TESTING FOR RANDOM EFFECTS IN COMPOUND RISK MODELS VIA BREGMAN DIVERGENCE

BY

HIMCHAN JEONG

ABSTRACT

The generalized linear model (GLM) is a statistical model which has been widely used in actuarial practices, especially for insurance ratemaking. Due to the inherent longitudinality of property and casualty insurance claim datasets, there have been some trials of incorporating unobserved heterogeneity of each policyholder from the repeated observations. To achieve this goal, random effects models have been proposed, but theoretical discussions of the methods to test the presence of random effects in GLM framework are still scarce. In this article, the concept of Bregman divergence is explored, which has some good properties for statistical modeling and can be connected to diverse model selection diagnostics as in Goh and Dey [(2014) *Journal of Multivariate Analysis*, **124**, 371–383]. We can apply model diagnostics derived from the Bregman divergence for testing robustness of a chosen prior by the modeler to possible misspecification of prior distribution both on the naive model, which assumes that random effects follow a point mass distribution as its prior distribution, and the proposed model, which assumes a continuous prior density of random effects. This approach provides insurance companies a concrete framework for testing the presence of nonconstant random effects in both claim frequency and severity and furthermore appropriate hierarchical model which can explain both observed and unobserved heterogeneity of the policyholders for insurance ratemaking. Both models are calibrated using a claim dataset from the Wisconsin Local Government Property Insurance Fund which includes both observed claim counts and amounts from a portfolio of policyholders.

KEYWORDS

Two-part frequency-severity model, compound random effects model, Bregman divergence, Bayesian sensitivity analysis, model selection, prior elicitation, generalized linear model (GLM), generalized Pareto (GP).

JEL codes: C11, C23, C52

Astin Bulletin 50(3), 777-798. doi:10.1017/asb.2020.19

TABLE 1

HYPOTHETICAL INFORMATION ON POLICYHOLDERS A AND B.

				Policyholder A		Policyholder B	
Year	Gender	Age	Vehicle size	No. of claim(s)	Claim amount	No. of claim(s)	Claim amount
2015	M	45	Medium	0	0	1	500
2016	M	46	Medium	0	0	2	4000
2017	M	47	Large	1	200	1	8000

1. INTRODUCTION

Traditionally, generalized linear models (GLMs) have been used as benchmarks in ratemaking of property and casualty (P&C) companies due to their interpretability and efficiency in modeling. In ratemaking with GLM, regression coefficients associated with the observable characteristics of policyholders (in other words, covariates) are estimated and used for future prediction of claims. But it is not possible to observe all the characteristics of policyholders which affect their risk profiles, such as driving habit. Since a policyholder can be observed repeatedly for many years by a P&C insurance company, one can try to capture the unobserved heterogeneity via random effects model.

Suppose we have the following information on policyholders A and B in Table 1, who are identical in terms of observable characteristics but show quite different patterns on their claims. This hypothetical example shows us that we might capture the unobserved heterogeneity in risk by observing the residuals after controlling for the effects of observed covariates, which can be explained in terms of random effects for policyholders A and B.

Because of the longitudinal property in most of P&C claim datasets, there have been some trials on the use of random effects model in actuarial science literature, which has a natural Bayesian interpretation. For example, Frangos and Vrontos (2001) tried to incorporate the random effects in bonus-malus system for automobile insurance and obtained a closed form formula for credibility premiums on compound loss, assuming the independence between the frequency and severity components. As an extension of their work, recently, Jeong *et al.* (2020) also explored a random effects model for auto insurance claims considering possible dependence between the frequency and severity components.

Although the presence of random effects in the hypothetical example is very clear, it can be less clear in real longitudinal datasets observed by an insurance company. Therefore, one should be careful to incorporate random effects in a ratemaking model because it may capture random noise as unobserved heterogeneity via random effects so that the model has unnecessary complexity. However, a theoretical approach has not been attempted for testing the presence of random effects in claim modeling.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Intuitively, assuming absence of random effects on the heterogeneity of risk profiles for policyholders is equivalent to set the multiplicative random effects for all policyholders as a constant, which means the use of a point mass prior for random effects. Therefore, one can see that it is possible to test the presence of random effects in a longitudinal dataset via prior elicitation in Bayesian statistics. Bayesian inference requires to have an assumed prior distribution, which represents any a priori beliefs or uncertainties about the parameters. According to Dubitzky *et al.* (2013), "Elicitation is the process of extracting knowledge, beliefs, and uncertainties about unknown quantities from the client so that these can be expressed as a prior probability distribution." In that sense, if a point mass prior is believed to be appropriate (or we have a strong belief that there is no presence of unobserved heterogeneity in the risks of policyholders), then it tells us that we can ignore the random effects in the modeling, whereas if a continuous prior is deemed to be suitable (or we have a strong belief that the impact of unobserved heterogeneity in the risks of policyholders is significant), then random effects are incorporated naturally in claim modeling.

It is possible to see some works on prior elicitation and Bayesian sensitivity analysis in actuarial literature though there is no previous direct work on testing the presence of random effects. For example, Gómez-Déniz *et al.* (1999) and Gómez-Déniz *et al.* (2000) performed Bayesian sensitivity analysis on Poisson-gamma frequency model to investigate how sensitive the posterior distribution of interest is to changes in prior distribution based on Esscher premium principle and variance premium principle, respectively.

In this article, Bregman divergence, proposed by Goh and Dey (2014), is used as a Bayesian model diagnostics for testing the robustness of a chosen prior. Since it is hardly possible to know the true prior distribution in general, we want the posterior distribution based on a chosen prior by the modeler would not deviate too much from the true posterior distribution regardless of the true prior. Therefore, if the posterior distribution based on a continuous prior shows less deviation from the true posterior distribution compared to the posterior distribution based on a point mass prior, then we can favor a continuous prior as the more robust one to possible misspecification of prior distribution and incorporate nonconstant random effects in our model accordingly. This idea is applied to actuarial science so that we can test the presence of random effects in a longitudinal claim dataset and suggest a sophisticated framework for ratemaking model selection.

This paper has been organized as follows. In Section 2, the two-part compound risk model is introduced and the models to be tested upon the presence of random effects are specified. In Section 3, the concept of Bregman divergence is introduced as well as the interpretation of Bregman divergence as a diagnostic for robustness of a chosen prior. In Section 4, description of the characteristics of the dataset and results of Bayesian sensitivity analysis are provided, which support the use of a continuous prior on the random effects rather than the use of a point mass. A conclusion is made in Section 5.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

2. LONGITUDINAL TWO-PART COMPOUND RISK MODEL

Suppose that we have available information on M policyholders for T years. Then we can define the number of claims for the policyholder $i \in \{1, 2, \ldots, M\}$ in year t as N_{it}. Likewise, the claim amount of the k^{th} accident (where $k \leq N_{it}$) for the policyholder i in year t can be defined as Y_{itk}. Furthermore, we can define the exposure $e_{it} \in [0, 1]$ which means the proportion of coverage period within the calendar year t for the policyholder i. Finally, we may define covariates $\mathbf{x}_{it}$, which often include age, gender, vehicle type, building type, building location, driving history, and so forth. Note that each policyholder is followed up to $T_i \leq T$. Here T_i means the number of insurance years for a specific policyholder i. Since it is not unusual for a policyholder to switch his/her insurance company once the automobile insurance contract expires, it is possible that T_i varies for each policyholder.

For ratemaking in P&C insurance, it is of our interest to predict the following total cost of claims for each policyholder i in year t:

$$S_{it} = \begin{cases} \sum_{k=1}^{N_{it}} Y_{itk}, & N_{it} \neq 0, \\ 0, & N_{it} = 0. \end{cases}$$

Then one can use two-part model to predict the number of claims N_{it} and the average claim amount C_{it} with the following decomposition of the joint density into the frequency part and conditional severity part:

$$f(N_{it}, C_{it}|\mathbf{x}_{it}) = f(N_{it}|\mathbf{x}_{it}) \times f(C_{it}|N_{it}, \mathbf{x}_{it}).$$

Note that C_{it} is defined as

$$C_{it} = \begin{cases} \frac{1}{N_{it}} \sum_{k=1}^{N_{it}} Y_{itk}, & N_{it} \neq 0, \\ \text{Undefined}, & N_{it} = 0. \end{cases}$$

Here "Undefined" can be understood as "NA" because without observing any accident (in order words, $N = 0$), there is no way to observe the average claim amount per claim.

2.1. Frequency part model

In actuarial practices, Poisson distribution has been used for the calibration of the number of claims with the presence of covariates as follows:

$$N_{it}|\mathbf{x}_{it}, e_{it} \overset{indep}{\sim} \mathcal{P}(v_{it}) \quad \text{where} \quad v_{it} = e_{it} \exp(\mathbf{x}_{it}\alpha), \tag{2.1}$$

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

which means N_{it} follows a Poisson distribution with mean ν_{it} and is independent of $N_{i't'}$ as long as $i \neq i'$ or $t \neq t'$ given the information on covariates and exposure.

Note that conditioning argument on $\mathbf{x}_{it}, e_{it}$ is suppressed afterward for notational convenience. Although this approach has been widely used due to its simplicity, the longitudinal property of usual claim datasets allows us to consider the unobserved heterogeneity of the policyholders via random effects as follows:

$$N_{it}|\theta_{N[i]} \overset{indep}{\sim} \mathcal{P}(\nu_{it}\theta_{N[i]}) \quad \text{where} \;\; \nu_{it} = e_{it}\exp(\mathbf{x}_{it}\alpha), \;\; \theta_{N[i]} \sim \pi_N(\theta), \tag{2.2}$$

which has been explored by some authors, such as Boucher *et al.* (2008).

We can see that this random effects approach has a good Bayesian interpretation because $\theta_{N[i]}$ is not observable so that we need to assume a prior on this. Furthermore, (2.1) can be interpreted as a special case of (2.2) with $\mathbb{P}(\theta_{N[i]} = \theta) = \mathbb{1}_{\{\theta=1\}}$ for all i where $\mathbb{1}_{\{\theta=1\}} = 1$ if $\theta = 1$ and $\mathbb{1}_{\{\theta=1\}} = 0$ otherwise.

Therefore, model selection between (2.1) and (2.2) is equivalent to the prior elicitation of $\pi_N(\theta)$. Since the impact of unobserved heterogeneity on claim frequency is usually unknown so that it needs to be assessed by the observed claim frequency, it is desirable to use noninformative prior on θ_N, which has less impact on our Bayesian analysis. Therefore, as in a lot of Bayesian literature including but not limited to Jeffreys (1946) and Berger (1985), it is natural to consider the use of the Jeffreys' prior as a candidate of noninformative prior on θ_N. Note that the Jeffreys' prior for a random variable Z with density $f(z|\theta)$ is defined as the square root of the Fisher information $I(\theta)$, where $I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f(z|\theta)\right]$.

Suppose $N|\theta_N \sim \mathcal{P}(\nu\theta_N)$ where $\theta_N > 0$. Then the Jeffreys' prior of θ_N is given as $\pi_N(\theta) = \theta^{-1/2}$, and the corresponding posterior distribution is gamma distribution with the following density:

$$\pi_N(\theta|N) \propto \theta^{N-1/2}\exp(-\nu\theta) \;\; \text{so that} \;\; \theta_N|N \sim \mathcal{G}(N + 0.5, \; \nu^{-1}),$$

because it is easy to see that

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log p(N|\theta)\right] = \nu/\theta \implies \pi_N(\theta) = \theta^{-1/2} \propto \sqrt{I(\theta)}$$

and

$$\pi_N(\theta|N) \propto \pi_N(\theta)p(N|\theta) \propto \theta^{N-1/2}\exp(-\nu\theta) \implies \theta_N|N \sim \mathcal{G}(N + 0.5, \; \nu^{-1}).$$

Note that although the Jeffreys' prior is improper, the corresponding posterior is still proper. However, even though we can have a prior with less information and proper posterior, we hope the mean of θ_N to be one because θ_N is a multiplicative random effect and it gives rise to an identifiability issue if $\mathbb{E}[\theta_N] \neq 1$. Therefore, we want to impose $\mathbb{E}[\theta_N] = 1$ as shown in Ng and Cook (2000) and Ding and Wang (2008) while we retain the same distribution on the posterior.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Hence, we may propose the following prior on θ_N, which satisfies $\mathbb{E}[\theta_N] = 1$ as well as has conjugate gamma posterior.

$$\pi_N(\theta) \propto \theta^{r-1} e^{-\theta r} \text{ so that } \theta_{N[i]} \overset{i.i.d.}{\sim} \mathcal{G}(r, 1/r),\ \mathbb{E}\left[\theta_{N[i]}\right] = 1, \text{ and } \mathrm{V}ar\left[\theta_{N[i]}\right] = \frac{1}{r}. \tag{2.3}$$

Now, according to the aforementioned arguments, we can formulate the model selection in case of frequency part as follows:

[Naive Frequency Model]

Data likelihood: $N_{it}|\theta_{N[i]} \overset{indep}{\sim} \mathcal{P}(v_{it}\theta_{N[i]})$.

Prior distribution: $\mathbb{P}(\theta_i^N = \theta) = \mathbb{1}_{\{\theta=1\}}$ for all i.

Posterior distribution:

$$\begin{aligned}\mathbb{P}(\theta_i^N = \theta | N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}) &\propto \mathbb{P}(\{\theta_i^N = \theta\} \cap \{N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}\}) \\ &\propto \mathbb{1}_{\{\theta=1\}}.\end{aligned}$$

Predictive distribution:

$$\begin{aligned}p(N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i}) &= \int p(N_{i,T_i+1}|\theta)\pi_N(\theta|N_{i1}, N_{i2}, \ldots, N_{iT_i})d\theta \\ &= p(N_{i,T_i+1}|\theta = 1).\end{aligned}$$

Therefore, we can see that $N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i} \sim \mathcal{P}(v_{i,T_i+1})$, which means that predictive distribution of N_{i,T_i+1} does not depend on the previous claim frequency observation due to the marginal independence among $N_{i,t}$.

[Proposed Frequency Model]

Data likelihood: $N_{it}|\theta_{N[i]} \overset{indep}{\sim} \mathcal{P}(v_{it}\theta_{N[i]})$.

Prior distribution: $\pi_N(\theta) \propto \theta^{r-1} e^{-\theta r}$ so that $\theta_{N[i]} \sim \mathcal{G}(r, 1/r)$ and $\mathbb{E}\left[\theta_{N[i]}\right] = 1, \mathrm{V}ar\left[\theta_{N[i]}\right] = \frac{1}{r}$. Therefore, as $r \to \infty$, $\pi_N(\theta)$ degenerates to the Dirac delta function at $\theta = 1$ which means that the naive frequency model is a merely limiting case of the proposed frequency model. According to Lemaire (1998), the observed number of claims from previous years has been widely used as an adjustment weight factor to penalize or provide bonus on policyholders, which is analogous to the empirical estimates of the values of random effects on claim frequency. Moreover, the range of adjustment weight factor on frequency premiums is usually from 54% to 200%. Therefore, it is natural to incorporate this knowledge on choosing the hyperparameter r for our proposed prior so that the 95% highest posterior density (HPD) interval of θ_N can include (0.54, 2.00). Thus, $r = 3.8$ is used as the hyperparameter so that 95% HPD interval of θ_N under the proposed prior can be around (0.16, 2.01).

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Posterior distribution:

$$\pi_N(\theta|N_{i1}, N_{i2}, \ldots, N_{iT_i}) \propto \pi_N(\theta) \prod_{t=1}^{T_i} p(N_{it}|\theta) \propto \theta^{r-1} e^{-\theta r} \left(\prod_{t=1}^{T_i} \theta^{N_{it}} \right) e^{-\sum_{t=1}^{T_i} v_{it}\theta}$$
$$\propto \theta^{\sum_{t=1}^{T_i} N_{it}+r-1} e^{-\theta(\sum_{t=1}^{T_i} v_{it}+r)},$$

so that $\theta_{N[i]}|N_{i1}, N_{i2}, \ldots, N_{iT_i} \sim \mathcal{G}(\sum_{t=1}^{T_i} N_{it} + r, [\sum_{t=1}^{T_i} v_{it} + r]^{-1})$.

Predictive distribution:

$$p(N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i}) = \int p(N_{i,T_i+1}|\theta)\pi_N(\theta|N_{i1}, N_{i2}, \ldots, N_{iT_i})d\theta$$
$$= \binom{\sum_{t=1}^{T_i+1} N_{it} + r - 1}{N_{i,T_i+1}} \left(\frac{\sum_{t=1}^{T_i} v_{it} + r}{\sum_{t=1}^{T_i+1} v_{it} + r} \right)^{\sum_{t=1}^{T_i} N_{it}+r}$$
$$\times \left(\frac{v_{i,T_i+1}}{\sum_{t=1}^{T_i+1} v_{it} + r} \right)^{N_{i,T_i+1}}.$$

Therefore, we can see that

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

so that $\mathbb{E}\left[N_{i,T_i+1}|N_{i1}, N_{i2}, \ldots, N_{iT_i}\right] = \frac{\sum_{t=1}^{T_i} N_{it}+r}{\sum_{t=1}^{T_i} v_{it}+r} v_{i,T_i+1}$.

Note that individual premium on the frequency component depends on random effect θ_N as well as the covariate information at time t, which is associated with the regression coefficient α so that we have the following:

$$\mathbb{E}[N_{i,T_i+1}|N_{i,1}, \ldots, N_{iT_i}] = \exp(\mathbf{x}_{i,T_i+1}\alpha)\, \mathbb{E}[\theta_{N[i]}|N_{i1}, \ldots, N_{iT_i}],$$

which means posterior mean of $\theta_{N[i]}$ is not the same as the predictive mean of N_{i,T_i+1} given $N_{i1}, \ldots, N_{iT}$. Furthermore, knowing predictive distribution of $N_{i,T_i+1}|N_{i1}, \ldots, N_{iT}$ could be useful since $\mathbb{E}[N_{i,T_i+1}e^{\gamma N_{i,T_i+1}}|N_{i1}, \ldots, N_{iT}]$ needs to be evaluated in order to obtain $\mathbb{E}[S_{i,T_i+1}|\, N_{i1}, \ldots, N_{iT_i}, C_{i1}, \ldots, C_{iT_i}]$ with dependence between the frequency and severity components.

2.2. Severity part model

Traditionally, gamma distribution has been used for the calibration of the average claim amount with the presence of covariates as follows:

$$C_{it}|\mathbf{x}_{it}, N_{it} \overset{indep}{\sim} \mathcal{G}(\psi_{it}, \mu_{it}/\psi_{it}) \quad \text{where } \mu_{it} = \exp(\mathbf{x}_{it}\beta + \gamma N_{it}),\ \psi_{it} = N_{it}/\phi. \tag{2.4}$$

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Note that conditioning argument on $\mathbf{x}_{it}, N_{it}$ is suppressed afterward for notational convenience. Again, the longitudinal property of usual claim datasets allows us to consider the unobserved heterogeneity of the policyholders via random effects as follows:

$$C_{it}|\theta_{C[i]} \overset{indep}{\sim} \mathcal{G}(\psi_{it}, \theta_{C[i]}\mu_{it}/\psi_{it}) \quad \text{where} \quad \mu_{it} = \exp(\mathbf{x}_{it}\beta + \gamma N_{it}),$$
$$\psi_{it} = N_{it}/\phi, \ \theta_{C[i]} \sim \pi_C(\theta). \tag{2.5}$$

Unlike the traditional approach for compound loss model which assumes independence between the frequency and severity components, here the observed frequency is also used as an explanatory variable for the average severity to capture the possible dependence between the frequency and the average severity. Although the independence assumption between the frequency and severity has been widely used due to its simplicity, recent research works in actuarial science show empirical evidences of dependence between the frequency and severity in various claim datasets. For the detailed approach, please see Garrido *et al.* (2016) and Jeong *et al.* (2020). We also have a good Bayesian interpretation in this case so that (2.4) can be interpreted as a special case of (2.5) with $\mathbb{P}(\theta_i^C = \theta) = \mathbb{1}_{\{\theta=1\}}$ for all i.

Therefore, model selection between (2.4) and (2.5) is equivalent to the prior elicitation of $\pi_C(\theta)$ and we again consider the use of the Jeffreys' prior as a candidate of noninformative prior on θ_C as follows. Suppose $C|\theta_C \sim \mathcal{G}(\psi, \mu\theta_C/\psi)$ where $\theta_C > 0$. Then the Jeffreys' prior of θ_C is given as $\pi_C(\theta) = \theta^{-1}$, and the corresponding posterior distribution is inverse gamma distribution with the following density:

$$\pi_C(\theta|C) \propto \left(\frac{1}{\theta}\right)^{-\psi-1} \exp\left(-\frac{\psi C\mu^{-1}}{\theta}\right) \ \text{ so that } \ \theta_C|C \sim \mathcal{IG}(\psi, \ \psi C\mu^{-1}),$$

since it is easy to see that

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f(C|\theta)\right] = \psi/\theta^2 \implies \pi_C(\theta) = \theta^{-1} \propto \sqrt{I(\theta)}$$

and

$$\pi_C(\theta|C) \propto \pi_C(\theta) f(C|\theta) \propto \left(\frac{1}{\theta}\right)^{-\psi-1} \exp\left(-\frac{\psi C\mu^{-1}}{\theta}\right)$$
$$\implies \theta_C|C \sim \mathcal{IG}(\psi, \ \psi C\mu^{-1}).$$

Note that although the Jeffreys' prior is improper, the corresponding posterior is still proper. However, again we hope the mean of θ_C to be one because θ_C is

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

a multiplicative random effect. Hence, we may propose the following prior on θ_C, which satisfies $\mathbb{E}[\theta_C] = 1$ as well as has conjugate inverse gamma posterior.

$$\pi_C(\theta) \propto \theta^{-k-2} e^{-k/\theta} \text{ so that } \theta_{C[i]} \overset{i.i.d.}{\sim} \mathcal{IG}(k+1, k),$$
$$\mathbb{E}\left[\theta_{C[i]}\right] = 1, \text{ and } \mathrm{V}ar\left[\theta_{C[i]}\right] = \frac{1}{k-1}. \tag{2.6}$$

Therefore, we can formulate the model selection in case of the average severity part as follows:

[Naive Severity Model]

Data likelihood: $C_{it}|\theta_{C[i]} \overset{indep}{\sim} \mathcal{G}(\psi_{it}, \theta_{C[i]}\mu_{it}/\psi_{it})$.

Prior distribution: $\mathbb{P}(\theta_i^C = \theta) = \mathbb{1}_{\{\theta=1\}}$ for all i.

Posterior distribution:

$$\mathbb{P}(\theta_i^C = \theta | C_{i1} = c_{i1}, \ldots, C_{iT_i} = c_{iT_i}) \propto \mathbb{P}(\{\theta_i^C = \theta\} \cap \{C_{i1} = c_{i1}, \ldots, C_{iT_i} = c_{iT_i}\})$$
$$\propto \mathbb{1}_{\{\theta=1\}}.$$

Predictive distribution:

$$f(C_{i,T_i+1}|C_{i1}, C_{i2}, \ldots, C_{iT_i}) = \int f(C_{i,T_i+1}|\theta)\pi_C(\theta|C_{i1}, C_{i2}, \ldots, C_{iT_i})d\theta$$
$$= f(C_{i,T_i+1}|\theta = 1).$$

Therefore, we can see that $C_{i,T_i+1}|C_{i1}, C_{i2}, \ldots, C_{iT_i} \sim \mathcal{G}(\psi_{i,T_i+1}, \mu_{i,T_i+1}/\psi_{i,T_i+1})$, which means that predictive distribution of C_{i,T_i+1} does not depend on the previous claim severity observations due to the marginal independence among $C_{i,t}$.

[Proposed Severity Model]

Data likelihood: $C_{it}|\theta_{C[i]} \overset{indep}{\sim} \mathcal{G}(\psi_{it}, \theta_{C[i]}\mu_{it}/\psi_{it})$.

Prior distribution: $\pi_C(\theta) \propto \theta^{-k-2} e^{-k/\theta}$ so that $\theta_{C[i]} \sim \mathcal{IG}(k+1, k)$ and $\mathbb{E}\left[\theta_{C[i]}\right] = 1, \mathrm{V}ar\left[\theta_{C[i]}\right] = \frac{1}{k-1}$. Therefore, as $k \to \infty$, $\pi_C(\theta)$ degenerates to the Dirac delta function at $\theta = 1$, which means that the naive severity model is a merely limiting case of the proposed severity model. According to the Lemaire (1998), most of countries except for South Korea do not use historically observed claim amounts for the construction of penalty or bonus on a policyholder, which supports the assertion that there is less variability on θ_C, the random effect of the severity component than on θ_N, which is the random effect of the frequency component. Therefore, $k = 11$ is used so that the 95% HPD interval of θ_C under the proposed prior can be around $(0.49, 1.61)$, which is narrower than the 95% HPD interval of θ_N under the proposed prior.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Indeed, if empirical Bayes method is applied by maximizing the marginal likelihood with respect to both β and k where initial value of k as 11, then the optimal k is given as 11.00226. However, since five digits of decimal might give a false feeling of precision and 11.00226 is not much different from 11, $k = 11$ is used as the value of hyperparameter throughout this article.

Posterior distribution:

$$\pi_C(\theta|C_{i1}, C_{i2}, \ldots, C_{iT_i}) \propto \pi_C(\theta)\prod_{t=1}^{T_i} f(C_{it}|\theta) \propto \theta^{-k-2}e^{-k/\theta}\left(\prod_{t=1}^{T_i}\theta^{-\psi_{it}}\right)e^{-\left(\sum_{t=1}^{T_i}\frac{\psi_{it}C_{it}}{\mu_{it}}\right)/\theta}$$
$$\propto \theta^{-\left(\sum_{t=1}^{T_i}\psi_{it}+k+2\right)}e^{-\left(\sum_{t=1}^{T_i}\frac{\psi_{it}C_{it}}{\mu_{it}}+k\right)/\theta},$$

so that $\theta_{C[i]}|C_{i1}, C_{i2}, \ldots, C_{iT_i} \sim \mathcal{IG}(k + \sum_{t=1}^{T_i}\psi_{it} + 1, \sum_{t=1}^{T_i}\frac{\psi_{it}C_{it}}{\mu_{it}} + k)$.

Predictive distribution:

$$f(C_{i,T_i+1}|C_{i1}, C_{i2}, \ldots, C_{iT_i}) = \int f(C_{i,T_i+1}|\theta)\pi_C(\theta|C_{i1}, C_{i2}, \ldots, C_{iT_i})d\theta$$
$$= \frac{(\psi_{i,T_i+1}C_{i,T_i+1}/\mu_{i,T_i+1})^{\psi_{i,T_i+1}}}{(k+\sum_{t=1}^{T_i+1}\psi_{i,t}C_{i,t}/\mu_{i,t})^{\sum_{t=1}^{T_i}\psi_{i,t}+k+1}}$$
$$\times \frac{\Gamma(\sum_{t=1}^{T_i}\psi_{i,t}+k+1)C_{i,T_i+1}^{-1}}{\Gamma(\psi_{i,T_i+1})\Gamma(\sum_{t=1}^{T_i}\psi_{i,t}+k+1)}.$$

Therefore, we can see that

$$C_{i,T_i+1}|C_{i1}, C_{i2}, \ldots, C_{iT_i} \sim \mathcal{GP}\left(k+\sum_{t=1}^{T_i}\psi_{it}+1, \left[k+\sum_{t=1}^{T_i}\psi_{i,t}\frac{C_{it}}{\mu_{it}}\right]\frac{\mu_{i,T_i+1}}{\psi_{i,T_i+1}}, \psi_{i,T_i+1}\right),$$

with predictive mean

$$\mathbb{E}\left[C_{i,T_i+1}|C_{i1}, C_{i2}, \ldots, C_{iT_i}\right] = \frac{\left[k+\sum_{t=1}^{T_i}\psi_{it}\frac{C_{it}}{\mu_{it}}\right]}{k+\sum_{t=1}^{T_i}\psi_{it}}\mu_{i,T_i+1}$$
$$= \frac{\left[k\phi+\sum_{t=1}^{T_i}S_{it}/\mu_{it}\right]}{k\phi+\sum_{t=1}^{T_i}N_{it}}\mu_{i,T_i+1},$$

since $\psi_{it}C_{it} = N_{it}C_{it}/\phi = S_{it}/\phi$.

Note that generalized Pareto (GP) distribution is defined with the following density as in Klugman *et al.* (2012):

$$f(y|a, \tau, c) = \frac{\Gamma(a+\tau)}{\Gamma(a)\Gamma(\tau)}\frac{c^a x^{\tau-1}}{(x+c)^{a+\tau}},$$

so that $\mathbb{E}[Y] = c\frac{\tau}{a-1}$ when $Y \sim \mathcal{GP}(a, \tau, c)$.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

One can see that the suggested compound model leaves the patterns that are usually used in the actuarial field, the underlying assumption of independence between the frequency and severity components by letting $\mathbb{E}[C_{it}|N_{it}, \theta_{C[i]}] = \theta_{C[i]}\mu_{it} = \theta_{C[i]} \exp(\mathbf{x}_{it}\beta + N_{it}\gamma)$, which is a flexible extension of traditional independent compound risk model so that we may capture possible dependence between the frequency and severity components via γ. It is an interesting topic by itself to understand a posteriori premium of S_{i,T_i+1} with both types of dependences, dependence between frequency and severity as well as dependence among the claims of the same policyholder. However, a thorough discussion on this topic is not covered in this article in order not to overwhelm the readers.

3. Bayesian sensitivity analysis with Bregman divergence

Bayesian sensitivity analysis, which is also known as robust Bayesian analysis, is an area which studies the impact on the posterior due to possible perturbations of prior distribution. According to the philosophy of Bayesian statistics, a modeler specifies a prior based on his/her insight or experience and updates it with observed data, which yields the posterior distribution used for inference about the parameter(s). It implies that misspecification of the prior distribution might affect the resulting posterior distribution and statistical inference based on the posterior. Thus, it is desirable to test whether a prior would be robust, or less sensitive to possible misspecification so that there can be more consistency on the Bayesian inference, which is based on the obtained posterior distribution of parameter(s) even if the used prior by the modeler is misspecified one.

There are some proposed methods for modeling misspecification of prior distribution which are described in Berger *et al.* (1994), but here we can use the comparison between given $\pi(\theta)$ and the ϵ-contaminated prior $\pi_\epsilon(\theta)$, which has been used in the actuarial literature such as Gómez-Déniz *et al.* (2002b), Gómez-Déniz and Vázquez-Polo (2005), and Gómez-Déniz *et al.* (2006) and is defined as follows:

$$\Pi = \{\pi_\epsilon(\theta) : \pi_\epsilon(\theta) = (1-\epsilon)\pi(\theta) + \epsilon\, q(\theta), q \in \mathbf{Q}, \epsilon \in [0,1]\}, \tag{3.1}$$

where $\mathbf{Q}$ is a certain class of prior distributions used as perturbations.

In this formulation, $\pi(\theta)$ is the prior chosen by the modeler, whereas $q(\theta)$ is a distinct prior irrelevant to $\pi(\theta)$. Therefore, by assuming that the "true" prior, $\pi^{true}(\theta)$, is given as $\pi_\epsilon(\theta)$, the amount of ϵ works as a degree of prior misspecification. For example, if $\epsilon = 0$, then $\pi^{true}(\theta) = \pi_\epsilon(\theta) = \pi(\theta)$ so that there is no misspecification of prior distribution chosen by the modeler. On the other hand, if $\epsilon = 1$, then $\pi^{true}(\theta) = \pi_\epsilon(\theta) = q(\theta)$ so that the modeler totally failed to specify the prior distribution in a correct way. In this regard, it is of our interest to measure the discrepancy between $\pi_\epsilon(\theta|z)$, the "true" posterior, and $\pi(\theta|z)$, the posterior based on the prior choice of the modeler.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

In the perspective of ratemaking with longitudinal data, we may apply this idea to test the robustness of the naive prior (equivalent to constant random effects, which is one of the industry benchmarks) and the proposed prior in both frequency and severity components by comparing the measured discrepancy between $\pi_\epsilon(\theta|z)$ and $\pi(\theta|z)$.

To measure the distance between the posterior densities, we can consider the concept of Bregman divergence proposed by Bregman (1967). Bregman divergence has some properties of usual metric but neither is symmetric nor satisfies the triangle inequality. After the introduction of its concept, it has been utilized in a variety of statistical learning.

For example, Gelfand and Dey (1991) used Kullback–Leibler divergence in Bayesian sensitivity analysis, while Peng and Dey (1995) applied f-divergence in the context of outlier detection. Note that both KL divergence and f-divergence can be explained in terms of Bregman divergence. Zhang (2004) used Bregman divergence to study statistical behavior and consistency of classification methods. Recently, Bregman divergence was used to obtain a class of loss function in order for robust Bayesian prediction according to Karimnezhad and Parsian (2018).

According to Goh and Dey (2014), the difference between $\pi_\epsilon(\theta|z)$ and $\pi(\theta|z)$ can be measured by using functional Bregman divergence which is defined as follows:

Definition 1. *Let h_1, h_2 be nonnegative measurable functions on σ-finite measure space $(\mathcal{Z}, \Omega, v)$ and $\psi : (0, \infty) \to \mathbb{R}$ be a strictly convex and differentiable function. Then the functional Bregman divergence D_ψ is defined as*

$$D(h_1, h_2) = \int \left\{ \psi(h_1(z)) - \psi(h_2(z)) - (h_1(z) - h_2(z))\psi'(h_2(z)) \right\} dv(z).$$

It is easy to check that $D(h, h) = 0$ for any nonnegative measurable function h. Therefore, it is desirable to minimize the following quantity, which measures the relative difference between $\pi_\epsilon(\theta|z)$ and $\pi(\theta|z)$:

$$\begin{aligned} D_\psi^R &= D_\psi \left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)}, 1 \right) \\ &= \int \left\{ \psi \left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)} \right) - \psi(1) - \left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)} - 1 \right) \psi'(1) \right\} \pi(\theta|z) d\theta \\ &= \int \left\{ \psi \left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)} \right) \pi(\theta|z) - \psi(1)\pi(\theta|z) - (\pi_\epsilon(\theta|z) - \pi(\theta|z))\, \psi'(1) \right\} d\theta \\ &= \int \psi \left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)} \right) \pi(\theta|z) d\theta - \psi(1). \end{aligned} \tag{3.2}$$

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

Note that if we can obtain the closed forms of both $\pi_\epsilon(\theta)$ and $\pi(\theta)$, and furthermore it is easy to evaluate the integral given in the end of (3.2), then it might be okay to directly use the closed form as in the following lemma.

Lemma 1. *Suppose $f(z|\theta)$ is data likelihood and $\mathbb{P}(\theta = 1) = 1$, in other words, θ has the point mass at 1. Then D_ψ^R is given as follows:*

$$\begin{aligned} D_\psi^R &= \psi\left(\frac{\pi_\epsilon(1|z)}{\pi(1|z)}\right) - \psi(1) \\ &= \psi\left(\frac{\pi_\epsilon(1)f(z|1)}{m_\epsilon(1|z)}\frac{1}{\pi(1|z)}\right) - \psi(1) \\ &= \psi\left(\frac{(1-\epsilon)\pi(1) + \epsilon\, q(1)}{(1-\epsilon)m(z) + \epsilon\, m_q(z)}\frac{f(z|1)}{\pi(1|z)}\right) - \psi(1) \\ &= \psi\left(\frac{(1-\epsilon) + \epsilon\, q(1)/\pi(1)}{(1-\epsilon) + \epsilon\, m_q(z)/f(z|1)}\right) - \psi(1), \end{aligned}$$

where $m(z) = \int f(z|\theta)\pi(\theta)d\theta$ and $m_q(z) = \int f(z|\theta)q(\theta)d\theta$.

Proof. Since θ has the point mass at 1, we have

$$\int \psi\left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)}\right)\pi(\theta|z)d\theta = \psi\left(\frac{\pi_\epsilon(1|z)}{\pi(1|z)}\right).$$

Furthermore, since $\mathbb{P}(\theta = 1|z) = 1$ as well, $\pi(\theta)$ and $\pi(\theta|z)$ are the same as the Dirac delta function at $\theta = 1$. Therefore,

$$m(z) = \int f(z|\theta)\pi(\theta)d\theta = f(z|1),\ \pi(1) = \pi(1|z).$$

□

Since $\pi(\theta)$ is the Dirac delta function, it has infinite value when $\theta = 1$. Therefore, in actual implementation, we may use $\tilde{\pi}(\theta) \sim \mathcal{N}(1, 10^{-12})$.

However, in most of cases, it might not be possible to obtain the closed form of either $\pi_\epsilon(\theta)$ or $\pi(\theta)$. Even though we have the closed forms, still we are not sure whether the integral is able to be evaluated in an analytic way. Therefore, by denoting $\delta(\theta) := \pi_\epsilon(\theta)/\pi(\theta)$, we may use the following equation as shown in Goh and Dey (2014), which is equivalent to (3.2) but enables us to implement MCMC algorithm to evaluate D_ψ^R numerically:

$$\begin{aligned} D_\psi^R + \psi(1) &= \int \psi\left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)}\right)\pi(\theta|z)d\theta \\ &= \int \psi\left(\frac{\pi_\epsilon(\theta)f(z|\theta)}{\pi(\theta|z)\int \pi_\epsilon(\theta)f(z|\theta)d\theta}\right)\pi(\theta|z)d\theta \\ &= \int \psi\left(\frac{\delta(\theta)\pi(\theta)f(z|\theta)}{\pi(\theta|z)\int \delta(\theta)\pi(\theta)f(z|\theta)d\theta}\right)\pi(\theta|z)d\theta \end{aligned}$$

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

$$
\begin{aligned}
&= \int \psi \left(\frac{\delta(\theta)\pi(\theta|z)}{\pi(\theta|z)\int \delta(\theta)\pi(\theta|z)d\theta} \right) \pi(\theta|z)d\theta \\
&= \int \psi \left(\frac{\delta(\theta)}{\int \delta(\theta)\pi(\theta|z)d\theta} \right) \pi(\theta|z)d\theta \\
&\simeq \frac{1}{J}\sum_{j=1}^{J} \left[\psi \left(\frac{\delta(\hat{\theta}_j)}{\frac{1}{J}\sum_{j=1}^{J}\delta(\hat{\theta}_j)} \right) \right],
\end{aligned} \tag{3.3}
$$

where $\hat{\theta}_j$ s are posterior samples derived from $\pi(\theta|z)$.

Finally, to perform the sensitivity analysis with the contaminated class of priors in (3.1), it is desirable to choose $\mathbf{Q}$ carefully so that it might neither be too broad nor narrow, as mentioned in Berger and Berliner (1986). Therefore, here we can consider the family $\mathbf{Q}$ which satisfies the usual assumption of multiplicative random effects, having 1 as the prior mean. Furthermore, since we are not sure whether the naive prior or the proposed prior represents true dynamics on θ, we consider the family of distribution $\mathbf{Q}$ which has the average of standard deviations of θ under the naive and proposed priors as the standard deviation of θ with $q(\theta)$. One can see that there are some research works which specified the class $\mathbf{Q}$ in terms of moments, including but not limited to Eichenauer *et al.* (1988), Young (1998), Insua *et al.* (1999), Gómez-Déniz *et al.* (2002a), Gómez-Déniz *et al.* (2005), Boratyńska (2017), and Sánchez-Sánchez *et al.* (2019).

Therefore, $\mathbf{Q}$ can be defined as follows:

$$
\mathbf{Q} = \left\{ q(\theta) : \mathbb{E}^q[\theta] = 1, \quad \mathbb{V}ar^q[\theta] = \frac{1}{4}(\mathbb{V}ar^p[\theta] + \mathbb{V}ar^n[\theta]) = \frac{\mathbb{V}ar^p[\theta]}{4} \right\}.
$$

Here $\mathbb{V}ar^n[\theta]$ means the variance of θ under the naive prior and $\mathbb{V}ar^p[\theta]$ means the variance of θ under the proposed prior. And it is also easy to see that $\mathbb{V}ar^n[\theta] = 0$ and $\sqrt{\mathbb{V}ar^q[\theta]} = \sqrt{\mathbb{V}ar^p[\theta]}/2$.

Therefore, in the following section, we are considering uniform, lognormal, and normal priors as possible perturbations for θ_N and θ_C, respectively, so that they can satisfy both mean and variance constraints as follows:

$$
\begin{aligned}
&\text{For } \theta_N : q_1(\theta) \sim \mathcal{U}(0.5557, 1.4443),\ q_2(\theta) \sim \mathcal{LN}(-0.0319, 0.2524), \\
&\qquad q_3(\theta) \sim \mathcal{N}(1, 0.0658), \\
&\text{For } \theta_C : q_1(\theta) \sim \mathcal{U}(0.7261, 1.2738),\ q_2(\theta) \sim \mathcal{LN}(-0.0123, 0.1571), \\
&\qquad q_3(\theta) \sim \mathcal{N}(1, 0.0250).
\end{aligned}
$$

4. Data analysis

For the empirical analysis, a public dataset on insurance claims provided by Wisconsin Local Government Property Insurance Fund (LGPIF) is used,

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

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	0.85	46.75
lnDeductIM	Log deductible amount for IM claim	0	5.34	9.21

TABLE 3

SUMMARY STATISTICS FOR CLAIM FREQUENCY.

		Minimum	Mean	Variance	Maximum
FreqIM	Number of IM claims in a year	0	0.06	0.1	6

which has been used in actuarial literature such as Frees *et al.* (2016). It consists of 5677 observations in training set and 1098 observations in test set. It is a longitudinal dataset with 1234 policyholders which can be tracked with a unique identifier, followed for 5 years on multiple lines of claims. Among the information on multiline insurance, only inland marine (IM) claim information was used. Given dataset has seven categorical explanatory variables, most of which are indicator variables on the types of location as described in Table 2. Note that "NoClaimCreditIM" is used in both frequency and severity modeling considering current practices in ratemaking, because a premium discount is followed by the absence of claim for three consecutive prior years, as a rule of thumb in practice.

As continuous variables, the coverage amount of IM claim and deductible amount for IM claim were used, which are expected to have positive and negative effects on the claims, respectively.

In order to apply the idea of capturing individual heterogeneity via random effect, we should assume that "The same person or object" is followed for many years by a unique identifier even though the characteristics of insurance contract change, and the source of individual heterogeneity is consistent for observed years. For inland marine insurance data described above, one can

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

TABLE 4

DISTRIBUTION OF IM FREQUENCY.

Count	0	1	2	3	4	5	6
FreqIM	5441	182	40	6	4	2	2

TABLE 5

SUMMARY STATISTICS FOR IM SEVERITY.

		Minimum	Mean	Variance	Maximum
log(yAvgIM)	(log) Average size of IM claim in a year	4.09	8.45	2.23	13.09

see that it satisfies both assumptions since a specific object can be observed repeatedly via a unique classifier. However, in case of automobile insurance, validity of the assumptions could be controversial. For example, it is possible a policyholder shares a car with his/her kid or his/her driving skills (one of the unobserved risk characteristics) might be improved over time. A thorough discussion on models with varying or multiple sources of random effects could be an interesting topic for future research.

In terms of frequency, IM has relatively moderate dispersion of the number of claims per year so that maximum number of claims per year is six as shown in Tables 3 and 4. Since the observed sample mean of the number of claims is much smaller than the observed sample variance, it is natural to consider the use of different types of frequency distribution on the modeling other than naive Poisson distribution. Moreover, it can be shown that marginal distribution of claim frequency follows a multivariate negative binomial (MVNB) distribution under the proposed prior so that it provides another rationale to consider a non-point mass prior on the random effect of the frequency component.

After we fixed the hyperparameters on the priors of each random effect component, the marginal likelihood of both frequency and the average severity components could be obtained with the naive and proposed priors, respectively. Upon the obtained marginal likelihood, as a type of empirical Bayes method, the regression coefficients can be estimated with the marginal likelihood and observed data. For the frequency component, use of the naive prior leads us to the marginal likelihood of independent Poisson distribution, whereas the use of proposed prior leads us to the marginal likelihood of MVNB distribution. Under each marginal likelihood, $\hat{\alpha}$ were obtained, which are hyperparameters in the frequency model associated with the explanatory variables. The estimated value of $\hat{\alpha}$ and loglikelihoods for both Poisson and MVNB models are shown in Table 6.

Note that one can try to apply either Wald test or likelihood ratio test (LRT) to test the presence of random effects in the frequency component,

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

TABLE 6

REGRESSION ESTIMATES FROM MARGINAL FREQUENCY LIKELIHOODS.

	Poisson		MVNB	
	Estimate	Standard error	Estimate	Standard error
(Intercept)	−6.9455	1.0211	−7.3601	1.1532
TypeCity	3.7219	1.0101	3.7844	1.1359
TypeCounty	4.5654	1.0124	4.6135	1.1409
TypeSchool	1.8423	1.0274	2.0799	1.1496
TypeTown	2.3378	1.0263	2.5526	1.1494
TypeVillage	2.7545	1.0139	2.9450	1.1383
CoverageIM	0.0647	0.0072	0.0946	0.0143
lnDeductIM	0.1531	0.0455	0.1732	0.0520
NoClaimCreditIM	−0.3697	0.1283	−0.1985	0.1326
Loglikelihood	−973.2993		−931.2171	

which is equivalent to test whether $r = \infty$ or not. If we let $H_0 : r = \infty$ and $H_1 : r < \infty$, then we have

$$\begin{aligned}\ell_0 =: \ell_N(\hat{\alpha}, \hat{r}|H_0) &= -973.2993 < -931.2171 = \ell_N(\hat{\alpha}, \hat{r}|r = 3.8) \\ &\leq \ell_N(\hat{\alpha}, \hat{r}|H_1) := \ell_1,\end{aligned}$$

so that $2(\ell_1 - \ell_0) \geq 2(-931.2171 + 973.2993) = 84.1644 > 15.1367 = \chi^2_{0.9999}(1)$. Therefore, one may conclude that MVNB model is more appropriate than naive Poisson model for given frequency data based on LRT. However, we should be careful to apply LRT in this case since we are testing the hypothesis at the boundary of the parameter space, and there is no assurance that the LRT test statistic would follow asymptotically χ^2 distribution, as mentioned in Andrews (2001) and Cameron and Trivedi (2013). Besides, the applicability of LRT depends on the nested structure between the naive and proposed prior, which may not hold when proposed prior follows, for example, a two-point mixture of gamma and normal distributions. In general, it is very unlikely to assume that $\pi_N(\theta)$ would follow the nested structure described as $H_0 : r = \infty$ and $H_1 : r < \infty$ for sure, but Bayesian sensitivity analysis can be used without assuming the nested structure, unlike LRT.

In case of the average severity component whose summary statistics is provided in Table 5, use of the naive prior leads us to independent gamma marginal likelihood, whereas use of the proposed prior leads us to marginal likelihood of multivariate generalized Pareto (MVGP) distribution. Again, under each marginal likelihood, $\hat{\beta}$ and $\hat{\gamma}$ were obtained, which are hyperparameters in the average severity model associated with the explanatory variables. The estimated values of $\hat{\beta}$, $\hat{\gamma}$, and loglikelihoods for both gamma and MVGP models are shown in Table 7.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

TABLE 7

REGRESSION ESTIMATES FROM MARGINAL AVERAGE SEVERITY LIKELIHOODS.

	Gamma		MVGP	
	Estimate	Standard error	Estimate	Standard error
(Intercept)	8.8954	2.2013	8.7666	1.5147
TypeCity	2.5348	2.0274	2.5222	1.3807
TypeCounty	2.5214	2.0309	2.4038	1.3865
TypeSchool	1.1701	2.0645	1.2718	1.4083
TypeTown	1.6037	2.0552	1.5909	1.4062
TypeVillage	1.2629	2.0284	1.2599	1.3793
CoverageIM	0.0326	0.0184	0.0346	0.0152
lnDeductIM	−0.1653	0.1442	−0.1533	0.1056
NoClaimCreditIM	−0.1095	0.2775	−0.0957	0.2000
FreqIM	−0.4632	0.1004	−0.4448	0.0717
Loglikelihood	−3256.036		−2416.886	

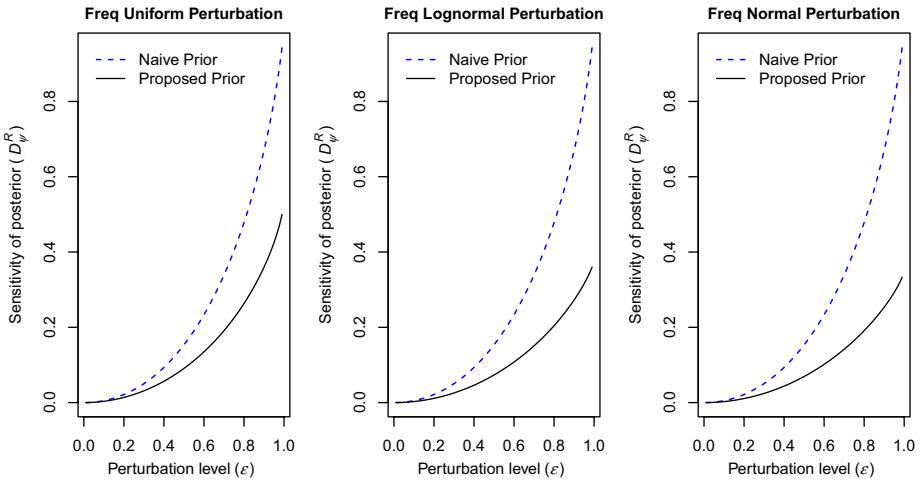

FIGURE 1: Sensitivities of frequency priors with various perturbation priors.

With the hyperparameters from the marginal likelihoods, we can perform Bayesian sensitivity analysis via Bregman divergence both for the frequency priors and the average severity priors. For calculating D_ψ^R, the following convex function $\psi(z) = z \log z - z + 1$ is used, which is a special case of a class of convex functions considered in Eguchi and Kano (2001).

Figure 1 shows us the result of Bayesian sensitivity analysis for the frequency component. Here the x-axis of each graph means the magnitude of perturbation given by $\epsilon \in [0, 1]$ and y-axis means the sensitivity of posterior

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

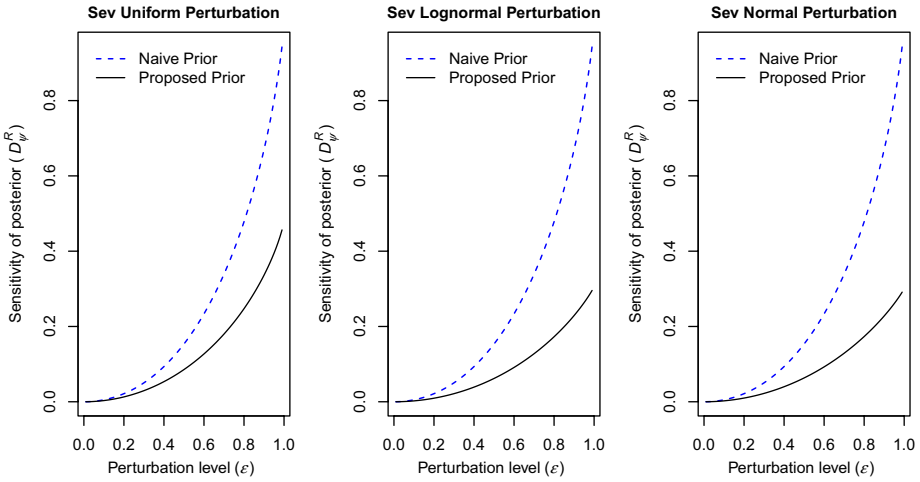

FIGURE 2: Sensitivities of severity priors with various perturbation priors.

for θ_N measured by $D_\psi^R = \int \psi\left(\frac{\pi_\epsilon(\theta|z)}{\pi(\theta|z)}\right)\pi(\theta|z)d\theta - \psi(1)$. Since $\pi_\epsilon(\theta_N|n) = \pi(\theta_N|n)$ if $\epsilon = 0$ and $\psi(1) = 0$ by definition; one can expect that $D_\psi^R = 0$ provided $\epsilon = 0$ and D_ψ^R increases as the value of ϵ gets larger because of the increasing deviation of $\pi(\theta_N|n)$ from $\pi_\epsilon(\theta_N|n)$, which is assumed to be the true posterior for θ_N. However, one can clearly see that the sensitivity of posterior D_ψ^R increases much faster when we use the naive prior than when the proposed posterior is used in all levels and distributions of perturbation. Therefore, Figure 1 tells us that use of the proposed prior for θ_N is less sensitive to possible misspecification of (usually unknown) the true prior than use of the naive prior for θ_N, which is the industry benchmark so that it would be desirable to consider the use of nonconstant random effects in the frequency component for given dataset.

In case of the average severity priors for θ_C, as shown in Figure 2, we can observe the similar results so that under perturbations with uniform, lognormal, and normal priors, again the naive prior, point mass shows higher sensitivity D_ψ^R in all perturbation levels ϵ. Therefore, we can claim that in both frequency and severity cases, use of the proposed priors is more robust from possible misspecification of the true prior distribution.

Finally, use of the proposed priors could be justified under out-of-sample validation. Using the predictive distributions in both frequency and the average severity, the expected total losses are calculated based on the observed characteristics of policyholders and compared with the actual total claims in the test set. As shown in Table 8, the combination of proposed models in both frequency and the average severity shows better performance on the prediction results of total claims in terms of both root-mean-square error (RMSE) and mean absolute error (MAE).

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

TABLE 8

VALIDATION MEASURES FOR THE PREDICTION OF TOTAL CLAIMS.

	Naive model	Proposed model
RMSE	6692.328	6443.930
MAE	1800.494	1541.881

5. CONCLUSION

In this paper, we explored a framework to test the presence of nonconstant random effects via prior elicitation and Bayesian sensitivity analysis. Use of a point mass prior on random effects might be too informative so that it might be vulnerable to possible misspecification of the true prior distribution compared to less informative priors proposed in this article. Upon the use of Bayesian sensitivity analysis with Bregman divergence, it was shown that the proposed priors might yield the more robustness than the naive priors both in frequency and the average severity components, respectively. Furthermore, the predicted values of total claims based on the estimates from the proposed marginal likelihood ended up with better performance than the predicted values of total claims based on the naive independent data likelihood. Therefore, this study provides a theoretical framework to test presence of nonconstant random effects in longitudinal insurance claim datasets as well as the empirical results.

ACKNOWLEDGEMENTS

I thank Benjamin Avanzi, the editor and anonymous referees for providing helpful comments and suggestions that have greatly improved the article. I also thank Dipak K. Dey and Gyuhyeong Goh for their inspiration. This work was supported by the James C. Hickman Scholar Doctoral Stipend funded by the Society of Actuaries.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2020.19.

REFERENCES

ANDREWS, D.W. (2001) Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, **69**(3), 683–734.

BERGER, J. and BERLINER, L.M. (1986) Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *The Annals of Statistics*, **14**(2), 461–486.

BERGER, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Berlin/Heidelberg, Germany: Springer Science & Business Media.

BERGER, J.O., MORENO, E., PERICCHI, L.R., BAYARRI, M.J., BERNARDO, J.M., CANO, J.A., DE LA HORRA, J., MARTÍN, J., RÍOS-INSÚA, D., BETRÒ, B., DASGUPTA, A., GUSTAFSON, P., WASSERMAN, L., KADANE, J.B., SRINIVASAN, C., LAVINE, M., O'HAGEN, A., POLASEK, W., ROBERT, C.P., GOUTIS, C., RUGGERI, F., SALINETTI, G. and SIVAGANESAN, S. (1994) An overview of robust Bayesian analysis. *Test*, 3, 5–124.

BORATYŃSKA, A. (2017) Robust Bayesian estimation and prediction of reserves in exponential model with quadratic variance function. *Insurance: Mathematics and Economics*, **76**, 135–140.

BOUCHER, J.-P., DENUIT, M. and GUILLÉN, M. (2008) Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance*, **2**(1), 135–162.

BREGMAN, L.M. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, **7**(3), 200–217.

CAMERON, A.C. and TRIVEDI, P.K. (2013) *Regression Analysis of Count Data*, Vol. 53. Cambridge, England: Cambridge University Press.

DING, J. and WANG, J.-L. (2008) Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**(2), 546–556.

DUBITZKY, W., WOLKENHAUER, O., YOKOTA, H. and CHO, K.-H. (2013) *Encyclopedia of Systems Biology*. Berlin/Heidelberg, Germany: Springer.

EGUCHI, S. and KANO, Y. (2001) *Robustifying maximum likelihood estimation*. Technical Report, Institute of Statistical Mathematics.

EICHENAUER, J., LEHN, J. and RETTIG, S. (1988) A gamma-minimax result in credibility theory. *Insurance: Mathematics and Economics*, **7**(1), 49–57.

FRANGOS, N.E. and VRONTOS, S.D. (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), 1–22.

FREES, E.W., LEE, G. and YANG, L. (2016) Multivariate frequency-severity regression models in insurance. *Risks*, **4**(1), 4.

GARRIDO, J., GENEST, C. and SCHULZ, J. (2016) Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, **70**, 205–215.

GELFAND, A. and DEY, D. (1991) On Bayesian robustness of contaminated classes of priors. *Statistics & Risk Modeling*, **9**(1–2), 63–80.

GOH, G. and DEY, D.K. (2014) Bayesian model diagnostics using functional Bregman divergence. *Journal of Multivariate Analysis*, **124**, 371–383.

GÓMEZ-DÉNIZ, E., BERMÚDEZ, L. and MORILLO, I. (2005) Computing bonus–malus premiums under partial prior information. *British Actuarial Journal*, **11**(2), 361–374.

GÓMEZ-DÉNIZ, E., HERNÁNDEZ, A., PÉREZ, J. and VÁZQUEZ-POLO, F. (2002a) Measuring sensitivity in a bonus–malus system. *Insurance: Mathematics and Economics*, **31**(1), 105–113.

GÓMEZ-DÉNIZ, E., HERNÁNDEZ-BASTIDA, A. and VÁZQUEZ-POLO, F.J. (1999) The Esscher premium principle in risk theory: A Bayesian sensitivity study. *Insurance: Mathematics and Economics*, **25**(3), 387–395.

GÓMEZ-DÉNIZ, E., HERNANDEZ-BASTIDA, A. and VÁZQUEZ-POLO, F.J. (2002b) Bounds for ratios of posterior expectations: Applications in the collective risk model. *Scandinavian Actuarial Journal*, **2002**(1), 37–44.

GÓMEZ-DÉNIZ, E., PEREZ-SANCHEZ, J.M. and VÁZQUEZ-POLO, F.J. (2006) On the use of posterior regret γ-minimax actions to obtain credibility premiums. *Insurance: Mathematics and Economics*, **39**(1), 115–121.

GÓMEZ-DÉNIZ, E. and VÁZQUEZ-POLO, F.J. (2005) Modelling uncertainty in insurance bonus–malus premium principles by using a Bayesian robustness approach. *Journal of Applied Statistics*, **32**(7), 771–784.

GÓMEZ-DÉNIZ, E., VÁZQUEZ POLO, F.J. and BASTIDA, A.H. (2000) Robust Bayesian premium principles in actuarial science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49**(2), 241–252.

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press

INSUA, S.R., MARTIN, J., INSUA, D.R. and RUGGERI, F. (1999) Bayesian forecasting for accident proneness evaluation. *Scandinavian Actuarial Journal*, **1999**(2), 134–156.
JEFFREYS, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **186**(1007), 453–461.
JEONG, H., AHN, J., PARK, S. and VALDEZ, E.A. (2020) Generalized linear mixed models for dependent compound risk models. *Variance*, accepted for publication.
KARIMNEZHAD, A. and PARSIAN, A. (2018) Bayesian and robust Bayesian analysis in a general setting. *Communications in Statistics-Theory and Methods*, **48**(15), 1–29.
KLUGMAN, S.A., PANJER, H.H. and WILLMOT, G.E. (2012) *Loss Models: From Data to Decisions*, Vol. 715. Hoboken, NJ: John Wiley & Sons.
LEMAIRE, J. (1998) Bonus-malus systems: The European and Asian approach to merit-rating. *North American Actuarial Journal*, **2**(1), 26–38.
NG, E. and COOK, R. (2000) A comparison of some random effect models for parameter estimation in recurrent events. *Mathematical and Computer Modelling*, **32**(1–2), 11–26.
PENG, F. and DEY, D.K. (1995) Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, **23**(2), 199–213.
SÁNCHEZ-SÁNCHEZ, M., SORDO, M., SUÁREZ-LLORENS, A. and GÓMEZ-DÉNIZ, E. (2019) Deriving robust Bayesian premiums under bands of prior distributions with applications. *ASTIN Bulletin: The Journal of the IAA*, **49**(1), 147–168.
YOUNG, V.R. (1998) Robust Bayesian credibility using semiparametric models. *ASTIN Bulletin: The Journal of the IAA*, **28**(2), 187–203.
ZHANG, T. (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**(1), 56–85.

HIMCHAN JEONG (Corresponding author)
Department of Statistics & Actuarial Science
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
E-Mail: himchan_jeong@sfu.ca

https://doi.org/10.1017/asb.2020.19 Published online by Cambridge University Press