risks

MDPI

Article

Approximation of Zero-Inflated Poisson Credibility Premium via Variational Bayes Approach

Minwoo Kim [1], Himchan Jeong [2,*] and Dipak Dey [3]

[1] Department of Statistics and Probability, Michigan State University, Wells Hall 619 Red Cedar Road, East Lansing, MI 48824, USA; kimminw3@msu.edu
[2] Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada
[3] Department of Statistics, University of Connecticut, 215 Glenbrook Rd. U-4120, Storrs, CT 06269, USA; dipak.dey@uconn.edu
* Correspondence: himchan_jeong@sfu.ca

Abstract: While both zero-inflation and the unobserved heterogeneity in risks are prevalent issues in modeling insurance claim counts, determination of Bayesian credibility premium of the claim counts with these features are often demanding due to high computational costs associated with a use of MCMC. This article explores a way to approximate credibility premium for claims frequency that follows a zero-inflated Poisson distribution via variational Bayes approach. Unlike many existing industry benchmarks, the proposed method enables insurance companies to capture both zero-inflation and unobserved heterogeneity of policyholders simultaneously with modest computation costs. A simulation study and an empirical analysis using the LGPIF dataset were conducted and it turned out that the proposed method outperforms many industry benchmarks in terms of prediction performances and computation time. Such results support the applicability of the proposed method in the posterior ratemaking practices.

Keywords: approximate credibility premium; claim frequency; posterior ratemaking; variational Bayes; zero-inflated Poisson distribution

Citation: Kim, Minwoo, Himchan Jeong, and Dipak Dey. 2022. Approximation of Zero-Inflated Poisson Credibility Premium via Variational Bayes Approach. *Risks* 10: 54. https://doi.org/10.3390/risks10030054

Academic Editors: Robin Van Oirbeek, Tim Verdonck, Florence Guillaume, Christopher Grumiau and Mina Mostoufi

Received: 21 January 2022
Accepted: 28 February 2022
Published: 3 March 2022

1. Introduction

Credibility premium has been widely used in actuarial practice to capture unobserved heterogeneity of policyholders via historical claim experiences. Traditionally, the Poisson-gamma random effects model has been used as a benchmark to model claim frequency with unobserved heterogeneity (Dionne and Vanasse 1989).

It is also well-known that the traditional Poisson-gamma random effects model can enjoy natural conjugacy between the underlying distribution and the prior distribution of the random effects so that both the posterior distribution of the random effects and predictive premiums are readily available in closed forms, which is quite effective to compute individual premium for millions of policyholders.

In spite of the aforementioned benefits, the traditional Poisson-gamma random effects model can be too restrictive due to natural indication of zero-inflation in claim frequency, which has been shown in many empirical studies including, but not limited to, Shi and Zhao (2020), Zhang et al. (2020), and Lee (2021).

Inspired by presence of zero-inflation in a longitudinal setting, Boucher and Denuit (2008) and Boucher et al. (2009) discussed possible use of zero-inflated Poisson models for panel data and subsequently derived credibility premiums under the proposed model. Zhao and Zhou (2012) used copula models to consider zero-inflation and time-dependence simultaneously. Lee and Shi (2019) is another example of using zero-inflated marginal distributions and copulas for longitudinal claims. Chen et al. (2019) considered a non-parametric approach to estimate the individual unobserved heterogeneity using zero-inflated Poisson likelihood

with fused LASSO penalty. Note that credibility premiums with these models are less computationally tractable compared to the traditional models such as Poisson-gamma random effects models.

While using a complicated model enables us to consider more realistic features of the observed data, calibration of such model may suffer from computational burden on the optimization and the opacity of the model, which makes the use of such model less attractive to the practitioners.

In this regard, there have been some attempts to approximate the Bayes credibility premium under a complicated model with relatively simpler form. Bühlmann credibility premium was proposed as a linear approximation of Bayes credibility premium (Bühlmann and Gisler 2006). Najafabadi (2010) and Najafabadi et al. (2012) considered a new approach to approximate the Bayes credibility premium with a simpler credibility premium via maximum entropy principle. However, their models do not include regression coefficients so that it may not capture observed heterogeneity in tariffication unlike the aforementioned literature. Oh et al. (2021) used similar approach to analyze impacts of historical frequency and severity on posterior ratemaking.

In this article, we propose a new approach to consider zero-inflation and time-dependence of claim frequency that provide a relatively simpler form of credibility premium via variational Bayes (VB) method. VB approach has received a lot of attention as a powerful alternative to Markov Chain Monte Carlo (MCMC) method. VB method is based on optimization, which provides the closest approximation to the true posterior (Jordan et al. 1999). Among a predetermined family of distributions, VB finds an optimal distribution using Kullback–Leibler (KL) divergence as a measure to characterize the dissimilarity. Ranganath et al. (2014) suggested a new optimization technique for VB using Monte Carlo (MC) estimates. Recently, Saha et al. (2020) proposed a geometric variational Bayes approach that uses L_2 distance instead of KL divergence.

This paper is organized as follows. In Section 2, we briefly review the concept of VB and specify our proposed model with details of optimization and premium calculation. Section 3 provides a simulation study to assess applicability of the proposed method. Section 4 presents estimation and validation results on an actual insurance dataset. We conclude the paper in Section 5 with a few remarks.

2. Proposed Methodology

2.1. Claim Frequency Model with Longitudinality and Zero-Inflation

Let N_{it} be the number of accidents for each policyholder i at time $t = 1, \ldots, T_i$. Insurance exposure $e_{it} \in [0, 1]$ and explanatory variables $\mathbf{x}_{it}$ are defined accordingly. Traditionally, claims frequency has been modeled by Poisson distribution with covariates as follows:

$$N_{it}|\mathbf{x}_{it}, e_{it} \overset{indep}{\sim} \mathcal{P}(\nu_{it}), \quad \text{where } \nu_{it} = e_{it} \exp(\mathbf{x}_{it}\alpha). \tag{1}$$

Based on usual longitudinality of Property and Casualty (P&C) insurance claim datasets, one can also consider the following extension by incorporating random effects as follows:

$$N_{it}|\theta_i \overset{indep}{\sim} \mathcal{P}(\nu_{it}\theta_i), \quad \text{where } \nu_{it} = e_{it} \exp(\mathbf{x}_{it}\alpha),\ \theta_i \sim \pi_N(\theta). \tag{2}$$

Note that we enforce a restriction on $\pi_N(\theta)$ so that $\mathbb{E}[\theta_i] = 1$, due to the identifiability issue.

By assuming $\pi_N(\theta) \propto \theta^{\gamma-1} \exp(-\theta\gamma)$, one can easily show that the predictive distribution of N_{i,T_i+1} is given as

$$N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i} \sim \mathcal{NB}\left(\sum_{t=1}^{T_i} N_{it} + \gamma, \frac{\nu_{i,T_i+1}}{\sum_{t=1}^{T_i+1} \nu_{it} + \gamma}\right),$$

so that

$$\mathbb{E}\left[N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i}\right] = \frac{\sum_{t=1}^{T_i} N_{it} + \gamma}{\sum_{t=1}^{T_i} \nu_{it} + \gamma} \nu_{i,T_i+1}, \tag{3}$$

which has been shown in actuarial literature including, but not limited to, Frangos and Vrontos (2001), Jeong (2020), and Jeong and Valdez (2020).

While use of the aforementioned Poisson-gamma model allows us to evaluate the individual predictive premium for a large portfolio at ease and naturally captures possible overdispersion, such a model cannot reflect the possibility of zero-inflation in claim frequency. Therefore, one can incorporate both zero-inflation and longitudinality of the claim frequency as follows:

$$N_{it}|\theta_i \overset{indep}{\sim} \mathcal{ZIP}(p_{it}, \nu_{it}\theta_i) \quad \text{where } p_{it} = \frac{\exp(\mathbf{x}_{it}\eta)}{1+\exp(\mathbf{x}_{it}\eta)}, \tag{4}$$

where $N \sim \mathcal{ZIP}(p, \nu)$ means

$$\mathbb{P}(N = n) = p \cdot \mathbb{1}_{\{n=0\}} + (1-p) \cdot \frac{\nu^n \exp(-\nu)}{n!}.$$

In spite of flexibility of the model in (4), we have some issues on the posterior analysis of θ with the proposed model. If we assume $\pi(\theta) \propto \theta^{\gamma-1}e^{-\theta\gamma}$, we cannot obtain the closed form expression of the posterior density $\pi(\theta_i|\mathcal{F}_{i,T_i}) := \pi(\theta_i|N_{i1}, \ldots, N_{i,T_i})$, which is defined as follows:

$$\pi(\theta_i|\mathcal{F}_{i,T_i}) = \frac{\pi(\theta_i)\prod_{t=1}^{T_i} p(N_{it}|\theta_i)}{\int_0^\infty \pi(\theta_i)\prod_{t=1}^{T_i} p(N_{it}|\theta_i)d\theta_i}. \tag{5}$$

By noting $\pi(\theta_i|\mathcal{F}_{i,T_i}) \propto \pi(\theta_i)\prod_{t=1}^{T_i} p(N_{it}|\theta_i)$, one can try to find posterior samples of θ_i by MCMC, which is usually quite time consuming and sometimes infeasible to be implemented for calculation of individual predictive premium in an insurance portfolio, which usually contains millions of policyholders. According to a numerical experiment in the Section 4 of Ahn et al. (2021a), it turns out that a use of MCMC for calculation of individual predictive premium requires excessive amounts of time compared to other methods.

In this regard, we have developed a new approach that are less computationally expensive and more accurate to approximate the true posterior $\pi(\theta_i|\mathcal{F}_{i,T_i})$ via a surrogate function using variational Bayes method. See the Section 2.2 for the details of our variational algorithm.

2.2. Variational Bayes

Variational Bayes (VB) is an optimization method for obtaining the best approximation to the true posterior among the predetermined family of distributions.

In the following, we propose a use of variational Bayes approach to approximate $\pi(\theta_i|\mathcal{F}_{i,T_i})$ defined in (5). Considering an insurance portfolio with M policyholders, we define prior distributions for the random effects $\theta = (\theta_1, \ldots, \theta_M)^\top$ as follows:

$$\begin{aligned} \pi(\theta) &= \prod_{i=1}^{M} \pi(\theta_i) \\ \pi(\theta_i) &\sim Gamma(\gamma, \gamma). \end{aligned} \tag{6}$$

Observe that we use the same values for both shape parameter and rate parameter in (6) such that $\mathbb{E}[\theta_i] = 1$ for $i = 1, \ldots, M$. Given θ, the likelihood of model (4) is written as:

$$L(N \mid \theta) = \prod_{i=1}^{M} L(N_i \mid \theta_i),$$
$$L(N_i \mid \theta_i) = \prod_{t=1}^{T_i} \left(\mathbb{1}_{\{N_{it}=0\}} p_{it} + (1 - p_{it}) \frac{(\nu_{it}\theta_i)^{N_{it}} e^{-\nu_{it}\theta_i}}{N_{it}!} \right),$$

where $N_i = (N_{i1}, \ldots, N_{i,T_i})^\top$. As the first step for using VB method, a family of distributions needs to be selected. We call the family of distribution **variational family (VF)** in which each distribution can be easily controlled by its own parameters. While there is no universal rule for the choice of VF family, one simple choice of VF is mean-field family (Blei et al. 2017), where the latent variables are mutually independent.

For the sake of containing observed or estimated information $(N_{it}, p_{it}, \nu_{it})$ and maintaining connection with the prior distribution, our choice of variational family is independent gamma family as follows:

$$\mathcal{Q} = \left\{ q(\theta; \gamma_q) = \prod_{i=1}^{P} q(\theta_i; \gamma_q) \mid q(\theta_i; \gamma_q) \sim Gamma\left(\gamma_q + \sum_{t=1}^{T} N_{it}, \gamma_q + \sum_{t=1}^{T} (1 - p_{it})\nu_{it} \right) \right\}.$$

Note that

$$\mathbb{E}_q[\theta_i] = \int_0^\infty \theta_i q(\theta_i; \gamma_q) d\theta_i = \frac{\gamma_q + \sum_{t=1}^{T_i} N_{it}}{\gamma_q + \sum_{t=1}^{T_i} (1 - p_{it})\nu_{it}},$$

which converges to the method of moment estimate $\widehat{\mathbb{E}\left[\theta_i | \mathcal{F}_{i,T_i}\right]} = \frac{\sum_{t=1}^{T_i} N_{it}}{\sum_{t=1}^{T_i}(1-p_{it})\nu_{it}}$ as $T_i \to \infty$ since $\theta_i = \frac{\mathbb{E}[N_{it}|\theta_i]}{(1-p_{it})\nu_{it}}$. As a special case of the proposed model, if p_{it}, the zero-inflation probability, equals 0 for all i and t, then the variational distribution $q(\theta; \gamma_q)$ coincides with the true posterior distribution.

We point out that a variational distribution $q(\theta; \gamma_q) \in \mathcal{Q}$ is characterized by a parameter γ_q referred to as the **variational parameter,** which will be updated in our VB algorithm. Although the variational family $\mathcal{Q}$ does not always contain the true posterior, we can find the optimal distribution $q(\theta; \gamma_q^*)$, which is closest to the true posterior in terms of Kullback–Leibler (KL) divergence:

$$KL(q \parallel \pi(\theta|\mathcal{F}_T)) = \mathbb{E}_q\left[\log q(\theta; \gamma_q) - \log \pi(\theta|\mathcal{F}_T)\right]. \tag{7}$$

Here, we define a function of the variational parameter:

$$\mathcal{L}(\gamma_q) := \mathbb{E}_q\left[\log \pi(\theta) + \log L(N \mid \theta) - \log q(\theta; \gamma_q)\right]. \tag{8}$$

The function (8) is called the Evidence Lower Bound (ELBO). One can easily show that minimizing (7) is equivalent to maximizing (8) (Wainwright and Jordan 2008) such that

$$q(\theta; \gamma_q^*) = \arg\max_{\gamma_q > 0} \mathcal{L}(\gamma_q).$$

In order to find γ_q^*, we employ a version of gradient descent algorithm, the stochastic optimization method in Ranganath et al. (2014). For this, it is needed to compute the gradient of the ELBO:

$$\begin{aligned}\nabla_{\gamma_q}\mathcal{L}(\gamma_q) &= \nabla_{\gamma_q}\int\{\log\pi(\theta)+\log L(N\mid\theta)-\log q(\theta;\gamma_q)\}q(\theta;\gamma_q)d\theta\\ &= \int\{\log\pi(\theta)+\log L(N\mid\theta)-\log q(\theta;\gamma_q)\}\nabla_{\gamma_q}q(\theta;\gamma_q)-q(\theta;\gamma_q)\nabla_{\gamma_q}\log q(\theta;\gamma_q)d\theta\\ &= \int\nabla_{\gamma_q}\log q(\theta;\gamma_q)\{\log\pi(\theta)+\log L(N\mid\theta)-\log q(\theta;\gamma_q)\}q(\theta;\gamma_q)d\theta\\ &\quad-\int q(\theta;\gamma_q)\nabla_{\gamma_q}\log q(\theta;\gamma_q)d\theta\\ &= \mathbb{E}_q\Big[\nabla_{\gamma_q}\log q(\theta;\gamma_q)\{\log\pi(\theta)+\log L(N\mid\theta)-\log q(\theta;\gamma_q)\}\Big],\end{aligned}\tag{9}$$

where ∇_{γ_q} means $\frac{\partial}{\partial\gamma_q}$. The third equality is due to the fact $\nabla_{\gamma_q}q(\theta;\gamma_q) = q(\theta;\gamma_q)\nabla_{\gamma_q}\log q(\theta;\gamma_q)$ and the last equality holds because the expected value of the score function is zero:

$$\begin{aligned}\int q(\theta;\gamma_q)\nabla_{\gamma_q}\log q(\theta;\gamma_q)d\theta &= \int\nabla_{\gamma_q}q(\theta;\gamma_q)d\theta\\ &= \nabla_{\gamma_q}\int q(\theta;\gamma_q)d\theta\\ &= \nabla_{\gamma_q}1=0.\end{aligned}$$

Below is the detailed derivation of (9):

$$\begin{aligned}\nabla_{\gamma_q}\log q(\theta;\gamma_q) &= \sum_{i=1}^{M}\nabla_{\gamma_q}\log q(\theta_i;\gamma_q),\\ \nabla_{\gamma_q}\log q(\theta_i;\gamma_q) &= \nabla_{\gamma_q}\left[\left(\gamma_q+\sum_{t=1}^{T}N_{it}\right)\log\left(\gamma_q+\sum_{t=1}^{T}(1-p_{it})\nu_{it}\right)-\log\Gamma\left(\gamma_q+\sum_{t=1}^{T}N_{it}\right)\right.\\ &\quad\left.+\left(\gamma_q+\sum_{t=1}^{T}N_{it}-1\right)\log\theta_i-\left(\gamma_q+\sum_{t=1}^{T}(1-p_{it})\nu_{it}\right)\theta_i\right]\\ &= \log\left(\gamma_q+\sum_{t=1}^{T}(1-p_{it})\nu_{it}\right)+\frac{\gamma_q}{\gamma_q+\sum_{t=1}^{T}(1-p_{it})\nu_{it}}+\frac{\sum_{t=1}^{T}N_{it}}{\gamma_q+\sum_{t=1}^{T}(1-p_{it})\nu_{it}}\\ &\quad-\frac{\Gamma'\left(\gamma_q+\sum_{t=1}^{T}N_{it}\right)}{\Gamma\left(\gamma_q+\sum_{t=1}^{T}N_{it}\right)}+\log\theta_i-\theta_i.\end{aligned}$$

With step sizes ρ_l, $l=0,1,2,\ldots$, which satisfy Robbins–Monro conditions (Robbins and Monro 1951), and using Monte Carlo (MC) estimate for (9), we iteratively update the given initial γ_q^0 until the ELBO converges:

$$\gamma_q^{l+1}\leftarrow\gamma_q^l+\rho_l\widehat{\nabla_{\gamma_q}}\mathcal{L}(\gamma_q),\ \ l=0,1,\ldots$$

where $\widehat{\nabla_{\gamma_q}}\mathcal{L}(\gamma_q)=\frac{1}{S}\sum_{s=1}^{S}\log\pi(\theta^s)+\log L(N\mid\theta^s)-\log q(\theta^s;\gamma_q)$ is the MC estimate for (9) based on the samples from the current variational distribution, $\theta^1,\ldots,\theta^S\sim q(\theta,\gamma_q^l)$.

3. Data and Results

To assess applicability of the proposed method, here we perform numerical studies using both simulated datasets and an actual insurance claim portfolio, which correspond to the aforementioned actuarial application where we have indication of zero-inflation in the claim counts observed over time. In this section, we first introduce the industry benchmarks for the ratemaking purpose. After that, both the proposed model and the benchmarks are calibrated with given data. Finally, the posterior premiums under each model are

computed and compared to the actual claim counts in the out-of-sample validation set to assess the prediction performances.

Note that all the calculations in this section and thereafter were performed using `R`, and a computer with Intel Core i7-8565U at 1.80 Ghz 4 cores, 16 GB memory.

3.1. Simulation Study

We generate $\{N_{it}\}_{i=1,\dots,5000,\, t=1,\dots,6}$ with the following hierarchical distributions:

$$N_{it}|\theta_i \sim \mathcal{ZIP}(p_{it},\, \nu_{it}\theta_i), \theta_i \sim \mathcal{G}(\gamma,\gamma),\ \text{and}\ X_{it} \sim \mathcal{N}(0,1),$$

where

$$p_{it} = \frac{\exp(\eta_0+\eta_1 X_{it})}{1+\exp(\eta_0+\eta_1 X_{it})},\ \nu_{it} = \exp(\alpha_0+\alpha_1 X_{it}),$$
$$\eta_0 = 1, \eta_1 = -2, \alpha_0 = -2.5, \alpha_1 = 2,\ \text{and}\ \gamma = 3.8.$$

We consider some frequency models and corresponding premium calculation for comparison. First, we use models without zero-inflation whose premium calculation are given as follows:

- Naive Poisson (NP): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = \hat{\nu}_{i,T_i+1}$.
- Poisson-Gamma (PG): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = \frac{\gamma^*+\sum_{t=1}^{T_i} N_{it}}{\gamma^*+\sum_{t=1}^{T_i}\hat{\nu}_{it}}\hat{\nu}_{i,T_i+1}$.

Note that NP and PG models are specified with the same mean structure so that we estimated $\hat{\alpha}_0$ and $\hat{\alpha}_1$ via `glm` function in `R`, which are still consistent regardless of possible misspecification in the working correlation structure. (Zeger et al. 1988) The estimation took around 0.07 s and the a priori premium $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}}\big] = \exp(\hat{\alpha}_0+\hat{\alpha}_1 X_{i,T_{i+1}})$ is the same in both NP and PG models while the posterior premiums vary.

Further, we also used models with zero-inflation whose premium calculations are given as follows:

- Naive ZIP (NZIP): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = (1-\hat{p}_{i,T_i+1})\hat{\nu}_{i,T_i+1}$.
- Proposed (VB): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = \frac{\gamma^*+\sum_{t=1}^{T_i} N_{it}}{\gamma^*+\sum_{t=1}^{T_i}(1-\hat{p}_{it})\hat{\nu}_{it}}(1-\hat{p}_{i,T_i+1})\hat{\nu}_{i,T_i+1}$.
- Bayes (BA): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = (1-\hat{p}_{i,T_i+1})\hat{\nu}_{i,T_i+1}\cdot\frac{1}{R}\sum_{r=1}^{R}\theta_i^{(r)}$ where $\{\theta_i^{(r)}\}_{r=1,\dots,R}$ are posterior samples of θ_i via MCMC. Note that the value of R should be large enough for the convergence of the posterior distribution while it also has a substantial impact on the computational time. To achieve a balance between the computational cost and prediction accuarcy, we set $R = 30{,}000$.
- True (TR): $\mathbb{E}\big[\widehat{N_{i,T_{i+1}}|\mathcal{F}_{i,T_i}}\big] = (1-\hat{p}_{i,T_i+1})\hat{\nu}_{i,T_i+1}\theta_i$.

Again, due to the same mean structure, $\hat{\eta}_0$, $\hat{\eta}_1$, $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are commonly estimated via `zeroinfl` function in `R` for ZIP, VB, BA, and PG models. γ^* is estimated via variational Bayes approach and used both in PG and VB models. The estimation took around 1.27 s and the a priori premium

$$\mathbb{E}\big[\widehat{N_{i,T_{i+1}}}\big] = \frac{\exp(\hat{\eta}_0+\hat{\eta}_1 X_{i,T_{i+1}})}{1+\exp(\hat{\eta}_0+\hat{\eta}_1 X_{i,T_{i+1}})}\exp(\hat{\alpha}_0+\hat{\alpha}_1 X_{i,T_{i+1}}),$$

is the same while the posterior premiums vary.

Note that our main interest is to establish a way to compute the predictive premium of $N_{i,T+1}$ given $N_{i,1},\dots,N_{i,T}$ with less computational cost. While one could estimate α and γ (note that θ_i are treated as random in our framework) via applying an EM algorithm as in Tzougas and Karlis (2020) or Tzougas and Jeong (2021), we chose a rather simpler way as in Pechon et al. (2019, 2020), to focus on different characterizations of $\mathbb{E}[\theta_i|N_{i,1},\dots,N_{i,T}]$ under different models. Note that the comparison of all models were done at the same

ground (the fixed effects were estimated in the same way as long as they have the same marginal mean structure) to assure that we make a fair comparison of the models.

By doing so, we focus on the efficiency of each model to incorporate the unobserved heterogeneity rather than estimation accuracy of the fixed effects. Note that the true model assumes perfect knowledge on the unobserved heterogeneity θ_i for each policyholder i (while it still allows for possible estimation errors in the fixed effects), which is not available in practice and only used as an (unattainable) benchmark.

After the benchmarks and the proposed model are specified, we assess the prediction performances of the models via root-mean squared error (RMSE), mean absolute error (MAE) that are defined as follows:

$$\begin{aligned} \text{RMSE} &: \sqrt{\frac{1}{M}\sum_{i=1}^{M}(N_{i,T_i+1} - \hat{N}_{i,T_i+1})^2}, \\ \text{MAE} &: \frac{1}{M}\sum_{i=1}^{M}|N_{i,T_i+1} - \hat{N}_{i,T_i+1}|. \end{aligned} \tag{10}$$

Note that RMSE and MAE measure the discrepancy between the actual values and predicted values in L_2 and L_1 norms, respectively, so that we prefer a model with lower values of RMSE and MAE. We also prefer a model with less computation time since it is required to evaluate individual posterior premium for a portfolio that consists of millions of policyholders in general. Out-of-sample validation results with the simulated data are provided in Table 1.

Table 1. Out-of-sample validation with simulated data.

	NP	**PG**	**NZIP**	**VB**	**BA**	**TR**
RMSE	5.1342	3.4719	4.1796	2.9404	3.3008	0.7636
MAE	0.4254	0.3940	0.4068	0.3768	0.3835	0.2822
Computation Time	0.07	1.57	1.27	380.33	6492.93	1.27

3.2. Case Study—Posterior Ratemaking with the LGPIF Data

For the empirical analysis, a public dataset on insurance claim provided by the Wisconsin Local Government Property Insurance Fund (LGPIF) is used. The dataset consists of claims information on multiple coverages and corresponding policy characteristics that have been observed from years 2006 to 2011. Among the information on multiple coverage, we only use inland marine (IM) claims information. Note that observations from years 2006–2010 are used to train the frequency models while observations from year 2011 are used to validate the trained models and compare their performance. Table 2 summarizes the distributions of covariates, which are used to determine rating factors for each policyholder via a regression model. For more detailed explanation and preliminary analysis of given dataset, see Frees et al. (2016). Note that the LGPIF dataset used here is a so-called traditional dataset while there has been emerging interest in new types of insurance claim datasets, which are high-dimensional and contain more information on the heterogeneity of the policyholders, such as telematics data (Gao et al. 2021). However, uses of such high-dimensional data and related models are still in a developing stage. Further, analyses of traditional datasets with limited range of features could be still meaningful, especially in the sense that the proposed method explores an efficient way to capture the unobserved heterogeneity of each policyholder that are not explainable by the available covariates.

Table 2. Observable policy characteristics used as covariates.

Categorical Variables	**Description**			**Proportions**
TypeCity	Indicator for city entity:	Y = 1		14%
TypeCounty	Indicator for county entity:	Y = 1		5.78%
TypeMisc	Indicator for miscellaneous entity:	Y = 1		11.04%
TypeSchool	Indicator for school entity:	Y = 1		28.17%
TypeTown	Indicator for town entity:	Y = 1		17.28%
TypeVillage	Indicator for village entity:	Y = 1		23.73%
NoClaimCreditIM	No IM claim in three consecutive prior years:	Y = 1		42.1%
Continuous Variables		**Minimum**	**Mean**	**Maximum**
CoverageIM	Log coverage amount of IM claim in mm	0	0.8483	46.7493
lnDeductIM	Log deductible amount for IM claim	0	5.340	9.210

Before we consider frequency models with zero-inflation, it is natural to test presence of zero-inflation from the data. To detect the zero-inflation in Poisson distribution, equivalently to test whether the zero-inflation parameter $p_{0i} = 0$ or not, Van den Broek (1995) showed that

$$S(\tilde{\alpha}) = \frac{(\sum_{i=1}^{M} \frac{\mathbb{1}_{\{N_i=0\}} - \tilde{p}_{0i}}{\tilde{p}_{0i}})^2}{(\sum_{i=1}^{M} \frac{1-\tilde{p}_{0i}}{\tilde{p}_{0i}}) - \sum_{i=1}^{M} N_i} \simeq \chi^2(1) \text{ under } H_0 : p_{0i} = 0,$$

where $\tilde{p}_{0i} = \exp(-\mathbf{x}_i\tilde{\alpha})$. With this data, the value of $S(\tilde{\alpha})$ for IM claims is 1.967 and corresponding p-value is 0.08038. Therefore, one can conclude that there is a not negligible level of overdispersion in IM claim frequencies.

Based on such indications, now we train the models that were used in Section 3 except for the true model, which assumes perfect knowledge on unobserved heterogeneity of the policyholders. As in the simulation study, we first estimate the fixed effects, only α for the models without zero-inflation via `glm` and η and α for the models with zero-inflation via `zeroinfl`, respectively. Table 3 summarizes the estimated coefficients for the fixed effects. Note that $\hat{\alpha}$ from the models without zero-inflation and $\hat{\alpha}$ from the ones with zero-inflation are not comparable due to the presence of covariate impacts on zero-inflation.

Table 3. Estimation results of the fixed effects.

	No ZI		**With ZI**			
	α		η		α	
	Estimate	***p*-Value**	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**
(Intercept)	−4.0315	0.0000	3.6900	0.0011	−0.7553	0.4308
TypeCity	0.9437	0.0000	1.8268	0.0903	1.7887	0.0080
TypeCounty	1.7300	0.0000	−0.2296	0.8584	1.3579	0.0485
TypeMisc	−2.7326	0.0071	0.9887	0.8437	−1.9120	0.6441
TypeSchool	-0.9172	0.0010	2.9200	0.0076	1.5831	0.0396
TypeTown	−0.3960	0.1531	1.4311	0.2196	0.7086	0.4187
CoverageIM	0.0664	0.0000	−0.2242	0.0002	0.0553	0.0000
lnDeductIM	0.1353	0.0031	−0.4826	0.0018	−0.2183	0.0509
NoClaimCreditIM	−0.3690	0.0049	−0.3249	0.4854	−0.5103	0.0746

After the fixed effects are estimated, we can compute the individual posterior premiums based on the covariates information in the out-of-sample validation set and claims history of each policyholder under the specified models as in Table 4.

Table 4. Out-of-sample validation result with actual dataset.

	NP	PG	NZIP	VB	BA
RMSE	0.3797	0.2334	0.2291	0.1794	0.1873
MAE	0.1267	0.1167	0.1174	0.1098	0.1133
Computation Time	0.02	1.32	0.64	71.54	1310.58

4. Discussion of the Results

According to the out-of-sample validation results of the simulation study in Table 1, it is observed that the proposed method, VB method outperforms NP, PG, NZIP, and BA models in terms of predictive performances. (Note that the true model is the best in terms of predictive performance as expected while it is not available in practice.) Even though the computation time under BA model was excessive, its performance improvement was not significant over other benchmarks. To assess the stability of the simulation results under the parameter changes, we also performed simulation studies with different sets of parameters and it showed the consistent pattern as in Table 1, while detailed results are omitted here. Note that the simulation studies conducted here are still restrictive due to pre-specified model structure and limited number of covariates so that there is no assurance that the proposed model outperforms the full Bayes model in general. As in the simulation study, the proposed model also shows the best performance on the prediction results with the LGPIF data in terms of both RMSE and MAE, and much less computation time compared to the BA model that uses MCMC to estimate the individual unobserved heterogeneity.

In that regard, the numerical illustrations given in Section 3 shows us the applicability of the proposed method as an acceptable approximation of the true unobserved heterogeneity that requires much less computation time compared to a naive use of MCMC in the presence of zero-inflation in claim counts.

5. Conclusions

As insurance companies are interested in better risk classification and tarrification by incorporating prevalent features of claims data such as indication of zero-inflation and the unobserved heterogeneity, computational costs in model calibration and individual premium calculation have been obstacles of using complicated models. To tackle this issue, we proposed a way to approximate the posterior density of the unobserved heterogeneity in risks with consideration of zero-inflation, which leads to an analytic form of the posterior premium. It was also shown that the proposed method can be used an alternative of full Bayes method due to its predictive performance and less computation cost.

The proposed approach is limited in the sense that the random effect θ, which captures the unobserved heterogeneity, is assumed to be static. It means there is no room for evolution of the unobserved risk characteristics of a policyholder over time under this model, which is somewhat unrealistic. While there are some research work focused on the use of dynamic random effects for determination of credibility premium (Ahn et al. 2021b; Pinquet 2020), calibration and prediction of dynamic random effects models are often computationally intensive and intractable. Therefore, as a direction for future research, one can expand the class of variational family so that impacts of dynamic random effects can be incorporated in the posterior premium calculation.

Author Contributions: Conceptualization, M.K. and H.J.; methodology, M.K. and H.J.; software, M.K. and H.J.; validation, M.K. and H.J.; resources, H.J.; data curation, H.J.; writing—original draft preparation, M.K. and H.J.; writing—review and editing, M.K., H.J. and D.D.; supervision, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: Himchan Jeong was supported from the SFU Faculty of Science start-up grant.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The LGPIF dataset that has been used in this article is publicly available at https://sites.google.com/a/wisc.edu/jed-frees/ (accessed on 10 January 2022). The codes used in the analyses of simulated dataset and the LGPIF dataset are available at https://github.com/ssauljin/VB_ZIP (accessed on 10 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LGPIF	Local Government Property Insurance Fund
KL	Kullback-Leibler
MCMC	Markov Chain Monte Carlo
VB	Variational Bayes
VF	Variational family

References

Ahn, Jae Youn, Himchan Jeong, and Yang Lu. 2021a. On the ordering of credibility factors. *Insurance: Mathematics and Economics* 101: 626–38.

Ahn, Jae Youn, Himchan Jeong, and Yang Lu. 2021b. A simple bayesian state-space model for the collective risk model. *arXiv* arXiv:2110.09657.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112: 859–77. [CrossRef]

Boucher, Jean-Philippe, and Michel Denuit. 2008. Credibility premiums for the zero-inflated poisson model and new hunger for bonus interpretation. *Insurance: Mathematics and Economics* 42: 727–35. [CrossRef]

Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillen. 2009. Number of accidents or number of claims? An approach with zero-inflated poisson models for panel data. *Journal of Risk and Insurance* 76: 821–46. [CrossRef]

Bühlmann, Hans, and Alois Gisler. 2006. *A course in Credibility Theory and Its Applications*. New York: Springer Science & Business Media.

Chen, Kun, Rui Huang, Ngai Hang Chan, and Chun Yip Yau. 2019. Subgroup analysis of zero-inflated poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics* 86: 8–18. [CrossRef]

Dionne, Georges, and Charles Vanasse. 1989. A generalization of automobile insurance rating models: The negative binomial distribution with a regression component. *ASTIN Bulletin: The Journal of the IAA* 19: 199–212. [CrossRef]

Frangos, Nicholas E., and Spyridon D. Vrontos. 2001. Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin: The Journal of the IAA* 31: 1–22. [CrossRef]

Frees, Edward W., Gee Lee, and Lu Yang. 2016. Multivariate frequency-severity regression models in insurance. *Risks* 4: 4. [CrossRef]

Gao, Guangyuan, Yanlin Shi, and He Wang. 2021. Telematics Car Driving Data Analytics. *SOA General Insurance Research Reports*. Available online: https://www.soa.org/globalassets/assets/files/resources/research-report/2021/telematics-driving-data.pdf (accessed on 10 January 2022).

Jeong, Himchan. 2020. Testing for random effects in compound risk models via bregman divergence. *ASTIN Bulletin: The Journal of the IAA* 50: 777–98. [CrossRef]

Jeong, Himchan, and Emiliano A. Valdez. 2020. Predictive compound risk models with dependence. *Insurance: Mathematics and Economics* 94: 182–95. [CrossRef]

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning* 37: 183–233. [CrossRef]

Lee, Gee Y., and Peng Shi. 2019. A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics* 87: 115–29. [CrossRef]

Lee, Simon C. K. 2021. Addressing imbalanced insurance data through zero-inflated poisson regression with boosting. *ASTIN Bulletin: The Journal of the IAA* 51: 27–55. [CrossRef]

Najafabadi, Amir T. Payandeh. 2010. A new approach to the credibility formula. *Insurance: Mathematics and Economics* 46: 334–38.

Najafabadi, Amir T. Payandeh, Hamid Hatami, and Maryam Omidi Najafabadi. 2012. A maximum-entropy approach to the linear credibility formula. *Insurance: Mathematics and Economics* 51: 216–21.

Oh, Rosy, Youngju Lee, Dan Zhu, and Jae Youn Ahn. 2021. Predictive risk analysis using a collective risk model: Choosing between past frequency and aggregate severity information. *Insurance: Mathematics and Economics* 96: 127–39. [CrossRef]

Pechon, Florian, Michel Denuit, and Julien Trufin. 2019. Multivariate modelling of multiple guarantees in motor insurance of a household. *European Actuarial Journal* 9: 575–602. [CrossRef]

Pechon, Florian, Michel Denuit, and Julien Trufin. 2020. Home and motor insurance joined at a household level using multivariate credibility. *Annals of Actuarial Science* 15: 82–114. [CrossRef]

Pinquet, Jean. 2020. Positivity properties of the arfima (0, d, 0) specifications and credibility analysis of frequency risks. *Insurance: Mathematics and Economics* 95: 159–65. [CrossRef]

Ranganath, Rajesh, Sean Gerrish, and David Blei. 2014. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Reykjavik: PMLR, pp. 814–22.

Robbins, Herbert, and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics* 22: 400–7. [CrossRef]

Saha, Abhijoy, Karthik Bharath, and Sebastian Kurtek. 2020. A geometric variational approach to bayesian inference. *Journal of the American Statistical Association* 115: 822–35. [CrossRef]

Shi, Peng, and Zifeng Zhao. 2020. Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims. *The Annals of Applied Statistics* 14: 357–80. [CrossRef]

Tzougas, George, and Dimitris Karlis. 2020. An em algorithm for fitting a new class of mixed exponential regression models with varying dispersion. *ASTIN Bulletin: The Journal of the IAA* 50: 555–83. [CrossRef]

Tzougas, George, and Himchan Jeong. 2021. An expectation-maximization algorithm for the exponential-generalized inverse gaussian regression model with varying dispersion and shape for modelling the aggregate claim amount. *Risks* 9: 19. [CrossRef]

Van den Broek, Jan. 1995. A score test for zero inflation in a poisson distribution. *Biometrics* 51: 738–43. [CrossRef]

Wainwright, Martin J., and Michael I. Jordan. 2008. Introduction to variational methods for graphical models. *Foundations and Trends in Machine Learning* 1: 1–103. [CrossRef]

Zeger, Scott L., Kung-Yee Liang, and Paul S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* 44: 1049–60. [CrossRef]

Zhang, Pengcheng, Enrique Calderin, Shuanming Li, and Xueyuan Wu. 2020. On the type i multivariate zero-truncated hurdle model with applications in health insurance. *Insurance: Mathematics and Economics* 90: 35–45. [CrossRef]

Zhao, Xiaobing, and Xian Zhou. 2012. Copula models for insurance claim numbers with excess zeros and time-dependence. *Insurance: Mathematics and Economics* 50: 191–99. [CrossRef]