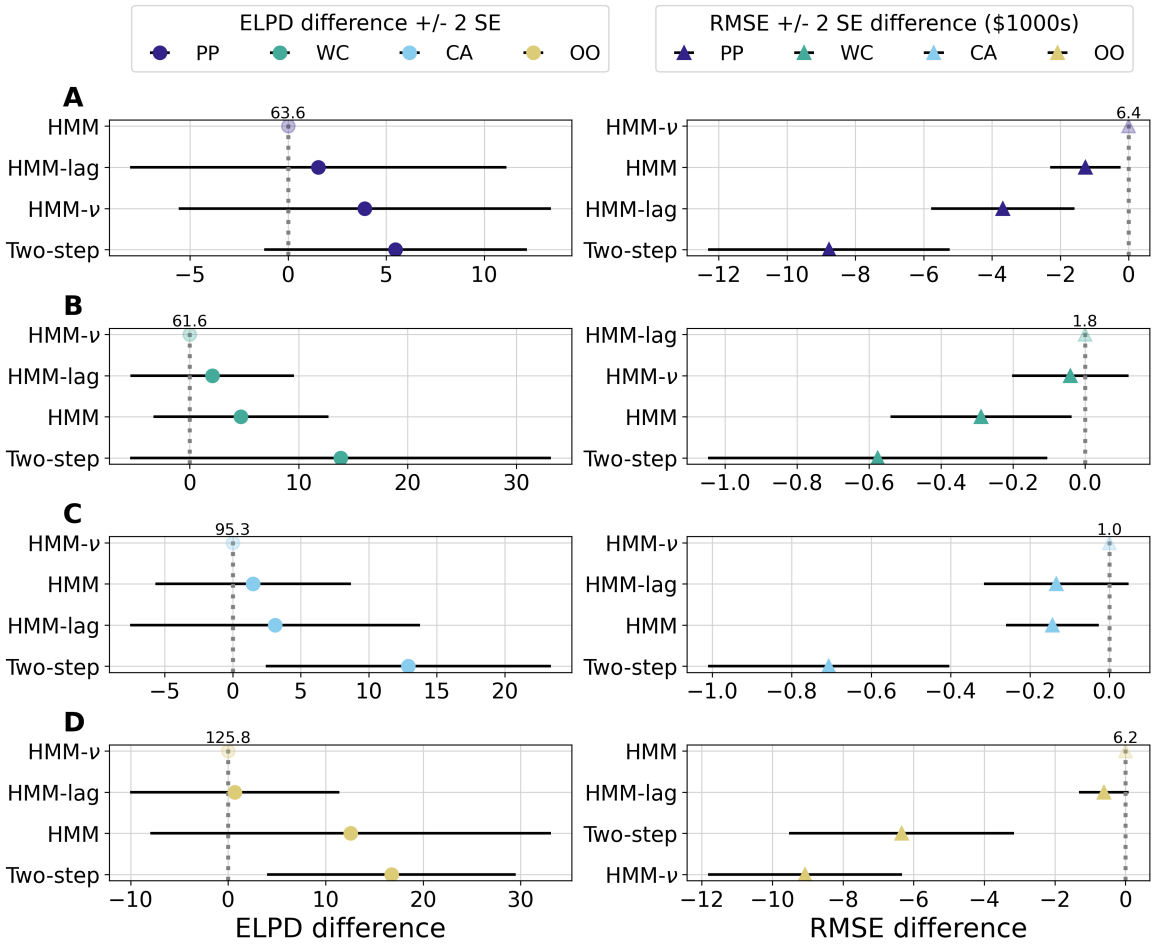

Figure 5: The ELPD (left column) and RMSE (in thousands of dollars; right column) differences (+/- 2 standard errors; SE) in order of performance for each model and line of business for the industry triangles. Rows A through D enumerate results for line of business separately. The best-performing model is shown at the top of each panel, with the absolute ELPD or RMSE value displayed above. Positive ELPD differences with an uncertainty interval that does not cross zero indicates a credible difference at the 95% level in favour of the top model. Negative RMSE differences with an uncertainty interval that does not cross zero indicates a credible difference in favour of the top model.

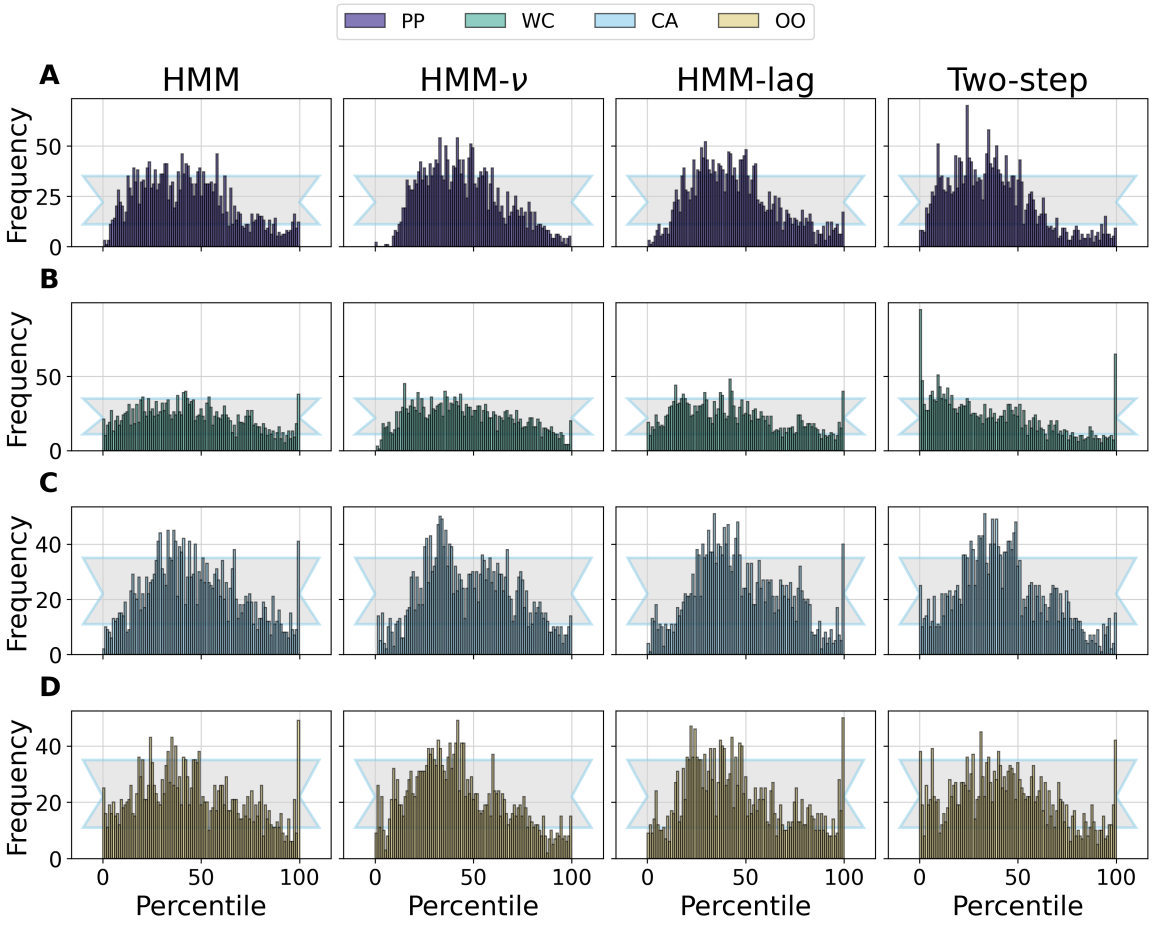

Figure 6: Percentiles of the true left-out values on the posterior distributions for each model and line of business (panels A through D) in the industry triangles. Grey shaded regions provide the 99% intervals of a discrete uniform distribution, for reference. Right-skewed histograms indicate under-estimation, left-skewed histograms indicate over-estimation, and inverted-U histograms indicate predictions that are uncertain.

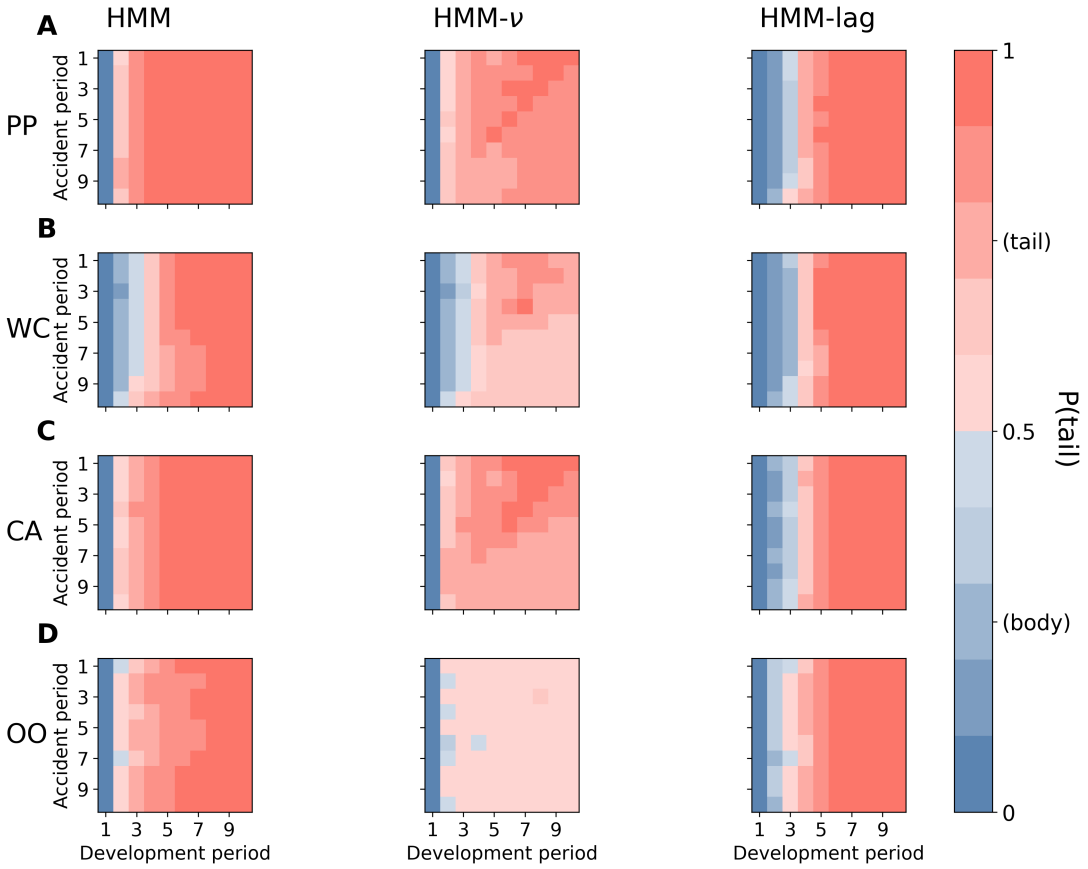

Figure 7: The average probability of being in the tail process for each hidden Markov model parameterization (columns) and line of business (rows A-D) across triangles in the industry data. Probabilities ≤ 0.5 *are coloured in blue whereas probabilities* > 0.5 *are coloured in orange. More faded squares indicating smaller probabilities of being in body and tail processes, respectively.*

lag 2 for the HMM and HMM-ν models, and by development lag 3 for most accident periods for the HMM-lag model. In contrast, WC stayed longest in the body state, followed by OO, and both WC and OO lines demonstrated relatively equitable probabilities of being in the body and tail at later development periods. This is particularly noticeable for the HMM-ν model, where the chance of returning to the body from tail process was allowed.

For the five literature triangles, the lack of hold-out data meant that the uncertainties around the ELPD and RMSE differences indicated more equal model performance

(Figure 8). The average ELPD and RMSE differences indicated that one of the hidden Markov models often performed better than the two-step approaches. The two-step approach out-performed the hidden Markov models for the Merz and Wüthrich (2015) triangle, and ranked second for RMSE for three of the five triangles. The manual selection of τ for the two-step process often closely aligned with the development lag from the best-performing HMM where the probability of being in the tail was > 0.5. The one exception was the long-tailed liability triangle of Balona and Richman (2022), where the chosen τ of 12 was much larger than the most likely switch-over lag from the HMM variant of $j = 3$. However, $j = 3$ did closely match the choice of $\rho_1 = 4$.

4 Discussion

This paper has proposed a hidden Markov loss development model for insurance loss triangles that combines body and tail development models, and automates the selection of body-to-tail switch-over points. Simulation-based calibration validated the hidden Markov model implementation as being unbiased, and across a range of different datasets, the hidden Markov model variants provided similar results to, and often out-performed, the traditional two-step approach. The hidden Markov models' automated detection of body and tail processes more gracefully captures loss development dynamics that may vary over triangle experience periods, as well as reduce analysts' degrees of freedom that make the traditional two-step approach reliant on difficult-to-reproduce and variable subjective decisions.

The hidden Markov development model posits a clear data-generating process for loss development dynamics. Although referred to here as 'body' and 'tail', these two latent states might equivalently be thought of as flexible and smooth periods of loss development, and can interchange depending on the context, as in the HMM-ν model variant here. In this way, the hidden Markov model is not a strictly analagous implementation of the two-step approach, as the two-step approach allows analysts to choose tail model training data windows that overlap the body-to-tail cutoff point. Thus, the same data points may be used in estimation of body and tail processes, rather than a discrete cutoff point between the two. Despite this flexibility, the two-step approach is not a single generative model, and should an analyst choose

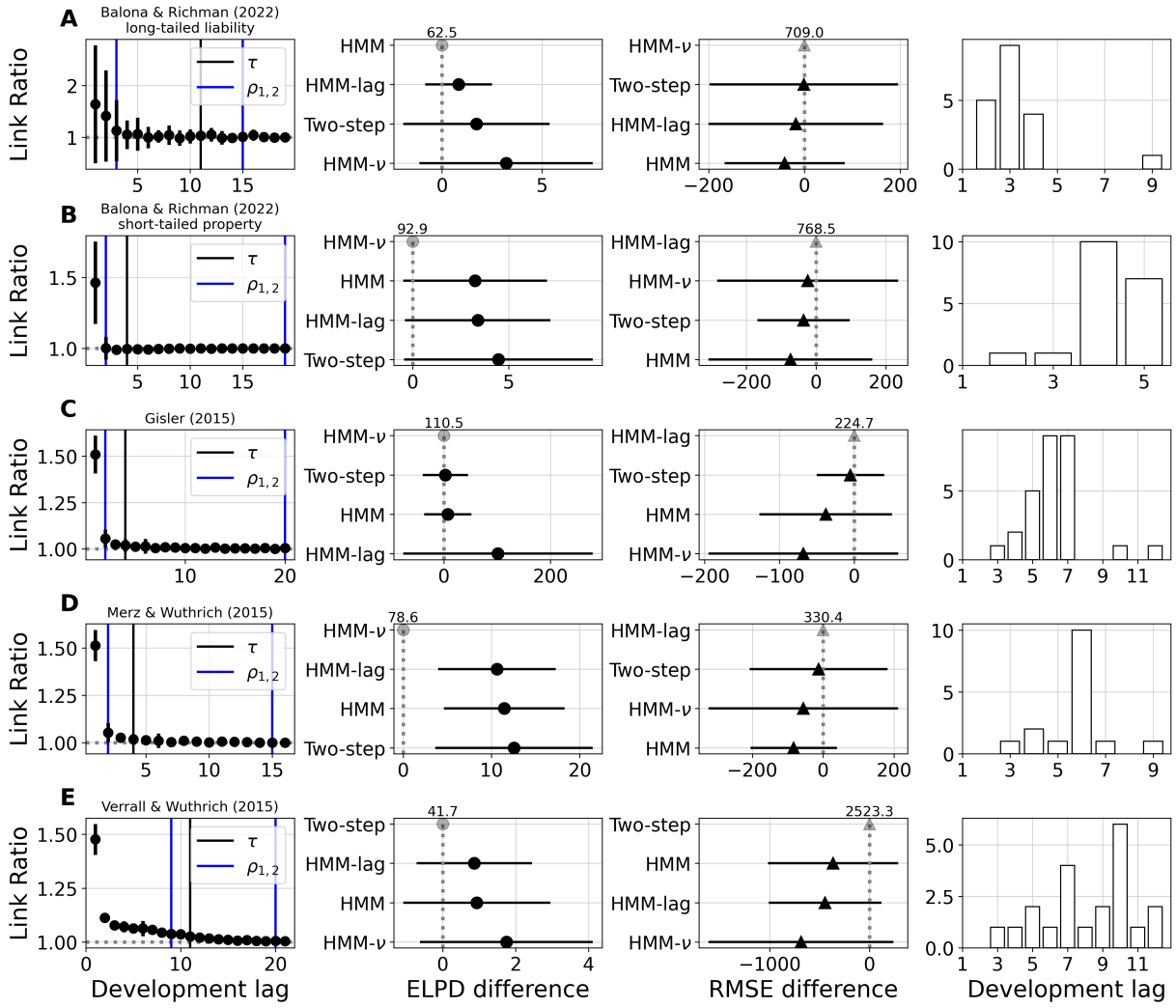

Figure 8: Results for the hidden Markov model and two-step approach on the five literature triangles (rows A - D). The first column shows the mean (+/- 2 standard deviations) of the empirical link ratios in the triangles. The black and blue vertical lines indicate τ and $\boldsymbol{\rho} = \rho_{1:2}$, the tail start and generalised Bondy model training windows, respectively (see model definitions, above). The second and third columns show the ELPD and RMSE differences (+/- 2 SE) from the best performing model (top model in each panel), calculated from predictions on the latest diagonal of data (out-of-sample) in the loss triangles. The absolute ELPD and RMSE of the best-performing models are shown above the top model, for reference. The final column shows the development lags where the probability of being in the tail was > 0.5 for the highest ELPD hidden Markov model variant.

values for τ and $\rho_{1,2}$ that are not representative of a particular experience period, the predictions from such a model could be extremely biased, as shown in the example of Figure 3. While the hidden Markov model may still make relatively poor predictions for those experience periods with little data (e.g. see the last column of panel A in Figure 3), uncertainty in the true latent state, $\boldsymbol{z}$, is more-accurately accounted for.

Although most of the hidden Markov model variants performed consistently better on average than the two-step approach on the curated industry datasets of Meyers (2015), the approaches were more similar on the five literature triangles. These two sets of data present different case studies. The industry triangles have been selected to encompass relatively large insurers with mostly stable loss dynamics (see Meyers, 2015, appendix A) over a period of 10 years. Due to the number of triangles, the two-step approach's cutoff points, τ and $\rho_{1,2}$, were chosen based on average empirical link ratios, which might not have been the best selection points for some triangles. By contrast, the literature triangles encompass more accident periods per triangle but also smaller books of business (e.g. the medium-sized triangles from Balona and Richman, 2022), and more variability in the tail than present in the industry triangles. Previous papers on loss development models combining body and tail dynamics have not considered the breadth of triangles and lines of business used here. For instance, England and Verrall (2001), Verrall and Wüthrich (2012), and Verrall and Wüthrich (2015) all used a single triangle to illustrate their approaches, and Zhang et al. (2012) used a dataset of 10 workers' compensation triangles and did not consider other lines of business or more volatile triangles. Moreover, the previous papers did not compare their approaches to the more common two-step approach applied in actuarial practice. The datasets used here are provided alongside this article in the repository for ease of access and comparison of other loss development modelling approaches.

Both the hidden Markov and two-step approaches demonstrated relatively poor calibration on the out-of-sample data from the Meyers (2015) dataset (Figure 6). Primarily, the out-of-sample predictions were often too uncertain, producing a predominance of percentiles falling within the central range of possible percentiles. Few articles have shown calibration plots from fully-Bayesian posterior distributions on out-of-sample loss development data, so this pattern of calibration requires further inspection in the literature. Meyers (2015) reports on calibration using the same data set, showing rel-

atively well-calibrated predictions. However, we note that Meyers (2015) calculated the percentile of the total ultimate losses in each triangle on a lognormal distribution with mean and variance informed by the total ultimate losses from their models. Thus, these are not directly comparable because the results above average percentiles over each posterior predictive distribution in the test data, not a distribution informed by just the average of posterior predictive distributions.

The hidden Markov models presented here could be extended in a number of useful ways. Notably, the transition matrix probabilities might be parametric or non-parametric functions of covariates, such as premium volume in each experience period or inflation levels in each calendar period, or include hierarchical effects for experience and development periods. The Bayesian framework, alongside the hidden Markov models implemented here and available in the supplementary material, make these extensions highly accessible. Additionally, the hidden Markov model framework is general enough to include any body or tail model, not just the chain ladder and generalised Bondy forms. For instance, there are a number of inverse power curves to use for tail modelling (CAS Tail Factor Working Party, 2013; Evans, 2015; Clark, 2017), and extensions and variations on the chain ladder model have been commonplace (England and Verrall, 2002). Although the analyses in this paper focused on paid losses, the same models could be applied to estimates of reported losses (i.e. paid loss plus estimates of reserve), or joint modelling of both paid and reported losses (e.g. see Zhang, 2010, for one approach).

5 Competing interests

The author declares no competing interests.

References

Al-Mudafer, M. T., Avanzi, B., Taylor, G., and Wong, B. (2022). Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics*, 105:144–174.

Balona, C. and Richman, R. (2022). The actuary and IBNR techniques: a machine learning approach. *Variance*.

Barnett, G. and Zehnwirth, B. (2000). Best estimates for reserves. In *Proceedings of the Casualty Actuarial Society*, volume 87, pages 245–321.

Beard, R. (1960). Three R's of insurance: risk, retention and reinsurance. *Journal of the Staple Inn Actuarial Society*, 15(6):399–421.

Bornhuetter, R. L. and Ferguson, R. E. (1972). The actuary and IBNR. In *Proceedings of the casualty actuarial society*, volume 59, pages 181–195.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(3):473–484.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of statistical software*, 76.

CAS Tail Factor Working Party (2013). The estimation of loss development tail factors: a summary report.

Clark, D. (2017). Estimation of inverse power parameters via GLM. *Actuarial Review, May-June 2017*.

Clarke, T. and Harland, N. (1974). A practical statistical method of estimating claims liability and claims cash flow. *ASTIN Bulletin: The Journal of the IAA*, 8(1):26–37.

De Alba, E. (2002). Bayesian estimation of outstanding claim reserves. *North American Actuarial Journal*, 6(4):1–20.

England, P. D. and Verrall, R. J. (2001). A flexible framework for stochastic claims reserving. In *Proceedings of the Casualty Actuarial Society*, volume 88, pages 1–38.

England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3):443–518.

England, P. D., Verrall, R. J., and Wüthrich, M. V. (2019). On the lifetime and one-year views of reserve risk, with application to IFRS 17 and Solvency II risk margins. *Insurance: Mathematics and Economics*, 85:74–88.

Evans, J. (2015). A continuous version of Sherman's inverse power curve model with simple cumulative development factor formulas. In *Casualty Actuarial Society E-Forum, Fall 2014-Volume*.

Friedland, J. (2010). Estimating unpaid claims using basic techniques. In *Casualty actuarial society*, volume 201.

Fröhlich, A. and Weng, A. (2018). Parameter uncertainty and reserve risk under Solvency II. *Insurance: Mathematics and Economics*, 81:130–141.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Gesmann, M. and Morris, J. (2020). Hierarchical compartmental reserving models. In *Casualty Actuarial Society*, page 4.

Gisler, A. (2009). The insurance risk in the SST and in Solvency II: modelling and parameter estimation. *Available at SSRN 2704364*.

Hesselager, O. (1994). A Markov model for loss reserving. *ASTIN Bulletin: The Journal of the IAA*, 24(2):183–193.

Kunce, J. and Chatterjee, S. (2017). A machine-learning approach to parameter estimation. *Virgina: CAS*.

Kuo, K. (2019). Deeptriangle: a deep learning approach to loss reserving. *Risks*, 7(3):97.

Lally, N. and Hartman, B. (2018). Estimating loss reserves using hierarchical Bayesian Gaussian process regression with input warping. *Insurance: Mathematics and Economics*, 82:124–140.

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T. A., Watanabe, Y. Y., Murgatroyd, M., and Papastamatiou, Y. P. (2017). Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, 8(2):161–173.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin: The Journal of the IAA*, 23(2):213–225.

Mack, T. (1994). Which stochastic model is underlying the chain ladder method? *Insurance: mathematics and economics*, 15(2-3):133–138.

Merz, M. and Wüthrich, M. V. (2015). Claims run-off uncertainty: the full picture. *Swiss Finance Institute Research Paper*, (14-69).

Meyers, G. (2015). Stochastic loss reserving using Bayesian MCMC models. Casualty Actuarial Society Arlington, VA.

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2023). Simulation-based calibration checking for Bayesian computation: the choice of test quantities shapes sensitivity. *Bayesian Analysis*, 1(1):1–28.

Munroe, D., Zehnwirth, B., and Goldenberg, I. (2018). Solvency capital requirement and the claims development result. *British Actuarial Journal*, 23:e15.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Scurfield, H. (1968). Motor insurance statistics. *Journal of the Staple Inn Actuarial Society*, 18(3):207–236.

Sherman, R. E. (1984). Extrapolating, smoothing and interpolating development factors. In *Proceedings of the Casualty Actuarial Society*, volume 71, pages 122–155.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Sivula, T., Magnusson, M., Matamoros, A. A., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv preprint arXiv:2008.10296*.

Stan Development Team (2024a). `CmdStan`: the command line interface to stan.

Stan Development Team (2024b). `CmdStanPy`: the python interface to cmdstan.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

Taylor, G., McGuire, G., and Greenfield, A. (2003). Loss reserving: past, present and future. *University of Melbourne Centre of Actuarial Studies Research Paper*, (109).

Taylor, G. C. (1977). Separation of inflation and other effects from the distribution of non-life insurance claim delays. *ASTIN Bulletin: The Journal of the IAA*, 9(1-2):219–230.

Taylor, G. C. and Ashe, F. R. (1983). Second moments of estimates of outstanding claims. *Journal of Econometrics*, 23(1):37–61.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Verrall, R. J. and Wüthrich, M. V. (2012). Reversible jump Markov chain monte carlo method for parameter reduction in claims reserving. *North American Actuarial Journal*, 16(2):240–259.

Verrall, R. J. and Wüthrich, M. V. (2015). Parameter reduction in log-normal chain-ladder models. *European Actuarial Journal*, 5:355–380.

Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*. John Wiley & Sons.

Zhang, Y. (2010). A general multivariate chain ladder model. *Insurance: Mathematics and Economics*, 46(3):588–599.

Zhang, Y., Dukic, V., and Guszcza, J. (2012). A Bayesian non-linear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 175(2):637–656.