

Mixed dynamic copulae for stochastic processes

[A practical motivation and illustration]

Michal Pešta

Robust 2016

Charles University in Prague



Table of contents

1. Introduction
2. Claim development as a stochastic process
3. Accident date
4. Reporting delay
5. Number of payments
6. Claim amounts
7. Several types of claims
8. Summary as utility for the insurance company

Introduction

Joint work with **Ostap Okhrin** (TU Dresden)

Sklar's theorem (1959), cf. yesterday's talk by Nešlehová & Genest

$$H(x, y) = C(F(x), G(y))$$

Many extensions:

- **Parametric** copula

$$H(x, y; \theta) = C(F_\theta(x), G_\theta(y)), \quad H(x, y; \theta) = C_\theta(F(x), G(y)), \\ H(x, y; \theta) = C_\theta(F_\theta(x), G_\theta(y))$$

- **Time-varying** (dynamic) copula $\theta \rightarrow \theta(t)$
- **Conditional** copula

$$F_{X,Y|W}(x, y|w) = C(F_{X|W}(x|w), G_{Y|W}(y|w)|w)$$

- **Mixed** copula models ... margins: continuous as well as discrete

Claims reserving in non-life insurance

A **non-life insurance** policy is a contract between the insurer and the insured. The insurer receives a deterministic amount of money, known as premium, from the insured in order to obtain a financial coverage against well-specified randomly occurred events. If such an event (claim) happens, the insurer is obliged to pay in respect of the claim a claim amount, also known as loss amount.

Claims reserving now means, that the insurance company puts sufficient provisions from the premium payments aside, so that it is able to settle all the claims (losses) that are caused by these insurance contracts. The main issue is how to determine or **estimate** these **claims reserves**, which should be held by the insurer so as to be able to meet all future claims arising from policies currently in force and policies written in the past.

Aggregated vs granular

Claims reserving methods based on **aggregated data** from run-off triangles are **predominantly used** to calculate the claims reserves, cd. [England and Verrall, 2002] or [Wüthrich and Merz, 2008].

Disadvantages of the conventional reserving techniques

- Loss of information from the policy and the claim's development due to the aggregation
- Zero or negative cells in the triangle
- Usually small number of observations in the triangle
- Only few observations for recent accident years
- Sensitivity to the most recent paid claims

To overcome deficiencies/imperfections of the aggregated methods

Granular loss reserving methods for individual claim-by-claim data need to be derived.

Claim development as a stochastic process

Timeline of a claim

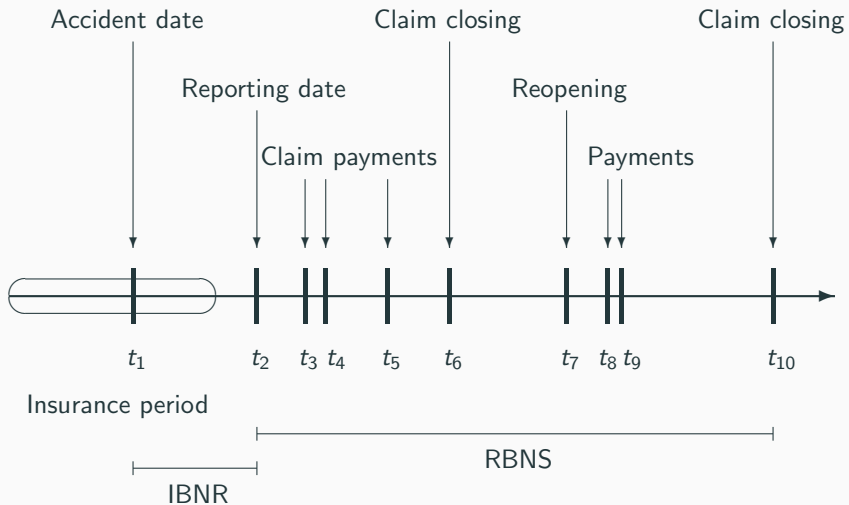


Figure 1: Time development of a non-life insurance claim.

IBNR

Incurred But Not Reported claims. These are the claims that occurred before the present moment, but will be reported in the future. Therefore, the present moment is somewhere between t_1 and t_2 in Figure 1.

RBNS

Reported But Not Settled claims. These are the claims that had occurred and were reported before the present moment, but their settlement would occur in the future. Hence in Figure 1, the present moment is somewhere between t_2 and t_6 or t_{10} .

Individual loss reserving methods, that are based on a position dependent marked Poisson process, involves work of [Norberg, 1993], [Haastrup and Arjas, 1996], [Larsen, 2007], and [Antonio and Plat, 2014]

Motivation and data structure

Nowadays, **modern databases and computer facilities** provide a foundation for loss reserving based on individual data.

The data consist of claims developments from the beginning of 2000 and are continuously updated. **Each record** in the data set contains:

- Claim ID. If one claim is associated with more payments, each payment is on a separate row.
- Type of claim. It can be either **bodily injury** or **material damage**.
- Accident date (occurrence).
- Reporting date (notification).
- Date of payment. It is a date when the payment is credited to the client's bank account.
- Amount of payment.

Accident date

Claim occurrence

Let us denote the **accident date** of the i th claims ($i = 1, 2, \dots$) by T_i and assume without losing of generality that these accident dates are chronologically ordered such that $T_{i_1} \leq T_{i_2}$ for $i_1 < i_2$. The **date differences** $V_i = T_i - T_{i-1}$ between two consecutive accident dates are supposed to be **iid** with **cdf** G and $T_0 \equiv 0$. The assumption of identically distributed accident date's differences seems reasonable, because an empirical analysis based on our database does not show any time effect on their distribution. Nevertheless, we aim at looking deeper into this question. Since we deal with approximately 50,000 claims within years 2000–2014, it is natural to assume that G is **zero-modified Poisson** or **zero-modified negative binomial**. The database of claim developments provides sufficient information such that the parameters of distribution G can be directly estimated using traditional approaches like maximum likelihood and its robustness will be investigated.

Reporting delay

Notification delay

The **reporting delay** (waiting time) of the i th claim occurred at fixed time t is denoted by $W_i(t)$, which is the time difference between the **occurrence epoch** (accident date) and the **observation epoch** (reporting date) given the occurrence time is fixed to t . For simplicity of the analysis, at first step processes $\{W_i(t)\}_t$ are supposed to be iid for all i and the conditional distribution of the reporting delay W_i with the beginning at the **random accident date** T_i given the beginning of the **time period** $T_i = t$ is H_t , i.e.,

$$P [W_i(T_i) \leq x | T_i = t] = H_t(x).$$

For instance, a Weibull distribution with time-varying parameters can be used for the conditional cdf H_t . The reasoning, why the conditional distribution of reporting delays should depend on time, comes from the exploratory analysis of the claims' database.

Accident date and reporting date

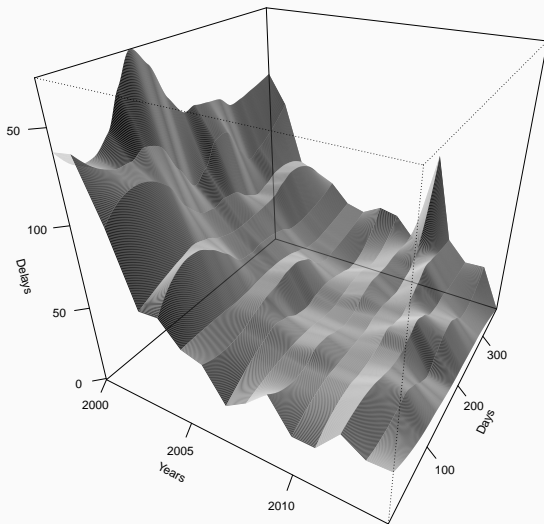


Figure 2: Reporting delays (daily medians) for different accident years and across days within the accident year.

Reporting delays are becoming shorter

The reporting delays are becoming shorter and shorter, which can be explained by a possibility of **reporting an accident over internet** and even by a **denser net of insurance company's branches**. Therefore, we restrict H_t in the way that the **conditional expectation of $W_i(T_i) | T_i = t$** should be **decreasing in t** . Conditional distribution H_t can be estimated by proposing some parametric form or by non-parametric smoothing. Nevertheless, the number of claims in the database is so high that the empirical cdf nicely serves the purpose of finding a suitable estimate for H_t .

Number of payments

Number of loss payments I

Suppose that $N_i(\tau)$ represents the number of loss payments corresponding to the i th claim during time τ after the reporting date, i.e., how many payments were carried out within the time window of length τ after the observation epoch. Time τ can be thought of an internal claim's time during which the claim is developed after being reported. In general, $\{N_i(\tau)\}_{\tau>0}$ are counting processes for all i and it is believed that they are iid. An inhomogeneous Poisson point process with intensity function $\lambda(\cdot)$ can be chosen as a candidate for such counting process. Its intensity can be consequently estimated from the historical (already reported) claims' occurrences assuming some parametric decay (e.g., polynomial or exponential).

Number of loss payments II

Since the behavior of insurance company in order to close a claim has not changed over time and empirical study from the database does not show the time effect of **accident date** on the **number of payments**, we assume that T_i is independent of $N_i(\tau)$.

For a fixed $t > 0$, $W_i(t)$ is a **continuous** random variable on the positive half of the real line and, for a fixed $\tau > 0$, $N_i(\tau)$ can be viewed as a **discrete** random variable. The dependence between the reporting delay and the number of payments is modeled by time-varying copula C_τ .

$$\begin{aligned} P [W_i(t) \leq x, N_i(\tau) \leq n | T_i = t] \\ = C_\tau \left(H_t(x), \sum_{k=0}^n \left[\int_0^\tau \lambda(s) ds \right]^k \frac{\exp\{-\int_0^\tau \lambda(s) ds\}}{k!} \right) \end{aligned}$$

Number of loss payments III

Let us assume that the actual (present) time is a and the aim is to model the number of payments in the **future time window** $(a, b]$. In order to stochastically model the number of payments within time horizon $(a, b]$, one needs to distinguish two cases according to the claim development (cf. Figure 1):

- reported but not settled (RBNS) claims,
- incurred but not reported (IBNR) claims.

Number of loss payments IV

In the **RBNS** case ($a \geq T_i + W_i(T_i)$), realizations of T_i and $W_i(T_i)$ are **observed**. Suppose that $T_i = t$ and $W_i(T_i) = w$. Then, the conditional probability of the number of payments for the i th already reported claim, given that the number of payments up to time moment a is k , is modeled as

$$\begin{aligned} & \mathbb{P} \left[N_i(b - t - w) - N_i(a - t - w) = n \mid N_i(a - t - w) = k \right] \\ &= \mathbb{P} \left[N_i(b - t - w) - N_i(a - t - w) = n \right] \\ &= \left[\int_{a-t-w}^{b-t-w} \lambda(s) ds \right]^n \frac{\exp \left\{ - \int_{a-t-w}^{b-t-w} \lambda(s) ds \right\}}{n!} \end{aligned}$$

for $n \in \mathbb{N}_0$, because of independent increments of the inhomogeneous Poisson point process.

In the **IBNR** case ($a < T_i + W_i(T_i) \leq b$), T_i and $W_i(T_i)$ are **not observable**. If $T_i + W_i(T_i) > b$, then no payment could appear in $(a, b]$. Moreover, a payment can only be proceeded if the claim has already been reported.

Number of loss payments V

Firstly, let us derive the conditional joint density of $[W_i(T_i), N_i(\tau)] | T_i$ in the similar fashion as [Krämer et al., 2013], where Q_τ stands for the cdf of $N_i(\tau)$, $C_\tau^{(1)}(u, v) = \frac{\partial}{\partial u} C_\tau(u, v)$ for $[u, v] \in (0, 1)^2$, $C_\tau^{(\tau)}(u, v) = \frac{\partial}{\partial \tau} C_\tau(u, v)$ for $\tau > 0$ and $\frac{\partial}{\partial x} H_t(x) = H'_t(x)$:

$$\begin{aligned} & \text{P} [W_i(T_i) = x, N_i(\tau(x)) = n | T_i = t] \\ &= \frac{\partial}{\partial x} \left\{ \text{P} [W_i(t) \leq x, N_i(\tau(x)) \leq n | T_i = t] \right. \\ & \quad \left. - \text{P} [W_i(t) \leq x, N_i(\tau(x)) \leq n-1 | T_i = t] \right\} \\ &= \frac{\partial}{\partial x} C_{\tau(x)}(H_t(x), Q_{\tau(x)}(n)) - \frac{\partial}{\partial x} C_{\tau(x)}(H_t(x), Q_{\tau(x)}(n-1)). \quad (1) \end{aligned}$$

Number of loss payments VI

Moreover, expressions from (1) can be further refined by assuming copula's differentiability

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} C_{\tau(\mathbf{x})}(H_t(\mathbf{x}), Q_{\tau(\mathbf{x})}(n)) &= C_{\tau(\mathbf{x})}^{(1)}(H_t(\mathbf{x}), Q_{\tau(\mathbf{x})}(n)) H_t'(\mathbf{x}) \\ &+ C_{\tau(\mathbf{x})}^{(\tau)}(H_t(\mathbf{x}), Q_{\tau(\mathbf{x})}(n)) \tau'(\mathbf{x}) \\ &+ \frac{\partial C_{\tau(\mathbf{x})}(H_t(\mathbf{x}), Q_{\tau(\mathbf{x})}(n))}{\partial Q_{\tau(\mathbf{x})}(n)} \frac{\partial Q_{\tau(\mathbf{x})}(n)}{\partial \tau(\mathbf{x})} \tau'(\mathbf{x}). \end{aligned} \tag{2}$$

Besides the practical goal, we study theoretical properties of the **mixed copulae** (i.e., copulae having continuous as well as discrete margins) like non-uniqueness or smoothness.

Number of loss payments VII

Thus, the conditional distribution of the **number of payments for the i th not reported claim**, given its accident date, its reporting delay, and the fact that it will be reported before the end of time horizon b , is

$$\begin{aligned} & P \left[N_i(b - T_i - W_i) - N_i(a - T_i - W_i) = n \right. \\ & \quad \left. | T_i = t, W_i(T_i) = w, a < T_i + W_i(T_i) \leq b \right] \\ &= P \left[N_i(b - t - w) = n | T_i = t, W_i(T_i) = w, a < T_i + W_i(T_i) \leq b \right] \\ &= \frac{P \left[N_i(b - t - w) = n | T_i = t, W_i(t) = w \right]}{P \left[a < T_i + W_i(T_i) \leq b \right]} \\ &= \frac{P \left[W_i(t) = w, N_i(b - t - w) = n | T_i = t \right]}{H'_t(w) P \left[a < T_i + W_i(T_i) \leq b \right]} \\ &= \frac{\left\{ \frac{\partial}{\partial w} C_{b-t-w}(H_t(w), Q_{b-t-w}(n)) - \frac{\partial}{\partial w} C_{b-t-w}(H_t(w), Q_{b-t-w}(n-1)) \right\}}{H'_t(w) P \left[a < T_i + W_i \leq b \right]} \end{aligned} \tag{3}$$

for $a < t + w \leq b$ and $n \geq 1$, because $N_i(a - t - w) = 0$ for the IBNR.

Claim amounts

Claim amounts

Let us denote the j th payment's amount for the i th claim by $X_{i,j}$, where $j = 1, \dots, J_i$. The $X_{i,j}$'s are iid over all j 's and i 's as well with common cdf F . This assumption is based on the empirical analysis of the pairwise relationships between the first, second, and third claim payment's amounts shown in Figure 3. Note that this is the simplest setup for the claim payment's amounts, which can be generalized for instance by assuming underlying regression model, where some covariates (e.g., time or age of the insured person) can be considered.

To sum up, identically distributed collections

$$\{T_i, W_i(T_i), N_i(\tau), \{X_{i,j}\}_j\}_i$$

are assumed for the claims.

Payment's amounts

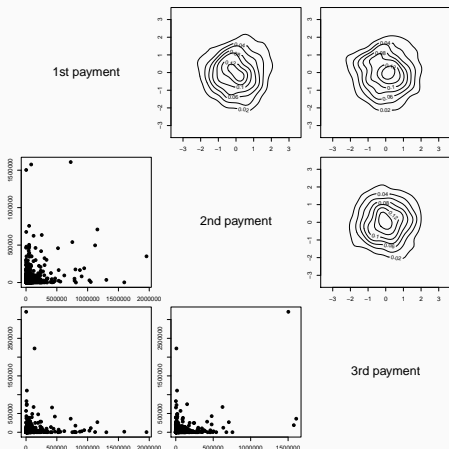


Figure 3: Pairwise relationship between claim payment's amounts (bodily injury claims). Subfigures below diagonal show $X_{i,j}$ versus $X_{i,k}$, where $j, k \in \{1, 2, 3\}$, $j \neq k$. Subfigures above the diagonal display $\Phi^{-1}\{\hat{F}_j(X_{i,j})\}$ against $\Phi^{-1}\{\hat{F}_k(X_{i,k})\}$, where \hat{F}_j and \hat{F}_k are the corresponding empirical cdfs.

Several types of claims

Several lines of business

Till this moment, we have restricted our modeling approach only to one type of claim. However, the **second type of claim** should also be taken into account and can be represented by identically distributed collections

$$\left\{ \tilde{T}_i, \tilde{W}_i(\tilde{T}_i), \tilde{N}_i(\tau), \{\tilde{X}_{i,j}\}_j \right\}_i.$$

Since there can be two types of claims on one policy, the **dependence between two types of claims** is modeled by copula D as

$$\begin{aligned} & \mathbb{P} \left[W_i(T_i) \leq x, N_i(\tau) \leq n, \tilde{W}_i(\tilde{T}_i) \leq \tilde{x}, \tilde{N}_i(\tilde{\tau}) \leq \tilde{n} \mid T_i = t, \tilde{T}_i = \tilde{t} \right] \\ &= D \left\{ C_\tau \left(\mathbb{P} \left[W_i(T_i) \leq x \mid T_i = t \right], \mathbb{P} \left[N_i(\tau) \leq n \right] \right), \right. \\ & \quad \left. \tilde{C}_{\tilde{\tau}} \left(\mathbb{P} \left[\tilde{W}_i(\tilde{T}_i) \leq \tilde{x} \mid \tilde{T}_i = \tilde{t} \right], \mathbb{P} \left[\tilde{N}_i(\tilde{\tau}) \leq \tilde{n} \right] \right) \right\}. \end{aligned}$$

Summary as utility for the insurance company

Conclusions

- Knowing the stochastic behavior of each component—**accident date, reporting delay, number of payments, claim amounts, type of claim**—and their joint relationship, one may predict future claim occurrences and payments (i.e., future cash-flows) . . . **granular loss reserving model**
- Simulate from the model using the estimated parameters and functionals many times (Monte Carlo based style) in order to obtain **simulated distributions** of the predictions

Questions?

Backup slide





Antonio, K. and Plat, R. (2014).

Micro-level stochastic loss reserving for general insurance.

Scand. Actuar. J., 2014(7):649–669.



England, P. D. and Verrall, R. J. (2002).

Stochastic claims reserving in general insurance (with discussion).

British Actuarial Journal, 8(3):443–518.



Haastrup, S. and Arjas, E. (1996).


Claims reserving in continuous time: A nonparametric bayesian approach.

ASTIN Bull., 26(2):139–164.

 Krämer, N., Brechmann, E. C., Silvestrini, D., and Czado, C. (2013).

Total loss estimation using copula-based regression models.

Insur. Math. Econ., 53(3):829–839.

 Larsen, C. (2007).


An individual claims reserving model.

ASTIN Bull., 37(1):113–132.

 Norberg, R. (1993).

Prediction of outstanding liabilities in non-life insurance.

ASTIN Bull., 23(1):95–115.

 Wüthrich, M. V. and Merz, M. (2008).

Stochastic claims reserving methods in insurance.

Wiley finance series. John Wiley & Sons.