

# A Markov-switching Generalized Additive Model for Compound Poisson Processes, with Applications to Operational Losses Models

J. HAMBUCKERS<sup>1,†</sup>, T. KNEIB<sup>1</sup>, R. LANGROCK<sup>2</sup> and A. SOHN<sup>1</sup>

September 21, 2016

## Abstract

We are interested in modeling the behavior of random sums over time. Such models are particularly useful in insurance and finance, to describe the behavior of total losses (like operational losses) over time, and to correctly estimate tail-related risk indicators. To this end, we formulate a generalized additive Markov-switching compound process combining Poisson and Generalized Pareto distributions. This flexible model takes into account two important features: on the one hand, we follow the idea of generalized additive models for location, scale and shape (GAMLSS), and thereby allow all parameters of the considered distributions to depend on economic covariates. On the other hand, we allow this dependence to vary over time, via a hidden state variable. A simulation study indicates that, even in the case of a short time series, the model is easily and well estimated with a standard maximum likelihood procedure. Relying on this approach, we analyze a dataset of 817 losses resulting from frauds in the Italian bank UniCredit. We show that our model improves the estimation of the total loss distribution over time, compared to standard alternatives. In particular, this model provides estimations of the 99.9% quantile that are never exceeded by the historical total losses.

Keywords: Generalized Pareto distribution, GAMLSS, distributional regression, Extreme events.

<sup>1</sup> Georg-August-Universität Göttingen, Chair of Statistics, Humboldtallee 3, 37073 Göttingen, Germany.

<sup>2</sup> Bielefeld University, Department of Business Administration and Economics, Universitätsstrasse 25, 33615 Bielefeld, Germany.

† corresponding author: julien.hambuckers@mathematik.uni-goettingen.de.

# 1 Introduction

In this paper, we are interested in modeling the distribution of random sums over time. The quantity of interest is given by

$$L_t = \sum_{i=1}^{N_t} Y_{i,t}, \quad (1)$$

for  $t = 1, \dots, T$ , where  $N_t$  is the number of events over the period  $t$ , and  $Y_{i,t}$  is the severity of the  $i^{\text{th}}$  event occurring during period  $t$ . In the actuarial literature, this model is used for modeling the total claim of an insurance company over time [Embrechts et al., 1997, p.9] whereas in the banking literature it is used to model the total operational loss suffered by a bank [Chernobai and Yildirim, 2008]. The number of events is therefore a number of losses (i.e. individual claims or operational events), whose intensity is measured in monetary units. Standard actuarial/financial models usually make the following assumptions on the behavior of  $N_t$  and  $Y_{i,t}$ :

$$\begin{aligned} N_t &\stackrel{iid}{\sim} \text{Poisson}(\lambda), \\ Y_{i,t} &\stackrel{iid}{\sim} \text{GPD}(\gamma, \sigma), \end{aligned}$$

with  $\lambda, \gamma, \sigma > 0$ , and where realisations of  $N_t$  and  $Y_{i,t}$ ,  $\forall \{t, i\}$  are independent.  $\lambda$  is the frequency parameter, whereas  $\text{GPD}(\gamma, \sigma)$  stands for the Generalized Pareto cumulative distribution function, given by

$$\text{GPD}(Y; \gamma, \sigma) = \begin{cases} 1 - \left(1 + \gamma \frac{Y}{\sigma}\right)^{-1/\gamma}, & \gamma \neq 0 \\ 1 - e^{-Y/\sigma}, & \gamma = 0 \end{cases} \quad (2)$$

where  $\gamma$  is the shape parameter,  $\sigma$  the scale parameter and  $Y > 0$ . In the remainder, we only focus on the case where  $\gamma > 0$  (i.e. the heavy tail case).

This model is known as a *compound Poisson process*, or a *double stochastic process* [Guillou et al., 2015] with GPD distributed intensity. In the financial literature, the distribution of  $N_t$  is referred to as the *frequency distribution* whereas the distribution of  $Y_{i,t}$  is termed the *severity distribution*. In practice, the total loss over a time period is characterized by a large number of small losses and a small number of very large (extreme) losses, with the latter being the main drivers of  $L_t$ . For example, over a year, a bank might suffer from many small transaction errors (e.g. employees making accounting mistakes) and then it records one huge loss (e.g. due to unauthorized trading). In this context, we use the GPD to correctly take into account the probabilities of these extremely large losses. This is particularly important in financial applications where the goal is to estimate a quantile of  $L_t$  far in the tail (e.g. a 99.9% quantile).

The basic model given by equation (1) can be modified in several ways to account for more realistic empirical features. For example, in a regression context, the parameters of the frequency and severity distributions might be expressed as functions of explanatory variables. Indeed, it is quite likely that the general economic conditions, the situation of the financial markets or firm-specific factors influence the loss process. Such relationships are suggested by a number of recent empirical studies [see, among others, Chernobai et al., 2011, Cope et al., 2012, Wang and Hsu, 2013, Chavez-Demoulin et al., 2016]. Mathematically speaking, we could assume that, for  $\theta \in \{\lambda, \sigma, \gamma\}$ , (i.e. one of the frequency or severity parameters in model (1)), the following relationship holds:

$$g(\theta) = \eta_\theta = X^\theta \beta_\theta + \sum_{j=1}^J h_{\theta,j}(X_j^\theta), \quad (3)$$

where  $g(\cdot)$  is a monotonic link function,  $X^\theta$  is a vector of linear predictor for parameter  $\theta$ ,  $X_j^\theta$ ,  $j = 1, \dots, J$ , is the  $j^{\text{th}}$  nonlinear predictor and  $h_{\theta,j}$  is an unspecified smooth function of the covariate (nonparametric part). This general setup is known as the semiparametric generalized additive model for location, scale, and shape (GAMLSS), which is a generalization of the generalized additive model (GAM) to response variables that are not of the exponential family type [Rigby et al., 2005]. It allows to consider linear and nonlinear dependences of covariates with the parameters of a distribution function. This approach has been successfully used in different applications. Among others, Rigby et al. [2005] consider a Box-Cox  $t$  distribution for the Body Mass Index (BMI) of Dutch girls, Stasinopoulos and Rigby [2007] study insurance claims data using the negative binomial distribution, Serinaldi [2011] models the price of electricity whereas Sohn et al. [2015] explore the characteristics of personal income in Germany with the Dagum distribution. More recently, Chavez-Demoulin et al. [2016] consider a GAMLSS framework for both Poisson and GPD distributions in a model of type (1). They investigate the effect of two covariates: the type of event (e.g. if a loss results from a fraud, employee malpractices or another failure of internal controls) and the year of occurrence. These models can easily be estimated via penalized likelihood maximization with cubic splines for the nonparametric parts. Model and smoothing parameters selection is usually performed using the Akaike Information criterion (AIC) and/or generalized cross-validation procedures [Fahrmeir et al., 2013].

On the other hand, a model as the one given by equation (1) often focuses on data that have a time series structure. In case of parameters instabilities (as it is often the case in economics and finance; see Hamilton [1989], Lamoureux and Lastrapes [1990] and Ang and Timmermann [2012]), equation (3) is not sufficient to account for abrupt changes in the dependence structure. In this situation, a popular modeling strategy is to rely on regime-switching models, where parameters vary over time according to an unobservable Markov chain. Such models are referred to the literature as Markov-switching (MS) or

regime-switching models [Hamilton, 1989, Zucchini et al., 2016]. They have been applied successfully to model changes in market volatility [Klaassen, 2002], in interest rates [Pesaran et al., 2006] or in business cycles [Goodwin, 1993]. Extensions of the basic MS model allows for multiple covariates and response functions from the generalized linear model (GLM) framework. Until recently, the literature focused mainly on parametric dependencies with the covariates, but Langrock et al. [2015] introduce a Markov-switching generalized additive model (MS-GAM) allowing for nonparametric functional forms of the effect of explanatory variables, depending on a latent Markov chain. Their approach allows for flexible dependences in the model when the distribution of the response variable belongs to the exponential family (i.e. a distribution considered in the GAM framework). They provide a simple estimation procedure, based on a forward recursion and a maximization of a penalized likelihood function. The applications that they consider rely on normally- and Poisson-distributed response variables. However, distributions like the GPD (that are not of the exponential type) or more complicated compound stochastic processes have not been considered yet, thus excluding models of type (1) to benefit from these advanced modeling approaches.

In this paper, we consider an extension of the framework of Langrock et al. [2015] to these more complex models, called Markov-switching GAMLSS (MS-GAMLSS) in the following. Then we focus on the particular case of model (1). Such an extension is of interest because taking into account the parameters' instability might drastically improves the goodness-of-fit of the models used for the sum of losses over time. Indeed, Guillou et al. [2013] and Guillou et al. [2015] argue that insurance claims frequency and severity distributions might vary over time, according to some (unobservable) environmental and economic factors (like competition intensity). They considered either discrete shock sizes or shock sizes drawn from a GPD with parameters depending on a state variable. However, they did not allow for other explanatory variables than the state variable. Chavez-Demoulin et al. [2014] as well as Chavez-Demoulin et al. [2016] consider the effect of time on sums of extreme events over time nonparametrically, but not in the MS framework. Moreover, it is likely that the effect of a covariate might vary according to the state of the economy or from the regulatory environment. For example, in the banking sector, Cope et al. [2012] suggest two potential opposite relationships between unemployment rate and the severity of operational losses: on one side an increase in the unemployment rate might indicate stronger criminal activities (e.g. robbing banks instead of robbing individuals), increasing the expected loss severity. On the other side, a decrease of the unemployment rates might be a sign for a thriving economy with incentives to commit larger felonies (e.g. robbing banks because there is more money on the table), leading to the same effect on the expected loss severity. One can, therefore, imagine that in period of economic expansion the latter relationship holds, whereas in economic recessions the other effect takes place. In addition, Cope

et al. [2012] outline a series of regulatory indicators (level of controls, rule of law, state of the corruption) that appear to be linked with the severity of operational losses. As noticed in Ang and Timmermann [2012], MS models in finance were initially used to distinguish changes in regulatory policies [Sims and Zha, 2006]. Therefore, a MS component in our model should help to take into account the parameters' instability induced by changes in (latent) economic and regulatory conditions that are not captured by the other explanatory variables. Finally, the internal controls set by banks to mitigate and manage their operational losses might change abruptly over time: a change of stakeholder or of the CEO might cause strategic changes that have an impact on the risk-taking process, as well as on internal controls. Such features, which are hard to quantify, might be captured by a MS component.

These considerations motivate the formulation of a model of type (1) where the parameters  $\lambda$ ,  $\gamma$  and  $\sigma$  depend (nonparametrically) on covariates, and where the functional relationships between covariates and parameters vary according to a (hidden) regime-switching process. Our contribution draws a bridge between the methodological approaches of Chavez-Demoulin et al. [2016] and Guillou et al. [2015] by combining semiparametric regression and regime-switching models for a compound Poisson-GPD response variable into a single model. It enables a flexible modeling of the behavior of random sums over time, jointly taking into account the time dependence as well as the dependence with other covariates. Contrary to the GAMLSS approach of Chavez-Demoulin et al. [2016], the Poisson part of the model cannot be estimated independently from the GPD part, and we show how to perform the joint estimation. An interesting feature of the considered model is that by combining information about the frequency and the severity of the process, we have several observations at each point in time. This allows for improving the estimation of the state probabilities and the regression functions, even if the time series is short. A simulation study supports this finding.

Lastly, we use our model and estimation technique to study a database of 817 large losses resulting from external frauds in the Italian bank UniCredit over a time period of 38 quarters. We assume that the parameters depend nonparametrically on three explanatory variables, and that the functional forms of the dependence vary according to a hidden state variable. We use the percentage of revenue coming from fees (PRF), the unemployment rate and the VIX index as covariates to account for the internal economic performance, the macroeconomic situation and the general state of the financial markets, respectively. We compare our model with three alternatives (constant, pure MS and pure GAMLSS models) and show the economic consequences (in term of requested capital) of using the different models. Our results suggest that the MS-GAMLSS model provides the best fit among all alternative considered, as it is the model with the best AIC criterion and the only model for which the estimated 99.9% quantiles are not exceeded by the observed total losses.

The structure of the remainder of the paper is as follows: in Section 2, we describe the model and discuss its estimation. In Section 3, we perform a simulation study to analyze the finite sample properties of the suggested estimation technique. In particular, we consider the situation where the number of available periods is small. In Section 4, we introduce the data and the considered covariates, then we perform the analysis using the MS-GAMLSS model. Additionally, we investigate the regulatory consequences arising from our approach, compared to a constant approach, a MS-constant approach or a GAMLSS approach. We also discuss the economic interpretation of our results. Lastly, we conclude with Section 5.

## 2 Methodology

### 2.1 MS-GAMLSS for a compound Poisson-GPD process

In this section, we begin by presenting the general framework of Markov-switching generalized additive models for location, scale and shape parameters [MS-GAMLSS, see Adam et al., 2016, preprint], before considering the special case of a compound Poisson-GPD process. We assume that the distribution of a response variable  $Z_t$  at time  $t$  given a vector of covariates  $X_t$  and an underlying  $M$ -state Markov chain  $S_t$  is given by

$$\mathbb{P}(Z_t < z | X_t, S_t) = F(z; \theta_t^{(S_t)}(X_t)), \quad (4)$$

for  $t = 1, \dots, T$ , where  $\theta_t^{(S_t)}(X_t)$  is the vector of associated parameters of size  $K$  at time  $t$  and  $F$  is the cumulative distribution function of  $L_t$ .  $\{z_t\}_{t=1, \dots, T}$ ,  $\{x_t\}_{t=1, \dots, T}$  and  $\{s_t\}_{t=1, \dots, T}$  denote the realized time series of the response variable, the vector of covariates and the latent state, respectively. Additionally, we suppose that, for  $k = 1, \dots, K$ ;  $t = 1, \dots, T$  and  $S_t = 1, \dots, M$ , we have

$$g_k(\theta_{k,t}^{(S_t)}(X_t^{\theta_k})) = \eta_{k,t}^{(S_t)}(X_t^{\theta_k}) = X_t^{\theta_k} \beta_{\theta_k}^{(S_t)} + \sum_{j=1}^{J_{\theta_k}} h_{\theta_k,j}^{(S_t)}(X_{j,t}^{\theta_k}), \quad (5)$$

where  $g_k$  is a monotonic link function for the  $k^{th}$  parameter  $\theta_{k,t}^{(S_t)}$ ,  $X_t^{\theta_k}$  is a vector of covariates at time  $t$  for the  $k^{th}$  parameter, considered in a linear way;  $\beta_{\theta_k}^{(S_t)}$  is a vector of parameters for these covariates in state  $S_t$ ;  $X_{j,t}^{\theta_k}$ ,  $j = 1, \dots, J_{\theta_k}$ , is the  $j^{th}$  covariate at time  $t$  considered in the nonparametric, whereas  $h_{\theta_k,j}^{(S_t)}$  is an unspecified smooth function that links this variable to  $\theta_{k,t}^{(S_t)}$ . Each function  $h_{\theta_k,j}^{(S_t)}$  can be expressed as a finite linear combination of B-spline basis functions  $B_1, \dots, B_q$ :

$$h_{\theta_k,j}^{(S_t)}(X_{j,t}^{\theta_k}) = \sum_{i=1}^q w_{k,j,i}^{(S_t)} B_i(X_{j,t}^{\theta_k}). \quad (6)$$

As in Langrock et al. [2015] and Chavez-Demoulin et al. [2016], we consider cubic B-splines. More details are given in the next section.

In the case of  $Z_t$  following a Poisson distribution, model (4) is fully characterized by a single parameter ( $K = 1, \theta = \lambda$ ) and equation (5) characterizes the conditional expectation of  $Z_t$ , meaning that the model can be reduced to the MS-GAM framework of Langrock et al. [2015]. However, if  $Z_t$  is of the type GPD, then  $K = 2$  and equation (5) can be used to describe the conditional joint behavior of  $\gamma$  and  $\sigma$ . Now, if  $Z_t = L_t$  and is given by equation (1), we have

$$N_t | X_t, S_t \stackrel{iid}{\sim} \text{Poisson}(\lambda_t^{(S_t)}(X_t^\lambda)), \quad (7)$$

$$Y_{i,t} | X_{i,t}, S_t \stackrel{iid}{\sim} \text{GPD}(\gamma_{i,t}^{(S_t)}(X_{i,t}^\gamma), \sigma_{i,t}^{(S_t)}(X_{i,t}^\sigma)). \quad (8)$$

In the framework of the MS-GAMLSS,  $\lambda_t^{(S_t)}(X_t^\lambda)$ ,  $\gamma_{i,t}^{(S_t)}(X_{i,t}^\gamma)$ , and  $\sigma_{i,t}^{(S_t)}(X_{i,t}^\sigma)$  can be fully characterized by the following equations

$$\log(\lambda_t^{(S_t)}(X_t^\lambda)) = \eta_\lambda^{(S_t)}(X_t^\lambda) = X_t^\lambda \beta_\lambda^{(S_t)} + \sum_{j=1}^{J_\lambda} h_{\lambda,j}^{(S_t)}(X_{j,t}^\lambda), \quad (9)$$

$$\log(\gamma_{i,t}^{(S_t)}(X_{i,t}^\gamma)) = \eta_{i,\gamma}^{(S_t)}(X_{i,t}^\gamma) = X_{i,t}^\gamma \beta_{\gamma,i}^{(S_t)} + \sum_{j=1}^{J_\gamma} h_{\gamma,j}^{(S_t)}(X_{i,j,t}^\gamma), \quad (10)$$

$$\log(\sigma_{i,t}^{(S_t)}(X_{i,t}^\sigma)) = \eta_{i,\sigma}^{(S_t)}(X_{i,t}^\sigma) = X_{i,t}^\sigma \beta_{\sigma,i}^{(S_t)} + \sum_{j=1}^{J_\sigma} h_{\sigma,j}^{(S_t)}(X_{i,j,t}^\sigma). \quad (11)$$

Notice that the indices  $i$  in equations (10) and (11) allow to account for loss-specific explanatory variables that might be different across losses at a given time (e.g. categorical variables). A similar feature might be considered for the frequency distribution. In that case, the total number of events in a time period is the sum of all events across categories during this period. For notational brevity, we omit this extension in the rest of Section 2. The density function of  $N_t$ ,  $Y_{i,t}$  and  $L_t$ , conditional on  $S_t$  and  $X_t$ , will be denoted by  $f_N(n_t; \lambda_t^{(s_t)})$ ,  $f_Y(y_{i,t}; \gamma_t^{(s_t)}, \sigma_t^{(s_t)})$  and  $f_L(l_t; \lambda_t^{(s_t)}, \gamma_t^{(s_t)}, \sigma_t^{(s_t)})$ , respectively, with  $l_t$  denoting a realization of  $L_t$ . The reference to  $X_t$  is not made explicit to simplify the notation. Thanks to the (conditional) independence assumption between frequency and severity, we have the following relationship between these densities:

$$f_L(l_t; \lambda_t^{(s_t)}, \gamma_t^{(s_t)}, \sigma_t^{(s_t)}) = f_N(n_t; \lambda_t^{(s_t)}) \prod_{i=1}^{n_t} f_Y(y_{i,t}; \gamma_t^{(s_t)}, \sigma_t^{(s_t)}). \quad (12)$$

The cumulative distribution function  $F_L$  of  $L_t$  can be obtained via standard numerical integration techniques (usually through Monte Carlo simulations).

Lastly, we need to characterize the transition process between the  $M$  Markov states. We make the assumption of a first-order Markov chain. Then, under the homogeneity assumption of the Markov chain, the probabilities of transiting from state  $i$  to state  $j$ ,  $\forall i, j = 1, \dots, M$  are constant over time. Consequently, we can summarize the transition probabilities in a transition probability matrix (tpm)  $\Gamma = (\pi_{ij})$  of size  $M \times M$ , where  $\pi_{ij} = \mathbb{P}(S_t = j | S_{t-1} = i)$ ,  $i, j = 1, \dots, M$ . The initial state probabilities are summarized in a vector  $\delta$ , where  $\delta_i = \mathbb{P}(S_1 = i)$ ,  $i = 1, \dots, M$ . Usually, the various states are strongly persistent, meaning that the diagonal elements of  $\Gamma$  are close to one.

In this model, we make the assumption that the Markov chain that drives the conditional frequency and the conditional severity is identical. This implies that the latent factors driving both processes are similar. This assumption could be relaxed, though, and we would be in the case of a particular multiple MS process.

## 2.2 Estimation of the model

We use the forward recursion procedure described in Langrock et al. [2015], adapted for the special case of our model described by equations (1) and (7) to (11). More specifically, we define the vector of forward variable,

$$\alpha_t = (\alpha_t(1), \dots, \alpha_t(M)), t = 1, \dots, T, \quad (13)$$

where  $\alpha_t(j) = f_L(l_1, \dots, l_t, s_t = j | x_1, \dots, x_t)$ , for  $j = 1, \dots, M$ . These forward variables are, at each time  $t = 1, \dots, T$  and for each state  $j = 1, \dots, M$ , the joint probability of observing our sample up to time  $t$  and to have  $s_t = j$ . Then the recursive scheme presented in Langrock et al. [2015] can be applied :

$$\alpha_1 = \delta Q(l_1), \quad (14)$$

$$\alpha_t = \alpha_{t-1} \Gamma Q(l_t) \quad (t = 2, \dots, T), \quad (15)$$

where  $Q(l_t) = \text{diag}(f_L(l_t; \lambda_t^{(1)}, \gamma_t^{(1)}, \sigma_t^{(1)}), \dots, f_L(l_t; \lambda_t^{(M)}, \gamma_t^{(M)}, \sigma_t^{(M)}))$ . This recursion follows from

$$\alpha_t(j) = \sum_{i=1}^M \alpha_{t-1}(i) \pi_{ij} f_L(l_t; \lambda_t^{(j)}, \gamma_t^{(j)}, \sigma_t^{(j)}). \quad (16)$$

Additional details can be found in Zucchini et al. [2016]. Eventually, the likelihood function of model (7)-(12) is given by

$$\mathcal{L}(\Theta) = \sum_{i=1}^M \alpha_T(i) = \delta Q(l_1) \Gamma Q(l_2) \dots \Gamma Q(l_T) \mathbf{1}, \quad (17)$$

where  $\mathbf{1} \in \mathbb{R}^M$  is a column vector of ones and  $\Theta$  is the set of all parameters of the model. In our case of interest, this expression is easily obtained from the conditional density given by equation (12).

If we consider all explanatory variables in a linear way, then an estimator  $\hat{\Theta}$  of  $\Theta$  is obtained by maximizing equation (17) with respect to the parameters  $\beta = (\beta_\lambda, \beta_\gamma, \beta_\sigma)$  in equations (9)-(11),  $\Gamma$  and  $\delta$ . Besides, if we consider several covariates nonparametrically (i.e. if  $J_{\theta_k} > 0$ ), the functions  $h_{\theta_k, j}^{(S_t)}$ ,  $k = 1, \dots, K$  and  $j = 1, \dots, J_{\theta_k}$  can be expressed as a finite linear combination of B-spline basis functions as in equation (6). Following Eilers and Marx [1996], the number  $q$  of basis functions that are used should be chosen sufficiently large to account for complex functional forms (throughout this paper, we use  $q = 11$ ). In order to avoid overfitting, the estimation is performed by maximization of a penalized likelihood, where the penalty term is a function of the integrated squared curvature of the nonparametric function estimate [Eilers and Marx, 1996]. This integral is approximated by the second order difference of (6). The penalizing term is given by equation (7) in Langrock et al. [2015]. For our model, it becomes

$$D = \sum_{k=1}^3 \sum_{m=1}^M \sum_{j=1}^{J_{\theta_k}} \frac{\kappa_{k,m,j}}{2} \sum_{i=3}^q (\Delta^2 w_{k,j,i}^{(m)})^2, \quad (18)$$

where  $\kappa_{k,m,j} \geq 0$  is the smoothing parameter in state  $m$  for the  $j^{\text{th}}$  functional form of the  $k^{\text{th}}$  parameter and  $\Delta^2 w_{k,j,i}^{(m)} = w_{k,j,i}^{(m)} - 2w_{k,j,i-1}^{(m)} + w_{k,j,i-2}^{(m)}$ , i.e. the second-order difference of  $w_{k,j,i}^{(m)}$ . Maximum likelihood estimators of the proposed model are obtained by maximizing the penalized log-likelihood function

$$\mathcal{L}_{pen.}(\Theta) = \log(\mathcal{L}(\Theta)) - \sum_{k=1}^3 \sum_{m=1}^M \sum_{j=1}^{J_{\theta_k}} \frac{\kappa_{k,m,j}}{2} \sum_{i=3}^q (\Delta^2 w_{k,j,i}^{(m)})^2, \quad (19)$$

for a particular vector of smoothing parameters  $\kappa$ .

The maximization of equation (19) is rapidly subject to numerical underflow when  $T$  is large. This issue is addressed by considering the log-likelihood function instead of the likelihood itself. However, because we are dealing with product of matrices, we cannot simply apply a log-transform to equation (17). Instead, we need to use a scaling algorithm, as described in Zucchini et al. [2016], pp. 48 and following. Theoretically speaking, an additional underflow issue might arise when some elements of  $\{n_t\}$  are too large, since the density in equation (12) goes to zero for large  $n_t$ . Nevertheless, in practice  $n_t$  is often quite small, and we did not come across this issue in the considered application.

Beside underflow issues, maximizing expressions (17) and (19) can be challenging numerical tasks. It is important to test several starting values to avoid local minima. We need also to take into account constraints on the parameters (especially positivity constraints and sum of the rows equal to 1 for the tpm). In a first step, Zucchini et al. [2016] suggest to use a re-parametrization. Subsequently, the maximization can be carried out with a regular optimization routine. In this work, we follow Langrock et al. [2015] and use a standard nonlinear optimizer after a suitable re-parametrization employing a multinomial logistic link function for the tpm.

To ensure the correct identifiability of this model, we fix the value of one of the basis coefficient  $w_{k,j,i}^{(S_t)}$  in each function of each regime and for each parameter. A common strategy consists in taking an odd number of B-spline basis  $q$  and to assume  $w_{k,j,\frac{q+1}{2}}^{(S_t)} = 0$ , for  $j = 1, \dots, J_{\theta_k}$ ,  $k = 1, \dots, 3$  and  $S_t = 1, \dots, M$ . Furthermore, regarding the transition probabilities, we make the assumption of a stationary transition process, implying that  $\delta$  is the solution to the linear system  $\delta = \delta\Gamma$  subject to  $\sum_{i=1}^M \delta_i = 1$ . This standard assumption is discussed, among others, in Zucchini et al. [2016].

### 2.3 Choice of the smoothing parameter and the number of states

In this paper, we use a  $C$ -fold cross-validation technique in the simulations and the application for the purpose of smoothing parameters selection. In this approach, the initial time series is randomly partitioned into  $C$  subsamples. Then the model is repeatedly fitted treating one of the subsamples as missing data, and the (negative) cross-validated likelihood score is computed on the subsample only, treating this time the calibration data as missing. The chosen smoothing parameter is the one that minimizes the average cross-validated score over the  $C$  partitions. Nevertheless, in case of small samples and the presence of very large discrepancies between partitions, the  $C$ -fold approach might be problematic. An alternative is to use repeated random subsamples: the initial dataset is repeatedly partitioned into two, randomly and without replacement. This method is appealing in our case, as the GPD can produce data that vary a lot (a feature potentially reinforced with the MS component of the model). For computational convenience, the selection is performed over a search grid  $\Lambda \subset \mathbb{R}_{\geq 0}^{M \times (J_\lambda + J_\sigma + J_\gamma)}$ .

Model comparison can be done with the AIC. The AIC for a particular model (i.e. with a particular smoothing parameters  $\kappa$  and a particular subset of covariates) is given by

$$AIC(\hat{\Theta}, \kappa) = -2 \log(\mathcal{L}(\hat{\Theta})) + 2 \cdot df(\kappa) \quad (20)$$

where  $df(\kappa)$  is the effective degrees of freedom of the final model obtained with smoothing parameter  $\kappa$  and  $\hat{\Theta}$  the estimated parameters of the model. In a parametric model, the degrees of freedom is given by the number of parameters whereas in a semiparametric context, it can be obtained by the trace of the product of the Fisher information matrix for the unpenalized likelihood and the inverse Fisher information matrix for the penalized likelihood [Gray, 1992]. Among several models, the one that minimizes the AIC will be considered as the best model. This approach is computationally more efficient than a cross-validation approach but it might also lead to significant undersmoothing problems [Hurvich et al., 2012]. Moreover, due to the complexity of the likelihood function, the Fisher information matrix needs to be numerically estimated. We use this criterion in the empirical study to compare several candidate models.

Lastly, regarding the number of states, Langrock et al. [2015] argue that in practice this choice is rather arbitrary. In the economic context, Ang and Timmermann [2012] explain that this choice is difficult to do ex-ante using the data, and should be based on theoretical arguments. They notice that it is not uncommon to fix the number of regime at some values, typically two. One of the main reasons is that related econometric tests are difficult to implement, due to non-standard distributions of the test statistics under the null. Besides, several information criteria (AIC, SBIC, Hannan-Quinn criterion or cross-validated likelihood) can be of some help, but they tend to favour a high number of states. A better practice seems to check a-posteriori the goodness-of-fit of the estimated models using different numbers of states, with the goal to detect if an additional complexity might improve the overall goodness-of-fit. This analysis can be carried out by examining the autocorrelation of the (pseudo-)residuals.

Another important criterion that needs to be taken into account is the length of the time series under study. It is intuitively clear that it will be very difficult to properly identify a large number of different states when the time series is short, due to the typical high persistence of the states and consequently to the possibility of not observing all states. In our particular case of random sums over time, however, this issue is partially reduced. Thanks to the time aggregation, we observe many loss realizations for a single point in time (we have a single realization of the Poisson process, but many realizations of the GPD random variable). Hence, fewer time periods are needed to estimate well the regression function in a given state compared to other models. In the present work, we restrict our attention to cases where two states are sufficient to describe the temporal structure, since in our application we only have few time periods.

### 3 Simulations

In this section, we study the finite sample behavior of the suggested estimation procedure, for the proposed model. We consider a two-states MS-GAMLSS compound Poisson-GPD process:

$$\begin{aligned} L_t &= \sum_{i=1}^{N_t} Y_{i,t}, \quad t = 1, \dots, T, \\ N_t &\sim \text{Poisson}(\lambda_t^{(s_t)}(X_t^\lambda)), \\ Y_{i,t} &\sim \text{GPD}(\gamma_t^{(s_t)}(X_t^\gamma), \sigma_t^{(s_t)}(X_t^\sigma)), \\ s_t &\in \{1, 2\}. \end{aligned}$$

This model is similar to the one described by equations (1) and (7) to (12). We assume that the severity at time  $t$  is only a function of time-dependent covariates, and not from loss-specific covariates

(i.e.  $X_{i,t}^\theta = X_{j,t}^\theta = X_t^\theta, \forall i, j$  and  $\theta \in \{\sigma, \gamma\}$ ). The superscripts  $\lambda, \gamma$  and  $\sigma$  indicate that we can consider different subsets of explanatory variables for each parameter. Furthermore, we assume that all covariates follow a uniform distribution between 0 and 1:

$$X_t^\theta \sim U(0, 1), \theta \in \{\lambda, \gamma, \sigma\}.$$

We consider the following regression equations for the parameters, in state  $s_t \in \{1, 2\}$ :

$$\begin{aligned} \log(\lambda_t^{(s_t)}(X_t^\lambda)) &= \beta_{\lambda,0}^{(s_t)} + h^{(s_t)}(X_t^\lambda), \\ \log(\sigma_t^{(s_t)}(X_t^\sigma)) &= \beta_{\sigma,0}^{(s_t)} + h^{(s_t)}(X_t^\sigma), \\ \log(\gamma_t^{(s_t)}(X_t^\gamma)) &= \beta_{\gamma,0}^{(s_t)}, \end{aligned}$$

with

$$h^{(1)}(x) = 0.6(x - 0.5)^2 + \sin(-2x - 1),$$

and

$$h^{(2)}(x) = -0.5 - 2.8(x - 0.5) + 0.2(x - 0.5)^2 + 0.6 \sin(-2x - 1) + 0.5 \cos(4x - 2).$$

We use the following values for the constant terms:  $\beta_{\sigma,0}^{(s_t)} = [1.3, 1.5]$  and  $\beta_{\gamma,0}^{(s_t)} = [-.2, -.7]$ . For  $\beta_{\lambda,0}^{(s_t)}$ , we assume  $\beta_{\lambda,0}^{(1)} = \beta_{\lambda,0}^{(2)} = \beta_{\lambda,0}$  and we use three different values:  $\beta_{\lambda,0} \in \{2, 3, 4\}$ . The different values of  $\beta_{\lambda,0}$  imply average numbers of observations at each time that are around 4, 12 and 25, respectively. It allows us to look at the effect of an increasing number of observations per time period on the quality of the estimation. Regarding the covariates, we consider the cases where  $x_t = x_t^\lambda = x_t^\sigma$  (Scenario 1) and where  $x_t^\lambda \neq x_t^\sigma$  (Scenario 2). For the tpm, we assume that  $\pi_{11} = \pi_{22} = .95$ . In both scenarios, we make the simplifying assumption that the behavior of  $\gamma$  is independent from any covariate (this assumption can be easily relaxed, though). The selection of the smoothing parameters is made using a 10-fold cross-validation procedure, considering that the smoothing parameter in a given state is the same for  $\lambda$  and  $\sigma$ . Therefore the selection is performed among 25 vectors of size 4, with their components belonging to  $\Lambda = \{0.5, 2, 8, 25, 50\}$ . We use two different values for the number of time periods:  $T \in \{50, 100\}$ . Regarding the number of basis functions  $q$  and the number of simulated samples  $B$ , we set them to 11 and 200, respectively. For identifiability, the weight associated to the 6<sup>th</sup> basis is set to 0. This simulation set-up is inspired by the one of Langrock et al. [2015].

To assess the quality of the estimation, we compute an estimator of the MISE between the true functional form  $h^{(j)}$  and the estimated one,  $\hat{h}^{(j)}$ , for each  $\theta \in \{\lambda, \sigma\}$  and each state  $j \in \{1, 2\}$ :

$$\widehat{MISE}_{\nu_\theta^{(j)}} = \frac{1}{B} \sum_{i=1}^B \left( \int_0^1 \left( \hat{h}_\theta^{(j)}(x) - h_\theta^{(j)}(x) \right)^2 dx \right). \quad (21)$$

In addition, we compute the average estimates  $\bar{\beta}_{\theta,0}^{(j)}$  of  $\beta_{\theta,0}^{(j)}$ , for  $j \in \{1, 2\}$  and  $\theta \in \{\lambda, \gamma, \sigma\}$ , as well as the average estimate  $\bar{\pi}_{ii}$  of the diagonal elements  $\pi_{ii}$ ,  $i = 1, 2$  of the tpm. Using the Viterbi algorithm detailed in Zucchini et al. [2016], we decode the estimated states  $\hat{s}_t$ . We compute the proportion of correctly classified state, for each sample:

$$R = \frac{1}{T} \sum_{t=1}^T \mathbb{1}(s_t = \hat{s}_t), \quad (22)$$

with  $\mathbb{1}(\cdot)$  being an indicator function taking value 1 when the condition in parenthesis is met, 0 otherwise. The average  $\bar{R}$  and median  $\tilde{R}$  across all samples give us measures of the accuracy with which we are able to identify the states over time.

Tables 1 and 2 display the various MISE and average estimates for Scenarios 1 and 2, respectively. Figure 1 shows the boxplots of the correct classification rates (mean and median ratio in each simulation can be found in Table 3). In general, we estimate quite well the dependence structure even for short time series when the average number of observations per time period is at least 12. Figure 2 shows the 95% coverage bands on the final parameters estimates for  $\lambda$  and  $\sigma$ , when  $T = 50$ ,  $\beta_{\lambda,0} = 4$  for Scenario 1. The coverage bands for the other scenario are available as supplementary materials. We see that in general, the functional forms are well estimated, but when the covariates take values close to their upper bound, the quality of the estimation deteriorates. It seems that in this region, both functional forms are closer one to another and thus are harder to distinguish. Overall, all other parameters of the model are well estimated and the quality of the estimation improves quite quickly when the number of observations during each time period increases. No big differences are observed between Scenarios 1 and 2. Regarding the correct classification ratio, even for  $T = 50$  we reach an average of 98.7% of correctly classified states when  $\beta_{\lambda,0} = 4$ . The median ratio is even better, varying between 94% and 100%. One should also notice that for the case where  $\beta_{\lambda,0} = 2$  and  $T = 50$  (i.e. samples with an average size of 202 observations), very bad estimated values of  $\pi_{ii}$ ,  $i = 1, 2$  and of the classification ratio are sometimes observed. This, in turn, has a negative effect on the estimation of the other parameters. Hence, when the time series is short, the frequency of the process (i.e. the number of losses over a time period) should compensate, in order to obtain a good final estimation. However, overall, the simulation study demonstrates the good performance of the estimation procedure, even for samples of reasonable sizes.

$x_t^\lambda = x_t^\sigma$	$\beta_{\lambda,0}$	$\widehat{MISE}_{h_\lambda^{(1)}}$	$\widehat{MISE}_{h_\lambda^{(2)}}$	$\widehat{MISE}_{h_\sigma^{(1)}}$	$\widehat{MISE}_{h_\sigma^{(2)}}$	$\bar{\pi}_{11}$	$\bar{\pi}_{22}$
$T = 50$	2	0.159	0.184	0.240	0.223	0.899	0.890
		(0.310)	(0.280)	(0.311)	(0.274)	(0.152)	(0.173)
	3	0.067	0.099	0.106	0.112	0.901	0.918
		(0.092)	(0.115)	(0.148)	(0.155)	(0.123)	(0.103)
	4	0.044	0.061	0.053	0.056	0.923	0.921
		(0.053)	(0.066)	(0.074)	(0.051)	(0.110)	(0.097)
$T = 100$	2	0.082	0.099	0.134	0.121	0.933	0.938
		(0.179)	(0.126)	(0.206)	(0.154)	(0.061)	(0.070)
	3	0.046	0.058	0.069	0.078	0.934	0.933
		(0.064)	(0.067)	(0.088)	(0.095)	(0.054)	(0.053)
	4	0.027	0.037	0.042	0.048	0.940	0.937
		(0.033)	(0.029)	(0.049)	(0.043)	(0.044)	(0.045)
	$\beta_0(\lambda)$	$\bar{\beta}_{\lambda,0}^{(1)}$	$\bar{\beta}_{\lambda,0}^{(2)}$	$\bar{\beta}_{\sigma,0}^{(1)}$	$\bar{\beta}_{\sigma,0}^{(2)}$	$\bar{\beta}_{\gamma,0}^{(1)}$	$\bar{\beta}_{\gamma,0}^{(2)}$
$T = 50$	2	1.973	1.948	1.456	1.496	-0.531	-1.119
		(0.343)	(0.367)	(0.520)	(0.381)	(0.828)	(1.285)
	3	3.033	3.047	1.347	1.561	-0.262	-0.790
		(0.228)	(0.276)	(0.363)	(0.332)	(0.243)	(0.330)
	4	3.996	3.988	1.286	1.513	-0.219	-0.739
		(0.175)	(0.147)	(0.223)	(0.167)	(0.131)	(0.145)
$T = 100$	2	1.981	1.953	1.43	1.498	-0.383	-0.864
		(0.231)	(0.204)	(0.379)	(0.233)	(0.488)	(0.546)
	3	2.986	3.007	1.307	1.505	-0.247	-0.771
		(0.167)	(0.155)	(0.233)	(0.195)	(0.144)	(0.329)
	4	3.989	3.999	1.304	1.505	-0.216	-0.722
		(0.128)	(0.118)	(0.168)	(0.143)	(0.138)	(0.102)

Table 1: Results of Scenario 1, with  $x_t^\lambda = x_t^\sigma$ . The MISE are computed with  $B = 200$  simulated samples, with the length  $T$  of the time series being either 50 or 100.  $\bar{\beta}_{\theta,0}^{(j)}$  denotes the mean estimate over all estimates  $\hat{\beta}_{\theta,0}^{(j)}$ , for  $j = 1, 2$  and  $\theta \in \{\lambda, \sigma, \gamma\}$ . The different values of  $\beta_{\lambda,0}$  correspond to (average) total sample sizes of 202, 565 and 1497 for  $T = 50$ , and 414, 1191 and 3025 for  $T = 100$ , respectively. Standard errors over the 200 samples are in parentheses.

$x_t^\lambda = x_t^\sigma$	$\beta_{\lambda,0}$	$\widehat{MISE}_{h_\lambda^{(1)}}$	$\widehat{MISE}_{h_\lambda^{(2)}}$	$\widehat{MISE}_{h_\sigma^{(1)}}$	$\widehat{MISE}_{h_\sigma^{(2)}}$	$\bar{\pi}_{11}$	$\bar{\pi}_{22}$
$T = 50$	2	0.1641	0.1563	0.1943	0.1954	0.907	0.906
		(0.231)	(0.223)	(0.238)	(0.275)	(0.119)	(0.131)
	3	0.094	0.103	0.107	0.118	0.924	0.908
		(0.186)	(0.145)	(0.186)	(0.176)	(0.087)	(0.105)
	4	0.036	0.0485	0.052	0.065	0.925	0.923
		(0.04)	(0.039)	(0.083)	(0.062)	(0.072)	(0.073)
$T = 100$	2	0.093	0.112	0.122	0.126	0.939	0.933
		(0.1314)	(0.188)	(0.154)	(0.161)	(0.052)	(0.084)
	3	0.040	0.055	0.065	0.066	0.934	0.936
		(0.047)	(0.061)	(0.081)	(0.067)	(0.057)	(0.058)
	4	0.026	0.039	0.034	0.046	0.939	0.935
		(0.028)	(0.027)	(0.043)	(0.042)	(0.047)	(0.05)
	$\beta_0(\lambda)$	$\bar{\beta}_{\lambda,0}^{(1)}$	$\bar{\beta}_{\lambda,0}^{(2)}$	$\bar{\beta}_{\sigma,0}^{(1)}$	$\bar{\beta}_{\sigma,0}^{(2)}$	$\bar{\beta}_{\gamma,0}^{(1)}$	$\bar{\beta}_{\gamma,0}^{(2)}$
$T = 50$	2	2.004	1.917	1.396	1.522	-0.545	-1.172
		(0.343)	(0.309)	(0.442)	(0.351)	(1.199)	(1.228)
	3	3.011	2.992	1.322	1.504	-0.279	-0.788
		(0.215)	(0.251)	(0.285)	(0.276)	(0.279)	(0.344)
	4	4.002	3.974	1.289	1.491	-0.225	-0.77
		(0.143)	(0.159)	(0.214)	(0.192)	(0.153)	(0.272)
$T = 100$	2	2.012	1.973	1.312	1.516	-0.295	-1.009
		(0.245)	(0.277)	(0.277)	(0.271)	(0.284)	(1.129)
	3	2.987	3.004	1.299	1.509	-0.226	-0.760
		(0.160)	(0.175)	(0.223)	(0.185)	(0.157)	(0.301)
	4	3.999	4.001	1.287	1.505	-0.207	-0.719
		(0.118)	(0.133)	(0.155)	(0.149)	(0.089)	(0.134)

Table 2: Results of Scenario 2, with  $x_t^\lambda \neq x_t^\sigma$ . The MISE are computed with  $B = 200$  simulated samples, with the length  $T$  of the time series being either 50 or 100.  $\bar{\beta}_{\theta,0}^{(j)}$  denotes the mean over all estimates  $\hat{\beta}_{\theta,0}^{(j)}$ , for  $j = 1, 2$  and  $\theta \in \{\lambda, \sigma, \gamma\}$ . The different values of  $\beta_{\lambda,0}$  correspond to (average) total sample sizes of 202, 565 and 1497 for  $T = 50$ , and 414, 1191 and 3025 for  $T = 100$ , respectively. Standard errors over the 200 samples are in parentheses.

$T = 50$	$x_t^\lambda = x_t^\sigma$			$x_t^\lambda \neq x_t^\sigma$			$T = 100$	$x_t^\lambda = x_t^\sigma$			$x_t^\lambda \neq x_t^\sigma$		
	$\beta_0^{(S_t)}(\lambda)$	2	3	4	2	3		4	$\beta_0^{(S_t)}(\lambda)$	2	3	4	2
$\bar{R}$	0.891	0.957	0.987	0.876	0.955	0.995	$\bar{R}$	0.942	0.976	0.992	0.953	0.982	0.995
$\tilde{R}$	0.94	0.98	1.00	0.94	0.980	1.00	$\tilde{R}$	0.95	0.98	0.99	0.97	0.99	1
$\tilde{\pi}_{11}$	0.951	0.935	0.947	0.948	0.950	0.943	$\tilde{\pi}_{11}$	0.947	0.948	0.951	0.953	0.95	0.95
$\tilde{\pi}_{22}$	0.943	0.945	0.947	0.946	0.943	0.942	$\tilde{\pi}_{22}$	0.953	0.947	0.949	0.949	0.953	0.945

Table 3: Mean and median correct classification ratio, as well as median of the estimated diagonal transition probabilities for Scenarios 1 and 2. Left (resp. right) panel:  $T = 50$  (resp.  $T = 100$ ).

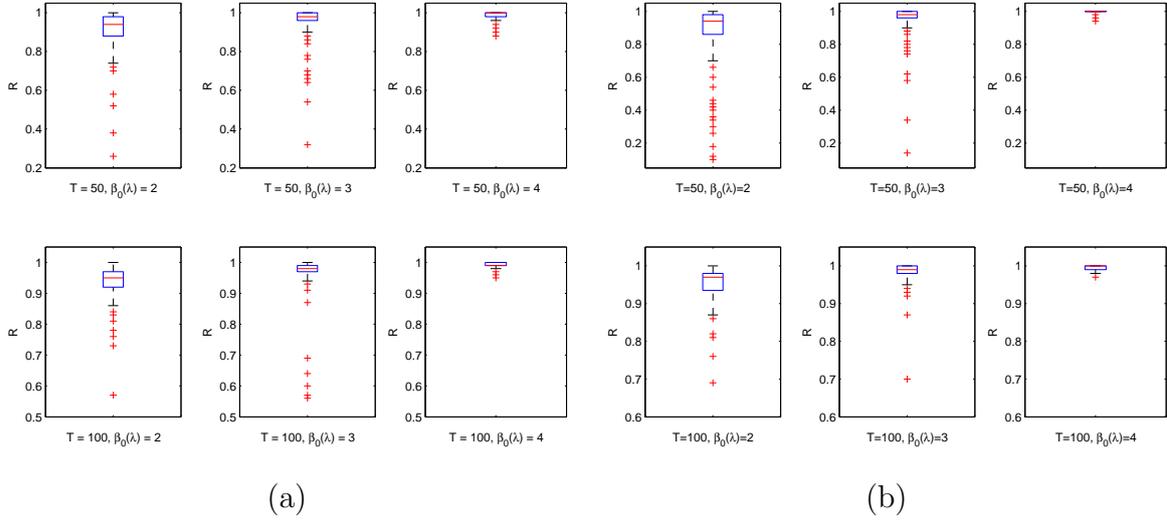


Figure 1: Boxplots for the proportion of correctly classified latent states, given by equation (22). 1 corresponds to 100% of the time periods correctly classified. (a):  $x_t^\lambda = x_t^\sigma$ . (b):  $x_t^\lambda \neq x_t^\sigma$ . Top:  $T = 50$ . Bottom:  $T = 100$ .

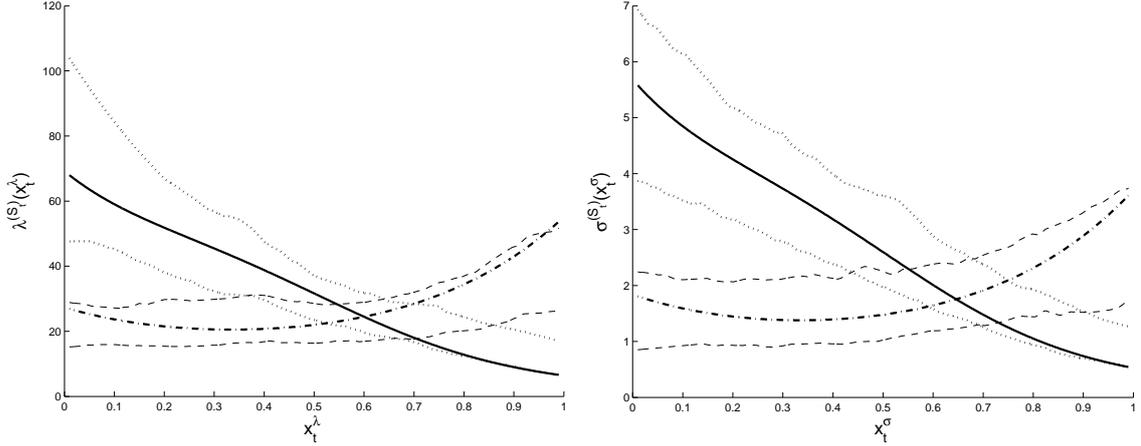


Figure 2: 95% coverage bands (dotted and dashed) for  $\lambda_t^{(j)}(x_t^\lambda)$  (left) and  $\sigma_t^{(j)}(x_t^\sigma)$  (right). Dashed dotted (resp. solid): true function for  $j = 1$  (resp.  $j = 2$ ).  $T = 50$  and  $\beta_0(\lambda) = 4$ .

## 4 Empirical Study

In this section, we apply the proposed procedure to real financial data. We consider a novel dataset of 817 operational losses at UniCredit, one of the largest European banks, resulting from external frauds. Operational losses are defined by the Basel Committee for Banking Supervision (BCBS) as *direct or indirect losses resulting from inadequate or failed internal processes, people and systems or from external events* [BCBS, 2004]. In this application, we focus on the particular class of operational losses resulting from external frauds. Over the past 15 years, impressive fraud losses regularly made the headlines of newspapers for their huge economical consequences. Famous examples are losses from rogue trading at the Barings bank (1995), rogue trading at Société Générale (2007) and JP Morgan (2012), the LIBOR scandal (2011) or more recently, massive fraudulent cyber-attacks of the SWIFT system in Bangladesh, Russia and Japan (2016).

### 4.1 Description of the data

UniCredit operates in 17 countries. In 2007, its share price reached €42.8, before a drastic decline following the subprime crisis in 2008 and the sovereign crisis in 2011 to reach €2.83. Its total asset at the end of the first quarter of 2016 reached €892.04bn. Around 50% of UniCredit total revenue comes from its Italian activities. In the context of the *Advanced Measurement Approach*, UniCredit is asked by its regulator to model the distribution of its total operational loss  $L_t$  over time. For operational losses, the capital reserve is derived from the estimated 99.9% quantile of  $L_t$  (i.e. from the estimation

of  $Q_{0.999}(L_t)$  defined s.t.  $\mathbb{P}(L_t \leq Q_{0.999}(L_t)) = 0.999$ ). The standard modeling approach relies on unconditional compound Poisson-GPD processes [see, among others, Moscadelli, 2004, Dutta and Perry, 2006, Soprano et al., 2009, Chapelle et al., 2008]. However, a recent stream of papers emphasizes the effect of some economic variables and the importance of regulatory changes on the distribution of  $L_t$ , suggesting that it might be more accurate to model the quantile of  $L_t$ , conditional on covariates and a latent state:  $Q_{0.999}(L_t; X_t, S_t)$ , defined by  $\mathbb{P}(L_t \leq Q_{0.999}(L_t; X_t, S_t)) = 0.999$ . Since our MS-GAMLSS approach handles these features, we will use it to study the variation over time of the total loss distribution, and in particular of  $Q_{0.999}(L_t; X_t, S_t)$ .

The collection period of the losses ranges between January 2005 and June 2014. Losses have been scaled by an unknown factor for anonymity reasons. The minimal amount considered is 25,000€ (so that we have a sample of extreme losses, where the GPD is a reasonable hypothesis for the severity distribution, see e.g. Embrechts et al. [1997], Beirlant et al. [2005]). We have access to the exact date of each loss, meaning that we can assign each loss to a specific year and a specific quarter, and compute the total loss for each quarter. We work on a quarterly basis as we wouldn't have enough time periods to estimate correctly the transition probabilities of the model otherwise. However, on a regulatory point of view, banks are expected to define the capital reserve for a horizon of one year. In this case, a simple addition of the forecast capital for 4 consecutive quarters could be used. As in Cope et al. [2012] and Chavez-Demoulin et al. [2016], we adjust the loss amounts for inflation, using the Italian consumer price index from the OECD website. The final loss amounts used to fit the model are obtained by subtracting the collection threshold (25,000€) to the adjusted loss amount (it can be reinstated later to shift all losses from this amount, the other parameters being not affected by this transformation). Figure 3 shows the evolution of the total loss per quarter, the number of losses per quarter as well as the distribution of the loss amounts over time (in the log scale). The number of losses per quarter ranges between 8 and 54. We observe a decrease of the number of losses starting in 2008. Regarding the total loss, we face a big spike in the second quarter of 2009, due to a single extremely large loss. The total loss during that quarter is twice as big as the second biggest observed total loss, taking place in 2006. Following this spike, the average total loss per quarter seems to remain at much lower levels until the end of the considered period.

Regarding the economic covariates, Chernobai et al. [2011] discover a strong link between firm-specific covariates and the intensity of the operational loss process. They conclude that a high level of financial distress is associated with more frequent operational events, especially frauds. Regarding the severity of the operational losses, Cope et al. [2012] suggest that the expected severity of the losses is positively linked with the economic well-being of a country. In this application, we use the following covariates:

- To model the frequency parameter  $\lambda$ , we use a firm-specific variable, namely the percentage of the total revenue coming from fees (PRF). The PRF might be seen as a measure of the economic well-being of the bank. The higher this ratio is, the less dependent the bank is from market interest rates. Simultaneously, the PRF could also measure the level of activity of the bank on behalf of clients. The higher the PRF, the more the bank provide services to clients. A high level of this activity can increase the average number of losses resulting from frauds (as more clients are likely to commit frauds).
- To model the scale parameter  $\sigma$ , we use the Italian unemployment rate. This quantity serves as a proxy of the overall economic performance of Italy, where UniCredit has its main activities. As explained in the introduction, the bad or good state of an economy might create incentives for people to commit larger frauds.
- To model the shape parameter  $\gamma$ , we consider the values of the VIX index. The VIX is a measure of the market volatility, based on put and call options of the S&P500. It is also considered as a barometer of market sentiments. It might help measuring the stability of the financial system. High values of the VIX indicate a high uncertainty on the financial markets, which can be translated in higher probabilities of extreme losses.

A natural extension of the present framework would be to use several covariates for each parameter. However, since our time series is quite short, we don't want too many parameters needing to be estimated. Hence we restrict our attention to the case where each parameter depends on a single covariate.

At each time, we consider the covariates' values from the previous quarter to perform the regression. The use of lagged values of the covariates allows to limit causality issues and to predict the distribution without also predicting these variables (in the case of a one-step ahead forecast). The evolution of these covariates, over the considered period, is displayed in Figure 4. We observe several important changes (at the end of 2007, in September 2009 or in 2011), all corresponding to some crisis (namely the subprime crisis and the sovereign crisis). These changes might indicate potential shifts between two different regimes or some structural breaks. Especially, notice that we observe a high peak of the (lagged) VIX index at the same time as the observed peak of the total loss (for the second quarter of 2009).

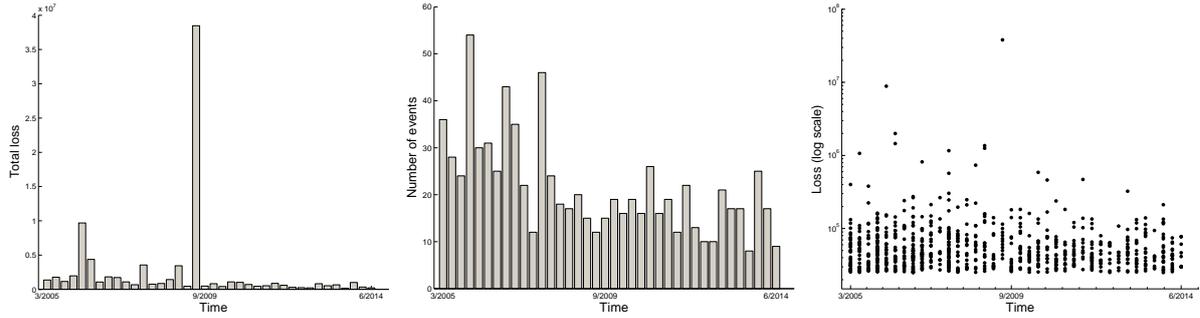


Figure 3: From left to right: total loss, number of losses per quarter, observed losses (grouped per quarter in log-scale).

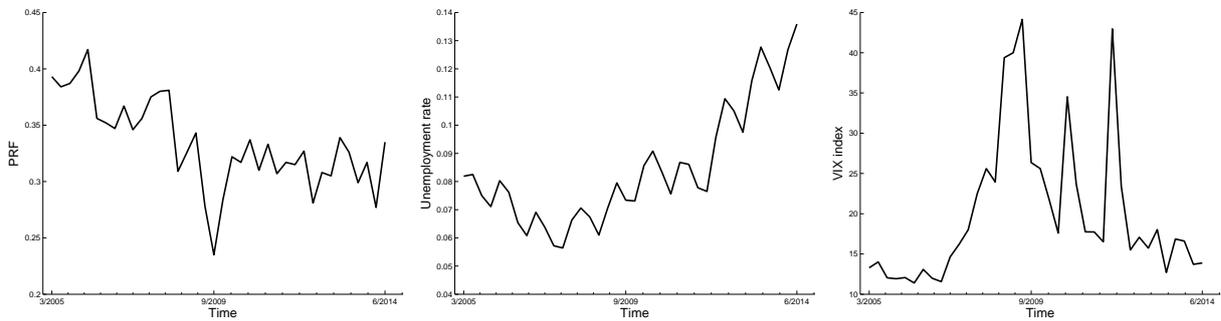


Figure 4: Value of the explanatory variables over the considered period. From left to right: lagged values (one quarter) of the PRF, the Italian unemployment rate and the VIX index.

## 4.2 Results

Figure 5 shows the estimated functional forms for the different parameters and in the two states (VIX values have been divided by 100 to vary between 0 and 1). The smoothing parameters have been chosen with a CV procedure (with repeated resamples), over a grid of vectors built on the values  $\{.5, 2, 8, 15, 25\}$ . The selected value of the smoothing parameter is identical for each covariate and in each state, and equal to 8. The nonparametric approach seems useful to handle some non-linearities (in particular for the unemployment rate and the VIX). We used 11 basis functions. Increases in the PRF are associated with increases in  $\lambda$ , in both states. This association is stronger in state 2 compared to state 1. For the unemployment rate, we observe a U-shape of the functional forms: up to a given point (around 10%), an increase in the unemployment rate is associated with a decrease in  $\sigma$ , and then with an increase in  $\sigma$  beyond this point (however, since the density of the covariate beyond this point is low, it can be simply a boundary effect). For the VIX index, we observe in one state a strong increase in  $\gamma$  when the VIX increases, whereas in the other state, the functional form is quite flat. Estimated values of the constant

parameters and of the transition probabilities, are given in Table 4. Both states are found to be highly persistent, as indicated by the diagonal elements of the tpm close to one (see Table 4).

Figure 6 displays the estimated parameters in both states, over time. Using the Viterbi algorithm, we assign the period January 2005 - March 2007 to the second state. Then we observe three shifts between states: from state 2 to state 1 during the 2<sup>nd</sup> and 3<sup>rd</sup> quarter of 2007, from state 1 to state 2 during the last quarter of 2007, and then from state 2 to state 1 during the first quarter of 2008, up to June 2014. Figure 7, left side, shows the state probabilities over time. The first shift corresponds to the period where a merger with another bank, Capitalia, takes place. Looking at UniCredit historical figures and yearly reports, we observe a huge increase in the leverage ratio for the 3<sup>rd</sup> quarter of 2007, following the buyout of Capitalia during the previous quarter. The second shift (during the last quarter of 2007) corresponds to a period of intense reduction of the leverage ratio, as well as a restructuring of UniCredit activities. Lastly, the third shift corresponds to the start of a progressive increase in the Tier-I capital ratio. It could indicate that following the crisis, UniCredit decided to operate a drastic change in its risk management. January 2008 sees also the enforcement of Basel II rules for the advanced internal rating-based approach. This change of regulation could be the source of changes in the risk management process, which in turn influence the frequency and severity of operational losses.

The relationships between parameters and covariates seem different among regimes. Especially, variations of the VIX appear to be linked with larger variations of  $\gamma$  in state 1, compared to state 2. It could be due to the fact that state 1 is a crisis regime, where the uncertainty is particularly important. At these times, the likelihood of an extreme event appears more tightly linked with market conditions, and increases drastically. Despite its strategical and/or structural changes, UniCredit seems to be heavily dependent from the volatility of the financial markets in this regime. However, when the VIX decreases,  $\gamma$  decreases more in state 1, compared to state 2. Consequently, when the VIX is low in state 1, we reach lower values of  $\gamma$  and lower probabilities of extreme events. This is also due to a difference of level between regimes: the constant parameter is lower in state 1, compared to state 2 (for a value of the VIX equal to zero, we would observe respectively values of  $\gamma$  equal to 0.39 and 0.50).

Regarding the scale parameter  $\sigma$ , we see that it decreases with the increase in the unemployment rate, up to a given value. The inflexion point seems different among regimes. We could give the following interpretation of these results: when the economic situation worsens (i.e. when the unemployment rate increases), there are less opportunities and incentives to commit large frauds (i.e. the scale of the frauds decreases). However, when the situation becomes too bad (e.g. when the unemployment rate reaches 10%), people have less to lose and we face an increased probability of large frauds. Over time,  $\sigma$  is larger in state 1 than in state 2, indicating that the change of regime is synonym of an increased variability of

the losses, possibly because of the increased macroeconomic instability in state 1.

For  $\lambda$ , no big differences among regimes are observed for the functional form of the dependence. A way of explaining the observed relationship could be that the PRF measures the level of activity of the bank. When the PRF increases, it indicates that the bank carries out more services for clients, leading to potentially more losses resulting from frauds of clients. Nevertheless, the frequency parameter seems mostly determined by the constant parameter, indicating that the relationship with the covariate is weak. This constant is quite different between regimes, and induces a major change of level: in state 2,  $\lambda$  fluctuates around 35, whereas in the other state,  $\lambda$  varies between 15 and 20. Once again, structural or regulatory changes might be the cause of this drop.

Bootstrap confidence intervals (based on 1000 resamples) are displayed in Figure 10 in the Appendix. They are fairly wide and hence do not let us firmly conclude that differences among functional forms in the two regimes are significant, nor do they allow us to say if the relationships with the covariates are significant. This is a concern with MS models, already reported by Zucchini et al. [2016], and that is emphasized by the complex nonparametric dependence in the model.

To study the regulatory implications of the proposed approach, we compute Monte Carlo estimators  $\hat{Q}_{0.999}(L_t; X_t, S_t)$  of  $Q_{0.999}(L_t; X_t, S_t)$ . Figure 7 displays these estimations in each state, computed for every quarter using 100,000 random draws from the convolution of the Poisson and GPD distributions. Looking at the quantiles' values conditional on the decoded states, we see that up to 2007,  $\hat{Q}_{0.999}(L_t; X_t, S_t)$  is quite stable. Then, after the regime change,  $\hat{Q}_{0.999}(L_t; X_t, S_t)$  increases drastically in 2008 and 2009, during the crisis. We never observe a breach of the quantile, i.e. total losses larger than the corresponding estimated quantiles (conditional on the decoded state), despite a huge total loss on the 2<sup>nd</sup> quarter of 2009. Thanks to the proposed dependence structure, we detect a large increase in the shape parameter at that time, and an increase in the 99.9% quantile. From a regulatory perspective, our model indicates correctly that more capital should be needed to cover potential operational losses during that period, compared to the previous periods. This is interesting economically speaking, as this model could prevent financial institution to set aside too much or too little capital.

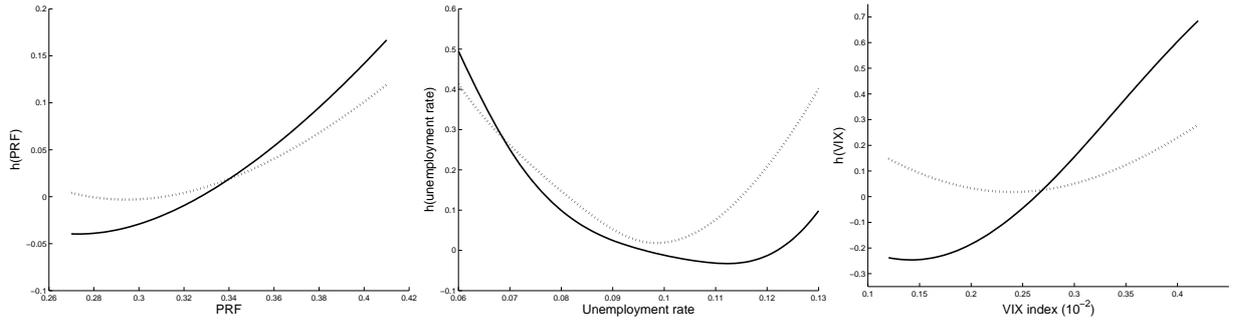


Figure 5: Estimated nonparametric functional forms between the covariates and the parameters of the model. Solid (resp. dotted) line: state 1 (resp. 2). From left to right: PRF, unemployment rate and VIX index.

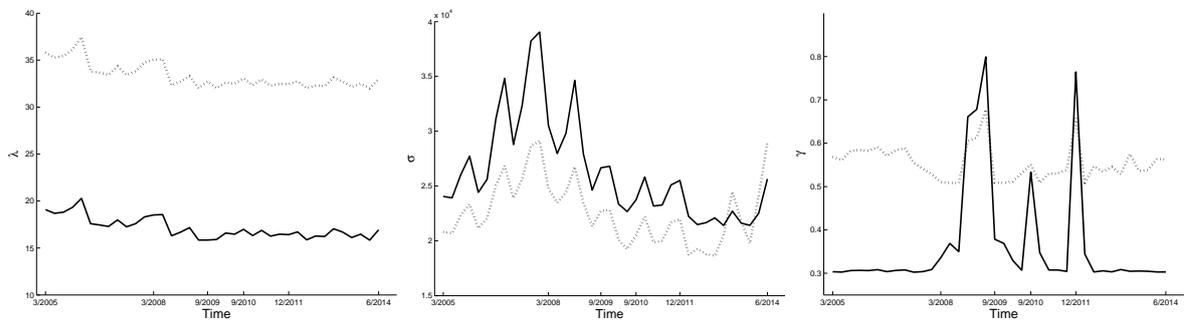


Figure 6: Estimated values of the parameters in both states (solid: state 1; dotted: state 2) over the covered period.

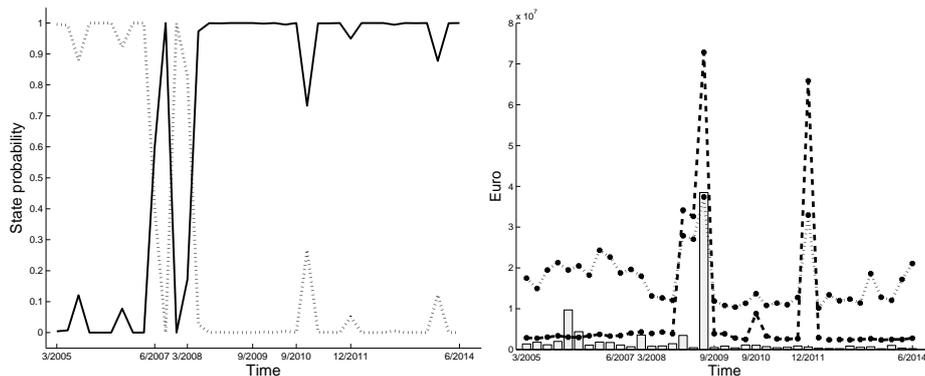


Figure 7: Left: state probabilities for the different regimes (thin solid: first state; dotted: second state). Right:  $\hat{Q}_{0.999}(L_t; X_t, S_t)$  over time, in both states (dashed: first state, dotted: second state). Grey bar: total loss

To benchmark the performance of our model, we compare our two-states MS-GAMLSS model with three others models. First, we consider the unconditional, constant model: the frequency and severity

distributions are assumed to have parameters constant over time and across covariates, as in Embrechts et al. [1997], Chapelle et al. [2008], Dutta and Perry [2006]. Second, we consider a GAMLSS model as in Chavez-Demoulin et al. [2016], where the dependence with covariates is taken into account (non-parametrically), but where the structure is assumed not to change over time. Third, we consider a MS-Poisson-GPD model without explanatory variables (MS-CST): the parameters of the frequency and the severity distributions change according to a latent state variable only (no additional dependence with economic variables is assumed). This model is close to the one considered in Guillou et al. [2013]. These three alternatives are, in fact, special cases of our more general MS-GAMLSS model introduced in Section 2. For the pure GAMLSS model, the shape of the functional forms for the different nonparametric components can be found in Appendix A (Figure 12). The estimated functional forms seem similar to the ones obtained with the MS-GAMLSS model in the first state. The values of the constants for these models can be found in Table 4.

We plot the estimations of  $Q_{0.999}(L_t; X_t, S_t)$  obtained with the various models (Figure 8). The constant model seems to be the worst: it does not detect structural changes, and leads to VaR estimations that appear to be too small during the first period, and too large in the second period. The MS-CST model detects the same structural changes as the MS-GAMLSS model, but it cannot adapt well to changes inside a particular regime (especially during the second quarter of 2009). Last, the GAMLSS models seems to better stick to the data than the two other models, but we observe a breach of the quantile for the first quarter of 2006. Thus, a simple GAMLSS model cannot handle structural changes as the one detected with the MS-GAMLSS models. In terms of AIC, the constant and GAMLSS models are less good than the MS models. The MS-GAMLSS model has the best AIC, but it is close to the AIC of the MS-CST model. However, in term of breach of the 99.9% quantile (i.e. the number of total loss realizations that are larger than the computed 99.9% quantile), the MS-GAMLSS model provides the best results among all models. This is surely an important criterion from a regulatory perspective, since no breach is expected from the regulators.

For an informal assessment of the global fit of the considered models, we draw conditional pseudo-residuals QQ-plots for the severity distribution. By *conditional*, we mean here pseudo-residuals conditional on the decoded state. More precisely, we use the Viterbi algorithm and the estimated regression function to assign at each time and each observation an estimated state  $\hat{s}_t$ , as well as estimated severity parameters  $\hat{\gamma}^{(\hat{s}_t)}(x_t)$  and  $\hat{\sigma}^{(\hat{s}_t)}(x_t)$ . Then, we compute  $e_{i,t} = (1/\hat{\gamma}^{(\hat{s}_t)}(x_t)) \log(1 + \hat{\gamma}^{(\hat{s}_t)}(x_t)(z_{i,t} - \tau)/\hat{\sigma}^{(\hat{s}_t)}(x_t))$ ,  $\forall t, i$ , which should be approximately i.i.d. realizations of a standardized exponential distribution [Chavez-Demoulin et al., 2016]. Figure 9 indicates a relative good fit for all considered models, except far in the tail. This is presumably due to a few losses (and the relative small size of our sample), for which none

of the models seem to estimate the distribution correctly.

Overall, the MS-GAMLSS model seems provide the best fit for the data (as we do not observe a breach of the quantile) and to adequately describe the historical macroeconomic scenario: in 2008, we faced an increased uncertainty on the financial markets, and on the solvency of the whole banking system. In this situation, extremely large losses become more likely (as the financial system could completely collapse at that time) and the shape parameter of the total loss distribution increases despite drastic modifications of the risk management. The VIX seems to capture this uncertainty. At the end of 2010 and 2011, a similar high uncertainty takes place during the sovereign crisis and raises (once again) the question of the survival of the global banking system. Hence, a model allowing for structural changes and dependence with economic covariates is able to handle this kind of scenario.

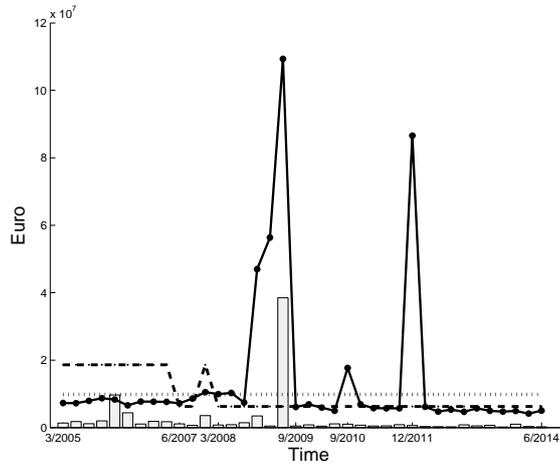


Figure 8:  $\hat{Q}_{0.999}(L_t; X_t, S_t)$  over time obtained with the alternative models: constant (dotted), MS-CST (dashed), GAMLSS (solid). Grey bar: total loss.

Final results	$\hat{\pi}_{11}$	$\hat{\pi}_{22}$	$\hat{\beta}_0^{(1)}(\sigma)$	$\hat{\beta}_0^{(2)}(\sigma)$	$\hat{\beta}_0^{(1)}(\gamma)$	$\hat{\beta}_0^{(2)}(\gamma)$	$\hat{\beta}_0^{(1)}(\lambda)$	$\hat{\beta}_0^{(1)}(\lambda)$	-LL	AIC
CST	-	-	10.11	-	-0.66	-	3.07	-	9674.8	19354
GAMLSS	-	-	9.94	-	-0.69	-	2.9	-	9631.4	19283
MS-CST	0.94	0.88	10.08	10.07	-0.74	-0.55	2.81	3.55	9631.4	19279
MS-GAMLSS	0.95	0.93	10	9.81	-0.95	-0.70	2.8	3.47	9619.2	19272

Table 4: Estimations of the constant parameters for the four models. CST stands for the unconditional constant model, whereas MS-CST stands for the MS model with no additional explanatory variables.

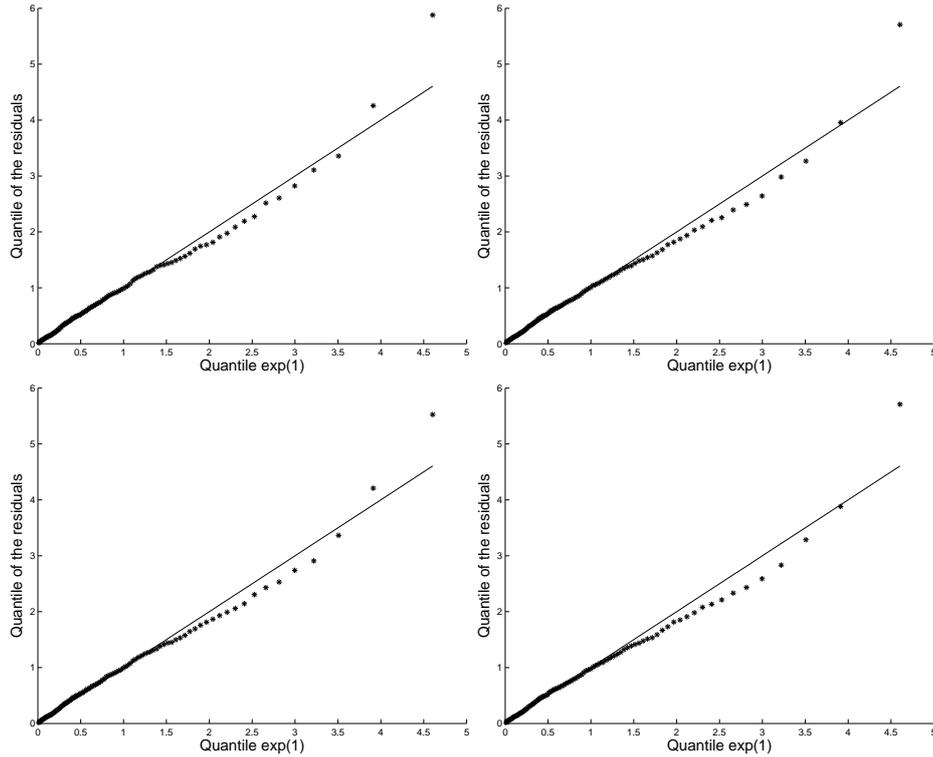


Figure 9: QQ-plots for the pseudo-residuals of the four models. Top, from left to right: MS-GAMLSS and MS-CST models. Bottom, from left to right: GAMLSS and CST models.

### 4.3 Economic implications and limitations

Economically speaking, a bank that uses a MS-GAMLSS model to establish its requested capital is able to set aside more capital in a high-risk period, and less capital in a low-risk period. Hence, the model facilitates a better allocation of financial resources. Also, it meets the expectations of the regulators for a better adequateness between the risk level and the requested capital. In particular, in this application, the MS-GAMLSS model is the only one that avoids a breach of the estimated 99.9% quantile. In contrast, the other models suffer at least a breach of the quantile which indicates structural issues, since the probability of a breach at a particular quarter, conditional on the current economic conditions, is of  $1/1000$ . If such an event happens, either the model suffers from a misspecification or we are extremely unlucky. Lastly, since we use lagged values for the explanatory variables in our application, it implies that we are able to provide good (in-sample) one-step ahead predictions of the total loss distribution.

Nevertheless, the use of covariates and switching regimes has the drawback of increasing the variability of the capital requirements over time. In our application, we use the VIX index, which is based on the derivative market and prone to brutal variations. With the MS-GAMLSS model, the required capital would be multiplied by 10 in a single period. Financially speaking, it might not be possible for an ALM

department to provision 9 million a quarter, then 90 million the next quarter, and 350 million more 6 months after (recall, however, that the losses have been multiplied by an unknown scaling factor for anonymity reasons, implying that the given numbers are in reality different). To reduce this variability, a solution would be to set an upper threshold for the VIX, limiting the consequences of unexpected large bursts. However, such a constraint is hardly justifiable from a theoretical perspective.

Moreover, in term of managerial actions, little can be done with explanatory variables not internal to the bank (e.g. the bank has no power -alone - on the VIX level, or on the unemployment rate). Our results suggest an important dependence with market volatility, but it is quite likely that the VIX itself is not a determinant of the severity distribution. An explanatory variable based on the interaction between the market conditions and the total exposure could be more suited and more useful in term of risk management. Thus, MS-GAMLSS models seem to be interesting tools to study the formation process of operational losses, but should be carefully justified from a theoretical perspective (especially regarding the set of covariates) if used to make predictions.

An additional issue might arise for the prediction and decoding of the states. Here, we consider the simplistic case of a stationary two-state model, but the transition probabilities might vary over time, or more states might exist. A misspecification of these aspects of the model could be damageable in term of predictive ability. Moreover, if we make a mistake in the prediction of the state, we could wrongly assume a low-risk state instead of a high-risk state and decide to set aside less capital than needed. Once again, it is a danger of MS models, and it needs to be analyzed by risk managers willing to use them. Lastly, notice that sometimes, the initialization of the estimation procedure might be problematic. Here, we use an initial solution that allows the minimization algorithm to converge to a good solution after a reasonable number of iterations (we test multiple sets of starting parameters). We came across several regions of the likelihood function where it does not happen, and where the final solution leads to different decoded sequences of states. Therefore, it seems quite important to select carefully this initial solution.

Lastly, this empirical study suggests that the distribution of operational losses is, indeed, non-stationary and varies with the economic conditions, since simpler models cannot avoid breaches of the 99.9% quantile. Therefore, this feature should be taken into considerations by the regulators. Otherwise we face the risk to set the requested capital to some kind of average value. Such a strategy would be clearly inefficient, because it would lead to set aside too much capital in low-risk periods (e.g. when market conditions are good and the risk management is effective) and too few capital in high risk periods (e.g. when the market conditions are unstable and the internal structure of the bank generates risks).

## 5 Conclusion

In this paper, we studied the particular case of a compound Poisson-GPD process, a model frequently used in insurance and finance for the behavior of random sums over time. We focused on a particular extension of this model, in the context of Markov-switching generalized additive models for location, scale and shape (MS-GAMLSS). The interest of this extension lays in the fact that it can properly take into account a dependence structure with covariates, as well as structural changes arising over time. We detailed how to estimate the parameters of this model, using a direct maximisation of the log-likelihood function. In a simulation study, we showed that even if the length of the time series is as short as 50 time periods, we can still correctly estimate the parameters of the model when the average number of events occurring in a period is at least 12. Subsequently, we applied our MS-GAMLSS model to the modeling of operational losses in the banking industry. We considered a novel database of 817 fraud losses, provided by the bank UniCredit. We studied the conditional distribution of the total losses per quarter, over a 10-year period. As explanatory variable for the frequency parameter, we used lagged values of the percentage of revenue coming from fees (PRF), whereas for the scale parameter we used lagged values of the Italian unemployment rate. For the shape parameter, we used lagged values of the VIX index.

We found increasing relationships between the PRF, the VIX index and the related parameters, whereas an increase in the unemployment rate up to 10% was linked with an decrease in the scale parameter, then with an increase beyond that point. Using the Viterbi algorithm, we found three shifts between states, two of them in 2007, and the third one on the 1<sup>st</sup> quarter of 2008. We suggested that the regimes can be labelled as non-crisis and in-crisis regimes, and are mostly characterized by different levels of the parameters. Especially, in the regime assigned to the period 2008-2014 (the in-crisis regime), we observed a strong dependence of the severity distribution with the market volatility (approximated by the VIX). Due to this increased uncertainty in 2009, the probability of an extreme event was found to increase drastically in 2009. However, when the market conditions go back to normal, the probability of extreme losses was found to be lower in this regime. We conjectured that these finding might be due either to structural changes undergone by UniCredit in 2007 following its merger with Capitalia, or by changes in its risk management, following the financial crisis and the enforcement of Basel II rules in 2008.

From a regulatory perspective, and compared to simpler models (without switching regimes or explanatory variables), the MS-GAMLSS model proved to be better since it was the only model where we did not observe a total loss larger than the 99.9% quantile. These results indicate that a Markov-switching

structure and a dependence with covariates help to better take into account the non-stationarity of the distribution of operational losses. In addition, this model allows to provision less capital during less risky periods, with the consequence that financial resources can be allocated more adequately. However, it comes to the cost of potentially big variations of the requested capital from one period to another.

Several extensions and improvements of the present work could be considered. First, we focused on a complex but single case of the MS-GAMLSS framework (namely the compound Poisson-GPD process). It would be quite easy to consider other distributions of the GAMLSS family that could be useful in other applications (e.g. for stock returns models with generalized hyperbolic distributions, or loss given default models in credit risk applications with Beta distributions). Second, we only considered a two-state case, motivated by the shortness of the time series in our application. Considering additional states (if sufficient data are available) might help to capture different features of the data. Third, we used a simplistic selection procedure of the smoothing parameter, searching over a grid of candidate vectors in  $\mathbb{R}_{\geq 0}^{M \times (J_\lambda + J_\sigma + J_\gamma)}$ . However, this search becomes rapidly unfeasible when the number of covariates or the number of states is high. A solution would be to use a shrinking penalizing method, similar to ridge or LASSO regression in the parametric context [Fahrmeir et al., 2013]. Adapted procedures have been proposed in the GAM case [Marra and Wood, 2011] and in the GAMLSS case [Mayr et al., 2012] but it does not exist yet in our context. This question should be undoubtedly the focus of future investigations. Fourth, we did not consider extensively issues related to confidence intervals and model selection. We restrict ourselves to the use of the AIC and bootstrap procedures with fixed covariates. Instead, one should consider e.g. extensive bootstrap procedures combined with cross-validation. Regarding the model selection, we believe that cross-validation or out-of-sample predictive ability should be the preferred selection procedures, and not information criteria which are often too lenient on overcomplex models. Due to the limited length of our time series we did not perform out-of-sample predictions but such prediction exercises should assess the practical usefulness of the proposed model. Lastly, we made the assumption that a single Markov chain was driving the behavior of the conditional frequency and severity distributions. However, assuming different Markov chains in equations (9) to (11) could ease the estimation of the model, because we could estimate both switching regimes independently. We did not consider this alternative but this is clearly a possible extension of the present work.

## Acknowledgements

J. Hambuckers acknowledges the support of the Research Training Group 1644 *Scaling Problems in Statistics* funded by the German Research Foundation (DFG). The authors acknowledge Dr. F. Piacenza

for providing us with the data.

## 6 Appendix A

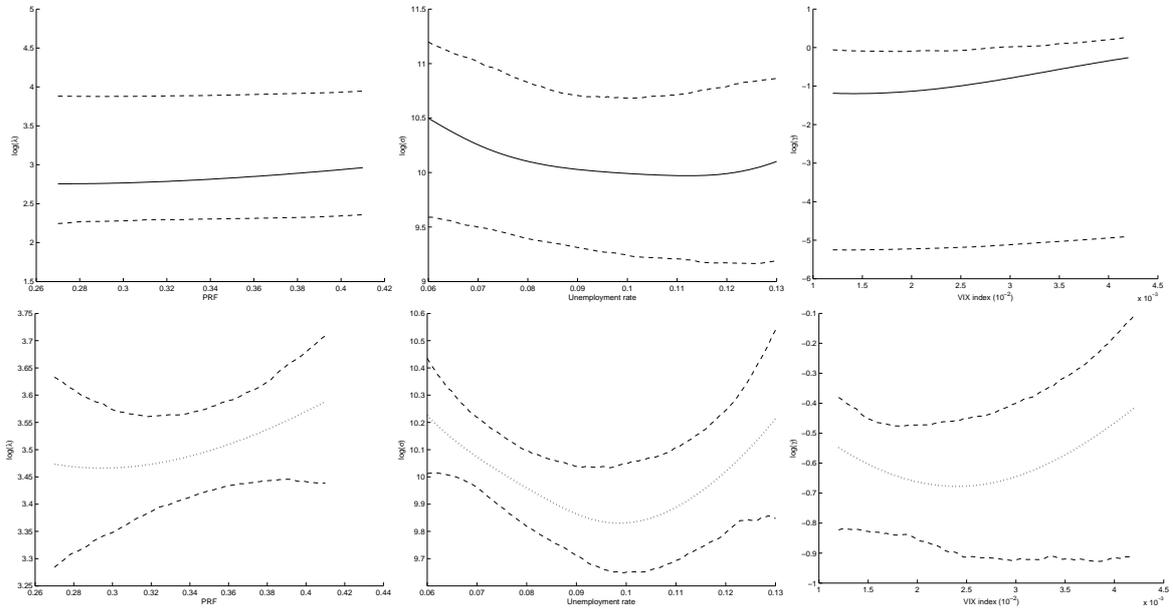


Figure 10: Estimated functional forms related to the three covariates considered, obtained with the MS-GAMLSS model (solid line: state 1, dotted: state 2). Dashed: 95% confidence intervals. X-axis: value of the covariate. Y-axis:  $\log(\hat{\theta}^{(j)}(X))$ ,  $\theta \in \{\lambda, \sigma, \gamma\}$  and  $j = 1, 2$ .

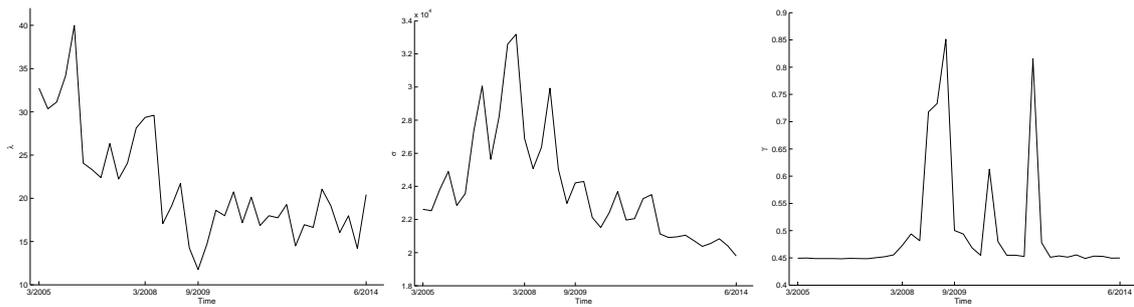


Figure 11: Estimated frequency, scale and shape parameters with the GAMLSS model, over time. Cross-validated smoothing parameters are respectively 2, 25 and 8.

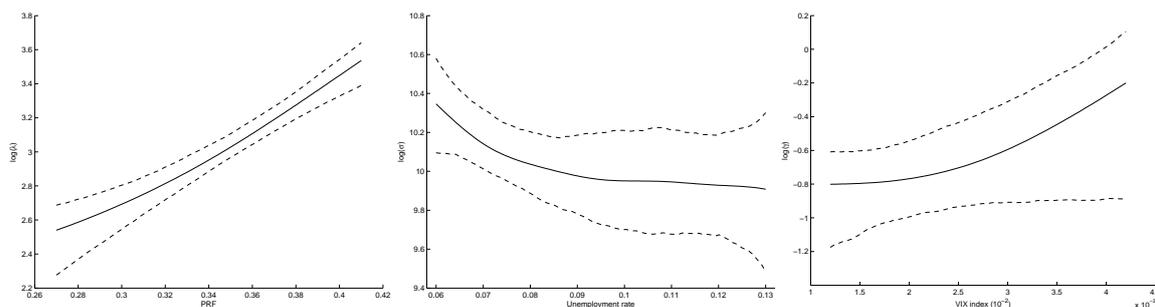


Figure 12: Estimated functional forms related to the three covariates considered, obtained with the GAMLSS model (solid line). Dashed: 95% confidence intervals. X-axis: value of the covariate. Y-axis:  $\log(\hat{\theta}^{(j)}(X))$ ,  $\theta \in \{\lambda, \sigma, \gamma\}$  and  $j = 1, 2$ .

## References

- T. Adam, R. Langrock, and T. Kneib. Flexible latent-state time series regression models for location, scale and shape. *preprint*, 2016.
- A. Ang and A. Timmermann. Regime changes and financial markets. *Annual Review of Financial Economics*, 4(1):313–337, 2012.
- Basel Committee on Banking Supervision BCBS. Basel II: international convergence of capital measurement and capital standards. A revised framework. Technical Report 251, Basel, Switzerland, 2004.
- J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, D. De Waal, and C. Ferro. *Statistics of extremes: Theory and applications*. 2005. doi: 10.1002/0470012382.
- A Chapelle, Y. Crama, G. Hübner, and J.-P. Peters. Practical methods for measuring and managing operational risk in the financial sector: a clinical study. *Journal of Banking & Finance*, 32:1049–1061, 2008.
- V. Chavez-Demoulin, P. Embrechts, and S. Sardy. Extreme-quantile tracking for financial time series. *Journal of Econometrics*, 181(1):44–52, 2014.
- V. Chavez-Demoulin, P. Embrechts, and M. Hofert. An extreme value approach for modeling Operational Risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2016.
- A. Chernobai and Y. Yildirim. The dynamics of operational loss clustering. *Journal of Banking & Finance*, 32(12):2655–2666, 2008.

- A. Chernobai, P. Jorion, and F. Yu. The derminants of Operational Risk in U.S. financial institutions. *Journal of Financial and Quantitative Analysis*, 46(8):1683–1725, 2011.
- E. Cope, M. Piche, and J. Walter. Macroenvironmental determinants of operational loss severity. *Journal of Banking & Finance*, 36(5):1362–1380, 2012.
- K. Dutta and J. Perry. A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital. *Research Review*, Jul-Dec(6):11–14, 2006.
- P. Eilers and B.D. Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2): 89–102, 1996.
- P. Embrechts, C. Klupperlberg, and T. Mikosch. *Modelling extremal events for insurance and finance*, volume 648. Springer - Verlag, Berlin, 1997.
- L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression : models, methods and applications*. 2013. doi: 10.1007/978-3-642-34333-9.
- T. Goodwin. Business-Cycle Analysis Model Markov-Switching. *Journal of Business & Economic Statistics*, 11(3):331–339, 1993.
- R. Gray. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.
- A. Guillou, S. Loisel, and G. Stupfler. Estimation of the parameters of a Markov-modulated loss process in insurance. *Insurance: Mathematics and Economics*, 53(2):388–404, 2013.
- A. Guillou, S. Loisel, and G. Stupfler. Estimating the parameters of a seasonal Markov-modulated Poisson process. *Statistical Methodology*, 26:103–123, 2015.
- J. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384, 1989.
- C. Hurvich, J. Simonoff, and C. Tsai. Information Criterion Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293, 2012.
- F. Klaassen. Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics*, 27(2):363, 2002.

- C. Lamoureux and W. Lastrapes. Persistence in variance, structural change, and the GARCH model. *Journal of Business & Economic Statistics*, 8(2):225–234, 1990.
- R. Langrock, T. Kneib, R. Glennie, and T. Michélot. Markov-switching generalized additive models. *Statistics and Computing*, 2015.
- G. Marra and S. Wood. Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55(7):2372–2387, 2011.
- A. Mayr, N. Fenske, B. Hofner, and T. Kneib. Generalized additive models for location, scale and shape for high dimensional data a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, pages 403–427, 2012.
- M. Moscadelli. The modelling of operational risk: experience with the analysis collected by the Basel Committee. Technical report, Roma, 2004.
- M. Pesaran, D. Pettenuzzo, and A. Timmermann. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084, 2006.
- R. Rigby, D. Stasinopoulos, and P. Lane. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.
- F. Serinaldi. Distributional modeling and short-term forecasting of electricity prices by Generalized Additive Models for Location, Scale and Shape. *Energy Economics*, 33(6):1216–1226, 2011.
- C. Sims and T. Zha. Were There Regime Switches in U.S. Monetary Policy ? *The American Economic Review*, 96(1):54–81, 2006.
- A. Sohn, N. Klein, and T. Kneib. A Semiparametric Analysis of Conditional Income Distributions. *Schmollers Jahrbuch*, 135(1):13–22, 2015.
- A. Soprano, B. Crielaard, F. Piacenza, and D. Ruspantini. *Measuring Operational and Reputational Risk: A Practitioner’s Approach*. Wiley, 2009.
- D. Stasinopoulos and R. Rigby. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007.
- T. Wang and C. Hsu. Board composition and operational risk events of financial institutions. *Journal of Banking & Finance*, 37(6):2042–2051, 2013.

W. Zucchini, I. MacDonald, and R. Langrock. *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition*. Boca Raton, 2016.