The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2679–2704
https://doi.org/10.1214/24-AOAS1902
© Institute of Mathematical Statistics, 2024

A COPULA MODEL FOR MARKED POINT PROCESS WITH A TERMINAL EVENT: AN APPLICATION IN DYNAMIC PREDICTION OF INSURANCE CLAIMS

BY LU YANG[1,a], PENG SHI[2,b] AND SHIMENG HUANG[2,c]

[1]*School of Statistics, University of Minnesota,* [a]*luyang@umn.edu*
[2]*Wisconsin School of Business, University of Wisconsin-Madison,* [b]*pshi@bus.wisc.edu,* [c]*shimeng.huang@wisc.edu*

Accurate prediction of an insurer's outstanding liabilities is crucial for maintaining the financial health of the insurance sector. We aim to develop a statistical model for insurers to dynamically forecast unpaid losses by leveraging the granular transaction data on individual claims. The liability cash flow from a single insurance claim is determined by an event process that describes the recurrences of payments, a payment process that generates a sequence of payment amounts, and a settlement process that terminates both the event and payment processes. More importantly, the three components are dependent on one another, which enables the dynamic prediction of an insurer's outstanding liability. We introduce a copula-based point process framework to model the recurrent events of payment transactions from an insurance claim, where the longitudinal payment amounts and the time-to-settlement outcome are formulated as the marks and the terminal event of the counting process, respectively. The dependencies among the three components are characterized using the method of pair copula constructions. We further develop a stagewise strategy for parameter estimation and illustrate its desirable properties with numerical experiments.

In the application we consider a portfolio of property insurance claims for building and contents coverage obtained from a commercial property insurance provider, where we find intriguing dependence patterns among the three components. The superior dynamic prediction performance of the proposed joint model enhances the insurer's decision-making in claims reserving and risk financing operations.

1. Introduction.

Property and casualty (a.k.a. nonlife) insurance, which protects individuals and businesses against financial losses due to damage to their properties, plays a vital role in modern economies. The U.S. nonlife insurance industry collected \$655.5 billion in net premiums and paid out \$450.8 billion in property losses in 2020, according to the Insurance Information Institute. Accurate prediction of an insurer's outstanding liabilities is essential to key insurance operations and thus the financial health of the insurance sector.

This paper focuses on the dynamic prediction of outstanding liabilities for nonlife insurance companies. We adopt a micro-oriented view of an insurer's liabilities and analyze the cash flows associated with individual claims from the insurer's book of business. Our goal is to develop a statistical model for insurers to dynamically forecast unpaid losses by leveraging the granular transaction data on individual claims.

A distinctive feature of nonlife insurance is that the settlement of a claim often involves a sequence of payments that could take months or even years to complete. Figure 1 exhibits the transactions associated with an insurance claim from the time it is reported to the insurer to the time it is closed. At the present time (denoted by t_c in the figure), the quantities of interest to an insurer are the time-to-settlement (denoted by d in the figure) and the ultimate

Received June 2023; revised March 2024.
Key words and phrases. Conditional copula, joint model, life cycle of insurance claims, longitudinal and survival outcomes.

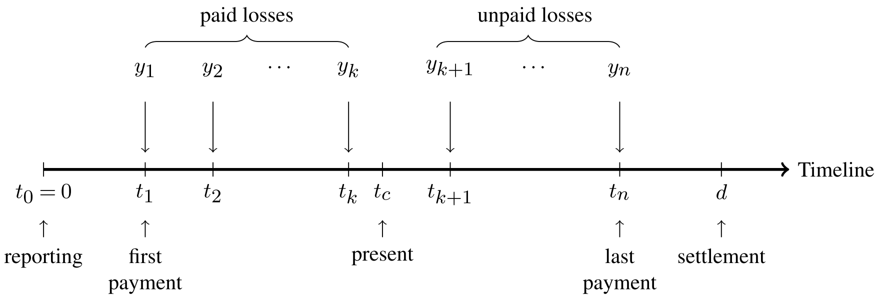

FIG. 1. *Illustration of transactions associated with an insurance claim.*

cost of the claim (i.e., $y_1 + y_2 + \cdots + y_n$). The prediction is dynamic because the insurer can continuously update the forecasts as the claim becomes more mature. As foreshadowed by Figure 1, the liability cash flow from a single claim is determined by an event process that describes the recurrences of payments, a payment process that generates a sequence of payment amounts, and a settlement process that terminates both the event and payment processes. More importantly, the three components are often dependent with one another, which is essential for the dynamic prediction of an insurer's outstanding liability.

We adopt the framework of marked point processes (see Cook and Lawless (2007)) for the joint modeling of longitudinal and survival outcomes. The joint modeling framework is critical to both inference and prediction. On the one hand, ignoring the association with the settlement outcome might introduce selection bias for the analysis of both recurrent events and longitudinal outcomes (Rizopoulos (2012)). On the other hand, the analysis of the time-to-settlement is subject to estimation bias when the repeated data on the occurrence and size of claim payments are considered as exogenous time-varying covariates (Prentice (1982)). Under the joint modeling framework, the dependence among the three components allows one to update the forecast of the claim settlement time with the most recent information on paid losses, which further feeds back into the prediction of unpaid losses.

We introduce a copula-based point process framework to model the recurrent events of payment transactions from an insurance claim, where the longitudinal payment amounts and the time-to-settlement outcome are formulated as the marks and the terminal event of the counting process, respectively. The dependencies among the three components are characterized using the method of pair copula constructions (Bedford and Cooke (2002)).

1.1. *Literature on joint models.*

The proposed copula-based approach fits in the broad literature on joint models for longitudinal and time-to-event data. Tackling the endogeneity of the longitudinal measurements and the informative drop-out of the subjects, the joint modeling framework has received extensive attention since the pioneering work of Gruttola and Tu (1994), Faucett and Thomas (1996), and Wulfsohn and Tsiatis (1997). The literature initially focused on a single longitudinal outcome and a single survival outcome and later expanded to multivariate settings (e.g., Brown, Ibrahim and DeGruttola (2005), Chi and Ibrahim (2006), Lin, McCulloch and Mayne (2002), Rizopoulos, Verbeke and Lesaffre (2009), Tang and Tang (2015), Zhu et al. (2012)). We refer readers to the comprehensive reviews of Tsiatis and Davidian (2004) and Papageorgiou et al. (2019) as well as two book-long monographs of Rizopoulos (2012) and Elashoff, Li et al. (2016).

The primary interest in the biostatistical literature is the survival outcome, and the most common framework for joint models is the shared parameter formulation. Under this framework the endogenous time-varying covariates in a survival model are typically modeled using linear mixed effects models (Laird and Ware (1982)), and the subject-specific predictors for the longitudinal outcome are also included in the relative risk model. The shared parameter model could be highly flexible in terms of nonlinear subject-specific profiles (for instance, see Brown, Ibrahim and DeGruttola (2005), Ding and Wang (2008), and Rizopoulos and Ghosh (2011)). However, parameter estimation is less straightforward as the dimension of the variance-covariance matrix for the random effects increases. Strategies to reduce computational burden include factorization of random effects (Li et al. (2012)) and use of autocorrelation structure (Proust-Lima, Dartigues and Jacqmin-Gadda (2016)).

1.2. *The copula approach.* The main difference between the copula and the shared parameter approach is the way dependence among the multivariate outcomes is induced. The shared parameter models (e.g., Kim et al. (2012), Król et al. (2016), Liu and Huang (2009), Liu, Huang and O'Quigley (2008)) use random effects to account for the association among the recurrent events, the longitudinal measurements, and the survival outcome. The strategy relies on a conditional independence assumption; that is, all outcomes are independent of each other conditioning on the random effects. In contrast, the proposed copula approach accommodates the dependence among the multiple outcomes of interest explicitly through parametric copulas. We represent the joint distribution of the multivariate outcomes in terms of a sequence of conditional bivariate distributions.

The proposed copula approach enjoys several advantages over models with random effects. First, because a copula separates the modeling of dependence from marginals, one has a wide range of strategies (for instance, linear vs. nonlinear and parametric vs. nonparametric) readily available for modeling the marginal distributions of the recurrent events, longitudinal measurements, and time-to-event data. Second, copulas are able to accommodate complex dependence structures (e.g., asymmetric and tail dependence) among the three components. Third, in absence of random effects, the proposed method is computationally efficient, which is critical to applications of high dimensions and large data. Last, the copula approach focuses on the predictive distribution of outcomes and thus provides a convenient analytical tool when the research interest goes beyond point predictions.

Copulas have previously been utilized to assist joint modeling of longitudinal and time-to-event data. In particular, Rizopoulos et al. (2008) and Rizopoulos, Verbeke and Molenberghs (2008) used a copula to specify the joint distribution of random effects in a shared parameter model. Ganjali and Baghfalaki (2015) employed a Gaussian copula to obtain the joint distribution of the survival outcome and the longitudinal measurements at fixed time points. More recently, Suresh, Taylor and Tsodikov (2021) proposed a bivariate copula model for the cross-sectional distribution of the conditional survival outcome and the longitudinal measurement at a given time while allowing this relationship to change smoothly over time. The existing copula methods heavily rely on the Gaussian copulas, which limits the dependence structure among outcomes. More importantly, they cannot be readily adapted to incorporate the third outcome, recurrent events, into the joint model for longitudinal and survival outcomes. Diao, Cook and Lee (2013) modeled the dependence between continuous marks and the event process without terminal events, while Zeller and Scherer (2022) assumed independence between the point process and its marks. In contrast to existing work, we propose a novel approach that utilizes pair copula constructions in a dynamic context, which allows for the joint modeling of the recurrent events, longitudinal measurements, and time-to-event data.

1.3. *Our contribution.* The paper makes a dual contribution. First, we introduce a copula approach for joint modeling of survival outcome and longitudinal data, specifically focusing on cases with informative observational times. In addition, we develop an efficient strategy for parameter estimation and dynamic prediction.

Second, we present an application of the proposed model in the domain of insurance operations, which stands out in two notable ways. First, it represents one of the pioneering attempts to apply joint models in business-focused areas, as far as our knowledge extends. Given the substantial size and significance of the insurance sector within the modern economy, the impact of our contribution is prominent. Second, our application underscores a distinct purpose of utilizing joint models to probabilistically forecast both the longitudinal and survival outcomes. This differs from the prevailing focus of dynamic prediction in the medical statistics literature, which is solely on survival outcomes.

The rest of the paper is organized as follows. Section 2 introduces the copula-based joint model. Section 3 discusses the strategies developed for stagewise parameter estimation and dynamic prediction. We provide comprehensive simulation studies in Section 4 to demonstrate the desirable properties of the proposed estimation strategies. Section 5 presents a detailed case study that showcases the practical application of the proposed model, utilizing a portfolio of claims from the building and contents coverage in commercial property insurance.

2. Joint models using copulas. We formulate the liability cash flow generated from an insurance claim as a marked point process. Specifically, for a given claim, the recurrence of payment transactions is characterized by a counting process, and the payment amounts and the settlement of claim are treated as the marks and the terminal event of the point process, respectively. We consider a portfolio of a large number of insurance claims from a pool of policies, where each claim generates a marked point process and the portfolio generates replications. The claims and thus the resulting marked point processes are assumed to be independent of each other.

The reporting of a claim to the insurer, which initiates the counting process for the recurrent payment event, is referred to as time origin. For the ith insurance claim, let $N_i(t)$ be the number of payments made by the insurer over time interval $[0, t]$ with convention $N_i(0) = 0$. The occurrences of payment transactions associated with the ith claim follow a counting process denoted by $\{N_i(t), 0 \le t\}$. Furthermore, we let t^+ and t^- denote the time points that are infinitesimally larger and smaller than t, respectively. The counting process is right continuous, that is, $N_i(t) = N_i(t^+)$. Define $\Delta N_i(t) = N_i((t + \Delta t)^-) - N_i(t^-)$ as the number of payments in the interval $[t, t + \Delta t)$, and denote $dN_i(t) = \lim_{\Delta t \downarrow 0} \Delta N_i(t)$. Hence, $dN_i(t) = 1$ indicates a payment occurs at t.

We denote the set of marks over time period $[0, t]$ by $\boldsymbol{Y}_i(t) = \{Y_{i1}, \dots, Y_{i,N_i(t)}\}$, where Y_{ij} is the amount of the jth payment for the ith claim. Let $S_i(t)$ denote the cumulative amount of payments by time t. Let T_{ij} be the occurrence time of the jth payment for the ith claim. We define the waiting time between payments by $W_{i1} = T_{i1}$ and $W_{ij} = T_{ij} - T_{i,j-1}$ for $j > 1$. One has the following relationships:

$$T_{ij} = W_{i1} + \cdots + W_{ij}, \quad j = 1, 2, \dots,$$

$$N_i(t) = \sum_{j=1}^{\infty} I(T_{ij} \le t),$$

$$S_i(t) = \sum_{j=1}^{N_i(t)} Y_{ij} = \sum_{j=1}^{\infty} Y_{ij} I(T_{ij} \le t),$$

where $I(\cdot)$ is the indicator function. Let $\boldsymbol{X}_i(t) = (X_{i1}(t), \dots, X_{ip}(t))'$ be the vector of external covariates, where $X_{ik}(t)$, for $k = 1, \dots, p$, is the measurement of the kth covariate at time t and can be either time-constant or time-varying. We denote the covariate process by $\boldsymbol{X}_i^{(t)} = \{\boldsymbol{X}_i(t), 0 \le t\}$ and the history of the marked point process at time t by $H_i(t) = \{N_i(s), \boldsymbol{Y}_i(s) : 0 \le s < t; \boldsymbol{X}_i^{(\infty)}\}$. In our model we assume that covariates are external, and the entire path of the time-varying covariates is known.

Now, we discuss the effects of the terminal event and denote the closing time of the ith claim by D_i. As the settlement terminates the payments for the claim, the counting process and its marks will only be observed during $[0, D_i]$. Let $D_i(t) = I(t \leq D_i)$ indicate whether the marked point process is under observation at time t. Define $d\bar{N}_i(t) = D_i(t)dN_i(t)$, and $d\bar{N}_i(t) = 1$ thereby indicates a payment occurs and is observed at time t. Furthermore, we define $\bar{N}_i(t) = \int_0^t d\bar{N}_i(t) = \int_0^t D_i(t)dN_i(t) = \sum_{j=1}^{\infty} 1(T_{ij} \leq \min(t, D_i))$ to represent the observed number of the counting process for the ith claim over $(0, t]$ and denote $\bar{\boldsymbol{Y}}_i(t) = \{Y_{i1}, \dots, Y_{i,\bar{N}_i(t)}\}$ as the corresponding observed payment amounts. The history of the observable process is included in $\bar{H}_i(t) = \{\bar{N}_i(s), \bar{\boldsymbol{Y}}_i(s), D_i(s) : 0 \leq s < t; \boldsymbol{X}_i^{(\infty)}\}$.

The probability for the observed process of claim i can be expressed in product integral notation as

$$\begin{aligned}(1)\qquad &\prod_{s\in[0,\infty)} \Pr\bigl(D_i(s)|\bar{H}_i(s)\bigr)\bigl\{\Pr\bigl(d\bar{N}_i(s)|\bar{H}_i(s), D_i(s) = 1\bigr)^{1-d\bar{N}_i(s)} \\ &\quad \times \Pr\bigl(d\bar{N}_i(s), Y_{i,\bar{N}_i(s)}|\bar{H}_i(s), D_i(s) = 1\bigr)^{d\bar{N}_i(s)}\bigr\}^{D_i(s)}.\end{aligned}$$

For convenience, we denote $N_i = N_i(D_i)$ as the total observed number of payments for claim i. Note that N_i is an induced random variable that can be fully derived using W_{ij}, $j = 1, \dots, \infty$, and D_i in our formulation, that is, $N_i = \sum_{j=1}^{\infty} I(\sum_{k=1}^{j} W_{ik} < D_i)$. Let $H_{i0} = \mathbf{X}_i^{(\infty)}$ and $H_{ij} = \{(W_{i1}, Y_{i1}), \dots, (W_{ij}, Y_{ij}), \mathbf{X}_i^{(\infty)}\}$ for $j \in \{1, \dots, N_i\}$. To jointly model the counting process, its marks, and the terminal event, we represent (1) as

$$\begin{aligned}&f\bigl(N_i(s), Y_i(s) : 0 \leq s \leq D_i|H_{i0}\bigr) \\ (2)\qquad &= f(D_i|H_{i0}) \prod_{j=1}^{N_i} f(W_{ij}, Y_{ij}|D_i, H_{i,j-1}) \\ &\quad \times \bigl\{1 - \Pr(W_{i,N_i+1} \leq D_i - T_{i,N_i}|D_i, H_{i,N_i})\bigr\}.\end{aligned}$$

We further make an assumption that conditional on D_i and covariates H_{i0}, the pairs observed at different time points $(W_{i1}, Y_{i1}), \dots, (W_{i,N_i+1}, Y_{i,N_i+1})$ are independent of one another. That is, the distribution function $F_{W_{ij},Y_{ij}|D_i,H_{i,j-1}}(w, y)$ does not explicitly depend on the history of $(W_{i1}, Y_{i1}), \dots, (W_{i,j-1}, Y_{i,j-1})$. As a result, (2) reduces to

$$(3)\qquad f(D_i|H_{i0}) \prod_{j=1}^{N_i} f(W_{ij}, Y_{ij}|D_i, H_{i0}) \cdot \bigl\{1 - \Pr(W_{i,N_i+1} \leq D_i - T_{i,N_i}|D_i, H_{i,N_i})\bigr\}.$$

Furthermore, under this assumption,

$$\Pr(W_{i,N_i+1} \leq D_i - T_{i,N_i}|D_i, H_{i,N_i}) = F_{W_{i,N_i+1}|D_i,H_{i0}}(D_i - T_{i,N_i}).$$

That is, the history of $(W_{i1}, Y_{i1}), \dots, (W_{i,N_i}, Y_{i,N_i})$ impacts the probability of terminating via the function argument. On the other hand, $F_{W_{i,N_i+1}|D_i,H_{i0}}(w|d)$, which we derive in the next section, does not depend on the history of waiting time.

2.1. *Copula model formulation.* We employ the method of pair copula constructions to develop the joint model in (3). In pair copula constructions, one constructs a multivariate distribution using bivariate copulas as building blocks. This method is flexible and applicable to data of different scales (see, for instance, Aas et al. (2009) and Shi and Yang (2018)).

We first examine the conditional distribution of the pairs (W_{ij}, D_i) and (Y_{ij}, D_i), given H_{i0}. Define marginal distributions $F_{\widetilde{W}_{ij}}(w) = F_{W_{ij}|H_{i0}}(w)$, $F_{\widetilde{Y}_{ij}}(y) = F_{Y_{ij}|H_{i0}}(y)$, and

$F_{\widetilde{D}_i}(d) = F_{D_i|H_{i0}}(d)$. We express the distributions for the two pairs using

$$F_{W_{ij},D_i|H_{i0}}(w,d) := F_{\widetilde{W}_{ij},\widetilde{D}_i}(w,d) = C_{(W,D)}\big(F_{\widetilde{W}_{ij}}(w), F_{\widetilde{D}_i}(d)\big), \tag{4}$$

$$F_{Y_{ij},D_i|H_{i0}}(y,d) := F_{\widetilde{Y}_{ij},\widetilde{D}_i}(y,d) = C_{(Y,D)}\big(F_{\widetilde{Y}_{ij}}(y), F_{\widetilde{D}_i}(d)\big), \tag{5}$$

where $C_{(W,D)}$ and $C_{(Y,D)}$ are the bivariate copulas associated with each pair. Using (4) and (5), we then construct the joint distribution for the pair (W_{ij}, Y_{ij}) conditional on D_i and H_{i0}. Define $h_{(W;D)}(u_1,u_2) = \partial C_{(W,D)}(u_1,u_2)/\partial u_2$ and $h_{(Y;D)}(u_1,u_2) = \partial C_{(Y,D)}(u_1,u_2)/\partial u_2$. We express the conditional joint distribution as

$$F_{W_{ij},Y_{ij}|D_i,H_{i0}}(w,y|d) := F_{\widetilde{W}_{ij},\widetilde{Y}_{ij}|\widetilde{D}_i}(w,y|d) = C_{(W,Y|D)}\big(F_{\widetilde{W}_{ij}|\widetilde{D}_i}(w|d), F_{\widetilde{Y}_{ij}|\widetilde{D}_i}(y|d)\big),$$

where

$$\begin{aligned} F_{\widetilde{W}_{ij}|\widetilde{D}_i}(w|d) &= h_{(W;D)}\big(F_{\widetilde{W}_{ij}}(w), F_{\widetilde{D}_i}(d),\big) \\ F_{\widetilde{Y}_{ij}|\widetilde{D}_i}(y|d) &= h_{(Y;D)}\big(F_{\widetilde{Y}_{ij}}(y), F_{\widetilde{D}_i}(d)\big), \end{aligned}$$

and $C_{(W,Y|D)}$ is the bivariate copula that joins the conditional distributions of W_{ij} and Y_{ij}, given D_i and H_{i0}.

Using above relations, we can express model (3) in terms of bivariate copulas. Specifically, for $j = 1, \dots, N_i$, we express the components of (3) as

$$\begin{aligned} & f(W_{ij}, Y_{ij}|D_i, H_{i0}) \\ &\quad = f_{\widetilde{W}_{ij}|\widetilde{D}_i}(W_{ij}|D_i) f_{\widetilde{Y}_{ij}|\widetilde{D}_i}(Y_{ij}|D_i) c_{(W,Y|D)}\big(F_{\widetilde{W}_{ij}|\widetilde{D}_i}(W_{ij}|D_i), F_{\widetilde{Y}_{ij}|\widetilde{D}_i}(Y_{ij}|D_i)\big) \\ &\quad = f_{\widetilde{W}_{ij}}(W_{ij}) f_{\widetilde{Y}_{ij}}(Y_{ij}) c_{(W,D)}\big(F_{\widetilde{W}_{ij}}(W_{ij}), F_{\widetilde{D}_i}(D_i)\big) c_{(Y,D)}\big(F_{\widetilde{Y}_{ij}}(Y_{ij}), F_{\widetilde{D}_i}(D_i)\big) \\ &\qquad \times c_{(W,Y|D)}\big(F_{\widetilde{W}_{ij}|\widetilde{D}_i}(W_{ij}|D_i), F_{\widetilde{Y}_{ij}|\widetilde{D}_i}(Y_{ij}|D_i)\big), \end{aligned} \tag{6}$$

where $c_{(W,D)}$, $c_{(Y,D)}$, and $c_{(W,Y|D)}$ are the corresponding copula densities for the three pairs. In addition, we have the probability of terminating as

$$\Pr(W_{i,N_i+1} \le D_i - T_{i,N_i}|D_i, H_{i0}) = h_{(W;D)}\big(F_{\widetilde{W}_{i,N_i+1}}(D_i - T_{i,N_i}), F_{\widetilde{D}_i}(D_i)\big). \tag{7}$$

Combining the marginal model of D_i, (6), and (7) yields model (3). Note that (6) features an employment of pair copula constructions in a low-dimensional setting, where one could relax the simplifying assumption. Specifically, one could allow the copula $C_{(W,Y|D)}(u_1, u_2|d)$ to be dependent of d.

2.2. *Copula model components.*

Recurrence of payments. The counting process for the recurrent payments is characterized by the intensity function, defined by

$$\lambda_i\big(t|H_i(t)\big) = \lim_{\Delta t \downarrow 0} \frac{\Pr(\Delta N_i(t) = 1|H_i(t))}{\Delta t},$$

which assumes that at most one event can occur at any given time. We specify

$$\lambda_i\big(t|H_i(t)\big) = h_i\big(B(t)|H_{i0}\big) = h_0\big(B(t)\big)\exp\big\{\eta\big(\boldsymbol{X}_i(t)\big)\big\}, \tag{8}$$

where $B(t) = t - T_{i,N_i(t^-)}$ is the recurrence time, that is, the time since the most recent payment and h_i is the hazard function for the waiting time. We consider a multiplicative model where h_0 is referred to as the baseline intensity and $\eta(\boldsymbol{X}_i(t))$ is a function of covariates, for instance, $\eta(\boldsymbol{X}_i(t)) = \boldsymbol{X}_i'(t)\boldsymbol{\alpha}$. We use the Weibull hazard as baseline in the application, that is, $h_0(B(t)) = p(B(t))^{p-1}$. The intensity (8) implies the conditional distribution function of the waiting time W_{ij} as

$$
\begin{aligned}
F_{\widetilde{W}_{ij}}(w) &= \Pr(W_{ij} \le w | H_{i0}) = 1 - \exp\left\{- \int_{T_{i,j-1}}^{T_{i,j-1}+w} \lambda_i(t|H_{i0}) dt\right\} \\
&= 1 - \exp\left\{- \int_0^w h_0(t) \exp\{\boldsymbol{X}_i(t + T_{i,j-1})' \boldsymbol{\alpha}\} dt\right\}.
\end{aligned} \tag{9}
$$

One further derives $f_{\widetilde{W}_j}(w) = \partial F_{\widetilde{W}_j}(w)/\partial w$, which, along with (9), is required in (6) and (7).

Amount of payments. The marks of the counting process, that is, the payment amounts, are modeled using a parametric regression based on the generalized beta of the second kind (GB2) distribution. Specifically, the density of Y_{ij} is parameterized as

$$
f_{\widetilde{Y}_j}(y) = \partial \Pr(Y_{ij} < y | H_{i0})/\partial y = \frac{\exp(\kappa_1 \omega_{ij})}{y|\sigma| B(\kappa_1, \kappa_2)[1 + \exp(\omega_{ij})]^{\kappa_1 + \kappa_2}},
$$

where $\omega_{ij} = (\ln y - \mu_{ij})/\sigma$. The GB2 distribution is defined by four parameters: μ_{ij} is the location parameter, σ is the scale parameter, and κ_1 and κ_2 are the shape parameters. With four parameters the distribution offers substantial flexibility to accommodate the skewness and heavy tails in the data. The GB2 distribution nests several well-known heavy-tailed distributions as special cases, including the generalized gamma and Burr XII distributions (McDonald and Xu (1995)), and it has been found useful particularly in modeling insurance claims (Shi (2014)). The location parameter is further specified as a function of covariates such that $\mu_{ij} = \mu(\boldsymbol{X}_{ij}) = \boldsymbol{X}'_{ij}\boldsymbol{\beta}$, where $\boldsymbol{X}_{ij} = \boldsymbol{X}(T_{ij})$.

Settlement time. The time-to-settlement outcome D_i is specified using a Cox model that can be extended to accommodate nonproportional hazards (Aalen, Borgan and Gjessing (2008)). The hazard function at time t is specified as

$$
\pi_i(t|H_{i0}) = \pi_0(t) \exp\{\boldsymbol{X}'_i(t)\boldsymbol{\gamma}\},
$$

where $\pi_0(t)$ is the baseline hazard and is left unspecified. The distribution function of settlement time is then given by

$$
F_{\widetilde{D}_i}(d) = \Pr(D_i \le d | H_{i0}) = 1 - \exp\left\{- \int_0^d \pi_0(t) \exp\{\boldsymbol{X}'_i(t)\boldsymbol{\gamma}\} dt\right\},
$$

which is required by the proposed copula model via (3), (6), and (7).

Associations. Bivariate copulas are employed to model the pairwise dependence between payment occurrence, payment amount, and settlement time, that is, $C_{(W,D)}$, $C_{(Y,D)}$, and $C_{(W,Y|D)}$ for pairs (W_{ij}, D_i), (Y_{ij}, D_i), and $(W_{ij}|D_i, Y_{ij}|D_i)$, respectively. We consider Gaussian copulas as well as their finite mixture due to their simplicity (Masarotto and Varin (2012)). Moreover, we assume that the associations for pairs (W_{ij}, D_i) and (Y_{ij}, D_i) are constant and do not depend on predictors; while for the pair $(W_{ij}|D_i, Y_{ij}|D_i)$, we relax the simplifying assumption and allow the association to vary by the value of D_i.

3. Inference and prediction.

Due to the parametric nature of the proposed copula-based joint model in Section 2, we design a likelihood-based method for its estimation and inference. Let $\boldsymbol{\theta}_W$, $\boldsymbol{\theta}_Y$, $\boldsymbol{\theta}_D$ denote the parameters in regression models for $F_{\widetilde{W}_{ij}}$, $F_{\widetilde{Y}_{ij}}$, and $F_{\widetilde{D}_i}$ respectively, and let $\boldsymbol{\theta}_\rho = (\boldsymbol{\theta}_{WD}, \boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D})$ denote the association parameters in the three bivariate copulas $C_{(W,D)}$, $C_{(Y,D)}$, and $C_{(W,Y|D)}$, respectively. All model parameters are collected into $\boldsymbol{\theta} = (\boldsymbol{\theta}_W, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_D, \boldsymbol{\theta}_\rho)$. Section 3.1 discusses a stagewise sequential maximum likelihood estimator for $\boldsymbol{\theta}$, and Section 3.2 investigates dynamic prediction based on the fitted model.

3.1. *Estimation.* Consider a portfolio of m insurance claims. We use lowercase letters to denote the realizations of the random variables defined in Section 2. For the ith ($i \in \{1, \dots, m\}$) claim, let d_i denote the realization of D_i, time-to-settlement from reporting, and n_i is the realization of N_i. During the time period $[0, d_i]$, there are n_i payments occurred at times $\{t_{ij} : j = 1, \dots, n_i\}$. For $j = 1, \dots, n_i$, let y_{ij} be the amount of payment at time t_{ij}, w_{ij} be the waiting time between payments $y_{i,j-1}$ and y_{ij}, and s_{ij} be the cumulative amount of payment at time t_{ij}. Namely, $w_{ij} = t_{ij} - t_{i,j-1}$ and $y_{ij} = s_{ij} - s_{i,j-1}$. Denote the external covariates by $\boldsymbol{x}_i^{(\infty)}$. Furthermore, define vectors $\boldsymbol{w}_i = (w_{i1}, \dots, w_{in_i})'$ and $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})'$. Given the observed data $\{\boldsymbol{w}_i, \boldsymbol{y}_i, d_i, \boldsymbol{x}_i^{(\infty)}\}_{i=1}^m$, the full log-likelihood function can be written as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^m l_i(\boldsymbol{\theta}) = \sum_{i=1}^m \log f\big(\boldsymbol{w}_i, \boldsymbol{y}_i, d_i | \boldsymbol{x}_i^{(\infty)}\big),$$

where we denote the log likelihood for the ith claim by $l_i(\boldsymbol{\theta}) = \log f(\boldsymbol{w}_i, \boldsymbol{y}_i, d_i | \boldsymbol{x}_i^{(\infty)})$ and

$$\begin{aligned}
&f\big(\boldsymbol{w}_i, \boldsymbol{y}_i, d_i | \boldsymbol{x}_i^{(\infty)}\big) \\
&\quad = f_{\widetilde{D}_i}(d_i) \prod_{j=1}^{n_i} f_{\widetilde{W}_{ij}|\widetilde{D}_i}(w_{ij}|d_i) f_{\widetilde{Y}_{ij}|\widetilde{W}_{ij},\widetilde{D}_i}(y_{ij}|w_{ij}, d_i) \\
&\quad\quad \times \big\{1 - h_{(W;D)}\big(F_{\widetilde{W}_{n_i+1}}(d_i - t_{in_i}), F_{\widetilde{D}_i}(d_i)\big)\big\} \\
&\quad = f_{\widetilde{D}_i}(d_i) \prod_{j=1}^{n_i} \big\{f_{\widetilde{W}_{ij}}(w_{ij}) c_{(W,D)}\big(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(d_i)\big)\big\} \\
&\quad\quad \times \prod_{j=1}^{n_i} \big\{f_{\widetilde{Y}_{ij}}(y_{ij}) c_{(Y,D)}\big(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(d_i)\big)\big\} \\
&\quad\quad \times \prod_{j=1}^{n_i} c_{(W,Y|D)}\big(h_{(W;D)}\big(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(d_i)\big), h_{(Y;D)}\big(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(d_i)\big)\big) \\
&\quad\quad \times \big\{1 - h_{(W;D)}\big(F_{\widetilde{W}_{n_i+1}}(d_i - t_{in_i}), F_{\widetilde{D}_i}(d_i)\big)\big\}.
\end{aligned}$$

In principle, one could estimate $\boldsymbol{\theta}$ by maximizing $L(\boldsymbol{\theta})$ directly. However, the evaluation of $L(\boldsymbol{\theta})$ can be computationally expensive. To improve computational efficiency, we propose a stagewise sequential estimation procedure as below:

(1) We estimate parameters $\boldsymbol{\theta}_D$ in $F_{\widetilde{D}_i}$ using the marginal distribution of D_i under a working independence assumption. Momentarily assuming all bivariate copulas are product copulas, the part of log-likelihood function involving $\boldsymbol{\theta}_D$ reduces to $L_1(\boldsymbol{\theta}_D) = L_D(\boldsymbol{\theta}_D)$, where $L_D(\boldsymbol{\theta}_D) = \sum_{i=1}^m \log f_{\widetilde{D}_i}(d_i)$. The estimator can be obtained via $\hat{\boldsymbol{\theta}}_D = \arg\max L_1(\boldsymbol{\theta}_D)$.

(2) We estimate parameters $\boldsymbol{\theta}_W$ in $F_{\widetilde{W}_{ij}}$ and parameters $\boldsymbol{\theta}_{WD}$ in $C_{(W,D)}$ using the distribution of $W_{ij}|D_i$, while fixing $\boldsymbol{\theta}_D = \hat{\boldsymbol{\theta}}_D$. The log-likelihood function to be maximized is equivalent to $L_2(\boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD}) = L_{W|D}(\hat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD})$, where

$$\begin{aligned}
L_{W|D}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD}) = {}& \sum_{i=1}^m \sum_{j=1}^{n_i} \log f_{\widetilde{W}_{ij}}(w_{ij}) + \sum_{i=1}^m \sum_{j=1}^{n_i} \log c_{(W,D)}\big(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(d_i)\big) \\
& + \sum_{i=1}^m \log\big\{1 - h_{(W;D)}\big(F_{\widetilde{W}_{n_i+1}}(d_i - t_{in_i}), F_{\widetilde{D}_i}(d_i)\big)\big\}.
\end{aligned}$$

One computes $(\hat{\boldsymbol{\theta}}_W, \hat{\boldsymbol{\theta}}_{WD}) = \arg\max L_2(\boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD})$.

(3) We then estimate parameters $\boldsymbol{\theta}_Y$ in the marginal of Y and the association parameters $(\boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D})$ using the distribution of $Y_{ij}|(W_{ij}, D_i)$, while fixing $\boldsymbol{\theta}_D = \hat{\boldsymbol{\theta}}_D$, $\boldsymbol{\theta}_W = \hat{\boldsymbol{\theta}}_W$, and $\boldsymbol{\theta}_{WD} = \hat{\boldsymbol{\theta}}_{WD}$. The log-likelihood function to be maximized equals to $L_3(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D}) = L_{Y|D}(\hat{\boldsymbol{\theta}}_D, \hat{\boldsymbol{\theta}}_W, \hat{\boldsymbol{\theta}}_{WD}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}) + L_{WY|D}(\hat{\boldsymbol{\theta}}_D, \hat{\boldsymbol{\theta}}_W, \hat{\boldsymbol{\theta}}_{WD}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D})$, where

$$\begin{aligned} &L_{Y|D}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}) \\ &\quad = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log f_{\widetilde{Y}_{ij}}(y_{ij}) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log c_{(Y,D)}\bigl(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(d_i)\bigr), \\ &L_{WY|D}(\boldsymbol{\theta}) \\ &\quad = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log c_{(W,Y|D)}\bigl(h_{(W;D)}\bigl(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(d_i)\bigr), h_{(Y;D)}\bigl(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(d_i)\bigr)\bigr). \end{aligned}$$

The estimator can be obtained by $(\hat{\boldsymbol{\theta}}_Y, \hat{\boldsymbol{\theta}}_{YD}, \hat{\boldsymbol{\theta}}_{WY|D}) = \arg\max L_3(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D})$.

Furthermore, one can choose copulas based on our sequential procedure. Given the marginal model, the copula $c_{(W,D)}$ should be chosen such that the likelihood in the stage (2), that is, $L_2(\boldsymbol{\theta}_W, \boldsymbol{\theta}_{WD})$ is maximized. Similarly, the copulas $c_{(Y,D)}$ and $c_{(W,Y|D)}$ should be chosen simultaneously such that $L_3(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_{YD}, \boldsymbol{\theta}_{WY|D})$ in stage (3) is maximized.

It is worth stressing that the above stagewise estimation procedure differs from the inference function for margins (IFM) (Joe (2005)), which is widely used for copula model estimation. The IFM estimates the parameters in the marginals and copulas in two separate steps. In contrast, our sequential method does not lead to complete separation between the estimation of parameters in the marginals and the copulas. Due to the dependence between the three processes, the IFM produces biased estimators for the parameters in the proposed copula model. We highlight this observation along with the desired properties of the proposed estimators in Section 4.

Under the regularity conditions in Newey and McFadden (1994), the stagewise estimator is consistent and asymptotically normal. The asymptotic covariance of $\hat{\boldsymbol{\theta}}$ admits a complicated Godambe form (Godambe (1960)). Though the standard plug-in estimator can be constructed for the asymptotic covariance of $\hat{\boldsymbol{\theta}}$, it can be quite cumbersome to implement. A practical solution to the estimation of the covariance is parametric bootstrap. The stagewise estimator is statistically less efficient than the simultaneous maximum likelihood estimator due to the working independence assumption. However, the gain in computational efficiency outweighs the loss in statistical efficiency in particular for large-scale data.

3.2. *Prediction.* One is able to make predictions for newly reported claims based on the fitted model. In practice, one could be interested in the timing and amount of individual payment transactions or, simply, the ultimate loss and the settlement time for a given claim. The probabilistic forecasts for an outcome of interest are characterized by a predictive distribution. Due to the complex structure of the model in our application, the outcome of interest might not have a predictive distribution of explicit form, for instance, the aggregate losses from a portfolio of claims. One strategy to obtain the predictive distribution is simulation. We summarize in Algorithm 1 the procedure for generating a random realization for a portfolio of m claims $\{\boldsymbol{w}_i, \boldsymbol{y}_i, d_i\}_{i=1}^{m}$. Replicating the algorithm, one obtains the predictive distribution for each individual claim as well as the portfolio.

One appealing feature of the proposed model is the dynamic prediction. That is, one is able to update the forecasts over time as new information arrives. Consider insurance claim i, and suppose there are $n_{ic} = N_i(t_c)$ payments made by an observation time t_c. We denote

Algorithm 1 Data Simulating from the Joint Model

```
For $i \in \{1, \dots, m\}$:
1. Generate covariates $\mathbf{X}_i(t) = (X_{i1}(t), \dots, X_{ip}(t))'$ for $t \geq 0$.
2. Generate a uniform variable $R_i$ and then the settlement time using $D_i = F^{-1}_{\widetilde{D}_i}(R_i)$.
3. For $j = 1$:
    3.1 Generate a bivariate uniform variable $(U_{ij}, V_{ij})$ from copula $C_{(W,Y|D)}$.
    3.2 Generate the waiting time and the corresponding payment amount using
            $W_{ij} = F^{-1}_{\widetilde{W}_{ij}}\big(g^{(W;D)}(U_{ij}, R_i)\big),$
            $Y_{ij} = F^{-1}_{\widetilde{Y}_{ij}}\big(g^{(Y;D)}(V_{ij}, R_i)\big),$
        where $g^{(W;D)}(u_1, u_2)$ is the inverse function of $h_{(W;D)}(u_1, u_2)$ with respect to the first
        argument, and $g^{(Y;D)}$ is defined similarly.
    3.3 Stop if $T_{ij} = \sum_{k=1}^{j} W_{ik} > D_i$. Otherwise set $j \leftarrow j + 1$, and go to Step 3.1.
```

the corresponding waiting times and payment amounts by $\boldsymbol{w}_{ic} = (w_{i1}, \dots, w_{i,n_{ic}})$ and $\boldsymbol{y}_{ic} = (y_{i1}, \dots, y_{i,n_{ic}})$, respectively. The (dynamic) predictive distribution for the settlement time is defined by

$$F_{D_i}\big(d|H_i(t_c), D_i > t_c\big) = \Pr\big(D_i \leq d|D_i > t_c, \boldsymbol{w}_{ic}, \boldsymbol{y}_{ic}, W_{i,n_{ic}+1} > t_c - t_{i,n_{ic}}, \boldsymbol{x}_i^{(\infty)}\big), \tag{10}$$

for $d > t_c$. The predictive distribution can be evaluated using

$$F_{D_i}\big(d|H_i(t_c), D_i > t_c\big) = 1 - \frac{\int_d p_i(s|H_i(t_c))ds}{\int_{t_c} p_i(s|H_i(t_c))ds}.$$

Here $p_i(s|H_i(t_c))$ is the conditional density of D_i and is proportional to the density of the joint distribution of D_i, $\boldsymbol{w}_{ic}$, $\boldsymbol{y}_{ic}$, and $W_{i,n_{ic}+1} > t_c - t_{i,n_{ic}}$,

$$\begin{aligned}
p_i\big(s|H_i(t_c)\big) &\propto f\big(s, \boldsymbol{w}_{ic}, \boldsymbol{y}_{ic}, W_{i,n_{ic}+1} > t_c - t_{i,n_{ic}}|\boldsymbol{x}_i^{(\infty)}\big) \\
&\propto f_{\widetilde{D}_i}(s)\big\{1 - h_{(W;D)}\big(F_{\widetilde{W}_{n_{ic}+1}}(t_c - t_{i,n_{ic}}), F_{\widetilde{D}_i}(s)\big)\big\} \\
&\quad \times \prod_{j=1}^{n_{ic}} \big\{c_{(W,D)}\big(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(s)\big)\big\} \prod_{j=1}^{n_{ic}} \big\{c_{(Y,D)}\big(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(s)\big)\big\} \\
&\quad \times \prod_{j=1}^{n_{ic}} c_{(W,Y|D)}\big(h_{(W;D)}\big(F_{\widetilde{W}_{ij}}(w_{ij}), F_{\widetilde{D}_i}(s)\big), h_{(Y;D)}\big(F_{\widetilde{Y}_{ij}}(y_{ij}), F_{\widetilde{D}_i}(s)\big)\big).
\end{aligned}$$

Note that certain terms in $p_i(s|H_i(t_c))$, such as $\prod_{j=1}^{n_{ic}} f_{\widetilde{W}_{ij}}(w_{ij}) f_{\widetilde{Y}_{ij}}(y_{ij})$, cancel out in the evaluation of (10), which simplifies the computation.

Algorithm 2 exhibits steps for generating random samples for $\{\boldsymbol{w}_i, \boldsymbol{y}_i, d_i\}$, given the history of the marked point process $H_i(t_c)$ at time t_c. The algorithm allows one to perform dynamic prediction and obtain predictive distributions for the outcomes of interest, be it either the settlement time or the outstanding payments, for individual claims as well as the portfolio of claims. In the simulation we generate D_i from (10), using the empirical supremum rejection sampling algorithm (Caffo, Booth and Davison (2002)) in which the normalizing constant $\hat{C}$ is chosen empirically (see, for instance, Peng (2018)). We use the conditional distribution of D_i, given $D_i > t_c$, denoted by $f_{\widetilde{D}_i}(\cdot|D_i > t_c)$, as the candidate

Algorithm 2 Data Simulating from the Dynamic Prediction

```
For $i \in \{1, \dots, m\}$, if $D_i \leq t_c$, stop. Otherwise:
1. Generate covariates $\mathbf{X}_i(t) = (X_{i1}(t); \dots; X_{ip}(t))'$ for $t \geq t_c$.
2. Generate the settlement time $D_i$ using rejection sampling. For $k = 1$:
    2.1 Initialize $\hat{C}_k$
    2.2 Draw $U_k \sim \text{Uniform}(0, 1)$.
    2.3 Draw $D_k^* \sim f_{\widetilde{D}_i}(\cdot | D_i > t_c)$.
    2.4 Set $D_i = D_k^*$ if $U_k \leq \frac{p_i(D_k^* | H_i(t_c))}{\hat{C}_k f_{\widetilde{D}_i}(D_k^* | D_i > t_c)}$, and stop.
        Otherwise, update $\hat{C}_k \leftarrow \max\{\hat{C}_k, \frac{p_i(D_k^* | H_i(t_c))}{f_{\widetilde{D}_i}(D_k^* | D_i > t_c)}\}$ and $k \leftarrow k + 1$, and go to Step 2.1.
3. Generate $(W_{i,n_{ic}+1}, \dots, W_{i,n_i})$ and $(Y_{i,n_{ic}+1}, \dots, Y_{i,n_i})$ using Algorithm 1. For $j = n_{ic} + 1$, go to Step 3.1–Step 3.3 in Algorithm 1. One can ensure $W_{i,n_{ic}+1} > t_c - t_{i,n_{ic}}$ by using the conditional distribution to generate
        $W_{i,n_{ic}+1} = F^{-1}_{\widetilde{W}_{i,n_{ic}+1}}\{g^{(W;D)}[U_{i,n_{ic}+1}(1 - h_{(W;D)}(F_{\widetilde{W}_{ij}}(t_c - t_{i,n_{ic}}), F_{\widetilde{D}_i}(D_i)))$
        $+ h_{(W;D)}(F_{\widetilde{W}_{ij}}(t_c - t_{i,n_{ic}}, F_{\widetilde{D}_i}(D_i)]\}$
   in 3.2 of Algorithm 1.
```

distribution. The starting value of the constant in the algorithm is chosen to be the ratio $p_i(d_i | H_i(t_c)) / f_{\widetilde{D}_i}(d_i | D_i > t_c)$ for a randomly generated $d_i \sim f_{\widetilde{D}_i}(\cdot | D_i > t_c)$.

Dynamic prediction for the survival outcome has been the center of interest in the current literature on the joint models for longitudinal and survival data. In contrast, the focus of dynamic prediction in our application is not only the settlement time but also, and often more importantly, the longitudinal payments. In Section 4 we provide additional numerical experiments to demonstrate the application of dynamic prediction and emphasize the effect of dependence misspecification on prediction.

4. Numerical experiments.

We perform two sets of numerical experiments to explore the operating characteristics of the proposed methodology. The first set is to explore the finite sample performance of the stagewise estimation for the copula-based joint model, and the second set is to highlight the dynamic prediction using the proposed joint model.

4.1. *Settings.*

This section describes the data-generating process for the numerical experiments. We set $\boldsymbol{X}_i(t) = (X_{i1}(t), X_{i2}(t))'$, where $X_{i1}(t)$ is assumed to be time constant and $X_{i2}(t)$ time-varying. Denote $X_{i1}(t) =: X_{i1}$, and let $X_{i1} \sim N(0, 1)$. For $X_{i2}(t)$, we consider a piecewise constant covariate process. Specifically, let $X_{i2}(t) = X_{i2}(T_{ij}) =: X_{i2,j}$ for $T_{i,j-1} < t \leq T_{ij}$, and assume $X_{i2,j} \sim N(0, 0.5^2)$ independently. In our model we assume that the entire path of the covariates is known.

For the occurrence of payments, we use a Weibull baseline intensity function, namely, $h_0(t) = pt^{p-1}$ with $p = 2$. For $t \in (T_{i,j-1}, T_{ij}]$, the intensity is $\lambda_i(t | \mathbf{X}_i^{(\infty)}) = pB(t)^{p-1} \times \exp\{\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2,j}\}$, and we let $(\alpha_0, \alpha_1, \alpha_2) = (-1, 1, 1)$. The conditional cumulative distribution of the waiting time W_{ij} has the form

$$F_{\widetilde{W}_{ij}}(w) = 1 - \exp\{-\exp\{\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2,j}\} w^p\}.$$

For the marginal model of payment amount Y_{ij}, we use a gamma regression model, which is a special case of the GB2 distribution. Its mean parameter is $\mu_{ij} = \exp\{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2,j}\}$, and we let $(\beta_0, \beta_1, \beta_2) = (1, 1, 1)$. The dispersion parameter ϕ is set to be 2.

TABLE 1
Summary statistics for simulation

	Low dependence				High dependence			
	D	N	W	Y	D	N	W	Y
Mean	10.120	5.942	1.531	6.455	10.120	6.049	1.502	5.982
SD	6.486	2.540	1.338	18.596	6.486	2.054	1.366	16.746

For the time-to-settlement outcome, we set the baseline hazard function to be piecewise constant. Denote the cutoff points by $a_0 = 0 < a_1 < a_2 < a_3 = \infty$ and use $a_1 = 5$, $a_2 = 10$. The baseline rate function takes the form

$$\pi_0(t) = \pi_1 I(a_0 < t \leq a_1) + \pi_2 I(a_1 < t \leq a_2) + \pi_3 I(t > a_2).$$

The distribution function of settlement time D_i is

$$F_{\tilde{D}_i}(d) = 1 - \exp\left\{-\exp\{X_{i1}\gamma_1\} \sum_{k=0}^{2} (\pi_{k+1} \max\{0, \min(a_{k+1} - a_k, d - a_k)\})\right\}.$$

We set the parameters to be $\gamma_1 = 1, \pi_1 = 0.01,\ \pi_2 = 0.1, \pi_3 = 0.8$.

We use bivariate Gaussian copulas in the simulation. Specifically, the copulas $C_{(W,D)}$, $C_{(Y,D)}$, and $C_{(W,Y|D)}$ are set to be Gaussian with association parameters θ_{WD}, θ_{YD}, $\theta_{WY|D}$, respectively. We consider three levels of dependence, varying from low ($\theta_{WD} = \theta_{YD} = \theta_{WY|D} = 0.2$), medium ($\theta_{WD} = \theta_{YD} = \theta_{WY|D} = 0.5$), to high ($\theta_{WD} = \theta_{YD} = \theta_{WY|D} = 0.8$). For illustration we assume constant association parameters in the copulas in the simulation study. Nonetheless, our model can easily accommodate more flexible dependence structures such as time-varying dependence and conditional dependence. Table 1 includes the summary statistics of the response variables in one of the replicates, under the low and high dependence.

4.2. *Finite sample performance.* In the experiment we let the number of claims m be 500 and 1000 and the association parameters be low, medium, and high. For each scenario we replicate our simulation 1000 times. The estimation results from the proposed stagewise procedure are summarized in Tables 2 and 3. We display the relative bias and standard deviation across the 1000 replicates. We also report the coverage rate of the 95% bootstrap confidence interval. We observe that, first, across all the scenarios with different sample sizes and dependence levels, our estimation procedure has excellent performance with a negligible bias, a small standard deviation, and a correct coverage level. Second, as the sample size increases, the bias and standard deviation reduce, as expected.

We further highlight that ignoring dependence can lead to biased estimators in our settings. Table 4 shows the estimation results of the IFM method. We can see from Table 4 that parameter estimates from this procedure are biased, in particular, the parameters in the marginal model of Y, as discussed in Section 3.1. The issue with IFM is also evidently reflected in the undesirable coverage rates of the bootstrap confidence intervals.

4.3. *Robustness against copula misspecification.* In this section we evaluate the robustness of the proposed method in situations where the copula family is misspecified. We simulate data with a Gumbel copula, which is featured with upper tail dependence, and a Clayton copula, which entails lower tail dependence. However, the data are mistakenly fit with Gaussian copulas. To ensure comparability, we set the parameters of different copulas such that the Kendall's tau is 0.5. Tables 5 and 6 present the results. Here we omit the results for the

TABLE 2
Stagewise estimation results

			Settlement time D				Waiting time W			
	Dependence	m	γ	π_1	π_2	π_3	α_0	α_1	α_2	p
True	Low		1	0.01	0.1	0.8	−1	1	1	2
Relative bias		500	0.001	0.003	0.002	0.004	0.001	0.001	0.001	0.001
SD			0.052	0.002	0.008	0.054	0.031	0.025	0.038	0.030
Coverage			0.956	0.948	0.948	0.942	0.938	0.938	0.947	0.940
Relative bias		1000	0.000	0.005	−0.000	0.004	0.000	−0.000	0.000	0.001
SD			0.037	0.001	0.005	0.037	0.021	0.017	0.027	0.021
Coverage			0.946	0.946	0.963	0.949	0.953	0.946	0.947	0.943
True	Medium		1	0.01	0.1	0.8	−1	1	1	2
Relative bias		500	0.001	0.003	0.002	0.004	0.001	0.001	0.001	0.001
SD			0.052	0.002	0.008	0.054	0.040	0.030	0.036	0.034
Coverage			0.952	0.950	0.949	0.942	0.941	0.941	0.943	0.944
Relative bias		1000	0.000	0.005	−0.000	0.004	0.000	0.000	0.000	0.000
SD			0.037	0.001	0.005	0.037	0.027	0.021	0.026	0.024
Coverage			0.946	0.945	0.961	0.948	0.955	0.947	0.938	0.953
True	High		1	0.01	0.1	0.8	−1	1	1	2
Relative bias		500	0.001	0.003	0.002	0.004	0.000	0.001	0.001	0.001
SD			0.052	0.002	0.008	0.054	0.053	0.041	0.033	0.047
Coverage			0.954	0.950	0.948	0.941	0.949	0.956	0.941	0.940
Relative bias		1000	0.000	0.005	−0.000	0.004	−0.000	0.000	0.001	0.001
SD			0.037	0.001	0.005	0.037	0.036	0.029	0.023	0.033
Coverage			0.948	0.943	0.962	0.950	0.951	0.949	0.947	0.940

TABLE 3
Stagewise estimation results (*continued*)

			Payment amount Y				Dependence		
	Dependence	m	β_0	β_1	β_2	ϕ	θ_{WD}	θ_{YD}	$\theta_{WY\|D}$
True	Low		1	1	1	2	0.2	0.2	0.2
Relative bias		500	−0.001	−0.001	−0.002	−0.001	0.002	−0.003	−0.003
SD			0.028	0.027	0.050	0.044	0.020	0.020	0.018
Coverage			0.946	0.942	0.951	0.958	0.953	0.945	0.943
Relative bias		1000	−0.001	0.001	−0.001	0.000	−0.000	0.000	−0.001
SD			0.019	0.020	0.036	0.032	0.014	0.014	0.013
Coverage			0.954	0.946	0.940	0.955	0.947	0.942	0.948
True	Medium		1	1	1	2	0.5	0.5	0.5
Relative bias		500	−0.001	−0.002	−0.002	−0.001	−0.001	−0.001	−0.001
SD			0.037	0.036	0.045	0.053	0.018	0.017	0.014
Coverage			0.945	0.946	0.959	0.947	0.943	0.950	0.950
Relative bias		1000	−0.002	0.000	−0.000	−0.000	−0.002	−0.000	−0.001
SD			0.024	0.025	0.033	0.036	0.012	0.012	0.010
Coverage			0.961	0.942	0.940	0.955	0.954	0.958	0.951
True	High		1	1	1	2	0.8	0.8	0.8
Relative bias		500	−0.001	0.000	−0.000	0.000	−0.001	−0.001	−0.000
SD			0.053	0.051	0.033	0.070	0.010	0.010	0.006
Coverage			0.936	0.953	0.950	0.946	0.948	0.943	0.953
Relative bias		1000	−0.002	0.001	−0.000	−0.000	−0.001	−0.001	0.000
SD			0.036	0.037	0.024	0.048	0.007	0.007	0.005
Coverage			0.955	0.944	0.939	0.946	0.946	0.942	0.949

TABLE 4
IFM estimation results

	Occurrence of payment W				Payments Y				Dependence		
	α_0	α_1	α_2	p	β_0	β_1	β_2	ϕ	θ_{WD}	θ_{YD}	$\theta_{WY\mid D}$
True	−1	1	1	2	1	1	1	2	0.800	0.800	0.800
Relative bias	0.016	0.013	0.044	0.028	−0.117	0.028	0.025	−0.014	−0.008	−0.009	−0.015
SD	0.054	0.046	0.044	0.050	0.053	0.052	0.053	0.073	0.010	0.010	0.007
Coverage	0.924	0.929	0.802	0.749	0.333	0.904	0.916	0.911	0.850	0.859	0.604

survival outcome, as they are not impacted by the dependence structure. Our estimator behaves similarly to the maximum likelihood estimators of misspecified models (White (1982)). We can see that, first, the parameters in the marginal models are still reasonably estimated, as evidenced by the small bias, small standard deviation, and correct coverage, despite the misspecification in the copula family. This stands in contrast to the results of IFM in Table 4. Second, the copula maintains the strength of dependence, even under misspecification. In the last three columns of Tables 5 and 6, we convert Kendall's tau into the dependence parameters in the corresponding Gaussian copulas. It is important to note that the value of 0.707 is not the true underlying dependence parameters, since the data are not generated from Gaussian copulas. Nevertheless, we can see that our method identifies the strength of dependence with a small bias. The low coverage rate of the dependence parameter is expected due to the misspecification.

4.4. *Predictive accuracy.* In this section we investigate prediction based on the proposed model. There are two quantities of particular interest to analysts in our application, the settlement time and the ultimate loss amount for individual claims. Specifically, for the ith claim, we are to forecast the settlement time D_i and total payment for the claim $S_i = \sum_{j=1}^{N_i} Y_{ij}$, given the history up to time t_c. We consider a portfolio of 500 insurance claims that follow the data-generating configuration outlined in Section 4.1. For each claim we obtain point forecasts and predictive distributions of the two outcomes using the simulation method in Algorithm 2. To demonstrate the dynamic prediction, we let t_c vary to be 5 and 8. Below we present the results in the high dependence scenario, that is, $\theta_{WD} = \theta_{YD} = \theta_{WY|D} = 0.8$, based on 500 replicates.

The point predictions for the settlement time and total payments for a claim are calculated using the means of the corresponding predictive distributions. To evaluate the overall accuracy, we employ the mean absolute prediction error (MAPE) and the mean squared prediction error (MSPE) to measure the closeness between the actual and the predicted values. The out-of-sample statistics are reported in Table 7. Comparing the first and third rows, we can see that as t_c increases, implying that we have more information at hand, the prediction

TABLE 5
Results of a Gumbel copula misspecified with a Gaussian copula

	Occurrence of payment W				Payments Y				Dependence		
	α_0	α_1	α_2	p	β_0	β_1	β_2	ϕ	θ_{WD}	θ_{YD}	$\theta_{WY\mid D}$
True	−1	1	1	2	1	1	1	2	0.707	0.707	0.707
Relative bias	−0.013	0.008	0.002	0.001	−0.036	−0.009	0.001	−0.0004	−0.004	0.0004	−0.043
SD	0.046	0.039	0.031	0.040	0.053	0.050	0.034	0.063	0.014	0.015	0.011
Coverage	0.944	0.945	0.962	0.957	0.874	0.942	0.956	0.943	0.948	0.954	0.227

TABLE 6

Results of a Clayton copula misspecified with a Gaussian copula

	Occurrence of payment W				Payments Y				Dependence		
	α_0	α_1	α_2	p	β_0	β_1	β_2	ϕ	θ_{WD}	θ_{YD}	$\theta_{WY\|D}$
True	-1	1	1	2	1	1	1	2	0.707	0.707	0.707
Relative bias	-0.017	-0.020	-0.006	-0.004	0.019	0.021	0.001	0.016	-0.033	-0.041	-0.035
SD	0.051	0.034	0.038	0.049	0.038	0.039	0.047	0.075	0.018	0.018	0.012
Coverage	0.937	0.916	0.960	0.938	0.920	0.917	0.956	0.932	0.746	0.619	0.466

accuracy improves in both settlement time and total payment. For comparison we also make predictions for the portfolio of claims assuming independence between waiting time, payment amount, and the settlement of claim. The MAPE and MSPE under the independence model are shown in the second and fourth rows of Table 7. The prediction error of the independence model is significantly higher than the dependence model in all cases, indicating that ignoring dependence could lead to suboptimal prediction.

Our copula-based method provides not only point forecasts but also, more importantly, the entire predictive distributions for the outcomes of interest. We obtain the predictive distributions for the settlement time and total payment for each claim in the portfolio, and we assess the prediction accuracy using a uniform test. Specifically, we first obtain the predictive distribution of D_i, denoted as $\hat{F}_{D_i}(\cdot|H_i(t_c), D_i > t_c)$, using Algorithm 2. Using the actual values of d_i for which $d_i > t_c$, we then construct a sequence of probability integral transforms $\hat{F}_{D_i}(d_i|H_i(t_c), D_i > t_c)$. If the predictive distribution is precise, we expect that $\hat{F}_{D_i}(d_i|H_i(t_c), D_i > t_c)$, for $i = 1, \dots, m$ and $d_i > t_c$, closely follow a uniform distribution on [0,1]. The same procedure is applied to S_i, whose probability integral transforms are denoted as $\hat{F}_{S_i,t_c}(s_i)$. The uniform tests are performed for $t_c = 5$ and $t_c = 8$, and the resulting QQ plots are displayed in Figure 2. To facilitate visualization, we conduct an inverse normal transformation, that is, $\Phi^{-1}(\hat{F}_{D_i}(d_i|H_i(t_c), D_i > t_c))$ and $\Phi^{-1}(\hat{F}_{S_i,t_c}(s_i))$, and normality is thereby the null pattern. All the plots in Figure 2 are reasonably close to the diagonal, implying our predictive distributions are accurate for both settlement time and total payment.

Finally, we look into the predictive distributions at the portfolio level. For illustration, we consider the distribution of the total losses $\sum_{i=1,\dots,m,d_i>t_c} S_i$ and the maximum settlement time $\max_{i=1,\dots,m,d_i>t_c} D_i$ of the portfolio. Figure 3 displays their distributions obtained under both independence and dependence models, where the actual values of outcomes are indicated by vertical lines. We observe that the dependence among component processes in the development of an insurance claim significantly impacts the predictive distributions of both quantities of interest. In the left panel for the maximum settlement time, both distributions under independence and dependence assumptions reasonably predict the outcome,

TABLE 7

Accuracy measures of out-of-sample point predictions

		Settlement time (D)		Total payment (S)	
	t_c	MAPE	MSPE	MAPE	MSPE
$\theta = 0.8$	5	1.241	9.117	6.227	221.344
Independent	5	2.486	16.565	23.667	2262.034
$\theta = 0.8$	8	0.971	4.795	3.916	145.489
Independent	8	2.110	17.765	18.214	1533.587

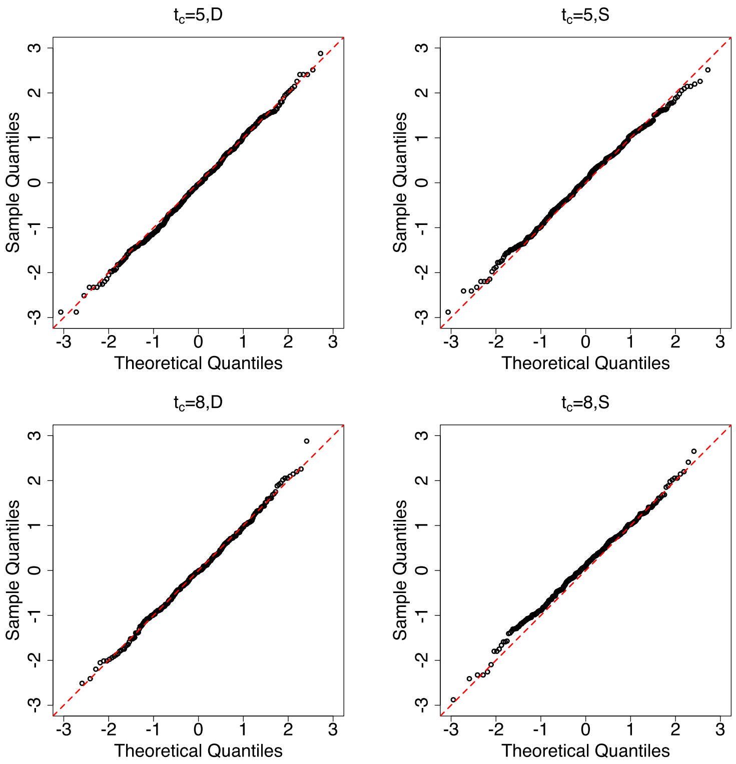

FIG. 2. *QQ plots of the uniform tests for out-of-sample validation. The left and right columns report for settlement time and total payment, respectively; The top and bottom rows report for $t_c = 5$ and $t_c = 8$, respectively.*

reflected by the fact that the realized value is covered by the prediction intervals. However, the dependence model leads to a sharper distribution with more concentration relative to the true observation than the independence model. In the right panel for the total losses, the actual observation lies in the middle of the predictive curve for which dependence is correctly characterized. In contrast, ignoring dependence leads to substantial underestimation of the total losses for the portfolio. This is particularly detrimental in claim management applications, as such prediction likely results in under reserving and hence causes insolvency of the insurer.

5. Application. This section demonstrates the application of the proposed method in insurance claims management. We show that the joint model leads to a dynamic prediction strategy that enhances the insurer's decision making in claims management operations. Section 5.1 summarizes the key data characteristics that motivate the joint model. Section 5.2 reports data analysis and estimation results. Section 5.3 summarizes the out-of-sample performance, emphasizing the implications of dependence in the proposed modeling framework.

5.1. *Data characteristics.* In the application we consider a portfolio of property insurance claims obtained from the Local Government Property Insurance Fund of Wisconsin. The property fund is viewed as a commercial property insurance provider that provides coverage for business, as opposed to homeowners. Our analysis focuses on the building and contents coverage for local government entities, including cities, counties, towns, villages, school districts, fire stations, and miscellaneous entities.

The portfolio consists of 8790 claims occurred between 2005 and 2014. For each claim we observe the time when it is reported to the insurer and the subsequent transactions since reporting. In particular, the data contain information on the settlement time of the claim as

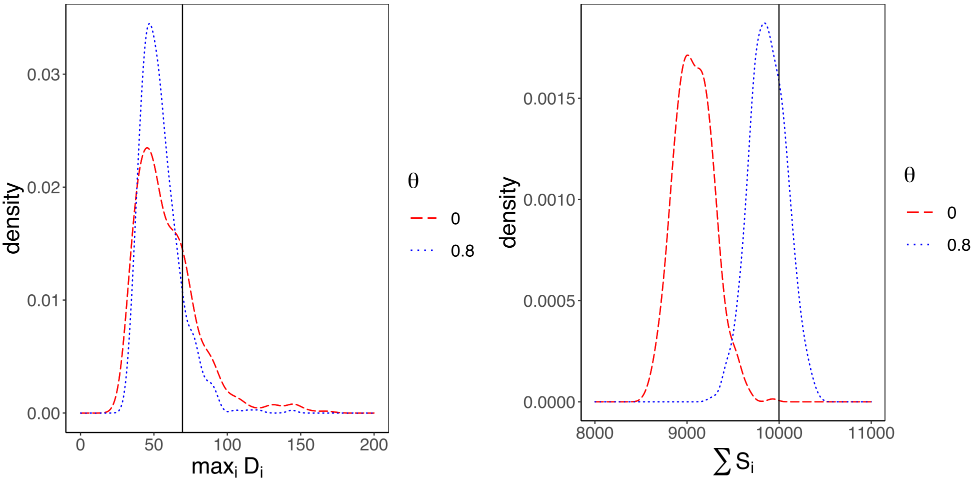

FIG. 3. *Predictive distributions of the maximum settlement time* (*left*) *and the total losses* (*right*) *for the portfolio of claims.*

well as the timestamps and amounts of payment transactions throughout the settlement. For illustration we exhibit in Figure 4 the payment and settlement transactions for two representative claims. In the first scenario, the claim is closed without any payments. This could happen when the coverage is denied (e.g., because of fraud) or when the loss is below the deductible amount. Zero-payment claims account for about 20% of our data. In the second scenario, the claim is settled with at least one payment, and the settlement occurs days after the last payment. The examples foreshadow several important features of the settlement process. First, a larger claim tends to be associated with more payment transactions and a longer settlement lag. Second, there is substantial variation in the lapse of time between the last payment and the closing of a claim. In our data, some claims are closed right after the last payment, whereas some take up to 510 days since the last payment.

Table 8 presents the descriptive statistics for the three outcome variables, the payment amount, Y (in dollars), the waiting time between successive payments, W (in days), and the settlement time of the claim, D (in days). We group the data by the payment frequency (zero or not). On average, the settlement times for zero-payment and positive-payment claims are comparable. Nevertheless, the latter shows more variation and skewness than the former. In addition, the waiting time and payment amount for claims with positive payments are skewed and heavy-tailed.

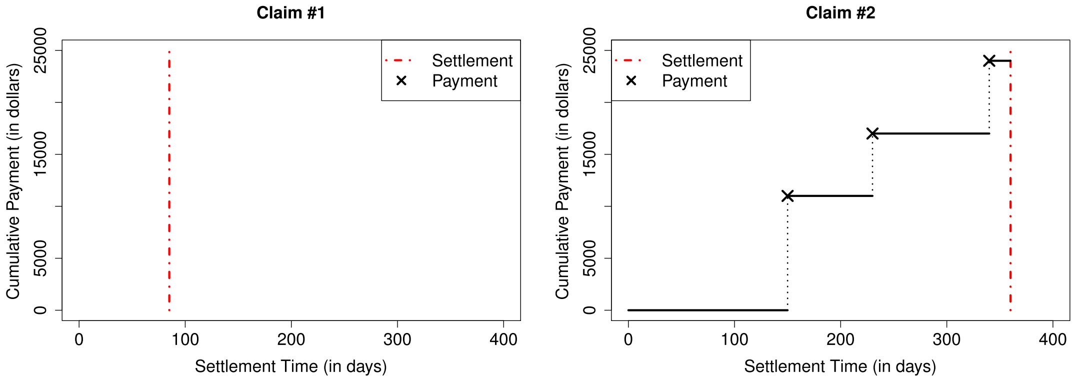

FIG. 4. *Payment and settlement transactions of two representative claims.*

TABLE 8
Descriptive statistics of outcomes of interests

			Percentiles				
Variable	Mean	SD	$10th$	$25th$	$50th$	$75th$	$90th$
Zero-payment Claims ($n = 1723$)							
Settlement time (D)	85.35	58.48	16.00	53.00	82.00	106.00	143.00
Positive-payment Claims ($n = 7067$)							
Settlement time (D)	82.60	95.82	9.00	19.00	57.00	106.00	186.00
Waiting time (W)	71.33	75.26	10.00	19.00	51.00	95.00	155.00
Payment amount (Y)	13,815.41	46,404.10	493.87	1049.00	2923.70	8168.02	25,700.00

Table 9 summarizes the set of predictors that are available in the data. We group the predictors into two categories, policy level and claim level. Policy level predictors describe the contractual features (such as deductible and amount of coverage) and policyholder characteristics (such as entity type). Claim level predictors are variables that are claim specific, including the cause of loss and the reporting delay. An average policy provides a coverage of \$332 million with a deductible of \$11,000. The average reporting delay is about one and half months, and vandalism is found to be the most frequent peril for the losses. We also notice a substantial amount of variation in both contractual features and claim-specific characteristics.

Lastly, we perform an exploratory analysis of the association among the outcomes. Figure 5 exhibits the box plots of the waiting time and payment amount, grouped by the value of settlement time. The plots suggest that a larger settlement time of a claim tends to be associated with both longer waiting times and larger incremental payment amounts. Figure 6 presents the scatter plot of average waiting time and average payment amount, with the data grouped by the range of settlement time, that is, < 6 months, $6 - 12$ months, and > 12 months. The data are clustered by unique values of settlement time, and each data point corresponds to the average waiting time and average payment amount from claims with the same settlement time. The size of the data point is proportional to the number of claims in each

TABLE 9
Description and summary statistics of covariates

Variable	Description	Mean	SD
Policy-level characteristics			
Deductible	per-occurrence deductible in thousand dollars	11.053	16.446
Coverage	amount of coverage in million dollars	332.715	574.996
City	$= 1$ if the policyholder is a city entity, $= 0$ otherwise	0.268	
County	$= 1$ if the policyholder is a county entity, $= 0$ otherwise	0.241	
School	$= 1$ if the policyholder is a school district, $= 0$ otherwise	0.332	
Town	$= 1$ if the policyholder is a town entity, $= 0$ otherwise	0.020	
Village	$= 1$ if the policyholder is a village entity, $= 0$ otherwise	0.116	
Miscellaneous	$= 1$ if other local government entities, $= 0$ otherwise	0.023	
Claim-level characteristics			
ReportDelay	time lapsed from loss occurrence to reporting in days	44.279	79.170
FireLightning	$= 1$ if loss is caused by fire and lightning, $= 0$ otherwise	0.188	
Vandalism	$= 1$ if loss is caused by vandalism, $= 0$ otherwise	0.292	
Vehicle	$= 1$ if loss is caused by vehicle, $= 0$ otherwise	0.199	
Water	$= 1$ if loss is caused by water damage, $= 0$ otherwise	0.136	
Weather	$= 1$ if loss is caused by weather related perils, $= 0$ otherwise	0.087	
Other	$= 1$ if loss is caused by other perils, $= 0$ otherwise	0.098	

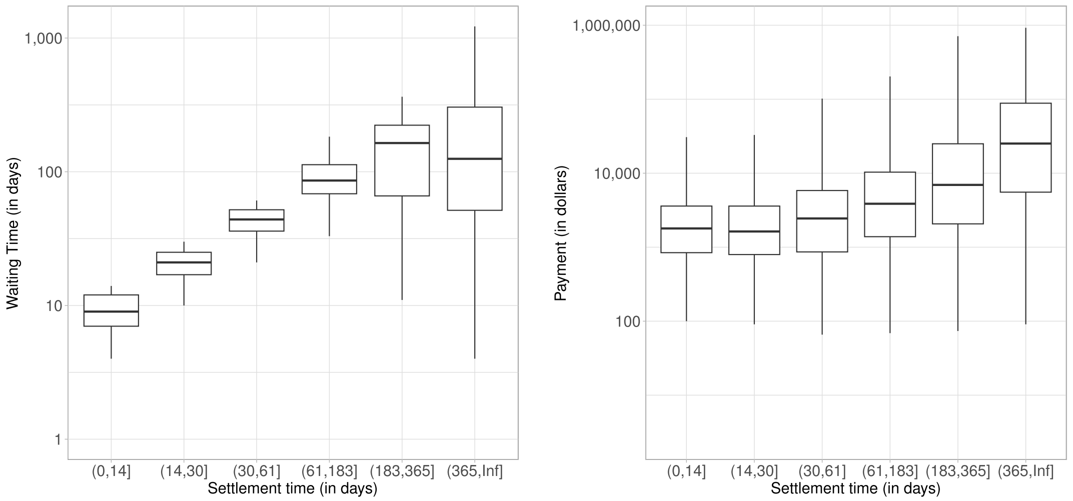

FIG. 5. *Box plots of waiting time (left panel) and incremental payment amount (right panel) by settlement time.*

cluster. The plot shows that the relationship between the two outcomes interestingly varies by the settlement time. The correlation coefficients are 0.16, −0.13, and −0.15, respectively, for the three cases. It suggests that the settlement time affects both the sign and the magnitude of the dependence between waiting time and payment amount.

5.2. *Empirical results.* We employ the proposed copula-based joint model to analyze the portfolio of property insurance claims described in Section 5.1. We follow an iterative process between model estimation and selection for model specification. In the final formulation, we use a Cox proportional hazard model for the settlement time of claims, a Weibull regression for the intensity of payment recurrences, and a GB2 regression for the amount of incremental payments. We follow the parsimony principle in model specification, balancing flexibility and interpretability. For instance, we find parametric models are sufficient for the waiting time and payment amount. In contrast, the Cox model with a nonparametric baseline demonstrates a satisfactory fit. Gaussian copulas are employed for the dependence. To provide flexibility, we consider variations, such as a mixture of copulas and conditional copulas, in the joint model. The parameters in the joint model are estimated using the stagewise procedure introduced in Section 3.1.

In marginal models, numerical predictors (deductible, coverage, and reporting delay) are used on a log scale to enhance stability. In the dependence model, we use a two-component Gaussian copula mixture for the dependence between W and D. The copula function is

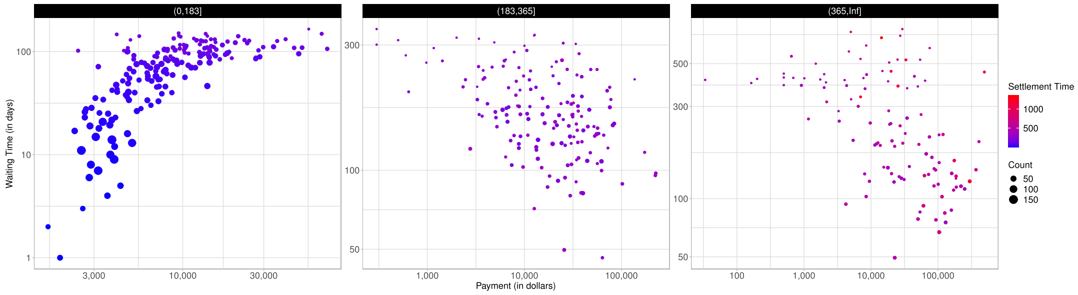

FIG. 6. *Scatter plot of average waiting time and average payment amount with data grouped by the range of settlement time.*

$C_{(W,D)} = qC_1 + (1-q)C_2$, where C_1 and C_2 are both Gaussian copulas with association parameters $\theta_{WD}^{(1)}$ and $\theta_{WD}^{(2)}$, respectively, and q is the weight on the first component C_1. The formulation is motivated by the fact that the waiting time and time-to-report could be highly correlated when a claim is closed with a single payment. The mixture specification accommodates this possibility by treating it as a latent class. Motivated by our observations of Figure 6, we employ a conditional copula approach for the dependence between Y and W given D, that is, to allow the association, denoted by $\theta_{YW|D}$, to vary with the value of D. Specifically, we consider

$$(11) \quad \theta_{YW|D=d} = \theta_{YW|D}^{(1)} I(d \le 183) + \theta_{YW|D}^{(2)} I(183 < d \le 365) + \theta_{YW|D}^{(3)} I(d > 365).$$

For comparison we also fit the model based on the simplifying assumption treating $\theta_{YW|D}$ as a constant. Estimated model parameters, using both the simplifying assumption and the conditional copula approach, are reported in Table 10. The dependence model between Y and D uses a Gaussian copula with association parameter θ_{YD}. Recall that the copulas $C_{(Y,D)}$ and $C_{(W,Y|D)}$ are estimated simultaneously in the last step of the stagewise estimation, as discussed in Section 3.1. Hence, the estimation of θ_{YD} is influenced by the assumption on $C_{(W,Y|D)}$. We report the estimates of θ_{YD} when $C_{(W,Y|D)}$ are formulated using both simplifying and conditional copula approaches.

Several noteworthy observations can be made from Table 10. First, the claim-level predictors (reporting delay and cause of losses) exhibit significant effect on all three outcomes. As anticipated, after accounting for the claim-specific covariates, the policy-level predictors are found to have less impact on the outcomes. Second, consistent with Figure 5, the settlement time is positively correlated with both the waiting time and payment amount. For the former pair (W, D), the Kendall's tau, calculated using the copula parameter, is 0.15 and 0.93 for the two mixing components. The component with high correlation focuses on claims with a single payment. The corresponding weight is about 0.76, which is consistent with that 72% of claims settled with a single payment. For the latter pair (Y, D), the estimated Kendall's tau is around 0.16 under both the simplifying assumption and the conditional copula approach. Third, the conditional association between the waiting time and the payment amount $(W, Y)|D$ varies with the settlement time. The conditional association is positive for claims that are settled within six months, whereas it is negative for claims that are settled after 12 months. For the claims with settlement time in between, the estimated association parameter is not statistically significant. Meanwhile, the conditional association is estimated to be 0.001 under the simplifying assumption. The plausible explanation is that the positive and negative dependencies average out when pooling across different values of settlement time. The estimated conditional dependence, which measures the residual association after accounting for the covariates effects, is largely consistent with the patterns revealed in the raw data, as shown in Figure 6.

5.3. *Comparison with independence model.*

The previous section provides compelling evidences for our hypothesis that the payment process and settlement process in claims management are interconnected. The proposed method demonstrates a flexible framework that can accommodate such complex relation. This section is dedicated to evaluating the accuracy of probabilistic forecasts on an out-of-sample dataset. Our focus is comparing the copula model with a benchmark independence model using both validation and cross-validation techniques. It is important to note that the independence model differs from the copula model in two aspects. First, the independence model does not consider the relationship among component processes in the claims management. Second, due to the misspecification of dependence, the marginal model for each component process in the independence model is subject to estimation bias, as we demonstrate in Section 4.

TABLE 10
Stagewise estimation results for model parameters

	Stage I		Stage II		Stage III			
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Marginal	Settlement Time		Waiting Time		Payment Amount			
Intercept			−4.984	0.024	8.288	0.228	8.326	0.228
log(Deductible)	−0.035	0.009	−0.041	0.001	−0.012	0.012	−0.015	0.012
log(Coverage)	−0.014	0.011	−0.010	0.001	−0.004	0.015	−0.007	0.014
City	0.044	0.075	0.057	0.009	−0.048	0.096	−0.024	0.096
County	−0.014	0.079	−0.010	0.009	0.091	0.101	0.111	0.100
School	−0.049	0.077	−0.044	0.009	−0.239	0.099	−0.211	0.099
Town	−0.021	0.105	−0.003	0.012	−0.193	0.133	−0.175	0.132
Village	−0.116	0.077	−0.111	0.009	−0.111	0.099	−0.085	0.098
log(ReportDelay)	0.197	0.008	0.199	0.001	−0.068	0.010	−0.061	0.010
FireLightning	0.033	0.042	0.060	0.006	0.304	0.056	0.291	0.056
Vandalism	0.246	0.041	0.245	0.005	−1.337	0.059	−1.324	0.059
Vehicle	0.296	0.043	0.304	0.006	−0.294	0.059	−0.303	0.059
Water	−0.201	0.045	−0.190	0.006	0.665	0.062	0.650	0.062
Weather	−0.364	0.050	−0.342	0.007	0.598	0.066	0.575	0.066
p (Weibull)			1.109	0.002				
σ (GB2)					0.761	0.062	0.778	0.064
κ_1 (GB2)					1.162	0.141	1.212	0.148
κ_2 (GB2)					0.903	0.105	0.925	0.109
Dependence					Simplifying		Conditional	
$\theta^{(1)}_{WD}$			0.232	0.018				
$\theta^{(2)}_{WD}$			0.995	0.000				
q			0.243	0.006				
θ_{YD}					0.251	0.010	0.244	0.010
$\theta_{YW\|D}$					0.001	0.013		
$\theta^{(1)}_{YW\|D}$							0.111	0.018
$\theta^{(2)}_{YW\|D}$							−0.037	0.023
$\theta^{(3)}_{YW\|D}$							−0.196	0.026

We validate our model using two different methods. The first is out-of-time validation, where we use data from years 2005–2011 as the training set and data from years 2012–2014 as the test set. The second is a 10-fold cross-validation where the claims are randomly split into 10 subsets. In both cases we estimate model parameters using the training data and calculate a score for each observation in the test set. The logarithmic score, a widely used local proper scoring rule (Parry, Dawid and Lauritzen (2012)), is employed to evaluate the models. Furthermore, we compare the probabilistic forecasts generated by the copula and the independence models using the Diebold–Mariano test (Diebold and Mariano (1995)). Let $S(F_i, y_i)$ denote the score for data y_i based on the predictive model F. The average score is calculated using the hold-out sample as

$$\bar{S}^F_{n_{\text{out}}} = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} S(F_i, y_i),$$

where n_{out} represents the number of observations in the hold-out sample. In the Diebold–Mariano test, we use $\bar{S}^{F_0}_{n_{\text{out}}}$ and $\bar{S}^{F_1}_{n_{\text{out}}}$ to denote the average score from the base model F_0 and

TABLE 11
Diebold–Mariano statistics using out-of-time validation and 10*-fold cross-validation*

		10-fold									
t_c	Out-of-time	1	2	3	4	5	6	7	8	9	10
0	−30.16	−16.38	−19.10	−18.61	−16.11	−17.64	−18.76	−20.22	−16.47	−20.30	−18.56
30	−26.97	−15.31	−16.83	−16.12	−13.46	−15.81	−16.81	−17.82	−14.80	−18.60	−16.33
91	−17.94	−12.11	−12.87	−11.92	−10.56	−12.29	−13.69	−13.80	−11.71	−14.41	−12.85

the competing model F_1. The test statistic is defined as

$$t_{n_{\text{out}}} = \sqrt{n_{\text{out}}} \frac{\bar{S}^{F_1}_{n_{\text{out}}} - \bar{S}^{F_0}_{n_{\text{out}}}}{\sigma_{n_{\text{out}}}},$$

where $\sigma^2_{n_{\text{out}}} = \frac{1}{n_{\text{out}}} \sum_{i=1}^{n_{\text{out}}} (S(F_{1i}, y_i) - S(F_{0i}, y_i))^2$. Based on this definition, a negative statistic implies a superior performance of the competing model.

To emphasize the notion of dynamic prediction, we perform the test at three time points: time of reporting ($t_c = 0$), one month since reporting ($t_c = 30$), and three months since reporting ($t_c = 91$). The results are summarized in Table 11. For example, using the independence model as benchmark, we observe a test statistic of −30.16 in the out-of-time validation at the reporting time, and the test statistics in the cross-validation ranges from −20.30 to −16.11. Both tests suggest that the copula model is preferred for prediction, and the conclusion is consistent across different time points.

We conclude this section with two additional numerical illustrations emphasizing the implications of dependence on insurance operations. The first example pertains to a single insurance claim. Figure 7 displays the dynamic predictive distributions for the time-to-settlement and outstanding payments for a randomly selected claim, at three time points ($t_c = 0, 30, 91$), from independence and copula models. Our analysis reveals that, for this particular claim, the dependence among the three underlying processes does not have a significant effect on the time-to-event prediction. However, the impact of dependence on outstanding payments is substantial. Specifically, the claim had one payment that occurred within 30 days since reporting, and no payments between 30 days and 90 days since reporting. The predictive distributions exhibit a probability mass at zero, indicating the possibility for the claim to be closed without any payment. At $t_c = 0$, the predictive distributions from the independence and copula models are similar. At $t_c = 30$, the distribution of outstanding payments from the copula model is higher than the distribution from independence model. This observation is explained by the positive dependence between the payment process and settlement process and the fact that the payment occurred early (relative to the average payment time) during the lifetime of the claim. At $t_c = 91$, there is a negligible update on the distribution from $t_c = 30$ because there were no payments between the two time points.

The second example concerns the dynamic prediction for the entire portfolio of insurance claims in the hold-out sample. Figure 8 shows the predictive distributions for the maximum time-to-settlement and the aggregate outstanding liability from the portfolio. As foreshadowed by Figure 7, the result suggests that the dependence is not essential for the prediction of the maximum settlement time but is critical to the prediction of outstanding liabilities for the portfolio. For instance, in reserving applications an insurer often uses 75th percentile of the distribution as a conservative approach to set the carried reserves. In this case the independence model will overestimate the carried reserves by 103%, 75%, and 45% at $t_c = 0, 30, 91$, respectively.

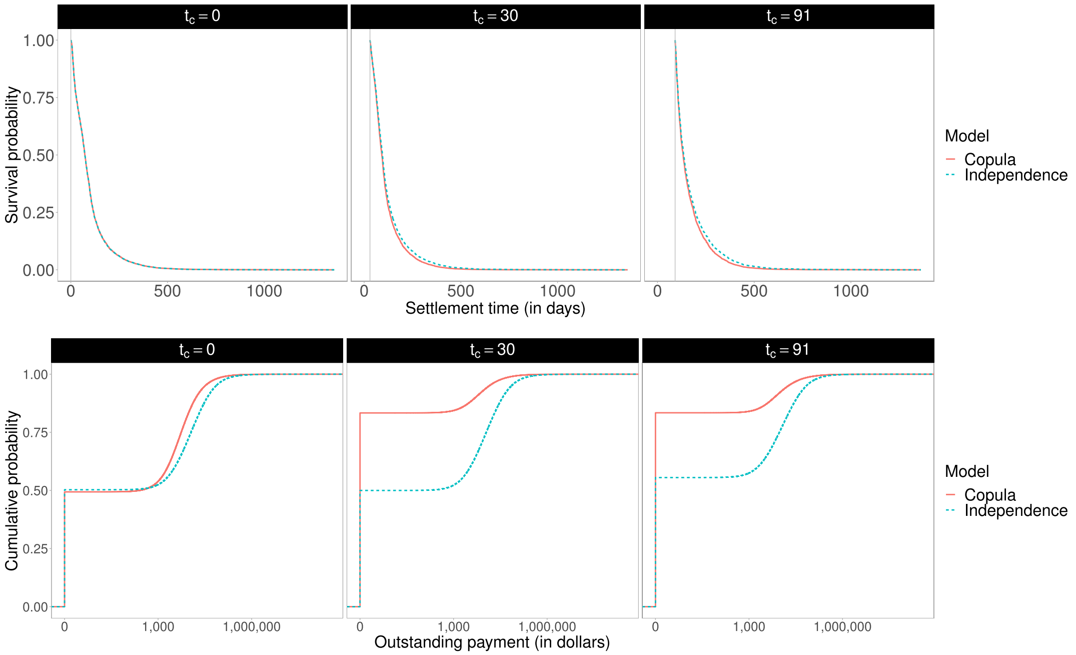

FIG. 7. *Dynamic predictive distributions of time-to-settlement and outstanding payments for a randomly selected claim. The upper panel shows the survival function of settlement time, and the lower panel shows the distribution function of the outstanding payments.*

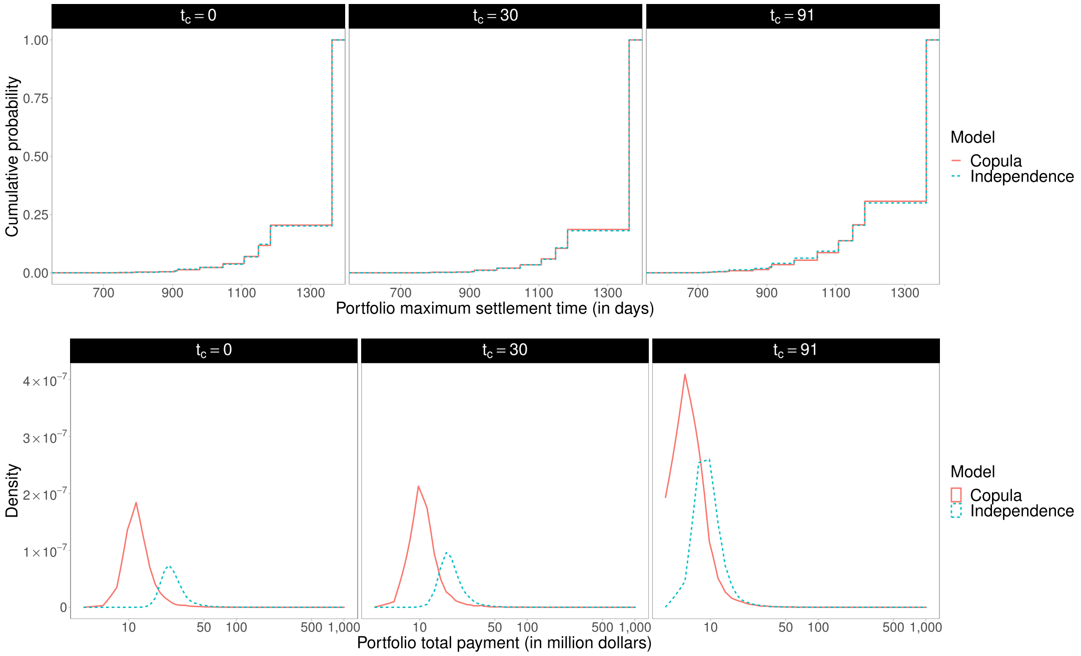

FIG. 8. *Dynamic predictive distributions of maximum time-to-settlement and outstanding liabilities for the insurance portfolio. The upper panel shows the distribution function of the maximum settlement time, and the lower panel shows the density of the outstanding liability.*

6. Conclusion. In this article we introduced a copula-based joint modeling framework to analyze the cash flows of insurance claims and performed dynamic prediction for outstanding payments. The work was motivated by the claims management operation in insurance companies where insurers make data-driven decisions on carried reserves and risk financing strategies. Specifically, we observed that a typical insurance claim is associated with a sequence of payments during its lifetime and the occurrence and size of payments are further correlated with the settlement time. In the proposed model, we adopted the viewpoint of a marked point process, where we treated the recurrence of payments transactions of a claim as a counting process, the payment amounts as the associated marks of the counting process, and the settlement of the claim as a terminal event for the marked point process. The dependence among the counting process, its marks, and the terminal event was accommodated using bivariate parametric copulas as building blocks. To improve computational efficiency, we further presented a sequential approach for the estimation of model parameters. The proposed joint model provides the predictive distributions of risk outcomes that could be obtained via Monte Carlo simulation when closed forms are not easily available.

In the case study, we analyzed a portfolio of insurance claims from a commercial property insurance provider. The empirical findings provided evidence of positive correlation between the payments (both waiting time and payment amount) and settlement of claims, which supports our hypothesis that, on average, it takes longer for insurers to close out larger and more complex claims. More importantly, we showcased how an insurer could leverage the granular transaction data with the joint model to improve operations in claims management.

Acknowledgments. The authors thank the anonymous reviewers and the Editor for their feedback and suggestions that improved the paper.

Funding. Yang was supported by NSF Grant DMS-2210712.

REFERENCES

AALEN, O. O., BORGAN, Ø. and GJESSING, H. K. (2008). *Survival and Event History Analysis*: *A Process Point of View*. *Statistics for Biology and Health*. Springer, New York. MR2449233 https://doi.org/10.1007/978-0-387-68560-1

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math*. *Econom*. **44** 182–198. MR2517884 https://doi.org/10.1016/j.insmatheco.2007.02.001

BEDFORD, T. and COOKE, R. M. (2002). Vines—a new graphical model for dependent random variables. *Ann*. *Statist*. **30** 1031–1068. MR1926167 https://doi.org/10.1214/aos/1031689016

BROWN, E. R., IBRAHIM, J. G. and DEGRUTTOLA, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61** 64–73. MR2129202 https://doi.org/10.1111/j.0006-341X.2005.030929.x

CAFFO, B. S., BOOTH, J. G. and DAVISON, A. C. (2002). Empirical supremum rejection sampling. *Biometrika* **89** 745–754. MR1946509 https://doi.org/10.1093/biomet/89.4.745

CHI, Y.-Y. and IBRAHIM, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62** 432–445. MR2227491 https://doi.org/10.1111/j.1541-0420.2005.00448.x

COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events*. *Statistics for Biology and Health*. Springer, New York. MR3822124

DIAO, L., COOK, R. J. and LEE, K.-A. (2013). A copula model for marked point processes. *Lifetime Data Anal*. **19** 463–489. MR3119993 https://doi.org/10.1007/s10985-013-9259-3

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J*. *Bus*. *Econom*. *Statist*. **13** 134–144.

DING, J. and WANG, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics* **64** 546–556. MR2432425 https://doi.org/10.1111/j.1541-0420.2007.00896.x

ELASHOFF, R., LI, N. et al. (2016). *Joint Models for Longitudinal and Time-to-Event Data*. CRC Press, Boca Raton.

FAUCETT, C. L. and THOMAS, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat. Med.* **15** 1663–1685.

GANJALI, M. and BAGHFALAKI, T. (2015). A copula approach to joint modeling of longitudinal measurements and survival times using Monte Carlo expectation-maximization with application to aids studies. *J. Biopharm. Statist.* **25** 1077–1099.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31** 1208–1211. MR0123385 https://doi.org/10.1214/aoms/1177705693

GRUTTOLA, V. D. and TU, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50** 1003–1014.

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.* **94** 401–419. MR2167922 https://doi.org/10.1016/j.jmva.2004.06.003

KIM, S., ZENG, D., CHAMBLESS, L. and LI, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Stat. Biosci.* **4** 262–281.

KRÓL, A., FERRER, L., PIGNON, J.-P., PROUST-LIMA, C., DUCREUX, M., BOUCHÉ, O., MICHIELS, S. and RONDEAU, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. *Biometrics* **72** 907–916. MR3545683 https://doi.org/10.1111/biom.12490

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LI, N., ELASHOFF, R. M., LI, G. and TSENG, C.-H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Stat. Med.* **31** 1707–1721. MR2947519 https://doi.org/10.1002/sim.4507

LIN, H., MCCULLOCH, C. E. and MAYNE, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Stat. Med.* **21** 2369–2382. https://doi.org/10.1002/sim.1179

LIU, L. and HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 65–81. MR2662234 https://doi.org/10.1111/j.1467-9876.2008.00641.x

LIU, L., HUANG, X. and O'QUIGLEY, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64** 950–958. MR2526647 https://doi.org/10.1111/j.1541-0420.2007.00954.x

MASAROTTO, G. and VARIN, C. (2012). Gaussian copula marginal regression. *Electron. J. Stat.* **6** 1517–1549. MR2988457 https://doi.org/10.1214/12-EJS721

MCDONALD, J. B. and XU, Y. J. (1995). A generalization of the beta distribution with applications. *J. Econometrics* **66** 133–152.

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu. Rev. Stat. Appl.* **6** 223–240. MR3939519 https://doi.org/10.1146/annurev-statistics-030718-105048

PARRY, M., DAWID, A. P. and LAURITZEN, S. (2012). Proper local scoring rules. *Ann. Statist.* **40** 561–592. MR3014317 https://doi.org/10.1214/12-AOS971

PENG, R. D. (2018). Advanced statistical computing. Work in Progress.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–342. MR0671971 https://doi.org/10.1093/biomet/69.2.331

PROUST-LIMA, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach. *Stat. Med.* **35** 382–398. MR3455508 https://doi.org/10.1002/sim.6731

RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC press, Boca Raton.

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959 https://doi.org/10.1002/sim.4205

RIZOPOULOS, D., VERBEKE, G. and LESAFFRE, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 637–654. MR2749911 https://doi.org/10.1111/j.1467-9868.2008.00704.x

RIZOPOULOS, D., VERBEKE, G., LESAFFRE, E. and VANRENTERGHEM, Y. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics* **64** 611–619. MR2432435 https://doi.org/10.1111/j.1541-0420.2007.00894.x

RIZOPOULOS, D., VERBEKE, G. and MOLENBERGHS, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95** 63–74. MR2409715 https://doi.org/10.1093/biomet/asm087

SHI, P. (2014). Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science*, *Volume I*: *Predictive Modeling Techniques* (E. W. Edward, G. Meyers and R. A. Derrig, eds.) 236–259. Cambridge Univ. Press, Cambridge.

SHI, P. and YANG, L. (2018). Pair copula constructions for insurance experience rating. *J. Amer. Statist. Assoc.* **113** 122–133. MR3803444 https://doi.org/10.1080/01621459.2017.1330692

SURESH, K., TAYLOR, J. M. G. and TSODIKOV, A. (2021). A Gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker. *Biostatistics* **22** 504–521. MR4287165 https://doi.org/10.1093/biostatistics/kxz049

TANG, A.-M. and TANG, N.-S. (2015). Semiparametric Bayesian inference on skew-normal joint modeling of multivariate longitudinal and survival data. *Stat. Med.* **34** 824–843. MR3326393 https://doi.org/10.1002/sim.6373

TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186 https://doi.org/10.2307/2533118

ZELLER, G. and SCHERER, M. (2022). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *Eur. Actuar. J.* **12** 33–85. MR4443594 https://doi.org/10.1007/s13385-021-00290-1

ZHU, H., IBRAHIM, J. G., CHI, Y.-Y. and TANG, N. (2012). Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics* **68** 954–964. MR3055200 https://doi.org/10.1111/j.1541-0420.2012.01745.x