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

J. Hambuckers[1,†], T. Kneib[1], R. Langrock[2] and A. Silbersdorff[1]

Abstract

This paper is concerned with modeling the behavior of random sums over time. Such models are particularly useful to describe the dynamics of operational losses, and to correctly estimate tail-related risk indicators. However, time-varying dependence structures make it a difficult task. To tackle these issues, we formulate a new Markov-switching generalized additive compound process combining Poisson and generalized Pareto distributions. This flexible model takes into account two important features: on the one hand, we allow all parameters of the compound loss distribution to depend on economic covariates in a flexible way. On the other hand, we allow this dependence to vary over time, via a hidden state process. A simulation study indicates that, even in the case of a short time series, this model is easily and well estimated with a standard maximum likelihood procedure. Relying on this approach, we analyze a novel dataset of 819 losses resulting from frauds at 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, a feature particularly desirable for banking regulators.

Keywords: GAMLSS, distributional regression, generalized Pareto distribution, hidden Markov model, operational losses, compound Poisson process.

[1] Georg-August-Universität Göttingen, Faculty of Economic Sciences, Chair of Statistics, Humboldtallee 3, 37073 Göttingen, Germany.

[2] Bielefeld University, Department of Business Administration and Economics, Statistics and Data Analysis Group, Universitätsstrasse 25, 33615, Bielefeld, Germany.

[†] Corresponding author: jhambuc@uni-goettingen.de.

1 Introduction

In the banking and actuarial literature, there has been great interest in modeling the time-varying distribution of the total operational loss (or the total claim) suffered by a financial institution. Especially with the recent financial crises, it has been emphasized that the tail of this distribution needs to reflect accurately the probabilities of extreme losses, since those probabilities are used for regulatory and risk management purposes. To achieve this goal, the total loss over time is usually modeled as a random sum. However, accounting for complex and possibly time-varying dependence structures with economic covariates has been neglected in existing model formulations. A corresponding model misspecification inevitably leads to errors in the computation of risk indicators, and can bias the risk management process.

In this paper, we propose a flexible random sum model that specifically takes into account these dependence structures. As a particular application, we study the case of quarterly operational fraud losses at the Italian bank UniCredit, over a 10-year period. Our quantity of interest, the total operational loss for period t, is given by

$$L_t = \sum_{i=1}^{N_t} Y_{i,t}, \tag{1}$$

for $t = 1, 2, \ldots$, where N_t is the number of events occurring over the period t, and $Y_{i,t}$ is the severity of the i^{th} event occurring during period t [Embrechts et al., 1997, p.9]. The number of events is therefore a number of losses (i.e. individual operational events), whose intensity is measured in monetary units. Standard models usually make the following assumptions on the behavior of N_t and $Y_{i,t}$:

$$N_t \overset{iid}{\sim} \text{Poisson}(\lambda), \tag{2}$$

$$Y_{i,t} \overset{iid}{\sim} \text{GPD}(\gamma, \sigma), \tag{3}$$

with $\lambda, \sigma > 0, \gamma \geq 0$ and where realizations of N_t and $Y_{i,t}$ are independent for all t and i. Here λ is the frequency parameter, while $\text{GPD}(\gamma, \sigma)$ stands for the generalized Pareto cumulative distribution function, given by

$$GPD(y; \gamma, \sigma) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{-1/\gamma}, & \gamma \neq 0 \\ 1 - \mathrm{e}^{-y/\sigma}, & \gamma = 0 \end{cases} \tag{4}$$

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

The process L_t 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, the GPD is used 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)–(3) 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)–(3) — the following relationship holds:

$$g(\theta) = \eta_\theta = \beta_{\theta,0} + \sum_{j=1}^{J} h_{\theta,j}(X_j^\theta), \tag{5}$$

where $g(\cdot)$ is a monotonic link function, $\beta_{\theta,0}$ a constant, X^θ is a vector of predictors for parameter θ, X_j^θ, $j = 1, \ldots, J$, is the j^{th} predictor and $h_{\theta,j}$ is some function of it (possibly parametric or nonparametric). This general setup is known as the 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 and Stasinopoulos, 2005]. It allows to consider linear and nonlinear dependencies of covariates with the parameters of a distribution function. This approach has been successfully used in different applications. Among others, Rigby and Stasinopoulos [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 whereas Serinaldi [2011] models the price of electricity. 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 equations (1)–(5) focuses on data that have a time series structure. In case of parameter instabilities (as it is often the case in

economics and finance; see Hamilton [1989], Lamoureux and Lastrapes [1990] and Ang and Timmermann [2012]), equation (5) 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 as Markov-switching (MS) or regime-switching models [Hamilton, 1989, Zucchini et al., 2016]. They have been successfully applied 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 allow 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. [2017] 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 dependencies 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 an 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. [2017] 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 improve 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 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 rate 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, an 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 an MS component.

These considerations motivate the formulation of a model of type (1)–(3) where the parameters λ, γ and σ depend (nonparametrically) on covariates, and where the functional relationships between covariates and parameters vary according to a (hidden) regime-switching process. Our methodological contribution draws a bridge between the 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 model considered 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 819 large losses resulting from external frauds in the Italian bank UniCredit over a time period of 38 quarters. Our goal is to shed some lights on the nature of the dependence between some economic factors and the distribution of the total operational loss, and to uncover the latent state process during that period. 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 Chicago Board Options Exchange volatility index (VIX) 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, an 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, 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

$$\mathbb{P}(Z_t < z | X_t, S_t) = F(z; \theta_t^{(S_t)}(X_t)), \tag{6}$$

for $t = 1, 2, \ldots$, where F is the (parametric) cumulative distribution function of Z_t with vector of associated parameters $\theta_t^{(S_t)}(X_t)$ of size K at time t. The realized time series of length T of the response variable, the vector of covariates and the realized sequence of the latent states are denoted by $\{z_t\}_{t=1,\ldots,T}$, $\{x_t\}_{t=1,\ldots,T}$ and $\{s_t\}_{t=1,\ldots,T}$, respectively. Additionally, we suppose the following structure for the parameters: for $k = 1, \ldots, K$, $t = 1, \ldots, T$ and $S_t = 1, \ldots, M$, we have

$$g_k(\theta_{k,t}^{(S_t)}(X_t^{\theta_k})) = \beta_{\theta_k,0}^{(S_t)} + \sum_{j=1}^{J_{\theta_k}} h_{\theta_k,j}^{(S_t)}(X_{j,t}^{\theta_k}), \tag{7}$$

where g_k is a monotonic link function for the k^{th} parameter $\theta_{k,t}^{(S_t)}$ ensuring proper restrictions, $\beta_{\theta_k,0}^{(S_t)}$ a constant parameter, $X_t^{\theta_k}$ is a vector of covariates at time t for the k^{th} parameter, $X_{j,t}^{\theta_k}$, $j = 1, \ldots, J_{\theta_k}$, is the j^{th} covariate at time t whereas $h_{\theta_k,j}^{(S_t)}$ is an unspecified smooth function that links this variable to $\theta_{k,t}^{(S_t)}$. We assume that each function $h_{\theta_k,j}^{(S_t)}$ can be approximated with a finite linear combination of B-spline basis functions $B_1, \ldots, B_q$

[Eilers and Marx, 1996, Langrock et al., 2017]:

$$h_{\theta_k,j}^{(S_t)}(X_{j,t}^{\theta_k}) = \sum_{i=1}^{q} w_{k,j,i}^{(S_t)} B_i(X_{j,t}^{\theta_k}). \tag{8}$$

As in Langrock et al. [2017] and Chavez-Demoulin et al. [2016], we consider cubic B-splines. More details are given in the next section. This structure allows considering complex nonlinear effects of the covariates on the distribution parameters (notice that a simplified version of this model is easily obtained by assuming a linear structure instead of smooth functions $h(\cdot)$).

In the case of Z_t following a Poisson distribution, model (6) is fully characterized by a single parameter ($K = 1$, $\theta = \lambda$) and equation (7) characterizes the conditional expectation of Z_t, meaning that the model can be reduced to the MS-GAM framework of Langrock et al. [2017]. However, if Z_t is of the type GPD, then $K = 2$ and equation (7) can be used to describe the conditional joint behavior of γ and σ. Now, if $Z_t = L_t$ and is given by equation (1), with t denoting a time period and not a point in time, we have

$$N_t | X_t, S_t \sim \text{Poisson}(\lambda_t^{(S_t)}(X_t^{\lambda})), \tag{9}$$

$$Y_{i,t} | X_t, S_t \sim \text{GPD}(\gamma_t^{(S_t)}(X_t^{\gamma}), \sigma_t^{(S_t)}(X_t^{\sigma})). \tag{10}$$

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

$$\log(\lambda_t^{(S_t)}(X_t^{\lambda})) = \beta_{\lambda,0}^{(S_t)} + \sum_{j=1}^{J_\lambda} h_{\lambda,j}^{(S_t)}(X_{j,t}^{\lambda}), \tag{11}$$

$$\log(\gamma_t^{(S_t)}(X_t^{\gamma})) = \beta_{\gamma,0}^{(S_t)} + \sum_{j=1}^{J_\gamma} h_{\gamma,j}^{(S_t)}(X_{j,t}^{\gamma}), \tag{12}$$

$$\log(\sigma_t^{(S_t)}(X_t^{\sigma})) = \beta_{\sigma,0}^{(S_t)} + \sum_{j=1}^{J_\sigma} h_{\sigma,j}^{(S_t)}(X_{j,t}^{\sigma}). \tag{13}$$

Equations (11)–(13) imply that all losses a drawn from the same distribution at time t. However, without loss of generality it is straightforward to include an additional dependence of the distribution parameter with the loss itself (i.e. a dependence in i) in equations (12) and (13). It would allow to account for loss-specific explanatory variables that might be different across losses occurring during the same time period. For example, in the context of operational losses, losses usually belong to different categories (called event types) related to their physical process (e.g. fraud, damage to physical assets, link with derivative products, etc.). This additional dependence could be included towards this mechanism. 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, $Y_{i,t}$, conditional on S_t and X_t, will be denoted by $f_N(n_t; \lambda_t^{(s_t)})$, $f_Y(y_{i,t}; \gamma_t^{(s_t)}, \sigma_t^{(s_t)})$, whereas the likelihood of $L_t = l_t$ is denoted $f_L(l_t; \lambda_t^{(s_t)}, \gamma_t^{(s_t)}, \sigma_t^{(s_t)})$, 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:

$$f_L(l_t; \lambda_t^{(s_t)}, \gamma_t^{(s_t)}, \sigma_t^{(s_t)}) = f_N(n_t; \lambda_t^{(s_t)}) \prod_{i=1}^{n_t} f_Y(y_{i,t}; \gamma_t^{(s_t)}, \sigma_t^{(s_t)}). \tag{14}$$

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, $\forall i, j = 1, \ldots, M$ 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 $\pi_{ij} = \mathbb{P}(S_t = j | S_{t-1} = i)$, $i, j = 1, \ldots, M$. The initial state probabilities are summarized in a vector δ, where $\delta_i = \mathbb{P}(S_1 = i)$, $i = 1, \ldots, M$. The stationary distribution implied by the estimated tpm gives the proportions of time the Markov chain spends in the different states. In several applications [e.g. Ang and Timmermann, 2012, Guillou et al., 2013], depending on the number of states assumed and the nature of the transition process, the diagonal elements are estimated to be close to one, but may also be much smaller (examples can be found in Langrock [2012] and King and Langrock [2016]). Diagonal elements close to one would usually indicate strongly persistent states and few state switches, but can also result from a misspecification of the underlying model or of the Markovian assumption, whereas small diagonal elements usually indicate less persistent or even transitory states (e.g. such that account only for isolated extreme events). The choice of the number of state is discussed in Section 2.3.

In our empirical study, we make the assumption that there is a single Markov chain that drives the conditional frequency and the conditional severity processes. It implies that the latent factors driving both processes are similar.

2.2 Estimation of the model

To calculate the likelihood, we use the forward recursion procedure described in Langrock et al. [2017], adapted for the special case of our model described by equations (1) and (9) to (13). We then maximize the likelihood *directly*, instead of using the expectation-maximization (EM) algorithm, the implementation of which is technically more involved than direct numerical likelihood maximization [MacDonald, 2014]. More specifically, we

define the vector-valued forward variable,

$$\alpha_t = (\alpha_t(1), \ldots, \alpha_t(M)), t = 1, \ldots, T, \tag{15}$$

where $\alpha_t(j) = f_L(l_1, \ldots, l_t, s_t = j | x_1, \ldots, x_t)$, for $j = 1, \ldots, M$. These forward variables are, at each time $t = 1, \ldots, T$ and for each state $j = 1, \ldots, M$, 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. [2017] can be applied:

$$\alpha_1 = \delta Q(l_1), \tag{16}$$

$$\alpha_t = \alpha_{t-1} \Gamma Q(l_t) \, (t = 2, \ldots, T), \tag{17}$$

where $Q(l_t)$ =diag$(f_L(l_t; \lambda_t^{(1)}, \gamma_t^{(1)}, \sigma_t^{(1)}), \ldots, f_L(l_t; \lambda_t^{(M)}, \gamma_t^{(M)}, \sigma_t^{(M)}))$ and δ is the vector of initial state probabilities (see equation (22) for more details). This recursion follows from

$$\alpha_t(j) = \sum_{i=1}^{M} \alpha_{t-1}(i) \pi_{ij} f_L(l_t; \lambda_t^{(j)}, \gamma_t^{(j)}, \sigma_t^{(j)}). \tag{18}$$

Additional details can be found in Zucchini et al. [2016]. Eventually, the likelihood function of model (9)-(14) is given by

$$\mathcal{L}(\Theta) = \sum_{i=1}^{M} \alpha_T(i) = \delta Q(l_1) \Gamma Q(l_2) \ldots \Gamma Q(l_T) \mathbf{1}, \tag{19}$$

where $\mathbf{1} \in \mathbb{R}^M$ is a column vector of ones and Θ is the set of all parameters of the model. In our case of interest, this expression is easily obtained from the conditional densities given by equation (14). In the fully parametric case, an estimator $\hat{\Theta}$ of Θ is easily obtained by maximizing equation (19) with respect to the parameters. Besides, if we consider several covariates nonparametrically, the functions $h_{\theta_k,j}^{(S_t)}$, $k = 1, \ldots, K$ and $j = 1, \ldots, J_{\theta_k}$ can be expressed as a finite linear combination of B-spline basis functions as in equation (8). 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 $q = 11$). 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 (8)[1]. The penalizing term is given by equation (7) in Langrock et al. [2017]. For our model, it becomes

$$D = \sum_{k=1}^{3} \sum_{m=1}^{M} \sum_{j=1}^{J_{\theta_k}} \frac{\kappa_{k,m,j}}{2} \sum_{i=3}^{q} (\Delta^2 w_{k,j,i}^{(m)})^2, \tag{20}$$

[1]This approximation is particularly convenient because it also ensures that a linear fit remains unpenalized.

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

$$\mathcal{L}_{pen.}(\Theta) = \log(\mathcal{L}(\Theta)) - \sum_{k=1}^{3}\sum_{m=1}^{M}\sum_{j=1}^{J_{\theta_k}} \frac{\kappa_{k,m,j}}{2} \sum_{i=3}^{q} (\Delta^2 w^{(m)}_{k,j,i})^2, \tag{21}$$

for a particular vector of smoothing parameters κ.

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 in equation (21) instead of the likelihood itself. However, because we are dealing with a product of matrices, we cannot simply apply a log-transform to equation (19). Instead, we use a simple scaling algorithm, as described in detail 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 (14) 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 (19) and (21) can be challenging numerical tasks. It is important to test several starting values to avoid local minima. We also need to take into account constraints on the parameters (especially positivity constraints and row sums 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[2]. To avoid optimizing under constraints, we follow Langrock et al. [2017] and use 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 $w^{(S_t)}_{k,j,i}$ 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 $w^{(S_t)}_{k,j,\frac{q+1}{2}} = 0$, for $j = 1, \ldots, J_{\theta_k}$, $k = 1, \ldots, 3$ and $S_t = 1, \ldots, M$. Furthermore, regarding the transition probabilities, we make the assumption of a stationary transition process, implying that δ is the solution to the linear system

$$\delta = \delta\Gamma, \tag{22}$$

subject to $\sum_{i=1}^{M} \delta_i = 1$. This standard assumption is discussed, among others, in Zucchini et al. [2016].

[2]In the present paper, we rely on the *fminunc* function of the Optimization Toolbox in MatLab, using the quasi-Newton *lm-line-search* algorithm. This algorithm is an implementation of the one presented in Fletcher [1987]. See also Nocedal and Wright [2006] for more details. Gradient and Hessian updates are based on finite difference procedures.

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 κ and a particular subset of covariates) is given by

$$AIC(\hat{\Theta}, \kappa) = -2\log(\mathcal{L}(\hat{\Theta})) + 2 \cdot df(\kappa) \tag{23}$$

where $df(\kappa)$ is the effective degrees of freedom of the final model obtained with smoothing parameter κ 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 is 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. [2017] argue that in practice this choice is often 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 instead. They notice that it is not uncommon to fix the number of regimes at some value, 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. Information criteria (AIC, SBIC, Hannan-Quinn criterion, or cross-validated likelihood) can be of some help, but they tend to favour models with higher numbers of states than usually seem adequate [Pohle et al., 2017]. 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 check if an additional state does indeed substantially improve the

overall goodness-of-fit, or if such additional complexity is instead mostly driven by the extra states capturing artifacts in the time series (e.g. outliers). Corresponding analyses could consider autocorrelation function and quantile-quantile plots for the (pseudo-)residuals. It can also be helpful to examine the structure of the tpm of the various models fitted, and to check the relevance of the various states by considering the Viterbi-decoded state sequences.

Another important criterion that needs to be taken into account is the length of the time series under study. It should be noted that very short time series complicate inference on the underlying Markov process, in particular regarding the selection of the number of different states. This is particularly problematic if the states exhibit high persistence such that only very few switches between the states are used to inform the estimation of the tpm. 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 to base the inference on.

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:

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

This model is similar to the one described by equations (1) and (9) to (14). 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 λ, γ and σ 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 start by making the simplifying assumption that the behavior of γ is independent from any covariate (this assumption is later relaxed).

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 all distribution parameters. Therefore the selection is performed among 25 vectors with their components belonging to $\Lambda = \{0.5, 2, 8, 25, 50\}$[3]. 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[4]. For identifiability, the weight associated to the 6^{th} basis is set to 0. The regression equations have been defined to mimic the shapes of the estimated nonparametric functions encountered in our empirical study. Similarly, the constant and the tpm have been chosen to obtain a data generating process close to the estimations obtained in the empirical application. Regarding λ, the different choices of the constant parameter ensure that the number of losses per time period are in the range of the empirical study. Regarding γ, the constant parameters where chosen such that the generated values are below one, avoiding infinite mean situations usually

[3]Values in the grid were based on the ones used in Langrock et al. [2017] and a trial-and-error phase.

[4]Here, we use cubic splines that are twice continuously differentiable, ensuring a visually smooth fit. The number of basis function has been chosen to obtain a balance between flexibility and computational feasibility. Indeed, since we have 3 distribution parameters and two states, the use of an additional basis function leads to 6 more basis weights to estimate. 11 is in line with, e.g., the default number of knots in the popular *mgcv* R package.

discarded by practitioners (see, e.g., the discussions in Neslehova et al. [2006] and Chavez-Demoulin et al. [2016]). More generally, this simulation set-up is inspired by the one in Langrock et al. [2017].

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 distribution parameter and each state $j \in \{1, 2\}$:

$$\widehat{\text{MISE}}_{\nu_\theta^{(j)}} = \frac{1}{B} \sum_{i=1}^{B} \left(\int_0^1 \left(\hat{h}_\theta^{(j)}(x) - h_\theta^{(j)}(x) \right)^2 dx \right). \tag{24}$$

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 π_{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:

$$R = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(s_t = \hat{s}_t), \tag{25}$$

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. When the sample size increases, the MISE as well as the bias and variance of the estimated constant parameters decrease. Figures 1 and 2 show 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 4 shows the 95% coverage bands on the final parameter estimates for λ and σ, 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 to one 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 poor estimates of π_{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.

In addition, we consider several scenarios with γ depending on covariates, as in our empirical applications. Due to the highly intensive computing task involved, as well as the high flexibility of the model, we considered a reduced number of configurations, with $\beta_\lambda = 4$, $T = \{50, 100\}$ and either $x_t = x_t^\lambda = x_t^\sigma = x_t^\gamma$ (Scenario 1) or $x_t^\lambda \neq x_t^\sigma \neq x_t^\gamma$ (Scenario 2). Thus, the equation for γ_t becomes

$$\log(\gamma_t^{(s_t)}(X_t^\gamma)) = \beta_{\gamma,0}^{(s_t)} + h^{(s_t)}(X_t^\gamma).$$

Constant parameters are chosen as $\beta_{0,\gamma}^{(s_t)} = [-.65, -.8]$. The function $h(\cdot)$ is the same as in the first simulation set-up. Results are given in rows labelled 4^* in Tables 1 to 3. Conclusions are qualitatively alike. Especially, diagonal elements and classification rates are as good as in the first configuration. The functional forms related to γ are less well estimated, though, and the estimations exhibit a higher MISE. Nevertheless, when the length of the time series increases, the variability of the estimation seems to decrease at an acceptable speed (see Figure 5 and Tables 1 and 2).

Last, since we are eventually interested in estimating the 99.9% quantile of the total loss over time, one might want to look at the proportion of exceedances associated with these estimates. In the supplementary material of this paper, we show in a Monte Carlo study that the proportion of exceedances is close to the expected 0.1% when the number of events per time period is sufficiently large (i.e. around 25 observations). This is definitely a by-product of a good estimation of the true model. Results concerning out-of-sample coverage are also available in the supplementary material.

Sc. 1	$\beta_{\lambda,0}$	$\widehat{MISE}_{h_\lambda^{(1)}}$	$\widehat{MISE}_{h_\lambda^{(2)}}$	$\widehat{MISE}_{h_\sigma^{(1)}}$	$\widehat{MISE}_{h_\sigma^{(2)}}$	$\widehat{MISE}_{h_\gamma^{(1)}}$	$\widehat{MISE}_{h_\gamma^{(2)}}$
$T = 50$	2	0.159	0.184	0.240	0.223	-	-
		(0.310)	(0.280)	(0.311)	(0.274)	-	-
	3	0.067	0.099	0.106	0.112	-	-
		(0.092)	(0.115)	(0.148)	(0.155)	-	-
	4	0.044	0.061	0.053	0.056	-	-
		(0.053)	(0.066)	(0.074)	(0.051)	-	-
	4*	0.061	0.062	0.078	0.082	0.161	0.272
		(0.13)	(0.11)	(0.192)	(0.114)	(0.197)	(0.218)
$T = 100$	2	0.082	0.099	0.134	0.121	-	-
		(0.179)	(0.126)	(0.206)	(0.154)	-	-
	3	0.046	0.058	0.069	0.078	-	-
		(0.064)	(0.067)	(0.088)	(0.095)	-	-
	4	0.027	0.037	0.042	0.048	-	-
		(0.033)	(0.029)	(0.049)	(0.043)	-	-
	4*	0.026	0.021	0.033	0.036	0.105	0.136
		(0.035)	(0.026)	(0.052)	(0.043)	(0.12)	(0.103)

Sc. 1	$\beta_{\lambda,0}$	$\bar{\beta}_{\lambda,0}^{(1)}$	$\bar{\beta}_{\lambda,0}^{(2)}$	$\bar{\beta}_{\sigma,0}^{(1)}$	$\bar{\beta}_{\sigma,0}^{(2)}$	$\bar{\beta}_{\gamma,0}^{(1)}$	$\bar{\beta}_{\gamma,0}^{(2)}$
$T = 50$	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)
	4*	3.987	4.018	1.309	1.514	−0.791	−1.004
		(0.176)	(0.184)	(0.232)	(0.236)	(0.485)	(0.446)
$T = 100$	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)
	4*	3.978	4.013	1.284	1.547	−0.719	−0.886
		(0.148)	(0.116)	(0.16)	(0.153)	(0.288)	(0.295)

Table 1: Results of Scenario 1, with $x_t^\lambda = x_t^\sigma = x_t^\gamma$. The MISE is 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. 4* indicates the results when γ depends on x_t^γ as well. Standard errors over the 200 samples are in parentheses.

Sc. 2	$\beta_{\lambda,0}$	$\widehat{MISE}_{h_\lambda^{(1)}}$	$\widehat{MISE}_{h_\lambda^{(2)}}$	$\widehat{MISE}_{h_\sigma^{(1)}}$	$\widehat{MISE}_{h_\sigma^{(2)}}$	$\widehat{MISE}_{h_\gamma^{(1)}}$	$\widehat{MISE}_{h_\gamma^{(2)}}$
$T = 50$	2	0.1641	0.1563	0.1943	0.1954	-	-
		(0.231)	(0.223)	(0.238)	(0.275)	-	-
	3	0.094	0.103	0.107	0.118	-	-
		(0.186)	(0.145)	(0.186)	(0.176)	-	-
	4	0.036	0.0485	0.052	0.065	-	-
		(0.04)	(0.039)	(0.083)	(0.062)	-	-
	4*	0.072	0.109	0.063	0.094	0.139	0.207
		(0.176)	(0.173)	(0.087)	(0.163)	(0.166)	(0.262)
$T = 100$	2	0.093	0.112	0.122	0.126	-	-
		(0.131)	(0.188)	(0.154)	(0.161)	-	-
	3	0.040	0.055	0.065	0.066	-	-
		(0.047)	(0.061)	(0.081)	(0.067)	-	-
	4	0.026	0.039	0.034	0.046	-	-
		(0.028)	(0.027)	(0.043)	(0.042)	-	-
	4*	0.083	0.094	0.079	0.085	0.114	0.171
		(0.17)	(0.19)	(0.156)	(0.195)	(0.172)	(0.254)

Sc. 2	$\beta_{\lambda,0}$	$\bar{\beta}_{\lambda,0}^{(1)}$	$\bar{\beta}_{\lambda,0}^{(2)}$	$\bar{\beta}_{\sigma,0}^{(1)}$	$\bar{\beta}_{\sigma,0}^{(2)}$	$\bar{\beta}_{\gamma,0}^{(1)}$	$\bar{\beta}_{\gamma,0}^{(2)}$
$T = 50$	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)
	4*	3.99	4.031	1.27	1.507	-0.751	-0.916
		(0.184)	(0.204)	(0.233)	(0.182)	(0.468)	(0.509)
$T = 100$	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)
	4*	4.006	3.991	1.323	1.489	-0.688	-0.875
		(0.151)	(0.123)	(0.199)	(0.176)	(0.344)	(0.341)

Table 2: Results of Scenario 2, with $x_t^\lambda \neq x_t^\sigma \neq x_t^\gamma$. The MISE is 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. 4^* indicates the results where γ depends on x_t^γ as well. Standard errors over the 200 samples are in parentheses.

$T = 50$							$T = 100$						
Sc. 1	$\bar{R}$	$\tilde{R}$	$\bar{\pi}_{11}$	$\tilde{\pi}_{11}$	$\bar{\pi}_{22}$	$\tilde{\pi}_{22}$	Sc. 1	$\bar{R}$	$\tilde{R}$	$\bar{\pi}_{11}$	$\tilde{\pi}_{11}$	$\bar{\pi}_{22}$	$\tilde{\pi}_{22}$
2	0.891	0.94	0.899	0.951	0.89	0.943	2	0.942	0.95	0.933	0.947	0.938	0.953
3	0.957	0.98	0.901	0.935	0.918	0.945	3	0.976	0.98	0.934	0.948	0.933	0.947
4	0.987	1	0.923	0.947	0.921	0.947	4	0.992	0.99	0.94	0.951	0.937	0.949
4*	0.985	1	0.906	0.946	0.912	0.949	4*	0.992	1	0.943	0.952	0.936	0.946
Sc. 2	$\bar{R}$	$\tilde{R}$	$\bar{\pi}_{11}$	$\tilde{\pi}_{11}$	$\bar{\pi}_{22}$	$\tilde{\pi}_{22}$	Sc. 2	$\bar{R}$	$\tilde{R}$	$\bar{\pi}_{11}$	$\tilde{\pi}_{11}$	$\bar{\pi}_{22}$	$\tilde{\pi}_{22}$
2	0.876	0.94	0.907	0.948	0.906	0.946	2	0.953	0.97	0.939	0.953	0.933	0.949
3	0.955	0.98	0.924	0.95	0.908	0.943	3	0.982	0.99	934	0.95	0.936	0.953
4	0.995	1	0.925	0.943	0.923	0.942	4	0.995	1	0.939	0.95	0.935	0.945
4*	0.981	1	0.896	0.94	0.892	0.938	4*	0.986	1	0.886	0.934	0.899	0.94

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. $T = 100$).

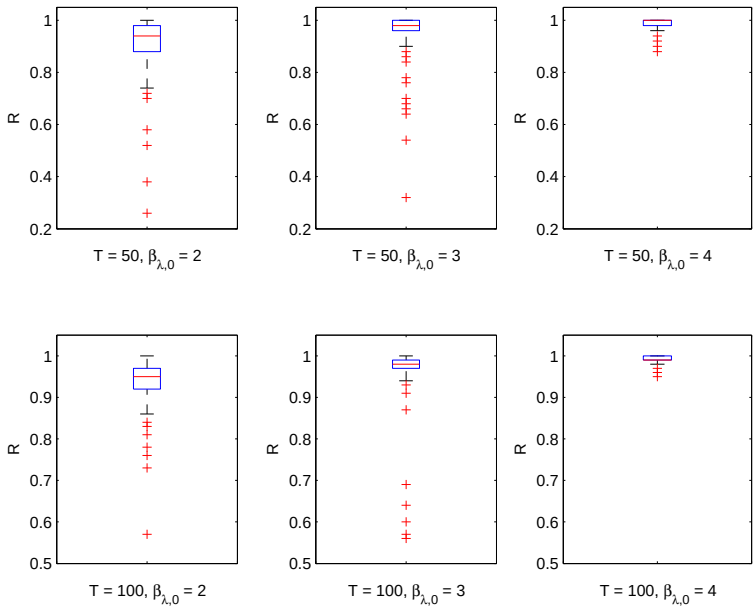

Figure 1: Boxplots for the proportion of correctly classified latent states, given by equation (25, for Scenario 1 ($x_t^\lambda = x_t^\sigma$). 1 corresponds to 100% of the time periods correctly classified. Top: $T = 50$. Bottom: $T = 100$.

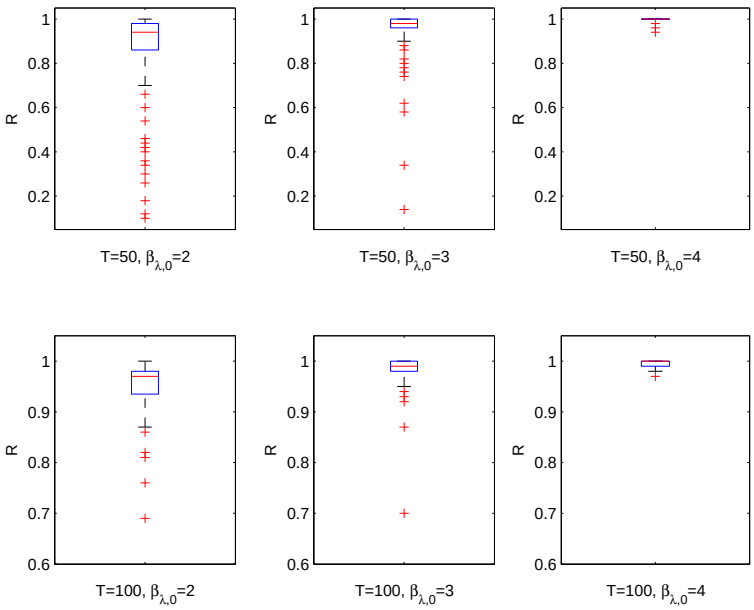

Figure 2: Boxplots for the proportion of correctly classified latent states, given by equation (25, for Scenario 2 ($x_t^{\lambda} \neq x_t^{\sigma}$). Top: $T = 50$. Bottom: $T = 100$.

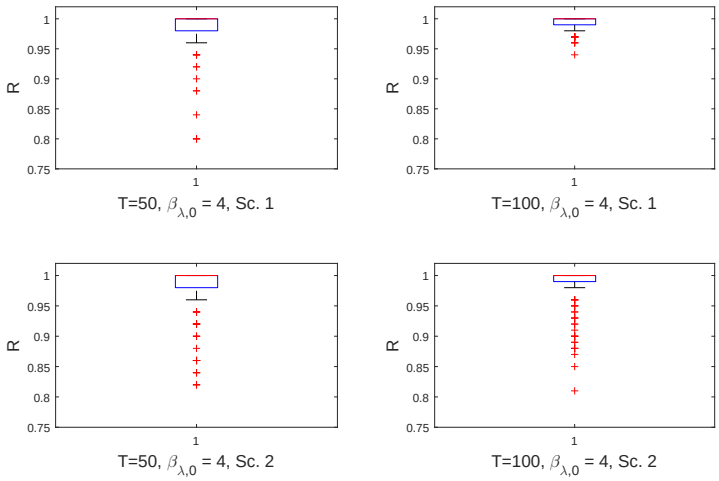

Figure 3: Boxplots for the proportion of correctly classified latent states, given by equation (25), when γ depends on covariates. Top (resp. bottom): Scenario 1. Left (resp. right): $T = 50$ (resp. $T = 100$).

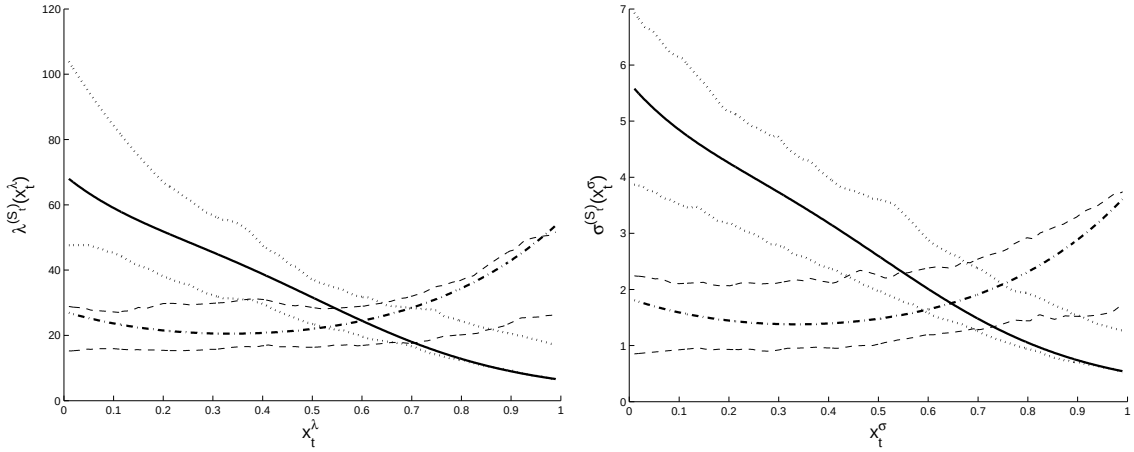

Figure 4: Dashed dotted (resp. solid): true function for $j = 1$ (resp. $j = 2$) between the parameters of the model and the covariate. Dotted and dashed: 95% Monte Carlo coverage bands of the maximum likelihood estimation of these functions. Left (resp. right) $\lambda_t^{(j)}$ (resp. $\sigma_t^{(j)}$). $T = 50$ and $\beta_{\lambda,0} = 4$ (Scenario 2).

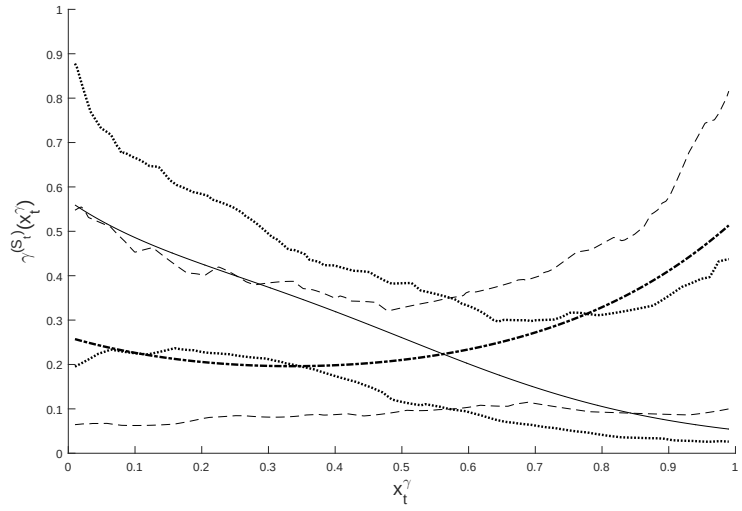

Figure 5: Dashed dotted (resp. solid): true function for $j = 1$ (resp. $j = 2$) between the parameters of the model and the covariate. Dotted and dashed: 95% Monte Carlo coverage bands of the maximum likelihood estimation of these functions, for $\gamma_t^{(j)}$, $T = 100$ and $\beta_{\lambda,0} = 4$ (Scenario 2).

4 Empirical Study

In this section, we apply the proposed procedure to a novel dataset of 819 operational losses at UniCredit, one of the largest European banks. 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]. 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. More recently, it made newspapers' headlines due to repeated liquidity and solvency issues. 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 [Moosa, 2011, Cope et al., 2012, Chavez-Demoulin et al., 2016] and the importance of regulatory changes [Cope et al., 2012, Dionne and Saissi Hassani, 2017] 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 here is 25,000€, so that we have a sample of extreme losses, where the GPD is a reasonable hypothesis for the severity distribution[5]. Such an approach can be related to the Peak-Over-Threshold technique [see e.g. Embrechts et al., 1997, Beirlant et al., 2005, Chavez-Demoulin et al., 2016]. 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

[5]Notice that this threshold is close to the 90% empirical quantile of all external fraud losses registered by the bank during the considered period.

total loss for each quarter. On a regulatory point of view, banks are expected to define the capital reserve for a horizon of one year but we work on a quarterly basis, as we would not have enough time periods to estimate correctly the transition probabilities of the model otherwise. If one want to obtain yearly measures, a simple addition of the forecast capital for four 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 6 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 correlated with the economic well-being of a country. In this application, we use the following covariates:

- To model the frequency parameter λ, 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 provides 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). Last, as noted in Povel et al. [2007] and Laeven and Levine [2007], a high level of non-interest incomes might also create incentives to commit frauds.

- To model the scale parameter σ, 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 [Povel et al., 2007, Cope et al., 2012].

- To model the shape parameter γ, 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 [Bekaert and Hoerova, 2014]. 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[6].

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 do not 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 7. 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).

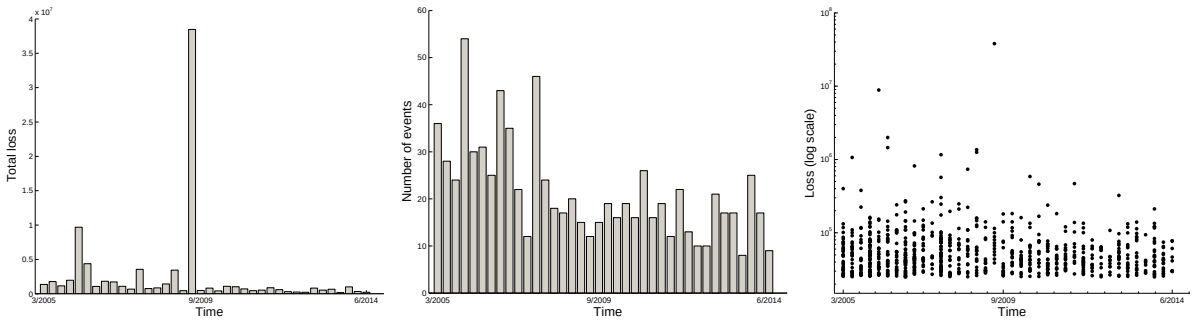

Figure 6: From left to right: total loss, number of losses per quarter, observed losses (grouped per quarter in log-scale).

[6]As suggested by a reviewer, an interesting alternative would be to consider e.g. VSTOXX instead of VIX as an explanatory variable for γ, since UniCredit has its core business in Europe. In the supplementary material, the interested reader can find the results obtained from estimating such a model on the present dataset. Results are almost identical and can be explained by today's financial globalization [Mendoza and Quadrini, 2010], causing VIX and VSTOXX time series to be highly similar: their correlation coefficient on the considered period is 0.96. To keep the discussion concise, and because VIX is a more popular measure of market uncertainty, only the results obtained with the VIX are discussed.

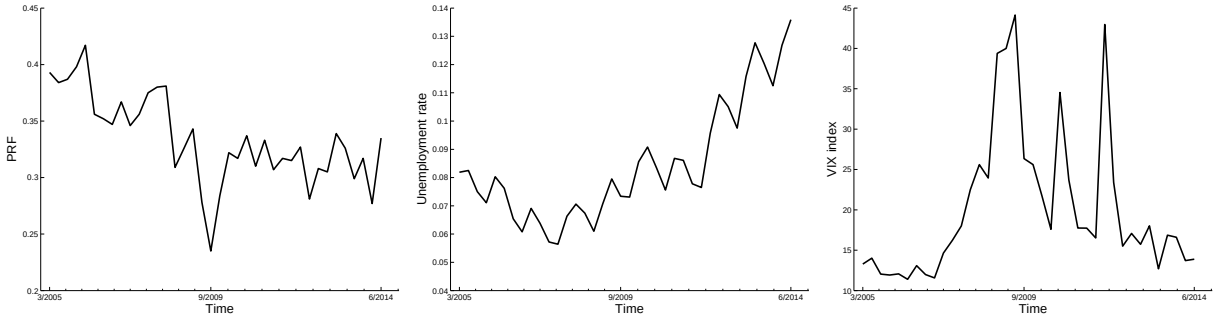

Figure 7: Value of the explanatory variables over the considered period. From left to right: lagged values (one quarter) of the PRF, the Italian unemployment rate and the VIX index.

4.2 Results

4.2.1 Estimated model and decoded states

Figure 8 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 when using 11 basis functions. The nonparametric approach seems useful to handle some non-linearities (in particular for the unemployment rate and the VIX). Increases in the PRF are associated with increases in λ, 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 σ, and then with an increase in σ 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 γ 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 9 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^{nd} and 3^{rd} 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 10, 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^{rd} 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 the suffered 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 γ 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 on the volatility of the financial markets in this regime. However, when the VIX decreases, γ decreases more in state 1, compared to state 2. Consequently, when the VIX is low in state 1, we reach lower values of γ 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 γ equal to 0.39 and 0.50).

Regarding the scale parameter σ, 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, σ 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 λ, 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, λ fluctuates around 35, whereas in the other state, λ varies between 15 and 20. Once again, structural or regulatory changes might be the cause of this drop.

Boostrap confidence intervals (based on 1000 resamples) are displayed in Figure 13 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.

4.2.2 Regulatory implications

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 10 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^{nd} 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.

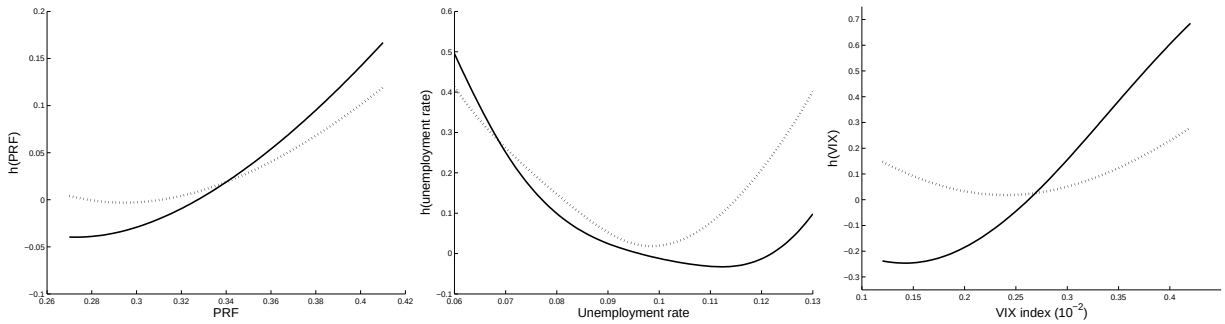

Figure 8: Estimated nonparametric functional forms between the covariates and the parameters of the model. Solid (resp. dotted) line: state 1 (resp. 2). From left to right: PRF, unemployment rate and VIX index.

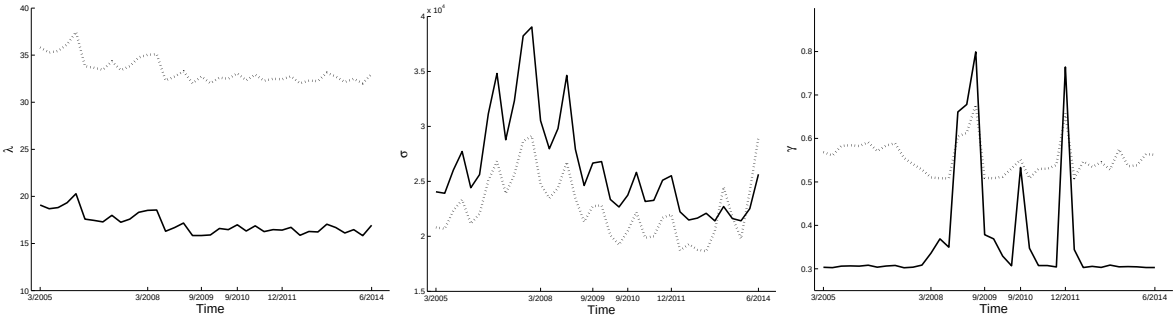

Figure 9: Estimated values of the parameters in both states (solid: state 1; dotted: state 2) over the covered period.

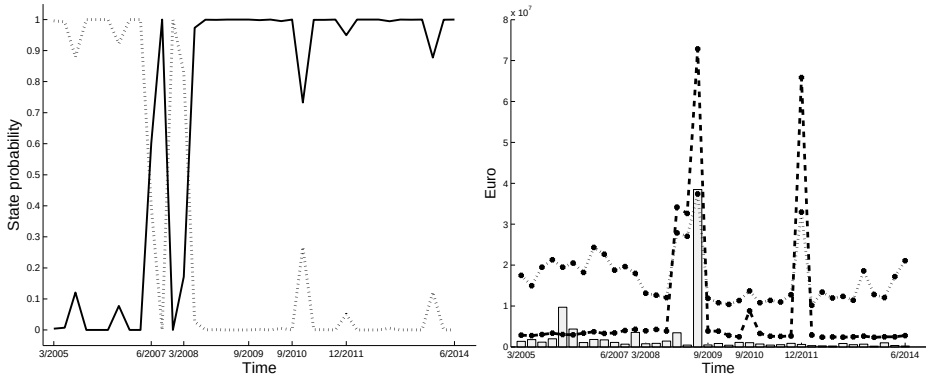

Figure 10: Left: state probabilities for the different regimes (thin solid: first state; dotted: second state). Right: $\hat{Q}_{0.999}(L_t; X_t, S_t)$ over time, in both states (dashed: first state, dotted: second state). Grey bar: total loss

4.2.3 Comparisons with simpler alternatives

To benchmark the perfomance 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 (nonparametrically), but where the structure is assumed not to change over time. Third, we consider an 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 15). 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 11). 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 worse 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 breaches of the 99.9% quantile (i.e. the number of total loss realizations that are larger than the estimated 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. Moreover, as suggested by the study of the in-sample coverage of the estimated quantile (presented in the supplementary material), the likelihood of suffering at least one breach under a correctly specified MS-GAMLSS model is quite low. Using a bootstrap procedure to estimate this quantity (for the estimated model and $T = 38$), we find a value of 1.9%. It implies that observing one or two breaches (in-sample) on a time series of length 38 is probably due to a misspecification of the estimated model.

For an 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 12 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 conditional distribution correctly.

Overall, the MS-GAMLSS model seems to 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.

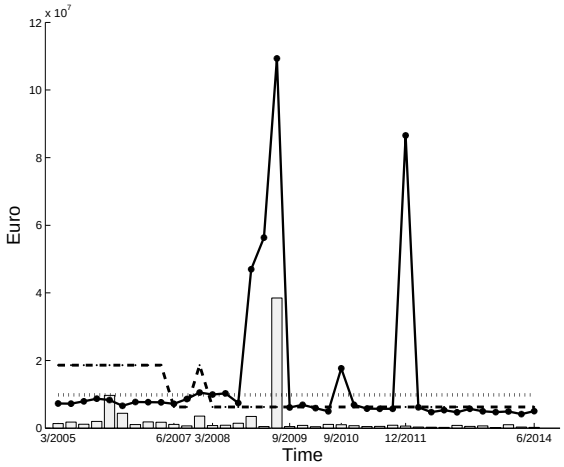

Figure 11: $\hat{Q}_{0.999}(L_t; X_t, S_t)$ over time obtained with the alternative models: constant (dotted), MS-CST (dashed), GAMLSS (solid). Grey bar: total loss.

Final results	$\hat{\pi}_{11}$	$\hat{\pi}_{22}$	$\hat{\beta}_0^{(1)}(\sigma)$	$\hat{\beta}_0^{(2)}(\sigma)$	$\hat{\beta}_0^{(1)}(\gamma)$	$\hat{\beta}_0^{(2)}(\gamma)$	$\hat{\beta}_0^{(1)}(\lambda)$	$\hat{\beta}_0^{(2)}(\lambda)$	-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.

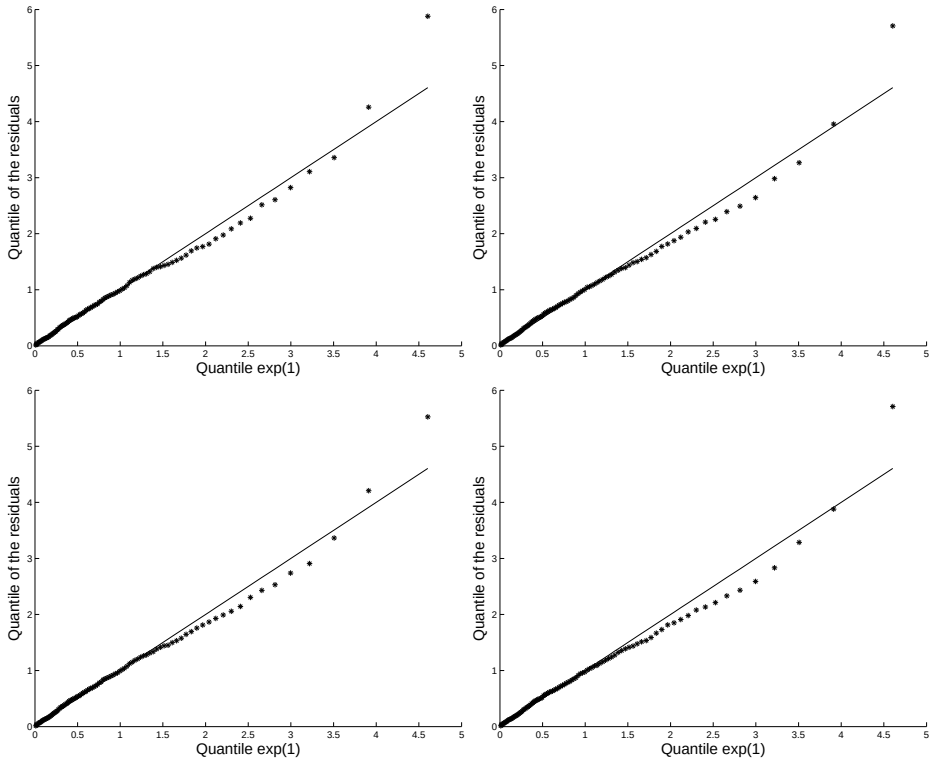

Figure 12: QQ-plots for the pseudo-residuals of the four models. Top, from left to right: MS-GAMLSS and MS-CST models. Bottom, from left to right: GAMLSS and CST models.

4.3 Economic implications and limitations

Economically speaking, a bank that uses an MS-GAMLSS model to establish its requested capital could retrospectively 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 particularly 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. This issue does not guarantee a superior out-of-sample performance. Once again, it is a danger of MS models, and it needs to be analyzed by risk managers willing to use them. As suggested by the out-of-sample analysis available in the supplementary material, large samples are important to ensure that the tpm is well estimated. In that situation, we observe good predictions of the state process and thus, quantile estimates providing the expected number of exceedances. 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. Despite the aforementioned limitations, this empirical study suggests that the distribution of operational losses is, indeed, not stable over time 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 little 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 819 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 a 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^{st} 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 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 (11) to (13) 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

The authors warmly thank an anonymous referee for her/his detailed comments and suggestions. It has led to improve the clarity of the present manuscript. J. Hambuckers acknowledges the support of the Research Training Group 1644 *Scaling Problems in Statistics* funded by the German Research Foundation (DFG). J. Hambuckers also thanks V. Chavez-Demoulin as well as the participants of HEC Lausanne seminars for the fruitful discussions that helped improving earlier versions of the present manuscript. T. Kneib

acknowledges the financial support of DFG, grant KN 922/4-2. The authors acknowledge F. Piacenza for providing them the data.

References

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*. Wiley, Chichester, 2005.

G. Bekaert and M. Hoerova. The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2):181–192, 2014.

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(6):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, 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.

G. Dionne and S. Saissi Hassani. Hidden markov regimes in operational loss data: Application to the recent financial crisis. *Journal of Operational Risk*, 12(1):23–51, 2017.

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. Klüppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer, Berlin, 1997.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression : models, methods and applications*. Springer, Berlin, 2013.

R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, 1987.

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 nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293, 2012.

R. King and R. Langrock. Semi-Markov Arnason-Schwarz models. *Biometrics*, 72(2): 619–628, 2016.

F. Klaassen. Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics*, 27(2):363, 2002.

L. Laeven and R. Levine. Is there a diversification discount in financial conglomerates? *Journal of Financial Economics*, 85:331–367, 2007.

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. Flexible latent-state modelling of Old Faithful's eruption inter-arrival times in 2009. *Australian & New Zealand Journal of Statistics*, 54(3):261–279, 2012.

R. Langrock, T. Kneib, R. Glennie, and T. Michelot. Markov-switching generalized additive models. *Statistics and Computing*, 27(1):259–270, 2017.

I.L. MacDonald. Numerical maximisation of the likelihood: a neglected alternative to EM? *International Statistical Review*, 82(2):296–308, 2014.

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)*, 61(3):403–427, 2012.

E. Mendoza and V. Quadrini. Financial globalization, financial crises and contagion. *Journal of Monetary Economics*, 57(1):24–39, 2010.

I. Moosa. Operational risk as a function of the state of the economy. *Economic Modelling*, 28(5):2137–2142, 2011.

M. Moscadelli. The modelling of operational risk: experience with the analysis collected by the Basel Committee. Technical report, Roma, 2004.

J. Neslehova, P. Embrechts, and V. Chavez-Demoulin. Infinite-mean models and the LDA for operational risk. *Journal of Operational Risk*, 1(1):3–25, 2006.

J. Nocedal and S. J. Wright. *Numerical Optimization, Second Edition*. Springer Series in Operations Research, Springer Verlag, 2006.

M. Pesaran, D. Pettenuzzo, and A. Timmermann. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084, 2006.

J. Pohle, R. Langrock, F.M. van Beest, and J. Nabe-Nielsen. Selecting the number of states in hidden Markov models pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):270–293, 2017.

P. Povel, R. Singh, and A. Winton. Booms, Busts, and Fraud. *Review of Financial Studies*, 20(4):1219–1254, 2007.

R. Rigby and D. Stasinopoulos. 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. Soprano, B. Crielaard, F. Piacenza, and D. Ruspantini. *Measuring Operational and Reputational Risk: A Practitioner's Approach*. Wiley, Chichester, 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*. Chapman & Hall, Boca Raton, 2016.

6 Appendix A

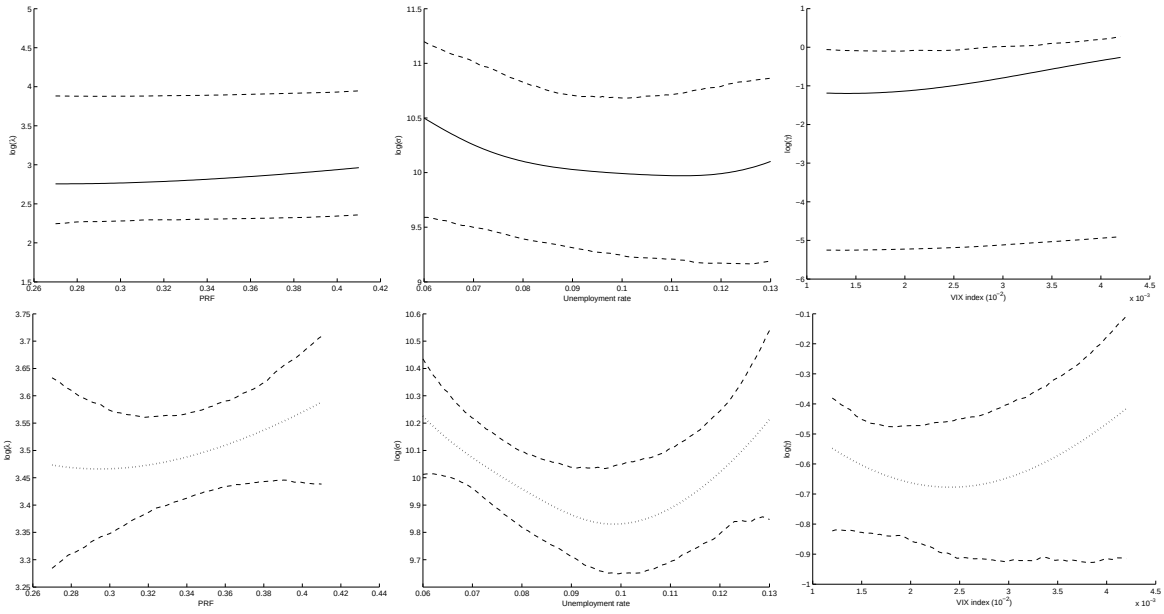

Figure 13: Estimated functional forms related to the three covariates considered, obtained with the MS-GAMLSS model (solid line: state 1, dotted: state 2). Dashed: 95% confidence intervals. X-axis: value of the covariate. Y-axis: $\log(\hat{\theta}^{(j)}(X)), \theta \in \{\lambda, \sigma, \gamma\}$ and $j = 1, 2$.

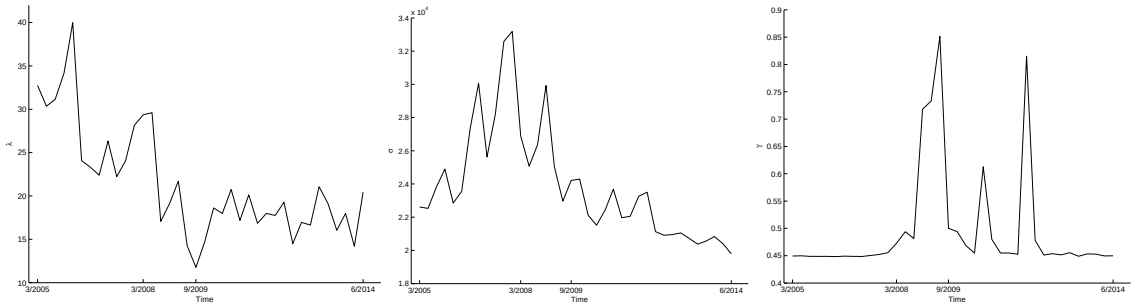

Figure 14: Estimated frequency, scale and shape parameters with the GAMLSS model, over time. Cross-validated smoothing parameters are respectively 2, 25 and 8.

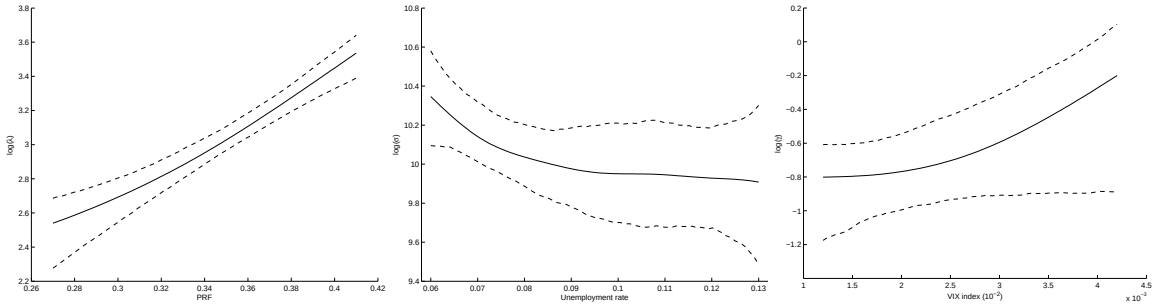

Figure 15: Estimated functional forms related to the three covariates considered, obtained with the GAMLSS model (solid line). Dashed: 95% confidence intervals. X-axis: value of the covariate. Y-axis: $\log(\hat{\theta}^{(j)}(X)), \theta \in \{\lambda, \sigma, \gamma\}$ and $j = 1, 2$.