Bayesian Analysis (2022) **17**, Number 1, pp. 67–93

# A Bayesian Approach to Modeling Multivariate Multilevel Insurance Claims in the Presence of Unsettled Claims

Marie-Pier Côté[*], Christian Genest[†,§], and David A. Stephens[‡]

**Abstract.** A Bayesian model for individual insurance claims is proposed which accounts for the multivariate multilevel features of the claims, including multiple claimants for the same event, each of whom may receive benefits under different coverages. A Bayesian approach makes it possible to account for missing values in the covariates and partial information contained in open files, thereby avoiding sampling bias induced when unsettled claims are ignored. For a given claim, the combination of coverages under which payments are made is modeled as a type with multinomial regression. The presence of legal and expert fees follows a logistic regression given the type. The non-zero claim amounts are then modeled with log skewed normal regressions linked by a Student $t$ copula. The Bayesian framework yields a predictive distribution for the amounts paid, including parameter risk and process risk, while handling missing covariates and open files. The approach is illustrated with Accident Benefits car insurance claims from a Canadian company.

**Keywords:** Bayesian model, censored data, copula, correlation, Fernández–Steel skewed normal, imputation, insurance claim, multinomial and logistic regression.

**MSC2020 subject classifications:** Primary 62F15, 62P05; secondary 62H05.

## 1 Introduction

General insurance companies are increasingly interested in statistical methods that take advantage of their detailed databases. One such application is the statistical modeling of paid amounts at the policyholder or claim level. With greater granularity come broader applicability and many challenges that can be overcome using a Bayesian perspective.

A claim amount can be viewed as a sum of losses relating to the various coverages in the insurance policy. A typical example is a car insurance product which may cover vehicle damage, medical and rehabilitation expenses, income replacement benefits, etc. A car accident can trigger payments under more than one coverage and for multiple claimants (e.g., the driver, passengers, pedestrians). As the total claim amounts are heterogeneous, it is natural to model the detailed components jointly before they are

[*] École d'actuariat, Université Laval, 2425, rue de l'Agriculture, Québec (Québec) Canada G1V 0A6, marie-pier.cote@act.ulaval.ca

[†] Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke ouest, Montréal (Québec) Canada H3A 0B9, christian.genest@mcgill.ca

[‡] Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke ouest, Montréal (Québec) Canada H3A 0B9, david.stephens@mcgill.ca

[§] Corresponding author

https://doi.org/10.1214/20-BA1243

aggregated. Both the probability of a payment for a component and its amount can depend on characteristics of the accident or the claimant. Dependence between the claim components naturally arises as they all relate to the same event.

A convenient way to construct multivariate distributions is to model the marginal behavior and the dependence structure separately through copulas. The usefulness of copulas in actuarial science and risk management is well documented; see, e.g., McNeil et al. (2005) or Denuit et al. (2006). This approach was adopted, for instance, in the multivariate claims model of Frees and Valdez (2008) and Frees et al. (2009), where the severity was split into two parts: the claim type, identifying the coverages with strictly positive payments, and the amounts given the type. The introduction of the claim type effectively takes care of the point mass at zero, due to the claims for which no payment is made. It also allows to model jointly the amounts that are strictly positive, conditionally on the claim type, in a copula model with margins that depend on covariates.

As an alternative, Shi et al. (2016) account for the mass at zero using Tweedie margins in a Gaussian copula model for policies involving more than one vehicle and multiple coverages. In this case, the copula describes the dependence in occurrences and amounts, thereby being somewhat less flexible than the claim type approach of Frees and Valdez (2008). In Yang et al. (2011), a copula links the components of bodily injury car insurance claims. In these papers, maximum likelihood, pairwise composite likelihood, or inference for margins are used for estimation. Thus, the resulting predictive distributions of claim amounts, approximated by simulation, do not incorporate the uncertainty relating to the large number of parameters.

We propose here a Bayesian model for multivariate multilevel claim amounts. As discussed in Shi et al. (2016), this data structure is frequent in insurance. Detailed car insurance claims, workers compensation insurance, group health insurance, and commercial auto insurance exhibit an unbalanced multilevel structure. Micro-level Accident Benefits insurance claims data from a large Canadian insurance company serve as justification for our approach. The proposed model has many applications; it could be used, e.g., to get the predictive distribution of the paid amount given the claim characteristics when a new claim gets reported. Such a model can complement (and maybe even replace in the long run) the costly initial assessment by a claims expert.

The dataset, described in Section 2, is complex: each claim may involve one or more claimants, and the payments for each claimant are broken down into four parts. Recent claims are the most relevant for the predictions, but because the settlement of Accident Benefits claims may take a few years, most of them are not yet settled and their total paid amount is unknown. Using only settled files for estimation causes sampling bias: the simpler cases close faster than the major ones. Ignoring the fact that open claims are censored also understates the paid amounts. To solve this problem, we propose an innovative Bayesian procedure that includes the open claims partial information.

The basic model setup, described in Section 3, is adapted from the hierarchical insurance claims model of Frees and Valdez (2008) and accounts for covariates known at the time of reporting. Claims are classified into five types depending on the combination of components with a strictly positive payment. These types are modeled with a multinomial regression. The presence of expenses in the form of subcontracted legal and claims expert fees is modeled conditionally on the type via logistic regression. Given the type and presence of expenses for each claimant in an accident, the strictly positive amounts are log skewed normal regression models linked by a Student $t$ copula.

The Bayesian framework facilitates the handling of missing covariates and partial information contained in open claims. Imputation models for the censored types, presence of expenses, and amounts are developed in Section 4 and are estimated using the settled claim development data. MCMC methods with a Metropolis-within-Gibbs algorithm can be set up, as explained in the Supplement (Côté et al., 2020), to obtain the posterior distribution of the parameters and the posterior predictive distribution of the multivariate claim, including parameter uncertainty. The good performance of the approach is illustrated in Section 5, where the posterior predictive distribution based on the proposed model and the data available in 2008 is compared to the posterior distribution based on the complete 2015 data. Predictive distributions and predictive intervals on paid amounts for individual claims or portfolios from the holdout sample are shown in Section 6. The paper ends with a short conclusion in Section 7.

## 2 A glimpse at the data

The dataset consists of claim characteristics and the associated inflation-adjusted paid amounts for an Accident Benefits car insurance product sold in the province of Ontario, Canada. In total, 54,281 accidents were reported between January 1, 2004 and December 31, 2015. The payments are known up to the latter date, at which point 15% of the claim files are open. A random sample of 60% of the claims is used for estimation; the remainder forms a holdout validation sample and is not discussed in this section.

Most claims involve one (81.6%) or two (13.9%) claimants, but in one case there were as many as eight claimants. For each claimant, the payment is broken down into four components: three insurance coverages and the legal and claims expert fees (Expenses). The coverages are medical and rehabilitation benefits (Medical), income replacement benefits (Income) and caregiver benefits (Caregiver). Each of these components has a mass at zero. To model them jointly, the claimant files were classified into the five types listed in Table 1; 0.4% of the files matched none of these categories and were discarded as they were deemed highly unlikely to occur ever again.

| Type | Coverages With Payment Exceeding \$100 | % of Claimants |
|---|---|---|
| 1 | None | 42.3% |
| 2 | Medical | 36.4% |
| 3 | Medical and Income | 6.9% |
| 4 | Medical and Caregiver | 9.4% |
| 5 | Medical, Income and Caregiver | 5.0% |

Table 1: Definition of types and distribution in (open and closed) claimant files.

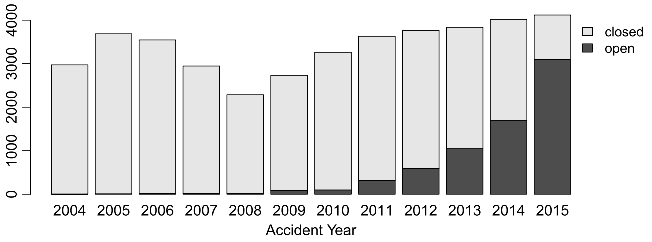


Figure 1: Breakdown of open and closed claimant files by accident year.

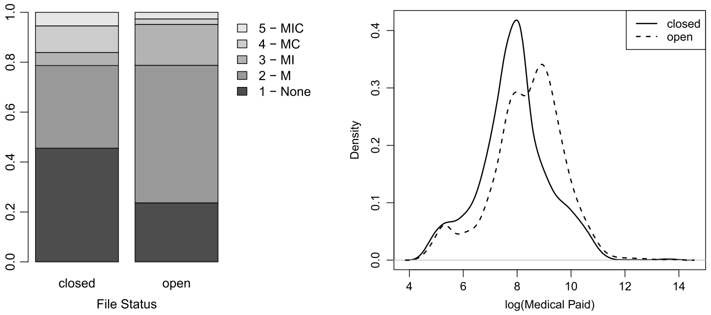


Figure 2: Distribution of type by claim status (left) and kernel density estimates (right) for after-reform log Medical amounts for closed (solid line) and open cases (dashed line).

As claim amounts below $100 can be regarded as immaterial for bodily injury claim modeling, they were treated as zeros when choosing the type. Such amounts hardly contain any information about types and their modeling would yield little benefit. Also, a claimant who receives income replacement or caregiver benefits must also incur medical expenses. Expenses were present in 28% of the files but were more likely for types 2–5.

Effective September 1, 2010, a reform of the Ontario Insurance Act (FSCO, 2010) affected the Accident Benefits product by modifying the allowed paid amounts depending on the claim gravity, especially for the caregiver benefits. Côté and Genest (2015) documented the impact of this reform in a higher-level model for multiple business lines including the aggregate version of the claim portfolio studied here. Figure 1 shows the number of claimants per accident year and by status. The recent accident years, which are post-reform and the most relevant for future predictions, contain many open files: the final settlement amount is unknown at evaluation date, i.e., December 31, 2015.

The left panel of Figure 2 contrasts the type distributions for closed and open files. These distributions clearly differ, and dropping the open claims would bias the type analysis towards minor claims such as those that close at zero. As will be seen in Section 4.2, the type of open claims can actually change until settlement.

The right panel of Figure 2 compares kernel density estimates of the strictly positive Medical paid amounts, on the log scale, for unsettled (dashed line) and closed (solid

| Component | 1: Medical | 2: Income | 3: Caregiver | 4: Expenses |
|---|---|---|---|---|
| Minimum | 4.605 | 4.611 | 4.619 | 4.608 |
| Median | 8.283 | 8.654 | 8.556 | 7.863 |
| Maximum | 14.058 | 13.122 | 13.805 | 12.681 |
| Mean | 8.361 | 8.562 | 8.420 | 7.731 |
| Standard deviation | 1.496 | 1.466 | 1.337 | 1.375 |

Table 2: Summary statistics of (non-zero) amounts paid by component, on the log scale.

line) claims incurred after the reform. It transpires from it that large claims are more likely to be unsettled at evaluation date: even though these amounts are censored, their distribution is shifted to the right compared to that of the settled claims amounts.

Therefore, the information conveyed in the open files ought to be exploited in the estimation procedure for the amounts. This issue was not relevant to the micro-level claim analyses of Frees and Valdez (2008) or Frees et al. (2009) as the claim status was unavailable and all claims were assumed to be settled.

Table 2 shows descriptive statistics of the log of the strictly positive paid amounts by component. At this stage, the dependence between amounts is not analyzed, as part of it is caused by common covariates. At the claim level, the covariates on the policy are: whether the vehicle is a motorcycle or a scooter, the number of years insured with the company (grouped), whether the policyholder also has property insurance, and whether there is optional income coverage. Two categories of covariates are given by claimant:

(a) *Claimant's demographics:* Age, income indemnity level (categorical), marital status (single, married, divorced or other), and rating score.

(b) *Claim characteristics:* Whether it occurred after the 2010 reform, gravity (minor, major, or catastrophic), number of injuries, pain level (scale of 1 to 10), presence of a legal counsellor, and initial case reserve (in thousands) for the three coverages.

Income indemnity level is used only in the Income paid amounts model and is missing in 8.1% of the type 3 and 5 cases. Age, which will be used in all the marginal amounts models, is missing in 8.9% of cases. Even though 47.8% of the pain levels are missing, this feature still improves both the type and amounts models.

# 3 Multivariate multilevel claims model

The basic setup of the proposed model is inspired by Frees and Valdez (2008). Given the covariates for each claimant involved in the claim, the type for each claimant, i.e., the combination of coverages with positive payment, is modeled by a multinomial regression. Conditional on the type, the presence of Expenses follows a logistic regression model, for each claimant. These two components take care of the masses at zero, and the log of the strictly positive amounts are modeled with skewed normal regressions linked in a Student $t$ copula of appropriate dimension, with a suitable correlation structure.

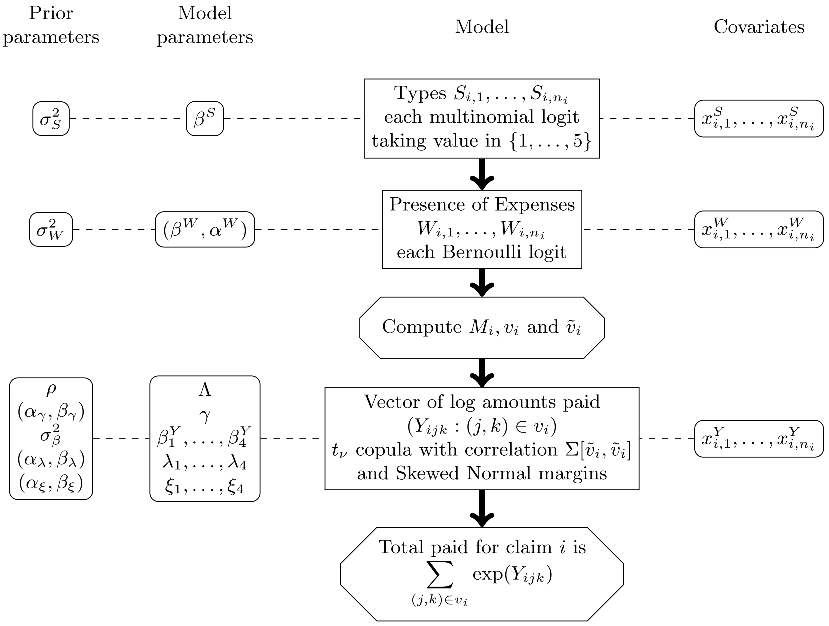

Prior parameters
Model parameters
Model
Covariates
$\sigma_S^2$
$\beta^S$
Types $S_{i,1}, \ldots, S_{i,n_i}$ each multinomial logit taking value in $\{1, \ldots, 5\}$
$x_{i,1}^S, \ldots, x_{i,n_i}^S$
$\sigma_W^2$
$(\beta^W, \alpha^W)$
Presence of Expenses $W_{i,1}, \ldots, W_{i,n_i}$ each Bernoulli logit
$x_{i,1}^W, \ldots, x_{i,n_i}^W$
Compute $M_i, v_i$ and $\tilde{v}_i$
$\rho$
$(\alpha_\gamma, \beta_\gamma)$
$\sigma_\beta^2$
$(\alpha_\lambda, \beta_\lambda)$
$(\alpha_\xi, \beta_\xi)$
$\Lambda$
$\gamma$
$\beta_1^Y, \ldots, \beta_4^Y$
$\lambda_1, \ldots, \lambda_4$
$\xi_1, \ldots, \xi_4$
Vector of log amounts paid $(Y_{ijk} : (j,k) \in v_i)$ $t_\nu$ copula with correlation $\Sigma[\tilde{v}_i, \tilde{v}_i]$ and Skewed Normal margins
$x_{i,1}^Y, \ldots, x_{i,n_i}^Y$
Total paid for claim $i$ is $\sum_{(j,k)\in v_i} \exp(Y_{ijk})$

Figure 3: Bayesian multivariate micro-level model for one claim $i$ involving $n_i$ claimants.

Given covariates and parameters, different claims are assumed to be mutually independent but amounts paid in relation to the same claim are related. This association may result from unmeasured covariates or the duration of the rehabilitation, which affects simultaneously the claimant's medical, income replacement and caregiver benefits, if any. The model is multivariate due to the joint distribution of the amounts for a claimant, and multilevel because of the dependence between amounts for different claimants involved in the same claim.

See Figure 3 for a schematic view of the model. Details on each submodel, the notation and the priors are given in this section and, for ease of understanding, are illustrated with the following running example.

**Example.** Claim $i = 5123$ involves two claimants: the first received \$2122.74 in Medical benefit; the second had a Medical benefit of \$9399.86 and a Caregiver benefit of \$3306.82. There were no Expenses for Claimant 1 but a \$6600.47 fee was allocated to Claimant 2.

## 3.1 Multinomial model for type

For each claim $i \in \{1, \ldots, N\}$ and claimant $j \in \{1, \ldots, n_i\}$, where $n_i$ is the number of claimant files relating to claim $i$, we model the type $S_{i,j}$, as defined in Table 1, using a multinomial logistic regression. Let $x_{i,j}^S$ be the row vector of relevant explanatory variables for claimant $j$ in claim $i$. In the present case, the components of $x_{i,j}^S$ were the policy age and accident year if prior to the reform, the claimant's marital status, number of injuries, gravity, pain level, and two interactions involving the reform indicator.

Denote by $\beta^S = (\beta_1^S, \ldots, \beta_5^S)^\top$ the complete parameter vector for the type model, where $\beta_1^S = \mathbf{0}$. Then, for each $s \in \{1, \ldots, 5\}$,

$$\Pr(S_{i,j} = s \mid x_{i,j}^S, \beta^S) = \frac{\exp(x_{i,j}^S \beta_s^S)}{\sum_{t=1}^5 \exp(x_{i,j}^S \beta_t^S)}.$$

As there is little prior information on the covariate impacts on the types, the parameters are assumed a priori independent with a diffuse prior $\beta^S \sim \mathcal{N}(0, \sigma_S^2 \mathbf{I})$, where $\sigma_S^2$ is large.

**Example** (cont'd)**.** Types are $s_{i1} = 2$ (Medical) and $s_{i2} = 4$ (Medical and Caregiver).

## 3.2 Binomial model for presence of Expenses

The indicator $W_{i,j}$ of the presence of Expenses takes value 1 whenever the amount for that component exceeds \$100. Given $S_{i,j}$, $W_{i,j}$ is modeled by a logistic regression with

$$\Pr(W_{i,j} = 1 \mid S_{i,j} = s, x_{i,j}^W, \beta^W, \alpha_s^W) = \frac{\exp(x_{i,j}^W \beta^W + \alpha_s^W)}{1 + \exp(x_{i,j}^W \beta^W + \alpha_s^W)},$$

where $x_{i,j}^W$ are the relevant covariates, $\beta^W$ is the vector of coefficients, and $\alpha_s^W$ is the effect of type $s$ with $\alpha_1^W = 0$. In the present case, the components of $x_{i,j}^W$ were policy age, a reform indicator, and gravity, as well as two interactions involving the reform indicator. The parameters are again assumed to have independent Gaussian priors centered around 0 and with large variance $\sigma_W^2$.

For notational simplicity, set $\alpha^W = (\alpha_1^W, \ldots, \alpha_5^W)$ and define $M_{ijk}$ as the indicator that the payment for component $k$ (Medical, Income, Caregiver and Expenses are numbered from 1 to 4) and claimant $j$ in claim $i$ is positive, i.e.,

$$M_{ij1} = \mathbf{1}(S_{i,j} > 1),\ M_{ij2} = \mathbf{1}(S_{i,j} \in \{3, 5\}),\ M_{ij3} = \mathbf{1}(S_{i,j} \in \{4, 5\}),\ M_{ij4} = W_{i,j},$$

and set the component vector for claim $i$ as $M_i = (M_{i11}, \ldots, M_{i14}, \ldots, M_{in_i1}, \ldots, M_{in_i4})$. For claim $i$, let $v_i$ be the index set of the claimant number and component with positive paid amounts, i.e.,

$$v_i = \big\{(j, k) : M_{ijk} = 1, j \in \{1, \ldots, n_i\}, k \in \{1, \ldots, 4\}\big\}$$

and let $\tilde{v}_i = \{k + 4(j-1) : (j, k) \in v_i\}$ be the index set of the position of the 1s in the vector $M_i$. If $v_i = \tilde{v}_i = \emptyset$, the total amount for claim $i$ is 0.

**Example** (cont'd)**.** As there were Expenses for Claimant 2 only, we have $w_{i1} = 0$ and $w_{i2} = 1$. Recall that $s_{i1} = 2$ and $s_{i2} = 4$, so we have $m_i = (1, 0, 0, 0, 1, 0, 1, 1)$ and the index set of positive amounts is $v_i = \{(1, 1), (2, 1), (2, 3), (2, 4)\}$ while $\tilde{v}_i = \{1, 5, 7, 8\}$.

## 3.3 Skewed normal regression models for the marginal amounts

Given the component vector $M_i$, the amounts that are exactly 0 are known, so it remains to model the strictly positive paid amounts. When $M_{ijk} = 1$, denote by $Y_{ijk}$ the natural logarithm of the payment made under component $k \in \{1, \ldots, 4\}$ (i.e., Medical, Income, Caregiver or Expenses) for the $j$th claimant in claim $i$.

**Example** (cont'd)**.** One has $Y_{i11} = \ln(2122.74)$, $Y_{i21} = \ln(9399.86)$, $Y_{i23} = \ln(3306.82)$ and $Y_{i24} = \ln(6600.47)$. The other $Y$s are undefined.

A preliminary analysis showed that the log amounts distribution is skewed but not heavy-tailed. Fernández and Steel (1998) showed how to build skewed versions of symmetric distributions and how a Metropolis-within-Gibbs algorithm can be used to estimate the parameters. Inspired by this work, we assume that

$$Y_{ijk} \mid M_{ijk} = 1 \sim \mathcal{SN}(x_{i,j}^Y \beta_k^Y, \sigma_k, \xi_k),$$

where $x_{i,j}^Y$ stands for the vector of covariates listed in Tables 9–10 in the Supplement.

Adapting the parametrization in Trottier and Ardia (2016), the density of the Fernández–Steel skewed normal can be written, for $y \in \mathbb{R}$, as

$$f(y \mid \mu, \sigma, \xi) = \frac{2\sigma^*}{\xi + 1/\xi}\, \phi \left\{ \frac{y\sigma^* + \mu^*}{\xi^{\operatorname{sign}(y\sigma^* + \mu^*)}} \right\},$$

where $\phi$ is the $\mathcal{N}(0,1)$ density, $\xi \in (0,\infty)$ is a skewness parameter, $\mu \in \mathbb{R}$, $\sigma \in (0,\infty)$,

$$\sigma^* = \sqrt{(1-2/\pi)(\xi^2 + 1/\xi^2) + 4/\pi - 1}/\sigma \quad \text{and} \quad \mu^* = (\xi - 1/\xi)\sqrt{2/\pi} - \mu\sigma^*.$$

In this parametrization, we have $\mathrm{E}(Y_{ijk}) = x_{i,j}^Y \beta_k^Y$ and $\operatorname{var}(Y_{ijk}) = \sigma_k^2$. The skewed normal with $\xi = 1$ reduces to the symmetric normal. The moment generating function obtained in the Supplement guarantees the existence of the mean and variance of the paid amounts on the original scale. Risk measures such as the Tail Value-at-Risk of the paid amounts can be expressed in terms of the distribution function $\Phi$ of the $\mathcal{N}(0,1)$ and its inverse, as can be derived with similar calculations as in Section 2 of the Supplement.

## 3.4 Student $t$ copula for the dependence between the paid amounts

To account for the residual dependence once covariates have been factored in, a copula links the marginal regressions. For example, for $m_i = (1,0,0,1)$, i.e., one claimant with Type 2 and Expenses, the joint distribution of the amounts can be expressed in terms of a bivariate copula $C : [0,1]^2 \to [0,1]$ as

$$F(y_{i11}, y_{i14}) = C\big\{F(y_{i11} \mid x_{i,1}^Y \beta_1^Y, \sigma_1, \xi_1), F(y_{i14} \mid x_{i,1}^Y \beta_4^Y, \sigma_4, \xi_4)\big\},$$

where $C$ is assumed to be the same across all values of the covariates. This flexible representation is widely used in multivariate analysis; see, e.g., Genest and Favre (2007) or Genest and Nešlehová (2012) for details on copulas and copula models.

The dimension of the copula for claim $i$ is the number of 1s in the component vector $M_i$, or $\sum_{j=1}^{n_i} \sum_{k=1}^{4} M_{ijk}$. We expect the shape of the dependence between two components to remain unchanged when another component is present. As the claimant labels convey no information, the within-claimant dependence should be the same for all claimants. We also expect the between-claimant association to be weaker than the within-claimant association.

These constraints can be embodied into a Student $t$ copula with correlation matrix $\Sigma_i$ built such that, for components $k_1, k_2 \in \{1, \ldots, 4\}$ and claimants $j_1, j_2 \in \{1, \ldots, n_i\}$,

$$\operatorname{corr}(Y_{i,j_1,k_1}, Y_{i,j_2,k_2}) = \begin{cases} \rho_{k_1,k_2} & \text{if } j_1 = j_2, \\ \gamma\rho_{k_1,k_2} & \text{if } j_1 \neq j_2, \end{cases}$$

where $\rho_{k_1,k_2} = \rho_{k_2,k_1}$, $\gamma \in [-1, 1]$ and $\rho_{k_1,k_2} = 1$ when $k_1 = k_2$. With the parameter $\gamma$, the dependence structure can be written using only seven parameters. This is particularly appealing as there are few cases of multiple claimants who all receive Income or Caregiver benefits, which would affect the estimation of between-claimant correlations. Let $n_{\max} = \max(n_1, \ldots, n_N)$ and define the positive definite matrices

$$\Gamma = \begin{pmatrix} 1 & \gamma & \cdots & \gamma \\ \gamma & 1 & & \vdots \\ \vdots & & \ddots & \\ \gamma & \cdots & \gamma & 1 \end{pmatrix}_{n_{\max} \times n_{\max}} \quad \text{and} \quad \Lambda = \begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \rho_{1,4} \\ \rho_{1,2} & 1 & \rho_{2,3} & \rho_{2,4} \\ \rho_{1,3} & \rho_{2,3} & 1 & \rho_{3,4} \\ \rho_{1,4} & \rho_{2,4} & \rho_{3,4} & 1 \end{pmatrix}.$$

For claim $i$, the largest possible correlation matrix necessary is the Kronecker product $\Sigma = \Gamma \otimes \Lambda$, which is necessarily positive definite. The actual copula parameter for claim $i$ is the submatrix $\Sigma[\tilde{v}_i, \tilde{v}_i]$ corresponding to rows and columns of the coverages with positive amount paid for each claimant. The joint density of the log positive amounts is

$$c_\nu\{u_i; \Sigma[\tilde{v}_i, \tilde{v}_i]\} \times \prod_{(j,k) \in v_i} f(y_{ijk} \mid x_{ij}^Y \beta_k^Y, \sigma_k, \xi_k),$$

where $c_\nu\{\cdot; \Sigma\}$ is the Student $t$ copula density with $\nu$ degrees of freedom and correlation $\Sigma$, and $u_i = (F(y_{ijk} \mid x_{ij}^Y \beta_k^Y, \sigma_k, \xi_k) : (j,k) \in v_i)$ is the vector of transformed positive amounts.

**Example** (cont'd)**.** Recall that $\tilde{v}_i = \{1, 5, 7, 8\}$, so the correlation matrix is

$$\Sigma[\{1,5,7,8\}, \{1,5,7,8\}] = \begin{pmatrix} 1 & \gamma & \gamma\rho_{1,3} & \gamma\rho_{1,4} \\ \gamma & 1 & \rho_{1,3} & \rho_{1,4} \\ \gamma\rho_{1,3} & \rho_{1,3} & 1 & \rho_{3,4} \\ \gamma\rho_{1,4} & \rho_{1,4} & \rho_{3,4} & 1 \end{pmatrix}.$$

To lighten notation, denote the conditional distribution function $F(\cdot \mid x_{ij}^Y \beta_k^Y, \sigma_k, \xi_k)$ by $F_{ijk}$. In this example, the joint density of the positive amounts is given by

$$\begin{aligned} c_\nu \{F_{i11}(y_{i11}), F_{i21}(y_{i21}), F_{i23}(y_{i23}), F_{i24}(y_{i24}); \Sigma[\tilde{v}_i, \tilde{v}_i]\} \\ \times f(y_{i11} \mid x_{i1}^Y \beta_1^Y, \sigma_1, \xi_1) f(y_{i21} \mid x_{i2}^Y \beta_1^Y, \sigma_1, \xi_1) \\ \times f(y_{i23} \mid x_{i2}^Y \beta_3^Y, \sigma_3, \xi_3) f(y_{i24} \mid x_{i2}^Y \beta_4^Y, \sigma_4, \xi_4). \end{aligned}$$

## 3.5 Prior specification in the model for the amounts

In such a large model, it is preferable to use proper priors for all parameters. For the marginal amounts model, the priors are given, with $\lambda_k = 1/\sigma_k^2$, by

$$\beta_k^Y \sim \mathcal{N}(0, \sigma_\beta^2 \mathbf{I}), \quad \lambda_k \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda), \quad \xi_k \sim \mathcal{G}(\alpha_\xi, \beta_\xi),$$

where $\mathcal{G}$ denotes the Gamma distribution. The covariate effects are taken to be independent a priori, and due to the different scales of the covariates, we use the diffuse variance $\sigma_\beta^2 = 100$. Weakly informative Gamma priors are used for the variance and skewness parameters, which are both positive. For the precision parameters $\lambda_k$, we choose $\alpha_\lambda = 4$ and $\beta_\lambda = 4.5$, so that $\mathrm{E}(\lambda_k) = 4/4.5$. Setting $\alpha_\xi = \beta_\xi$ implies that $\mathrm{E}(\xi_k) = 1$, so the prior expectation is that the normal is symmetric. It is known a priori that the skewness is moderate, so a reasonable range for $\xi_k$ is $(0.5, 2)$, as seen in Figure 1 of the Supplement. Using $\alpha_\xi = \beta_\xi = 10$ leads to $\Pr(\xi_k \leq 1) = 0.54$ and $\Pr(0.5 < \xi_k \leq 2) = 0.96$.

We detail below the choice of priors for the parameters of the Student $t$ copula.

### Prior for the correlation matrix parameters

The between-claimant correlation parameter $\gamma$ and within-claimant correlations are assumed independent a priori. Because of the size of the dataset, the prior settings for both $\Lambda$ and $\gamma$ are not the primary drivers of the posterior dependence structure, but it is interesting to show how to set informative priors for these parameters. For the between-claimant parameter, many options exist: we assume $(\gamma + 1)/2 \sim \mathcal{B}(8, 4)$, i.e., a Beta distribution with shape parameters 8 and 4, so that the prior expected correlation between claimants is 1/3.

Few distributions allowing for informative inputs exist for random correlation matrices. As explained in Smith (2013) and references therein, authors use priors based on Cholesky factorization, partial correlations, or restricted independent normals (Liechty et al., 2004). In many applications, the large dimension of the matrix motivates the use of shrinkage priors, as in Pitt et al. (2006), where the off-diagonal elements of the matrix may equal 0. Barnard et al. (2000) propose two priors, the marginally uniform prior derived from the Inverse Wishart distribution on the covariance, and the jointly uniform prior. These distributions are either non-informative or shrink the correlations towards 0. When higher correlations are riskier, shrinkage priors are inappropriate.

It is unintuitive to set prior expectations on a Cholesky decomposition, so we extend the marginally uniform prior of Barnard et al. (2000) to be informative on $\Lambda$. This prior exploits the Inverse Wishart distribution for symmetric positive definite matrices and the relationship between covariance and correlation matrices. Suppose $W$ is a $k \times k$ Inverse Wishart random covariance matrix with positive definite scale matrix $\Psi$ and degrees of freedom $\nu > k$. Then one can write $W = ARA$, where $A$ is a diagonal matrix with $i$th diagonal element $a_i > 0$, and $R$ is a correlation matrix. The distribution of $R$ when $W$ has a standard Inverse Wishart distribution can be used as a prior. The derivation is shown in the Supplement, along with our choice of parameters.

### Choice of degrees of freedom

Given the correlation matrix, the parameter $\nu \in (1, \infty)$ influences the tendency of variables to vary together in high quantiles, as measured by the Tail Dependence Coefficient (TDC): low values of $\nu$ imply strong tail dependence and beyond 12–15 degrees of freedom, the difference between the Student $t$ and the Gaussian copula (which has no asymptotic dependence) becomes practically negligible. Furthermore, it is common to use a degenerate prior for the parameter $\nu$ to avoid convergence issues due to the flatness of the $t$ likelihood.

| Pair | MI | MC | ME | IC | IE | CE |
|---|---|---|---|---|---|---|
| $\hat{\nu}$ | 9.1 | 12.6 | 17.1 | 8.0 | 12.8 | 18.5 |
| TDC($\hat{\nu}$) | 0.09 | 0.07 | 0.01 | 0.08 | 0.03 | 0.01 |
| TDC(7) | 0.14 | 0.17 | 0.09 | 0.10 | 0.10 | 0.08 |
| TDC(10) | 0.08 | 0.11 | 0.05 | 0.06 | 0.05 | 0.04 |
| TDC(13) | 0.05 | 0.07 | 0.02 | 0.03 | 0.05 | 0.02 |

Table 3: Pairwise Student $t$ copula maximum pseudo-likelihood fit based on residuals from closed claims; in the table, TDC($\nu$) refers to the fitted value of the tail dependence coefficient when the degrees of freedom are set to $\nu$. The letters M, I, C, and E respectively stand for the Medical, Income, Caregiver, and Expenses component.

To inform the choice of $\nu$, we explored the dependence structure in the closed claims with rank-based inference tools for copula regression developed in Côté et al. (2019). Bivariate Student $t$ copulas were fitted by maximum pseudo-likelihood based on the ranks of the residuals from each of the within-claimant pair. The results are shown in Table 3. For the Student $t$ copula, a single value of the parameter $\nu$ must be used for all pairs. As the sampling bias may reduce the tail dependence strength in this preliminary analysis, and considering that stronger tail dependence is more conservative (thus preferable) in insurance applications, we set $\nu = 10$. Other values of $\nu$ near 10 would yield very similar posterior distributions for the other parameters (not shown).

## 4 Estimation with open claims and missing covariates

In the model described in Section 3, the full likelihood factorizes into two parts, namely

a) the likelihood relating to the types $s \equiv (s_{1,1}, \ldots, s_{N,n_N})$ and presence of Expenses $w \equiv (w_{1,1}, \ldots, w_{N,n_N})$:

$$p(s, w\,|\beta^S, \beta^W, \alpha^W, x) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\exp(x_{i,j}^S \beta_{s_{i,j}}^S)}{\sum_{t=1}^{5} \exp(x_{i,j}^S \beta_t^S)} \frac{\{\exp(x_{i,j}^W \beta^W + \alpha_{s_{i,j}}^W)\}^{\mathbf{1}(w_{i,j}=1)}}{1 + \exp(x_{i,j}^W \beta^W + \alpha_{s_{i,j}}^W)};$$

b) the likelihood for $y \equiv (y_{ijk} : i \in \{1, \ldots, N\}$ and $(j,k) \in v_i)$ given the types

$m \equiv (m_1, \ldots, m_N)$:

$$p(y \mid \beta^Y, \sigma, \xi, \gamma, \Lambda, x, m) = \prod_{i=1}^{N} c_\nu\{u_i; \Sigma[\tilde{v}_i, \tilde{v}_i]\} \prod_{(j,k)\in v_i} f(y_{ijk} \mid x_{ij}^Y \beta_k^Y, \sigma_k, \xi_k).$$

Posterior distributions for the vectors of parameters $(\beta^S, \beta^W, \alpha^W)$ and $(\beta^Y, \sigma, \xi, \gamma, \Lambda)$ can thus be computed separately in parallel, because it is assumed that the parameters for the type and presence of Expenses are a priori independent of those for the amounts.

In a case where all covariates are observed and all claims are settled, inference proceeds with classical MCMC techniques. The posterior distribution of the type parameters is simply

$$\pi(\beta^S \mid s, x) \propto p(s \mid \beta^S, x)\pi_0(\beta^S),$$

where $\pi_0(\beta^S) = \prod_\ell \phi(\beta_\ell^S; 0, \sigma_S)$, and the posterior for the parameters $(\beta^W, \alpha^W)$ is

$$\pi(\beta^W, \alpha^W \mid s, w, x) \propto p(w \mid \beta^W, \alpha^W, s, x)\pi_0(\beta^W, \alpha^W),$$

where $\pi_0$ is again the independent Gaussian prior. The posterior distribution of the amounts parameters given the component vectors $m$ is

$$\pi(\beta^Y, \sigma, \xi, \gamma, \Lambda \mid y, x, m) \propto \int p(y \mid \beta^Y, \sigma, \xi, \gamma, \Lambda, x, m)\pi_0(\beta^Y, \sigma, \xi, \gamma)\pi_0(a, \Lambda)\mathrm{d}a,$$

where

$$\pi_0(\beta^Y, \sigma, \xi, \gamma) = \prod_\ell \phi(\beta_\ell^Y; 0, \sigma_\beta) \prod_{k=1}^{4} \pi_0(\sigma_k)\pi_0(\xi_k)\pi_0(\gamma_k),$$

as specified in Section 3.5 and $\pi_0(a, \Lambda)$ is given in (1) in the Supplement.

In the present case and in many actuarial applications, however, these posteriors cannot be used directly because claim information is incomplete at evaluation date. This section describes how to handle these issues by integrating into the fully Bayesian model the fact that some covariates may be missing and unsettled claims are censored.

In Section 4.1, we detail how the missing covariates $x_{\mathrm{mis}}$ can be treated as any other parameter in the Bayesian procedure. In Section 4.2, we handle the type and presence of Expenses for the open claims. Both are censored in an unconventional way due to the fact that open claims may change type before settlement. We model the transition probabilities before settlement conditional on the gravity and the time elapsed since claim occurrence.

In Section 4.3, we model the multivariate amounts given the information known at evaluation date. In the dataset, we observed that many claims are classified as open for multiple development periods before they close without further payments. Therefore, the settlement amounts for open claims are modeled with a continuous component and a point mass at the current paid amounts. The probability associated with this point mass depends on the number of consecutive periods without payment, gravity, and the time between claim occurrence and evaluation date.

Details on the MCMC algorithm based on the procedure described in this section are provided in the Supplement, along with the resulting posterior mean and standard errors of the model parameters.

## 4.1 Treatment of missing covariates

The pain level is a covariate in the type model and is missing whenever the claimant did not answer the question "Rate your pain on a scale of 1 to 10." This may happen, e.g., when the reporting delay is long and the question is no longer relevant, or when the claimant is unable to answer questions.

As the pain level is missing in a variety of claims ranging from simple to extreme, missingness is assumed to happen at random. In that case, and in view of the relatively low percentage of missingness, the analysis can still be performed without modeling the inclusion mechanism (Gelman et al., 2004). For the type model, the posterior distribution of the parameters, given the observed covariates $x_{\text{obs}}$, is

$$\begin{aligned}\pi(\beta^S \mid s, x_{\text{obs}}) &= \int \pi(\beta^S, x_{\text{mis}} \mid s, x_{\text{obs}})\,\mathrm{d}x_{\text{mis}} \\ &\propto \int p(s \mid \beta^S, x_{\text{obs}}, x_{\text{mis}})\pi_0(\beta^S)\pi_0(x_{\text{mis}} \mid x_{\text{obs}})\,\mathrm{d}x_{\text{mis}}, \qquad (1)\end{aligned}$$

where $x_{\text{mis}}$ is the vector of missing covariates. The prior distribution of the missing values is taken to be a multinomial proportional odds model fitted with the complete open and closed records. It was designed to yield a reasonable distribution of generated pain levels given the relevant covariate values (number of injuries, reform indicator, claim gravity, civil status, and policy age). Details are given in Section 5 of the Supplement.

Although integral (1) is high-dimensional, it can be computed with data augmentation in a Gibbs sampler; see the Supplement. The conditional posterior of the missing covariate $x_{\text{mis},ij}$ for claimant $j$ in claim $i$ is

$$\pi(x_{\text{mis},ij} \mid \beta^S, s_{ij}, x_{\text{obs}}) \propto p(s_{ij} \mid \beta^S, x_{\text{obs},ij}, x_{\text{mis},ij})\pi_0(x_{\text{mis},ij} \mid x_{\text{obs}}).$$

The claimant's age, income indemnity and pain levels appear in the amounts model, and they may be missing. Assuming that they are missing at random, the posterior distribution of the amounts parameters $(\beta^Y, \sigma, \xi, \gamma, \Lambda)$ given the observables $(y, x_{\text{obs}}, m)$ is

$$\begin{aligned}\pi(\beta^Y, \sigma, \xi, \gamma, \Lambda \mid y, x_{\text{obs}}, m) \propto \int\!\!\int & p(y \mid \beta^Y, \sigma, \xi, \gamma, \Lambda, x_{\text{obs}}, x_{\text{mis}}, m) \\ & \times \pi_0(\beta^Y, \sigma, \xi, \gamma)\pi_0(a, \Lambda)\pi_0(x_{\text{mis}} \mid x_{\text{obs}}, m)\,\mathrm{d}a\,\mathrm{d}x_{\text{mis}}, \qquad (2)\end{aligned}$$

where the nuisance parameter $a$ is integrated out; see the Supplement. The pain level was found to be unrelated to age or income, so a multinomial proportional odds model was used again as a prior, but this time only the records with strictly positive paid amounts were used to fit the parameters and the claim type was included as a covariate. As pain level is involved in both the type and amounts models, the updates could in principle

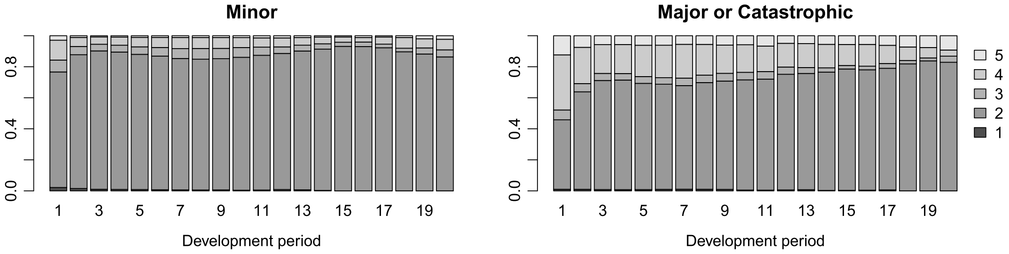


Figure 4: Distribution of ultimate type for open files in Type 2 (M) at end of period, for minor (left) or non-minor (right) claims.

be done once for both parts. However, doing so would prevent one from running the two chains in parallel and would complicate the programming with little added value.

The income indemnity level is also a multinomial proportional odds model involving the reform indicator, the number of initial injuries, the policy age, the claim type, whether the policyholder has home insurance, and the rating score. Given the income level, age is modeled by a linear regression model truncated from below at 15 and from above at 85. Details are provided in Section 5 of the Supplement.

## 4.2 Treatment of types and presence of Expenses for open files

At evaluation date, 15% of the claimant files are unsettled, and their type may change before they close. To study the behavior of the type from claim reporting to settlement, we use the development data for closed claims, available by quarter (3-month period). These data contain the history of the total payment made (thus the type) in each quarter. The period of claim occurrence is called the first development period (DP), the quarter following this is the second DP, and so on. The DP at evaluation date is the number of periods since the occurrence of the claim.

Figure 4 shows the distribution of the type at settlement for files that were open and in Type 2 (Medical) at the end of DPs 1–20. The left panel is for minor injuries, and the right one combines major and catastrophic injuries. These graphs illustrate that an open file may change type, and the probability of moving to another category depends on the gravity and the DP. Also, there are a few reimbursements (moves from 2 to 1).

Therefore, the types for open files are censored, although in an unusual way due to the reimbursement possibility and the low probability of transition to another type. This feature can be incorporated into the inference by imputation of the ultimate types for open files, given the censoring point. Let $s_{\text{op}}$ and $s_{\text{cl}}$ be the sets of ultimate types for files that are respectively open or closed at evaluation date. If the complete set $s = s_{\text{cl}} \cup s_{\text{op}}$ of types at settlement were known, the posterior for $\beta^S$ would be as in (1). As $s_{\text{op}}$ is unknown at evaluation date, however, one can only rely on the set $s^\star_{\text{op}}$ of current types for open files, yielding the following posterior for $\beta^S$:

$$\pi(\beta^S \mid s_{\text{cl}}, s^\star_{\text{op}}, x_{\text{obs}}) = \int \pi(\beta^S \mid s, x_{\text{obs}}) p(s_{\text{op}} \mid s^\star_{\text{op}}, s_{\text{cl}}, x_{\text{obs}}) \, \mathrm{d}s_{\text{op}}.$$

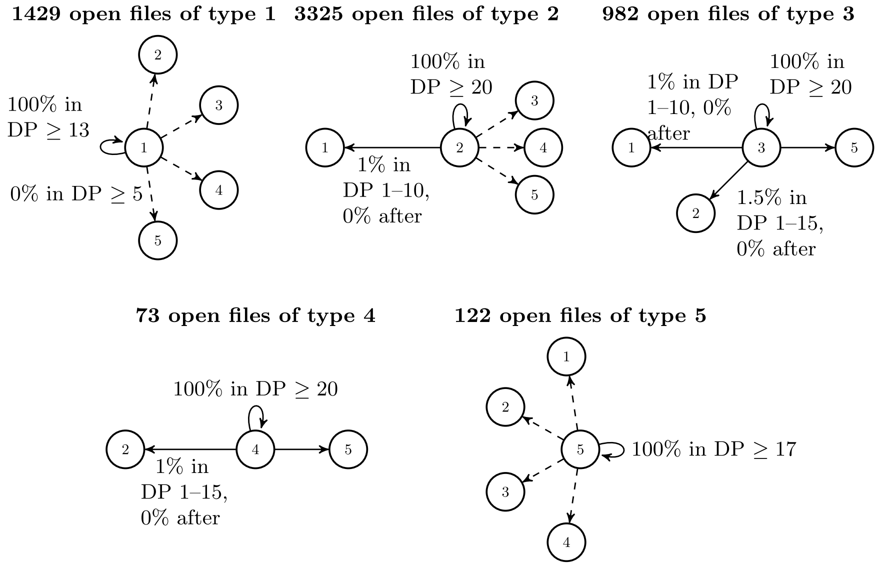


Figure 5: Transition probabilities for open types: solid lines represent fixed probabilities while dashed lines are proportional to the multinomial model probabilities.

With this posterior, we can then use data augmentation (Tanner and Wong, 1987) to simplify the estimation by imputing the possible values of $s_{\text{op}}$.

Here, the censoring is very informative and we must design an imputation model to describe the probabilities of transition from $s^\star_{\text{op}}$ to categories $\{1, \ldots, 5\}$, considering gravity and the development period at evaluation. This model is constructed separately from the multinomial model on the ultimate claim type, because the fact that the type is $s^\star_{\text{op}}$ after $t$ DPs contains more information on $s_{\text{op}}$ than the covariates known at the time of file opening. That the type changes after DP $t$ and before settlement is also plausibly unrelated to any reporting-time covariate $x_{\text{obs}}$ other than gravity. Details on the estimation of the transition probabilities are given in the Supplement. In the imputation procedure, these probabilities were treated as deterministic.

A schematic representation of the transition probabilities is depicted in Figure 5. The solid lines represent fixed probabilities, while the dashed lines are those that vary proportionately with the multinomial probabilities in the outer Bayesian model. This helps to retain the covariate effects. There are no lines when the probability is null, such as for going from type 3 (MI) to 4 (MC).

A similar analysis is performed for the presence of Expenses. In this case, there is no possibility of reimbursement, so $w^\star_{\text{op},ij} = 1$ implies $w_{\text{op},ij} = 1$ almost surely. However, if there are no Expenses at evaluation for an open file, the probability of transitioning to 1 before settlement is non-zero and is estimated using the development data, varying by DP and gravity. Estimated values are given in the Supplement.

Let $w = w_{\mathrm{op}} \cup w_{\mathrm{cl}}$ be the union of the sets of ultimate Expenses indicators for files that are respectively open or closed at evaluation date. Let also $x = x_{\mathrm{obs}} \cup x_{\mathrm{mis}}$ denote the completed covariate information. Then the posterior distribution of the type and expense parameters $\pi(\beta^S, \beta^W, \alpha^W \mid \; s_{\mathrm{cl}}, s^{\star}_{\mathrm{op}}, w_{\mathrm{cl}}, w^{\star}_{\mathrm{op}}, x_{\mathrm{obs}})$ is proportional to

$$\begin{aligned}
&\int\!\!\int \pi(\beta^S, \beta^W, \alpha^W \mid s, w, x_{\mathrm{obs}}) p(s_{\mathrm{op}} \mid s_{\mathrm{cl}}, s^{\star}_{\mathrm{op}}, x_{\mathrm{obs}}) p(w_{\mathrm{op}} \mid w_{\mathrm{cl}}, w^{\star}_{\mathrm{op}}, x_{\mathrm{obs}}) \,\mathrm{d}s_{\mathrm{op}}\,\mathrm{d}w_{\mathrm{op}} \\
&\quad \propto \int\!\!\int\!\!\int p(s \mid \beta^S, x_{\mathrm{mis}}, x_{\mathrm{obs}}) p(w \mid s, \beta^W, \alpha^W, x_{\mathrm{mis}}, x_{\mathrm{obs}}) \pi_0(\beta^S, \beta^W, \alpha^W) \\
&\qquad \times \pi_0(x_{\mathrm{mis}} \mid x_{\mathrm{obs}}) p(s_{\mathrm{op}} \mid s_{\mathrm{cl}}, s^{\star}_{\mathrm{op}}, x_{\mathrm{obs}}) p(w_{\mathrm{op}} \mid w_{\mathrm{cl}}, w^{\star}_{\mathrm{op}}, x_{\mathrm{obs}}) \mathrm{d}x_{\mathrm{mis}}\,\mathrm{d}w_{\mathrm{op}}\,\mathrm{d}s_{\mathrm{op}}.
\end{aligned}$$

To compute this integral, we update the missing covariates $x_{\mathrm{mis}}$ and the ultimate responses for open claims $w_{\mathrm{op}}$ and $s_{\mathrm{op}}$ sequentially in a Gibbs sampler using the imputation models. For details about the algorithm and the posterior means and standard errors of the parameters, consult the Supplement.

## 4.3 Treatment of the amounts for open files

Let $y_{\mathrm{op}}$ and $y_{\mathrm{cl}}$ be the sets of ultimate log amounts for files that are respectively open or closed at evaluation date. Conditional on the complete response data $y = y_{\mathrm{cl}} \cup y_{\mathrm{op}}$, the posterior for the amounts parameters would be as given in (2). As $y_{\mathrm{op}}$ is unknown at evaluation date, however, one can only rely on the set $y^{\star}_{\mathrm{op}}$ of log amounts for open files at evaluation date, which are treated as censored observations because other payments relating to these files can be made before settlement. We can then compute the posterior $\pi(\beta^Y, \sigma, \xi, \gamma, \Lambda \mid y_{\mathrm{cl}}, y^{\star}_{\mathrm{op}}, x_{\mathrm{obs}}, m)$ with the integral

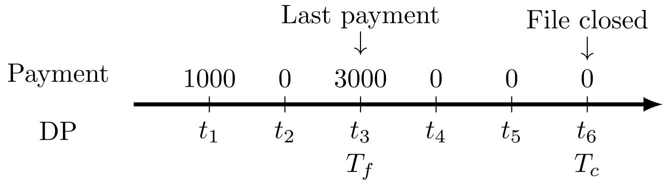
$$\int \pi(\beta^Y, \sigma, \xi, \gamma, \Lambda \mid y, x_{\mathrm{obs}}, m) p(y_{\mathrm{op}} \mid y^{\star}_{\mathrm{op}}, x, m) \,\mathrm{d}y_{\mathrm{op}}.$$

Note that if the type aggravates before settlement, then the paid amount for that component would be non-zero, but that does not inform on the amounts parameters. Accordingly, the analysis is performed conditional on $m$, and only the amounts that are strictly positive at the evaluation date are allowed to aggravate before settlement.

Many open files actually close later on without further payments, so for these claims the censored value is in fact equal to the settlement amount. It is thus necessary to account for the probability of settling at current value; otherwise the unknown settlement amounts will be overstated. The imputation model for the censored amounts has two parts: a mass at $y_{\mathrm{op},i} = y^{\star}_{\mathrm{op},i}$ representing the probability of closing without other payments, and a continuous component left truncated at $y^{\star}_{\mathrm{op},i}$.

As in Section 4.2, the development data for settled claims are used to assess the probability of closing without more payments. Let $t_{\mathrm{now}}$ be the evaluation date and $P_t$ be the payment made in DP $t$. Further, let $T_c$ be the *file closure* time, i.e., the period at the end of which the company releases the reserves for that file and closes it. If $T_c > t_{\mathrm{now}}$, then the file is open at evaluation date and the amount is censored. If $T_c \leq t_{\mathrm{now}}$, then the file is closed and the complete development is known.

Define $T_f$ as the time of *financial closure*, i.e., the period in which the last positive payment was made for a claim. We have $T_f = \sup\{t : t \le T_c, P_t > 0\}$. This is different from $T_c$ and in general, $T_f \le T_c$. For all claims that are open, i.e., for which $T_c > t_{\text{now}}$, the focus lies in whether the financial closure already occurred, i.e., whether $T_f \le t_{\text{now}}$. For example, consider the time line:



This sketch represents the cash flows of a claim that closed after six development periods, so $T_c = t_6$, but for which the time of financial closure is $T_f = t_3$, as the last payment was made in Period 3. At the end of DPs 3, 4 and 5, the file was open, but the financial closure had already happened. If $T_{\text{last}}$ is the last period, before and including $t_{\text{now}}$, where there was a strictly positive payment, i.e., $T_{\text{last}} = \sup\{t : t \le t_{\text{now}}, P_t > 0\}$, the probability of interest is

$$\Pr(T_f \le t_{\text{now}} \mid T_{\text{last}}, T_c \ge t_{\text{now}}) = \Pr(T_f = T_{\text{last}} \mid T_{\text{last}}, T_c \ge t_{\text{now}}).$$

This probability is modeled by a Bernoulli GLM with logistic link depending on three covariates: the development year $d_i$, the number of periods since last positive payment $t_{\text{now}} - T_{\text{last}}$, and the gravity. Using the development period instead of the development year yields very similar results.

There are more than one observation per claimant, and for the current estimation purpose, all observations for which the file was open at end of period, but ultimately closed before evaluation, are considered. For example, in the previous time line, the useful observations are the periods $t_1, \ldots, t_5$ for which the claim is still open at end of period, i.e., $t < T_c$. For period $t$, the response variable is the indicator $\mathbf{1}(T_f \le t)$.

Based on the development data for all claims that were settled as of December 31, 2015, the following model, selected with AIC, BIC and the optimal ROC curve, is used to simulate whether an open claim closes without other payments. The linear predictor is

$$\hat{\eta}_i = -0.5677 + 0.3512\, z_i - 0.3444\, d_i + 0.0384\, d_i^2 - 0.5749 \times \mathbf{1}_{\text{maj},i} - 0.7675 \times \mathbf{1}_{\text{cat},i}, \quad (3)$$

where $z_i = t_i - T_{\text{last},i}$ is the number of consecutive periods without payment, $d_i$ is the development year, and $\mathbf{1}_{\text{maj},i}$ and $\mathbf{1}_{\text{cat},i}$ indicate whether the gravity is major or catastrophic, respectively. The fitted probabilities are

$$\widehat{\Pr}(T_f \le t \mid T_c \ge t, T_{\text{last}}) = e^{\hat{\eta}_i}/(1 + e^{\hat{\eta}_i}).$$

Finally, the imputation distribution is the full conditional posterior, so that

$$\begin{aligned} &p(y_{\text{op}} \mid y^{\star}_{\text{op}}, x, m, \beta^{Y}, \sigma, \xi, \gamma, R, T_f) \\ &\quad \propto \left\{\mathbf{1}(y_{\text{op}} = y^{\star}_{\text{op}})\right\}^{\mathbf{1}(T_f \leq t_{\text{now}})} \left\{\mathbf{1}(y_{\text{op}} > y^{\star}_{\text{op}}) p(y_{\text{op}} \mid \beta^{Y}, \sigma, \xi, \gamma, R, x, m)\right\}^{\mathbf{1}(T_f > t_{\text{now}})}. \end{aligned}$$

For each open file, $\mathbf{1}(T_f \leq t_{\text{now}})$ can be simulated from the logistic model defined in (3). If it is 1, then $y_{\text{op}}$ is set equal to the current value $y^{\star}_{\text{op}}$, and if it is zero, $y_{\text{op}}$ is simulated using the full conditional Bayesian copula model for the amounts, truncated from below at the censoring values. This model preserves the covariate effects and the dependence. Pitfalls of this approach are the assumptions that the censoring point and the development year are uninformative on the levels of the future payments.

Details of the MCMC algorithms, starting values, proposal distributions and posterior distributions for the marginal parameters are given in the Supplement and the posterior distribution of the dependence parameters is discussed in Section 6.3.

## 5 Comparison of the proposed procedure with others using closed claims or ignoring the claim status

A good posterior predictive distribution based on the available data should resemble one that would be obtained if the complete response data were known, say the “target” posterior distribution. In this section, we illustrate how our Bayesian procedure performs better in this regard than simpler strategies. We compare three approaches:

***Closed only:*** The estimation is based solely on the data for files that were settled at the evaluation date. There is no imputation model as the complete data $(t, w, y)$ are known for this sample.

***No censoring:*** The estimation is based on the data for all claims that were reported at the evaluation date. We assume (rather optimistically) that all files are settled, as in Frees and Valdez (2008) or Frees et al. (2009) when the claim status information is not available. There is no imputation model as the complete data $(t, w, y)$ are assumed to be known.

***Imputation:*** The estimation is performed using our proposed method based on all claims that were reported at the evaluation date. The open claims are treated as censored and the imputation models presented in Sections 4.2–4.3 are used.

At the evaluation date of December 31, 2008, there are 5306 open files and sufficient data available (25,473 claimants) for a comparison. For this section, all the claims prior to December 2008 are used and the holdout sample is formed by those occurring after this point. Restricting attention to this subset of the claims allows us to obtain a nearly perfect target posterior distribution, based on the data view in December 2015. At that date, only 91 of these files were still unsettled, which represents 0.4% of the files and 3% of the total paid in 2015. As there are at least seven development years observed, the estimation of the target distribution is performed as if all claims were closed in 2015.

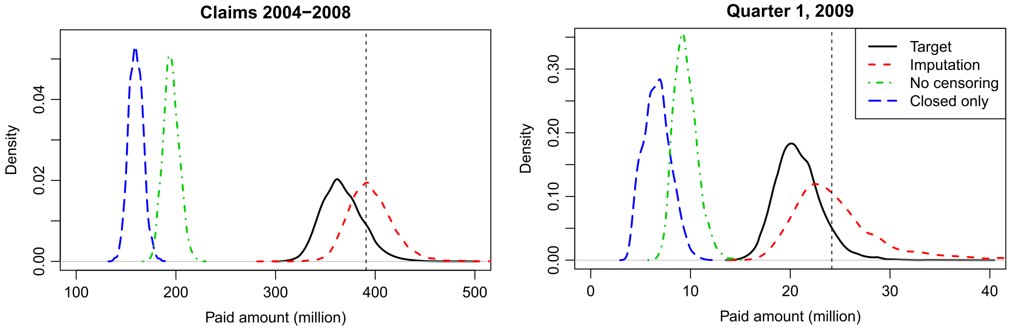


Figure 6: Posterior predictive distributions of paid amount for claims incurred in the period from 2004 to 2008 (left) and in the first quarter of 2009 (right), under the four estimation procedures.

Figure 6 shows the posterior predictive distributions of the total amount for all claims incurred from 2004 to 2008 (left), and for those incurred in the first quarter of 2009 (right). The solid (black) line is the target posterior, obtained using the data known as of 2015, and the dotted vertical line is the actual observation. These graphs illustrate the gross risk understatement arising from using only closed claims (blue curves) or including open claims but ignoring their status (green curves). These methods also produce inconsistencies in the posterior distribution of types (not shown).

As shown in Figure 6, the imputation of open claims (red curves) leads to more realistic distributions. In both cases, the observed amount falls close to the median of the red curve, but in the right tail of the (unachievable) target distribution. Indeed, while the Accident Benefits claims were known to spiral up before the legal reform, the extent of this effect had yet to be observed in the censored data as of 2008. In contrast, the black curve takes into account the (future) claim development until 2015. For the first quarter of 2009, the observed amount falls close to the median of both posterior distributions. The distribution with imputation overstates the risk slightly compared to the target distribution, as any conservative approach should.

# 6 Applications of the model and predictions

There are various actuarial applications for a claim amounts model. An insurance company may use the predictive distribution of the multivariate insurance claims for setting the initial reserves and capital, instead of basing them exclusively on the claims expert's assessment. The insurer can then decide to closely monitor claims that seem particularly risky at reporting. The posterior distributions of parameters relating to explanatory variables may be interpreted to identify the major risk factors (and maybe establish a prevention campaign) or to understand their impact, such as the effectiveness of the 2010 legislative reform. The model can also be used to determine reserves for individual claims that are "reported but not paid," i.e., those for which the insurer knows the claim characteristics but hasn't made any payments yet.

| | None | E | M | ME | MI |
|---|---|---|---|---|---|
| Observed | 15 | 0 | 6 | 1 | 0 |
| Predicted | 12.6 | 0.8 | 5.7 | 1.1 | 0.6 |
| 95% CI on Mean | $(12.0, 13.2)$ | $(0.7, 0.9)$ | $(5.2, 6.1)$ | $(1.0, 1.3)$ | $(0.5, 0.7)$ |
| | MIE | MC | MCE | MIC | MICE |
| Observed | 0 | 1 | 1 | 0 | 0 |
| Predicted | 0.3 | 1.0 | 1.0 | 0.3 | 0.5 |
| 95% CI on Mean | $(0.3, 0.4)$ | $(0.9, 1.2)$ | $(0.9, 1.2)$ | $(0.2, 0.4)$ | $(0.4, 0.6)$ |

Table 4: Observed count, mean of posterior predictive count and 95% credible interval on mean count for a given covariate vector for which all claims are closed.

## 6.1 Posterior distribution for types and presence of expenses

The multinomial regression for type, together with the binomial model for presence of expenses given the type, fits the holdout-sample data well. To confirm this, predicted and observed counts in each category can be compared, given a combination of covariates. For each unique combination $x^S$ of covariates in the multinomial model, the predicted counts are obtained as follows:

(i) Count the number $N$ of observations with this combination of covariates.

(ii) Sample $(\beta^S, \beta^W, \alpha^W)$ from their joint posterior distribution.

(iii) For each $s \in \{1, \ldots, 5\}$, compute $p(s \mid x^S, \beta^S)$ and $p(w = 1 \mid s, x^W, \beta^W, \alpha^W)$.

(iv) For type $s$, the mean counts are $Np(s \mid x^S, \beta^S)p(w = 1 \mid s, x^W, \beta^W, \alpha^W)$ with Expenses and $Np(s \mid x^S, \beta^S)\{1 - p(w = 1 \mid s, x^W, \beta^W, \alpha^W)\}$ without.

The predicted counts are the mean of the posterior distribution of the counts, estimated by the average of the counts obtained when repeating the steps (i)–(iv) many times.

For example, the following covariate combination appeared 24 times in the sample, and all files were closed: major claim in 2005, policy in force since more than five years, claimant is married, has two injuries and pain level is 4/10. In that case, the predicted mean and the 95% credible interval on the mean count are shown in Table 4; the fit seems adequate. For all combinations of covariates where $N > 10$, the predicted versus observed counts in the different categories are plotted in Figure 7 for holdout-sample observations. On average, the observed counts are greater than predicted counts for types 1-None and 2-Medical. This is expected as open claims can aggravate.

## 6.2 Posterior predictive distribution of the multivariate paid amount

To approximate the posterior predictive distribution of the multivariate paid amount for a claim or a portfolio of claims, perform the following steps many times:

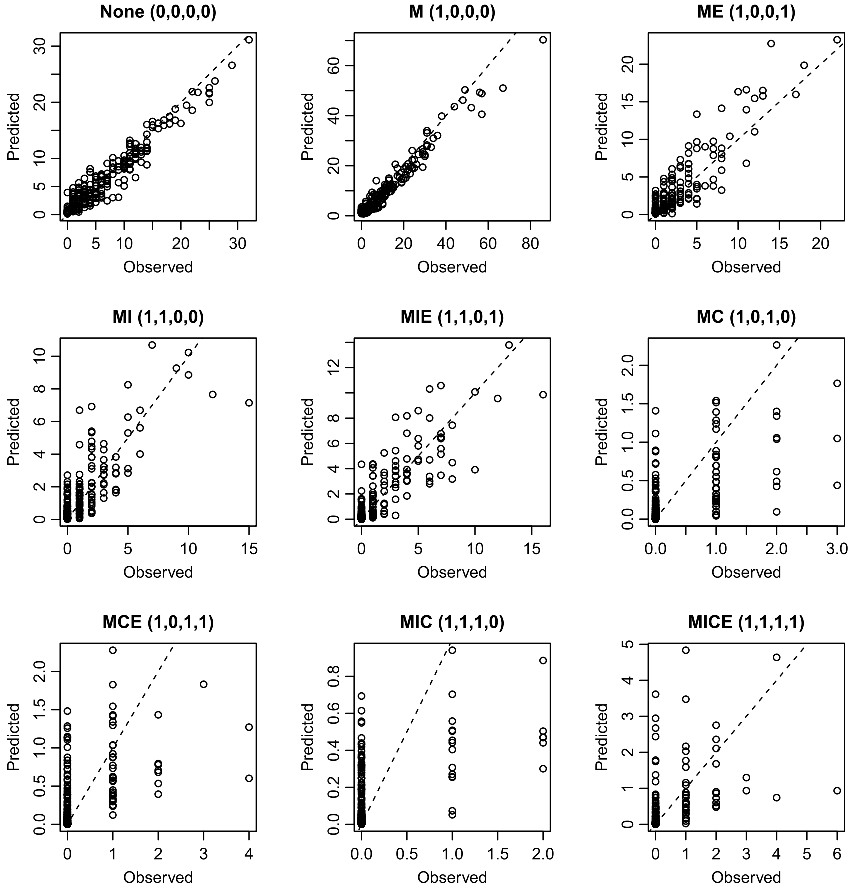


Figure 7: Predicted (posterior mean) versus observed counts in categories for holdout-sample data. A dot represents a unique level of the covariates involved in the multinomial model for which there were more than 10 (open or closed) observations.

1. Simulate $(\beta^S, \beta^W, \alpha^W, \beta^Y, \sigma, \xi, \Lambda, \gamma)$ from their joint posterior distribution.

2. If there are missing covariates, simulate $x_{\text{mis}}$ from $\pi_0(x_{\text{mis}} \mid x_{\text{obs}})$.

3. For each $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$, generate the types $S_{i,j}$ from the multinomial model.

4. Generate $W_{i,j}$ from the binomial model and compute $M_i$. Let $\Delta_i$ be the number of 1s in $M_i$.

5. If $\Delta_i = 0$, set the amount for claim $i$ to 0. If $\Delta_i = 1$, generate the log amount for claim $i$ using the relevant univariate skewed normal distribution.

| # Claimants | Amount | 50% Interval | 80% Interval | 99% Interval |
|---|---:|---|---|---|
| 1 | 0 | (0, 1355) | (0, 6465) | (0, 103,492) |
| 1 | 0 | (0, 0) | (0, 0) | (0, 8681) |
| 1 | 0 | (0, 19,793) | (0, 56,307) | (0, 358,904) |
| 1 | $\geq$ 0 | (1333, 9613) | (455, 26,656) | (0, 238,922) |
| 1 | $\geq$ 0 | (0, 0) | (0, 668) | (0, 16,581) |
| 1 | $\geq$ 367 | (1932, 19,845) | (534, 53,564) | (0, 404,716) |
| 1 | 1897 | (0, 2450) | (0, 5723) | (0, 60,443) |
| 2 | 2209 | (2181, 32,268) | (0, 76,083) | (0, 383,238) |
| 1 | 3566 | (0, 7578) | (0, 27,530) | (0, 227,052) |
| 1 | $\geq$ 11,499 | (1553, 11,027) | (483, 29,542) | (0, 247,822) |
| 1 | $\geq$ 13,183 | (1789, 14,323) | (527, 38,176) | (0, 307,987) |
| 1 | 13,491 | (11,503, 83,571) | (3642, 162,987) | (0, 721,433) |
| 2 | 48,040 | (411, 18,890) | (0, 50,073) | (0, 325,278) |
| 2 | 54,247 | (9274, 58,466) | (4163, 128,325) | (608, 651,732) |
| 1 | 65,270 | (0, 14,766) | (0, 50,264) | (0, 354,095) |

Table 5: Number of claimants, paid amount and corresponding predictive intervals for 15 randomly selected claims in the holdout set.

> If $\Delta_i > 1$, first simulate from the Student $t$ copula with correlation matrix $\Sigma[\tilde{v}_i, \tilde{v}_i]$, then use the probability inverse transform of the skewed normal to get observations of the log payments. The total amount for claim $i$ is $\sum_{(j,k)\in v_i} \exp(Y_{ijk})$.

A predictive interval for the paid amount at level $1-\alpha$ spans from the empirical quantiles $\alpha/2$ to $1-\alpha/2$ of the generated sample. Examples are shown in Table 5, where the first column is the number of claimants in the claim, the second column is the observed amount, where the sign $\geq$ designates the claims that are still open in the dataset. In this table, 80% of the closed claims fall in the 50% intervals, but 67% of all the claims fall in this range (which can change with the development of open files).

Checking model fit with censored data is a challenge: when using only settled claims, the sampling bias causes an overstatement of the paid amount on average but when the open files are included, amounts are then censored. Many 50% intervals include the case where the claim closes at 0. Note that when the upper bound of the 50% interval is 0, the posterior predictive probability of a null paid amount is in fact at least 75%.

Figure 8 shows the predictive distribution for three portfolios of 25 randomly selected claims from the holdout sample; the dashed vertical line is the censored observed paid amount. Figure 9 shows the predictive distribution of a portfolio of 500 randomly selected claims from the holdout sample. These claims comprise 607 claimant files, 70 of which are still open at evaluation. As it aggregates many claims, the posterior is closer to, but more heavy-tailed than, the normal.

For the entire holdout sample, consisting of 21,589 independent claims, the Central Limit Theorem kicks in and the predictive distribution is nearly Gaussian, as shown in

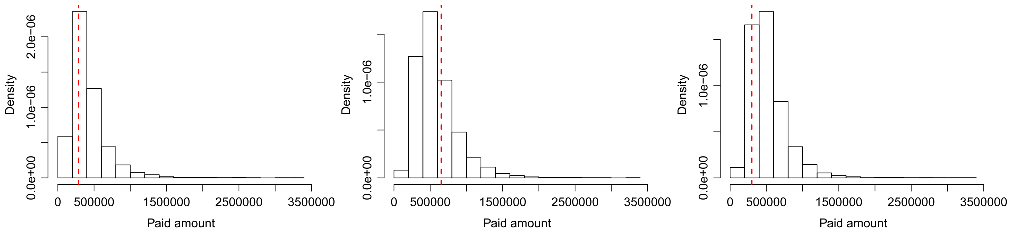


Figure 8: Predictive distributions of total paid for three portfolios of 25 randomly selected claims from the holdout sample, comprising (from left to right) 31, 37 and 34 claimants, from which 1, 3 and 6 are open, respectively.

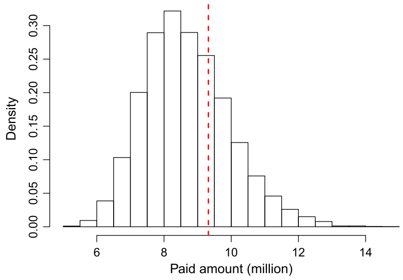


Figure 9: Predictive distribution of total paid for 500 randomly selected claims from the holdout sample, comprising 607 claimants, from which 70 are open.

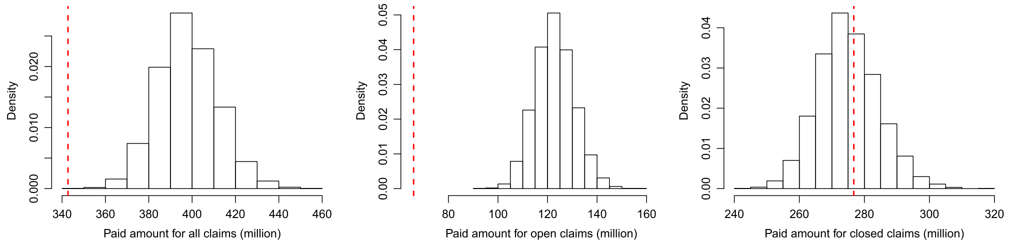


Figure 10: Predictive distribution of total paid for all (left), open (middle) and closed (right) claims in the holdout sample.

Figure 10. In the left panel, the censored total paid amount is 343M and 4063 claimant files are open. In the right panel, the observation for the closed claims falls in the middle of the predictive distribution. A 50% range for the paid amount is (388M, 407M), which translates in the interval (46M, 64M) for the outstanding amount to be paid after December 31, 2015 for the claims represented in the middle panel.

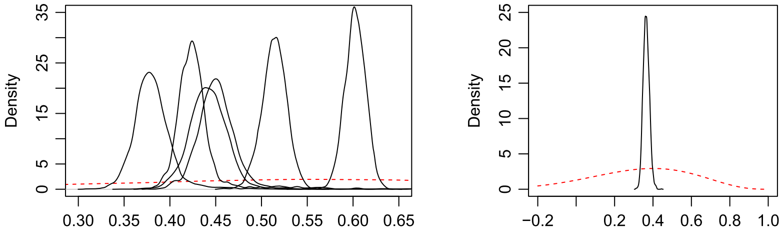


Figure 11: Prior (dashed line) versus posterior (solid line) distributions of the dependence parameters in $\Lambda$ (left) and of the between claimant correlation $\gamma$ (right).

### 6.3 Posterior distributions of the correlation parameters

The marginal posterior distributions of the correlation parameters are shown in the left panel of Figure 11, along with the common diffuse marginal prior (dashed line). The within-claimant dependence is clearly significant. Modeling of the between-claimant dependence is also justified, as the posterior distribution of $\gamma$ is narrow around 0.364, as shown in the right panel of Figure 11.

### 6.4 Joint posterior distribution of claim components

Finally, to visualize the joint predictive distribution of the components of a given claim, 10,000 realizations were drawn from the posterior distribution of Claim 5123 described in the running example. Approximately 41% of these realizations closed at 0. Figure 12 shows scatter plots of the non-negative payments (on the log scale) for the components that were ultimately observed for that claim. The points where the red lines cross correspond to the actual observation. The posterior association between the Medical and Caregiver payments for Claimant 2 is large, with the observed value of Kendall's tau equal to 39%.

## 7 Conclusion

In this article, we develop a sophisticated Bayesian model for multivariate insurance claim amounts. Major advantages of the Bayesian approach in this context are (i) the availability of the posterior predictive distribution of the paid amount, including process risk and parameter risk; (ii) the opportunity to account for expert opinion in the priors; and (iii) the possibility of taking advantage of the information in open claims and in records with missing covariates. It is crucial to appropriately include the unsettled claims in the inference, as extreme claims are more likely to be open at evaluation date.

Direct extensions of our approach include modeling the severity dispersion with covariates and using an asymmetric copula to account for the fact that high Income payments often imply high Medical payments, but not conversely. The ultimate goal is to construct a reserving model for the joint development of multivariate individual

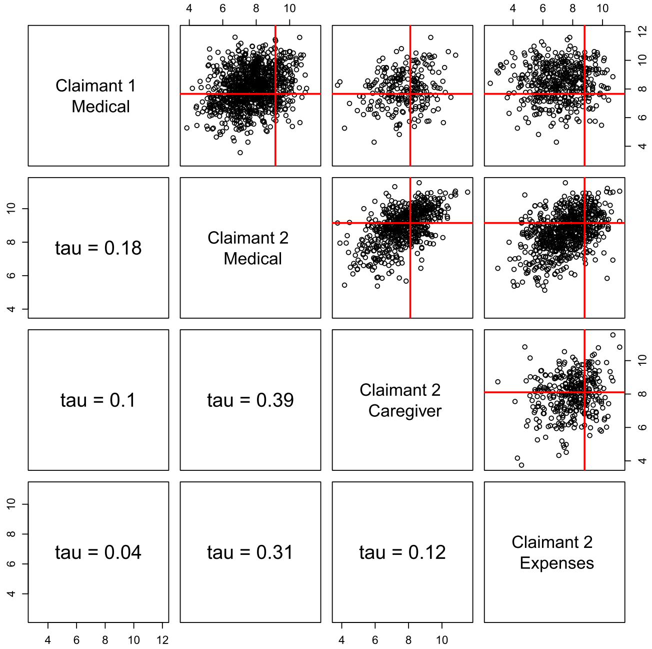


Figure 12: Realizations of the positive payments (log scale) for the claim in Example 1.

claims. While challenging, it would be advantageous to use a Bayesian approach for this task: it is ideally suited to the quantification of reserve variability.

## Supplementary Material

Supplement to "A Bayesian approach to modeling multivariate multilevel insurance claims in the presence of unsettled claims." (DOI: 10.1214/20-BA1243SUPP; .pdf). Contents: Expression for the moment generating function of the Fernández–Steel skewed normal; derivation of an informative prior on the correlation matrix and a prior specification; transition probabilities between types; regression models for missing covariates; MCMC algorithms; summary of the posterior distributions for model parameters.

## References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 10: 1281–1311. MR1804544. 76

Côté, M.-P. and Genest, C. (2015). "A copula-based risk aggregation model." *The Canadian Journal of Statistics*, 43: 60–81. MR3324428. doi: https://doi.org/10.1002/cjs.11238. 70

Côté, M.-P., Genest, C., and Omelka, M. (2019). "Rank-based inference tools for copula regression, with property and casualty insurance applications." *Insurance: Mathematics & Economics*, 89: 1–15. MR4007254. doi: https://doi.org/10.1016/j.insmatheco.2019.08.001. 77

Côté, M.-P., Genest, C., and Stephens, D. A. (2020). "Supplementary Material of "A Bayesian Approach to Modeling Multivariate Multilevel Insurance Claims in the Presence of Unsettled Claims"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1243SUPP. 69

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Chichester: Wiley. 68

Fernández, C. and Steel, M. F. J. (1998). "On Bayesian modeling of fat tails and skewness." *Journal of the American Statistical Association*, 93: 359–371. MR1614601. doi: https://doi.org/10.2307/2669632. 74

Financial Services Commission of Ontario (2010). *Changes to Automobile Insurance Regulations*. www.fsco.gov.on.ca/en/auto/autobulletins/2010/Pages/a-01_10.aspx. Last accessed on August 23, 2020. 70

Frees, E. W., Shi, P., and Valdez, E. A. (2009). "Actuarial applications of a hierarchical insurance claims model." *ASTIN Bulletin*, 39: 165–197. MR2749883. doi: https://doi.org/10.2143/AST.39.1.2038061. 68, 71, 84

Frees, E. W. and Valdez, E. A. (2008). "Hierarchical insurance claims modeling." *Journal of the American Statistical Association*, 103: 1457–1469. MR2655723. doi: https://doi.org/10.1198/016214508000000823. 68, 71, 84

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press, 3rd edition. MR3235677. 79

Genest, C. and Favre, A.-C. (2007). "Everything you always wanted to know about copula modeling but were afraid to ask." *Journal of Hydrologic Engineering*, 12: 347–368. doi: https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347). 74

Genest, C. and Nešlehová, J. (2012). "Copulas and copula models." In El-Shaarawi, A. H. and Piegorsch, W. W. (eds.), *Encyclopedia of Environmetrics*, 541–553. Chichester: Wiley, 2nd edition. 74

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). "Bayesian correlation estimation." *Biometrika*, 91: 1–14. MR2050456. doi: https://doi.org/10.1093/biomet/91.1.1. 76

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton, NJ: Princeton University Press. MR2175089. 68

Pitt, M., Chan, D., and Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models." *Biometrika*, 93: 537–554. MR2261441. doi: https://doi.org/10.1093/biomet/93.3.537. 76

Shi, P., Feng, X., and Boucher, J.-P. (2016). "Multilevel modeling of insurance claims using copulas." *The Annals of Applied Statistics*, 10: 834–863. MR3528362. doi: https://doi.org/10.1214/16-AOAS914. 68

Smith, M. S. (2013). "Bayesian approaches to copula modelling." In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (eds.), *Bayesian Theory and Applications*, 336–358. Oxford: Oxford University Press. MR3221171. doi: https://doi.org/10.1093/acprof:oso/9780199695607.003.0017. 76

Tanner, M. A. and Wong, W. H. (1987). "The calculation of posterior distributions by data augmentation." *Journal of the American Statistical Association*, 82: 528–540. MR0898357. doi: https://doi.org/10.1080/01621459.1987.10478458. 81

Trottier, D.-A. and Ardia, D. (2016). "Moments of standardized Fernández–Steel skewed distributions: Applications to the estimation of GARCH-type models." *Finance Research Letters*, 18: 311–316. doi: https://doi.org/10.1016/j.frl.2016.05.006. 74

Yang, X., Frees, E. W., and Zhang, Z. (2011). "A generalized beta copula with applications in modeling multivariate long-tailed data." *Insurance: Mathematics & Economics*, 49: 265–284. MR2811994. doi: https://doi.org/10.1016/j.insmatheco.2011.04.007. 68

### Acknowledgments

The first author acknowledges support through fellowships from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Society of Actuaries James C. Hickman Scholar program. This work was also financed through the Canada Research Chairs, the James McGill Research Chairs, and NSERC's Discovery Grant programs.