A semiparametric model for Generalized Pareto regression based on a dimension reduction assumption

Julien Hambuckers, Cédric Heuchenne, Olivier Lopez

A semiparametric model for Generalized Pareto regression based on a dimension reduction assumption

Julien Hambuckers[1], Cédric Heuchenne[2,3], Olivier Lopez[4]

September 8, 2016

Abstract

We consider a regression model in which the tail of the conditional distribution of the response can be approximated by a Generalized Pareto distribution. Our model is based on a semiparametric single-index assumption on the conditional tail index, while no further assumption on the conditional scale parameter is made. The underlying dimension reduction assumption allows the procedure to be of prime interest in the case where the dimension of the covariates is high, in which case the purely nonparametric techniques fail while the purely parametric ones are too rough to correctly fit to the data. We derive asymptotic properties of the resulting parameter estimators, and propose an iterative algorithm for their practical implementation. We study the finite sample behavior of our methodology through simulations. To exhibit the interest of the proposed approach in practice, the method is applied to a new database of operational losses from the bank UniCredit.

Key words: Curse-of-dimensionality; Extreme events; Semiparametric regression; Operational loss;

Short title: Semiparametric Generalized Pareto Regression

[1] Georg-August-Universität Göttingen, Chair of Statistics, Humboldtallee 3, 37073 Göttingen, Germany. E-mail: julien.hambuckers@mathematik.uni-goettingen.de.
[2] University of Liege, HEC Liege, Center for Quantitative Methods and Operation management, 14 rue Louvrex, 4000 Liège, Belgium. E-mail: c.heuchenne@ulg.ac.be.
[3] Université Catholique de Louvain (Louvain-la-Neuve), Institute of Statistics, 20 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.
[4] Sorbonne Universités, UPMC Université Paris VI, CNRS FRE 3684, LSTA, 4 place Jussieu, 75005 Paris, France. E-mail: olivier.lopez0@upmc.fr.

1 Introduction

Generalized Pareto regression models are adapted to the study of extreme events depending on covariates. Indeed, in presence of heavy tailed distributions, the exceedance of a random variable over a sufficiently high threshold can be approximated by a Generalized Pareto Distribution (GPD in the following), with its parameters denoted by γ for the tail index, and σ for the scale parameter. Among these two parameters, γ is of prime importance since it is related to the tail heaviness of the GPD, see e.g., Beirlant et al. (1999), Csörgő and Viharos (1998) or Beirlant et al. (2004), while σ can be seen as a nuisance parameter.

In a regression framework, one assumes that the parameters of the GPD depend on the covariates $X \in \mathbb{R}^d$, according to some specific model. Estimation of the parameters of the model can then be used to infer the tail index of the underlying conditional distribution of the response variable. A parametric approach, such as the one described in Beirlant and Goegebeur (2003), assumes that $(\gamma(X), \sigma(X)) = f(\theta_0, X)$, where $\theta_0 \in \mathbb{R}^k$, and f is a known function. Parametric models usually provide nice convergence rates of the estimators. Nevertheless, they rely on strong assumptions that may not hold in practice, resulting in poor fitting properties. At the contrary, a nonparametric approach as the one developed in Beirlant and Goegebeur (2004) relies on fewer assumptions, since it assumes that $(\gamma(X), \sigma(X)) = f(X)$, where f is unspecified, and only requires to satisfy some standard regularity conditions. In this framework, Beirlant and Goegebeur (2004) study a local polynomial estimator of the regression function f in the case where the covariate is one-dimensional. However, the convergence rate of this estimator is expected to decrease considerably when the dimension d of the covariates increases, which is known as the "curse of dimensionality".

In the present paper, we propose a new methodology based on a semiparametric regression model which can be seen as a convenient compromise between the two approaches that we mentioned. In view of applying the model to the statistical study of conditional extremes, we do not assume that the conditional distribution of the response is strictly GPD. The model that we consider is based on a dimension reduction assumption, which is related to single-index regression models, see e.g. Ichimura (1993) or Härdle et al. (1993).

Single-index regression models consist in assuming that $m(X) = g(\theta_0^T X)$, where m is a regression function, $\theta_0 \in \mathbb{R}^d$ an unknown parameter, and g an unknown link function. The idea is that, if we knew the true parameter θ_0, we would be back to a fully nonparametric model, but this time with an one-dimensional covariate. Therefore, we would not suffer from the so-called curse of dimensionality. The advantage of such a model stands in the fact that it requires less assumptions on the parametric regression model - and therefore is probably a better approximation of the real model - while it avoids specific failures of the nonparametric approach. In most papers related to single-index estimation, authors focus mostly on the special cases of m being a conditional mean (see Härdle et al. (1993), Delecroix et al. (2006), Xia and Li (1999), Xia et al. (1999), Xia et al. (2011)), a conditional density (see Delecroix et al. (2003), Bouaziz and Lopez (2010)), or a conditional quantile (see Wu et al. (2010)).

Relying on a similar idea, we propose to consider a Generalized Pareto (GP) regression model in which $\gamma(x) = \gamma_{\theta_0}(\theta_0^T x)$, with no additional assumption on the nuisance parameter $\sigma(x)$. We provide a $n^{1/2}-$consistent estimator of θ_0, and discuss its use to the estimation of the conditional tail index $\gamma(x)$. We prove the convergence of the estimator in a framework where the conditional distribution is not exactly a GPD, which shows that our technique can be successfully applied to regression models for extreme events. We also provide an iterative algorithm to perform the maximization of the pseudo-maximum likelihood criterion that we use, in order to choose the different parameters involved in the procedure. The performance of this method is discussed through a simulation study. Lastly, as a real data example, the method is applied to a new database of operational losses provided by the bank UniCredit.

The rest of this paper is organized as follows. In Section 2, we define the semiparametric GP-regression model thatwe consider in the following, and propose a way to estimate the parameters and regression functions through a pseudo-likelihood approach. Section 3 presents the asymptotic properties of this estimators, while Section 4 disscusses the practical implementation, through an iterative algorithm. In Section 5, we investigate on the finite sample behavior of the procedure through a simulation study and provide a practical example on a dataset of financial operational losses. The Appendix section presents some technical computations that are needed to prove the results of Section 3.

2 Semiparametric Generalized Pareto Regression Model

In this section, we first briefly present the GPD in Section 2.1, emphasizing its properties that are linked with extreme value theory. In Section 2.2 we define the semiparametric GP-regression model that will be studied throughout this paper. To describe the logic behind our estimation procedure, which is done in Sections 2.3 to 2.4, we first elude the fact that our responses may not follow exactly a GPD given the covariates. Then, we discuss, in Section 2.5, the modifications that should be introduced in case of applying a POT technique, which introduces a misspecification error.

2.1 Generalized Pareto Distribution

The GPD(γ, σ) is defined by the following cumulative distribution function,

$$G(y; \sigma, \gamma) = 1 - \frac{1}{\left(1 + \gamma \frac{y}{\sigma}\right)^{1/\gamma}},$$

for $y > 0$, $\gamma > 0$, and $\sigma > 0$. The parameter γ is the extreme-value index, while the parameter σ can be interpreted as some scale parameter. Indeed, the parameter γ is related to the following fundamental result in extreme value analysis. If we consider an i.i.d. random vector $(Z_i)_{1 \leq i \leq n}$, and if we denote by $Z_{(k)}$ the $k-$th order statistic, and if, for two sequences $(a_n)_{n \geq 0}$ and $(b_n)_{n \geq 0}$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{Z_{(n)} - b_n}{a_n} \leq z\right) = F(z), \tag{2.1}$$

where F is a distribution function, then F is necessarily of the following form,

$$F(z) = F_\gamma(z) = \begin{cases} \exp\left(-(1+\gamma z)^{-1/\gamma}\right), \ if \ (1+\gamma z) > 0, \ \gamma \neq 0, \\ \exp\left(-\exp(-z)\right), \ if \ \gamma = 0. \end{cases}$$

In the following, we will focus on the case where $\gamma > 0$. A proof of this result can be found in Fisher and Tippett (1928) and Gnedenko (1943). In view of this result, the parameter γ describes the heaviness of a distribution, and is of prime importance if one whishes to investigate on extreme events. A review of the main estimation methods of γ can be found in Beirlant et al. (2004).

Among them, the Peak Over Threshold (POT) technique proposed by Smith (1987) relies on the following result: if the distribution of Z_1 satisfies equation (2.1) with some $\gamma > 0$, then, if we denote $F_u(z) = \mathbb{P}(Z - u > z | Z \geq u)$, we have

$$\lim_{u \to z_F} \sup_{0<y<z_F-u} |F_u(z) - G(y; \sigma(u), \gamma)| = 0, \tag{2.2}$$

where $z_F = \inf\{z : \mathbb{P}(Z_1 \leq z) = 1\}$, for some $\sigma(u)$ (see Pickands (1975)). Hence, the GPD can be seen as an approximation of the distribution of the excesses over some threshold. Using this approximation, likelihood techniques can be used to estimate the parameter γ.

2.2 GP-regression

In the Generalized Pareto regression model that we consider, we assume that, given a set of covariates $X \in \mathcal{X} \subset \mathbb{R}^d$, the response variable Y follows a GPD with conditional tail index $\gamma(x)$, and conditional scale parameter $\sigma(x)$, when $X = x$. Our semiparametric regression model consists in assuming that the function γ depends on the covariates only through an unknown linear combination of these covariates, that is,

$$\gamma(x) = \gamma_{\theta_0}(\theta_0^T x), \tag{2.3}$$

where θ_0 is an unknown parameter belonging to $\Theta \subset \mathbb{R}^d$ (parametric part), and γ_{θ_0} is an unknown link function (nonparametric part). Compared to a fully nonparametric model, this approach performs dimension reduction, since the relevant information on the covariates needed to compute $\gamma(x)$ is summarized by $\theta_0^T x$. The study of the nonparametric part then becomes a problem of estimating a function of a one-dimensional random vector. To ensure the identifiability of the model, we need to add a constraint on θ_0, e.g., $\|\theta_0\| = 1$, where $\|\cdot\|$ denotes some norm in $\mathbb{R}^d$.

In our approach, we do not focus on the nuisance parameter σ. This choice is motivated by the fact that, from (2.2), the most important parameter for inference in the GPD is γ. Alternatively, if one wishes to make more precise inference on this scale parameter, it could be possible to also assume that a dimension reduction assumption holds for this function, such as $\sigma(x) = \sigma_{\beta_0}(\beta_0^T x)$. We can easily extend our estimation procedure to this framework but it would lead to a higher complexity of the procedure due to selection of the two vectors (θ_0, β_0). A way to simplify this issue would be to assume that $\beta_0 = \theta_0$, but

this assumption may not be realistic. In practice, the single-index assumption can be seen as an approximation of the real (nonparametric) model. Additional dimension reduction assumption could create some additional model bias, without necessarily providing a significant improvement of the result.

In the following, we assume that we have at our disposal a consistent estimator $\hat{\sigma}(x)$ of $\sigma(x)$ (obtained through parametric, semiparametric or nonparametric modeling and at least satisfying Assumption 11 in Section 3).

2.3 Estimation procedure for θ_0

In this section, we consider observations made of a random vector $(Y_i, X_i^T)_{1\leq i\leq n}$ which are i.i.d. replications of a random vector (Y, X^T) for which model (2.3) holds. In this first approach, we proceed as if the conditional law of Y_i given X_i were exactly a GPD.

As pointed out before, in studying a model such as (2.3), the key problem stands in the estimation of the parameter θ_0. Indeed, if this parameter were known, we would be back to a purely nonparametric problem which has already received satisfactory developments in the literature. To estimate θ_0, our procedure consists in adapting a maximum likelihood strategy. Let $\Gamma = \{\gamma_\theta : \theta \in \Theta\}$ denote a family of functions such that $\gamma_{\theta_0}(\theta_0^T x) = \gamma(x)$. We will discuss the choice of a proper family in detail in Section 2.4. If this family of functions were known, and if σ were known, the maximum likelihood approach would consist in maximizing

$$M_n(\gamma_\theta, \sigma; \theta) = \frac{1}{n}\sum_{i=1}^{n} l\left(\gamma_\theta(\theta^T X_i); \frac{Y_i}{\sigma(X_i)}\right), \tag{2.4}$$

with respect to θ, where

$$l(\gamma; y) = -\left(\frac{1}{\gamma} + 1\right)\log\left(1 + y\gamma\right).$$

Given a a family of nonparametric estimators $\hat{\gamma}_\theta$, and a nonparametric estimator $\hat{\sigma}$ of σ, we can define the following estimator of the single-index parameter,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} M_n(\hat{\gamma}_\theta, \hat{\sigma}; \theta). \tag{2.5}$$

In the following, we will show asymptotic properties of $\hat{\theta}$ which do not depend on the particular type of nonparametric estimator that is used for $\gamma_\theta(\theta^T x)$ (for a given θ), provided that this estimator satisfies some relatively standard conditions.

2.4 The family Γ

In Section 2.3, we focus solely on the estimation of the single-index parameter. This leaves us some latitude in view of choosing the family of functions Γ that we will use in our procedure. Let

$$\gamma_\theta^{(1)}(u) := \arg\max_\gamma E\left[l\left(\gamma; \frac{Y_i}{\sigma(X_i)}\right)|\theta^T X_i = u\right]; \tag{2.6}$$

the family $\Gamma^{(1)} = \{\gamma_\theta^{(1)} : \theta \in \Theta\}$ can represent a natural choice. However, this family has an important disadvantage: there is no closed form for $\gamma_\theta^{(1)}(u)$ defined as the optimum of some nonlinear function. If we wish to estimate $\gamma_\theta^{(1)}(u)$, we can use a local version using kernel weights, i.e.,

$$\hat{\gamma}_{\theta,h}^{(1)}(u) \quad = \quad \arg\max_a \sum_{i=1}^n l\left(a; \frac{Y_i}{\hat{\sigma}(X_i)}\right) K\left(\frac{\theta^T X_i - u}{h}\right), \tag{2.7}$$

where K is a kernel function (satisfying $\int K(u)du = 1$), and h a smoothing parameter. Alternatively, we can use a local polynomial estimator (see also Beirlant and Goegebeur (2004)),

$$(\hat{\gamma}_{\theta,h}(u), \hat{\gamma}'_{\theta,h}(u)) = \arg\max_{a,b} \sum_{i=1}^n l\left(a + b(\theta^T X_i - u); \frac{Y_i}{\hat{\sigma}(X_i)}\right) K\left(\frac{\theta^T X_i - u}{h}\right) \tag{2.8}$$

which presents the advantage to estimate not only $\gamma(x)$ but also its derivative $\gamma'(x)$ by $\hat{\gamma}'(x)$. Independently of the estimator we consider, no closed form exists for the solution of an optimization problem of the type of (2.7). This may lead to prefer other choices of sets of functions Γ, whose elements can be approached through closed form estimators.

Indeed, observe that

$$E\left[\frac{Y_i}{\sigma(X_i)}|X_i\right] = E\left[\frac{Y_i}{\sigma(X_i)}|\theta_0^T X_i\right] = \frac{1}{1 - \gamma_{\theta_0}(\theta_0^T X_i)},$$

provided that $\gamma_{\theta_0}(\theta_0^T X_i) < 1$. If we assume that this condition holds for all values of $\theta_0^T X_i$, one can use the family $\Gamma^{(2)}$ constituted of the following functions,

$$\gamma_\theta^{(2)}(u) = 1 - \frac{1}{E\left[\frac{Y_i}{\sigma(X_i)}|\theta^T X_i = u\right]}.$$

The main advantage is that $m_\theta(u) := E[Y_i\sigma(X_i)^{-1}|\theta^T X_i = u]$ can be used by the following kernel estimator

$$\hat{m}_\theta(u) = \sum_{i=1}^n \frac{K\left(\frac{\theta^T X_i - u}{h}\right)}{\sum_{j=1}^n K\left(\frac{\theta^T X_j - u}{h}\right)} \frac{Y_i}{\hat{\sigma}(X_i)},$$

leading to

$$\hat{\gamma}_\theta^{(2)}(u) = 1 - \frac{1}{\hat{m}_\theta(u)}.$$

This estimator is easier to compute, and will facilitate the implementation of our estimator of θ_0.

Note that choosing Γ^2 instead of Γ^1 impacts the estimation of $\gamma_{\theta_0}(\theta_0^T X)$ but has less impact (at least asymptotically) on the final estimation of θ_0 (see Theorem 3.2). In our procedure, we separate the problem of estimating θ_0 from the estimation of $\gamma_{\theta_0}(\theta_0^T X)$: at each iteration of the proposed methodology, we could use different nonparametric estimators for $\gamma_{\theta_0}(\theta_0^T X)$.

2.5 Case of a misspecified distribution in the POT technique

In view of Section 2.1, Gnedenko (1943) showed that there was some equivalence between the fact that a distribution $H(z) = \mathbb{P}(Z_1 \leq z)$ has a tail behavior described by F_γ, and the fact that

$$1 - H(z) = z^{-\frac{1}{\gamma}} \delta(z), \tag{2.9}$$

where $\delta(z)$ is a slowly-varying function, that is

$$\lim_{z \to \infty} \frac{\delta(\lambda z)}{\delta(z)} \to 1,$$

for all $\lambda > 0$.

Consider that we observe $(Z_i, X_i^T)_{1 \leq i \leq m}$ i.i.d., where Z_i has a conditional distribution $F(z|x) = \mathbb{P}(Z \leq z | X = x)$ of the type (2.9) for some $\delta(z|x)$, and some $\gamma(x) > 0$ which satisfies (2.3). For the sake of simplicity, we assume that X_i belongs to the compact $\mathcal{X} \subset \mathbb{R}^d$. Moreover, defining some threshold function u_x, we have

$$\lim_{u_x \to z_{F,x}} \sup_{0 < y < z_{F,x} - u_x} |F_{u_x}(z|x) - G(z; \sigma(u_x, x), \gamma_{\theta_0}(\theta_0^T x))| = 0,$$

where $z_{F,x} = \inf\{z : F(z|x) = 1\}$, and $F_u(z|x) = \mathbb{P}(Z - u \geq y | X = x, Z \geq u)$. In the POT approach, one considers a threshold function $u_x(m)$ (with $u_x(m) \to z_{F,x}$ when m tends to infinity), taking $u_x(m)$ large enough so that the distribution of the excesses over $u_x(m)$ is sufficiently close to a GPD. In this case, the contrast (2.4) can still be used to estimate θ_0, but now with n replaced by $n(m) = \sum_{i=1}^m \mathbf{1}_{Z_i \geq u_{X_i}(m)}$, and Y_i replaced by

$Y_{i,m} = Z_{j(i)} - u_{X_{j(i)}}(m)$, where $j(i) = \inf\{k : \sum_{j=1}^{k} \mathbf{1}_{Z_j \geq u_{X_j}(m)} = i\}$. In this situation, we introduce some misspecification in the model, since the conditional distribution of Y_i given X_i is not exactly a GPD.

In the following, we will denote for the compact $\Theta \subset \mathbb{R}^d$,

$$\theta_0(u_x(m)) = \arg\max_{\theta \in \Theta} M_{u_x}(\gamma_\theta, \sigma; \theta),$$

where

$$M_{u_x}(\gamma_\theta, \sigma; \theta) = E\left[l\left(\gamma_\theta(\theta^T X); \frac{Z - u_X(m)}{\sigma(X)} \right) | Z \geq u_X(m) \right].$$

Because of this misspecification error, $\theta_0(u_x(m)) \neq \theta_0$, but the difference will be small provided that one adds conditions on the thresholds, see the next section. Finally, the empirical version of M_{u_x} will be denoted

$$M_n(\gamma_\theta, \sigma; \theta) = \frac{1}{n} \sum_{i=1}^{m} l\left(\gamma_\theta(\theta^T X_i); \frac{Z_i - u_{X_i}(m)}{\sigma(X_i)} \right) \mathbf{1}_{Z_i \geq u_{X_i}(m)}.$$

3 Asymptotic properties

In this section, we discuss the theoretical properties of the estimator described in the previous section. We first derive consistency in Section 3.1, then asymptotic normality in Section 3.2. We express the results under general assumptions on the nonparametric estimators $\hat{\gamma}_\theta$ and $\hat{\sigma}$ used in the procedure, so as to be adapted to different estimation strategies of the nonparametric part. Some of these nonparametric estimators and their applicability in the present context are discussed hereunder but since these do not consist in the main methodological issue, we prefer to focus on the proofs for $\hat{\theta}$. Some more details about the estimator $\hat{\gamma}_\theta^{(2)}(u)$ are however given in the Appendix. The considered Donsker classes are supposed to have a bracketing entropy of order ε^{-v} where ε is the length of bracket and $v < 2$.

3.1 Consistency of $\hat{\theta}$

Assumptions needed to ensure the consistency of $\hat{\theta}$ can be decomposed into three categories: assumptions on the criterion to maximize, assumptions on the used nonparametric

estimators and assumptions on the regression functions γ_θ and σ.

Assumptions on the asymptotic criterion.

The following assumption ensures that the true value of the parameter θ_0 is uniquely defined.

Assumption 1 *Let* $M(\gamma_\theta, \sigma; \theta) = E\left[l\left(\gamma_\theta(\theta^T X_i); \frac{Y_i}{\sigma(X_i)}\right)\right]$, *where* Y_i *given* X_i *exactly follows a GPD. Assume that,*

$$\forall \theta \in \Theta,\ M(\gamma_\theta, \sigma; \theta) = M(\gamma_{\theta_0}, \sigma; \theta_0) \Longrightarrow \theta = \theta_0.$$

Assumption 2 below ensures that some class of functions naturally linked to our problem satisfies a uniform law of large numbers property.

Assumption 2 *Define*

$$f_{\theta,m}(x, z) = l(\gamma_\theta(\theta^T x); (z - u_x(m))/\sigma(x)).$$

Assume that $\{(x, z) \to f_{\theta,m}(x, z) : \theta \in \Theta, m > 0\}$ *is a Glivenko-Cantelli class, that is*

$$\sup_{\theta,m} \left| \int f_{\theta,m}(x, z) d(\mathbb{P}_n - \mathbb{P}_{X,Z})(x, z) \right| = o_P(1),$$

where $\mathbb{P}_n$ *denotes the empirical distribution of* $(X_i, Z_i)_{1 \le i \le m}$, *for which* $Z_i \ge u_{X_i}(m)$, *and* $\mathbb{P}_{X,Z}$ *the true distribution of* (X, Z) *given* $Z_i \ge u_{X_i}(m)$.

In practice, this assumption can be replaced by imposing a regularity condition on the function γ_θ. Indeed, if we consider for example a constant $u_x(m) = u$ (the threshold does not depend on either X or m), this condition will hold if the function $\theta \to \gamma_\theta(\theta^T \cdot)$ is Lipschitz with respect to θ.

Assumptions on the nonparametric estimators.

We need the nonparametric estimators involved in the procedure to be uniformly consistent.

Assumption 3 *The nonparametric estimators $\hat{\gamma}_\theta$ and $\hat{\sigma}$ are uniformly consistent, i.e.*

$$\begin{aligned} \sup_{u\in\mathcal{U},\theta\in\Theta} |\hat{\gamma}_\theta(u) - \gamma_\theta(u)| &= o_P(1), \\ \sup_{x\in\mathcal{X}} |\hat{\sigma}(x) - \sigma(x)| &= o_P(1), \end{aligned}$$

where $\mathcal{U} = \{\theta^T x; x \in \mathcal{X}, \theta \in \Theta\}$.

Assumptions on the regression functions.

For technical reasons, we will also impose that the parameters in the GPD stay in a compact subset which does not include 0.

Assumption 4 *There exist some strictly positive constants c_γ, c_σ, C_γ and C_σ such that*

$$\begin{aligned} c_\gamma &\leq \gamma_\theta(\theta^T x) \leq C_\gamma, \\ c_\sigma &\leq \sigma(x) \leq C_\sigma \end{aligned}$$

for all $\theta \in \Theta$ and $x \in \mathcal{X}$.

We now state our consistency Theorem.

Theorem 3.1 *Under Assumptions 1 to 4,*

$$\hat{\theta} - \theta_0 = o_P(1)$$

as $n \to +\infty$.

The proof is postponed to Section 7.1 in the Appendix.

3.2 Asymptotic normality of $\hat{\theta}$

To obtain the asymptotic normality, we need additional assumptions. Basically, they arise from the need to have some rates of convergence and differentiability properties in order to complete the proofs. Additionally, a condition on the threshold (see Theorem 3.2 hereunder) is required as well as on the type of slowly varying function (with the so-called slow variation with remainder condition, see Goldie and Smith (1987)).

Assumptions on the asymptotic criterion.

Assumption 5 *θ_0 is an interior point of Θ.*

The above assumption will ensure that we can use differentiation at the point θ_0, since θ_0 is not located on the boundary.

Assumption 6 *Let $\Sigma = \nabla^2_\theta M(\gamma_{\theta_0}, \sigma; \theta_0)$, where $\nabla^2_\theta M$ denotes the Hessian matrix of M and the differentiation is performed with respect to all the occurrences of θ in $M(\gamma_\theta, \sigma; \theta)$. Assume that Σ is invertible.*

Assumption 7 *Denote $\mathbb{Z}_0^+$, the set of positive integers, $\partial^j_\gamma l$, the partial derivative of order j, $j = 1, 2 \ldots$, of $l(a; \cdot/b)$ with respect to a and $\partial^{jk}_{\gamma\sigma} l$, the partial derivative of order $j + k$, $j, k = 1, 2 \ldots$, of $l(a; \cdot/b)$ with respect to a (order j) and b (order k). Assume that there exist Donsker classes*

i) $\{(x, z) \to \partial^{11}_{\gamma\sigma} l(\gamma_{\theta_0}(\theta_0^T x); (z - u_x(m))/\sigma(x)) : m \in \mathbb{Z}_0^+\} \subset \mathcal{F}_{\partial^{11}}$;

ii) $\{(x, z) \to \partial^j_\gamma l(\gamma_{\theta_0}(\theta_0^T x); (z - u_x(m))/\sigma(x)) : m \in \mathbb{Z}_0^+\} \subset \mathcal{F}_{\partial^j}, j = 1, 2.$

Notice that in the practical case of a constant threshold strategy, Assumption 7 is automatically verified.

Assumptions on the nonparametric estimators and on the regression functions.

We next assume some differentiability and consistency properties for the nonparametric estimators (and their partial derivatives) that we consider.

Assumption 8 *Let $\nabla_\theta m$ denote the gradient vector of partial derivatives with respect to θ of a function m. Assume that $\gamma_\theta(\theta^T x)$ is twice continuously differentiable with respect to θ, and that*

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \left| \hat{\gamma}_{\theta_0}(\theta_0^T x) - \gamma_{\theta_0}(\theta_0^T x) \right| + \sup_x |\hat{\sigma}(x) - \sigma(x)| &= O_P(\eta_n), \\
\sup_{x \in \mathcal{X}} |\nabla_\theta \hat{\gamma}_{\theta_0}(\theta_0, x) - \nabla_\theta \gamma_{\theta_0}(\theta_0, x)| &= O_P(\eta'_n), \\
\sup_{\theta \in \Theta, x \in \mathcal{X}} \left| \nabla^j_\theta \hat{\gamma}_\theta(\theta, x) - \nabla^j_\theta \gamma_\theta(\theta, x) \right| &= o_P(1), \quad \textit{for} \quad j = 1, 2,
\end{aligned}$$

where $\nabla^1_\theta \equiv \nabla_\theta$. Assume that $\eta_n \eta'_n + \eta_n^2 = o(n^{-1/2})$.

These convergence rates are usually required for nonparametric estimators and in the single-index model literature. They are obtained for most of the kernel estimators.

Assumption 9 *Assume that there exist Donsker classes such that*

i) $\gamma_{\theta_0}(\theta_0^T\cdot) \in \mathcal{F}_\gamma$ *and* $\hat{\gamma}_{\theta_0}(\theta_0^T\cdot) \in \mathcal{F}_\gamma$ *with probability tending to one;*

ii) $\nabla_\theta\gamma_{\theta_0}(\theta_0,\cdot) \in \mathcal{F}_\nabla$ *and* $\nabla_\theta\hat{\gamma}_{\theta_0}(\theta_0,\cdot) \in \mathcal{F}_\nabla$ *with probability tending to one.*

Assumption 9 is often checked by adding regularity conditions on $\gamma_{\theta_0}(\theta_0^T\cdot)$ and $\nabla_\theta\gamma_{\theta_0}(\theta_0,\cdot)$ and by proving consistency properties of the estimators involved in these conditions.

Assumption 10 *Assume* $\nabla_\theta\gamma_{\theta_0}(\theta_0,x) = U(\theta_0^Tx)(x-E[X|\theta_0^Tx])$ *for some function* $U(\theta_0^Tx)$.

Assumption 10 is easily checked for $\gamma_\theta^{(2)}$ described above using Lemma 7.1 in the Appendix. This lemma is easily adapted to $\gamma_\theta^{(1)}$ by replacing $m_\theta(\theta^TX)$ by $E[\partial_\gamma l(\gamma_\theta(\theta^TX);Y/\sigma(X))|\theta^TX]$ and by assuming that the function $u \to \gamma_{\theta_0}(u)$ is bounded away from zero and has a continuous first derivative.

The next assumption is required for the part of the asymptotic representation coming from the estimation of $\sigma(\cdot)$. It is usually verified for nonparametric kernel estimators. In our context, we consider that this part, related to $\hat{\sigma}(x)-\sigma(x)$, is of minor interest and do not focus on it. It could be modeled in several different ways (for example, with another single-index assumption); here we simply require the general Assumption 11 hereunder.

Assumption 11 *Assume* $\hat{\sigma}(\cdot)-\sigma(\cdot)$ *belongs to a Donsker class with probability tending to one and admits the following representation*

$$\hat{\sigma}(x)-\sigma(x) = \frac{1}{n}\sum_{j=1}^{m}\nu_n(x,X_j,Z_j-u_{X_j}(m))\mathbf{1}_{Z_j\geq u_{X_j}(m)} + R_n(x),$$

where $\sup_x|R_n(x)| = o_P(n^{-1/2})$. *In addition, the function* $\nu_n(x,X,Z-u_X(m))$ *satisfies*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{m}\int \nu_n(x,X_i,Z_i-u_{X_i}(m))\mathbf{1}_{Z_i\geq u_{X_i}(m)}\varphi(x,w)d\mathbb{P}_{X,W}(x,w)\\
&= \frac{1}{n}\sum_{i=1}^{m}\Big[\nu(X_i,Z_i-u_{X_i}(m))\mathbf{1}_{Z_i\geq u_{X_i}(m)}\\
&\qquad\qquad \times E[\varphi(X_i,W)|X_i,Z\geq u_X(m)]f_{X|Z\geq u_X(m)}(X_i|Z\geq u_X(m))\Big] + R_n^*(x),
\end{aligned}$$

for some function $\nu(X, Z - u_X(m))$ *with* $E[\nu(X, Z - u_X(m))|X = x, Z \geq u_X(m)] = o_P(n^{-1/2})$ *(uniformly in* $X = x$*) and* $E[|\nu(X, Z - u_X(m))|^3|Z \geq u_X(m)] = O(1)$ *and where* $\sup_x |R_n^*(x)| = o_P(n^{-1/2})$, $\mathbb{P}_{X,W}(\cdot,\cdot)$ *is the joint distribution of* $(X, (Z - u_X(m))/\sigma(X))$ *(given* $Z \geq u_X(m)$*),* $f_{X|Z\geq u_X(m)}(\cdot|Z \geq u_X(m))$ *is the probability density function of* X *(given* $Z \geq u_X(m)$*),* $\varphi(x, w)$ *is a uniformly bounded function, differentiable with respect to all the components of* x *up to order two and all its derivatives (with respect to the components of* x*) up to order two are uniformly bounded.*

Assumption 12 i) and iii) is also required when constructing the part of the asymptotic representation related to $\hat{\sigma}(x) - \sigma(x)$. Assumption 12 ii) is required for the convergence of the Hessian of the log-likelihood function (to Σ defined in Assumption 6). This set of assumptions is purely technical and usually assumed in this context.

Assumption 12 *Assume*

i) $\sigma(\cdot)$, $\gamma_{\theta_0}(\theta_0^T\cdot)$ *and* $\nabla_\theta\gamma_{\theta_0}(\theta_0, \cdot)$ *are differentiable with respect to all the components of* x *up to order two and all their derivatives (with respect to the components of* x*) up to order two are bounded.*

ii) For each component $D_l(\theta, x)$ *of* $\nabla_\theta^j\gamma_\theta(\theta, x)$, $j = 0, 1, 2$ *(*$l = 1, \ldots, L$*, where* $L = 1, d, d^2$ *according to the value of* j*),*

$$\sup_{x\in\mathcal{X}} \frac{|D_l(\theta, x) - D_l(\theta', x)|}{\|\theta - \theta'\|} < C < \infty, \tag{3.1}$$

for all $\theta, \theta' \in \Theta$ *and some* $C > 0$ *(*$\|\cdot\|$ *denotes an appropriate norm).*

iii) $f_{X|W}(x|w)$*, the density of* X *given* $(Z - u_X(m))/\sigma(X)$ *(and* $Z \geq u_X(m)$*) is uniformly bounded, differentiable with respect to all the components of* x *up to order two and all its derivatives (with respect to the components of* x*) up to order two are uniformly bounded.*

Assumptions related to the tail behavior.

Finally, assumptions about the behavior of the random variables X_i and Z_i, $i = 1, \ldots, n$, distributions are added; these concern the right tail of the distribution of the Z_i and are usual in the extreme value theory context.

Assumption 13 *Assume that for* $f_{X|Z\geq u_X(m)}(x|Z \geq u_X(m))$,

$$\lim_{u_x(m)\to z_{F,x}} \sup_{x\in\mathcal{X}} |f_{X|Z\geq u_X(m)}(x|Z \geq u_X(m)) - \tilde{f}_X(x)| = 0$$

for a given cumulative distribution function $\tilde{F}_X(x)$ *(with bounded density* $\tilde{f}_X(x)$*).*

Assumption 14 *Assume that*

$$\frac{\delta(\lambda z|x)}{\delta(z|x)} = 1 + \phi(z|x)c(x)\int_1^\lambda u^{\rho(x)-1}du + o(\phi(z|x)),$$

as $z \to \infty$*, for each* $\lambda > 0$*, with* $\phi(z|x) > 0$*,* $\phi(z|x) \to 0$ *as* $z \to \infty$ *and* $\rho(x) \leq 0$*.*

Assumption 15 *Assume that* $\sup_x \phi(u_x|x) = o(n^{-1/2})$ *and* $\sup_x |c(x)| < \infty$*, where* $\phi(\cdot|\cdot)$ *and* $c(\cdot)$ *are defined in Assumption 14.*

We now state the main result of this section. The proof is presented in Section 7.2.

Theorem 3.2 *Under the assumptions of Theorem 3.1, Assumptions 5 to 15 and if, additionally,* $u_x(m)$ *is a strictly increasing function of* m*, for all* x*, and* $m^{-1}p_m^{(v+2)/(v-2)} = o(1)$ *for* $mp_m = m\mathbb{P}(Z \geq u_X(m))$*, a strictly increasing function of* m*, we have*

$$n^{1/2}(\hat{\theta} - \theta_0) \Longrightarrow \mathcal{N}\left(0, \Sigma^{-1}V\Sigma^{-1}\right)$$

as $n \to +\infty$*, where*

$$V = E[\tilde{\eta}_{\theta_0}(X,Y)\tilde{\eta}_{\theta_0}^T(X,Y)],$$

$$\begin{aligned}\eta_{\theta_0}(X, Z - u_X(m)) \;=\; & \nabla_\theta l\left(\gamma_{\theta_0}(\theta_0^T X); \frac{Z - u_X(m)}{\sigma(X)}\right)\mathbf{1}_{Z\geq u_X(m)} \\ & - \frac{\nu(X, Z - u_X(m))\mathbf{1}_{Z\geq u_X(m)}\nabla_\theta\gamma_{\theta_0}(\theta_0, X)f_{X|Z\geq u_X(m)}(X|Z \geq u_X(m))}{\sigma(X)(1 + \gamma_{\theta_0}(\theta_0^T X))(1 + 2\gamma_{\theta_0}(\theta_0^T X))},\end{aligned}$$

and $\tilde{\eta}_{\theta_0}(X,Y)$ *corresponds to* $\eta_{\theta_0}(X,Y)$ *where* $X \sim \tilde{F}_X(x)$ *and* $f_{X|Z\geq u_X(m)}(X|Z \geq u_X(m))$ *is replaced by* $\tilde{f}_X(x)$*.*

In the above Theorem, the variance can be estimated consistently in order to provide asymptotic confidence intervals. For example, if we consider $\sigma(x) = \sigma$ (and a global estimator for σ), the second term of $\eta_{\theta_0}(X, Z - u_X(m))$ disappears. Σ can then be estimated by $\nabla_\theta^2 M_n(\hat{\gamma}_{\hat{\theta}}, \hat{\sigma}; \hat{\theta})$ and V by

$$\frac{1}{n}\sum_{i=1}^m \nabla_\theta l(\hat{\gamma}_{\hat{\theta}}(\hat{\theta}^T X_i); \frac{Z_i - u_{X_i}(m)}{\hat{\sigma}})\nabla_\theta^T l(\hat{\gamma}_{\hat{\theta}}(\hat{\theta}^T X_i); \frac{Z_i - u_{X_i}(m)}{\hat{\sigma}})\mathbf{1}_{Z_i\geq u_{X_i}(m)}.$$

4 Practical implementation

In practice, maximizing M_n is a hard task that requires numerical procedures. In addition, the most classical nonparametric estimators used here (kernel or local polynomial estimators) rely on smoothing parameters, whose choice can strongly impact the practical behavior of the procedure. In Section 4.1, we develop an iterative algorithm to estimate θ_0 and to select an adequate smoothing parameter. Since the initialization of this algorithm is an important step, we provide a simple way to compute a preliminary estimator of θ_0 in Section 4.2. The choice of the threshold $u_X(m)$ in practice is left to Section 5.2.

4.1 An iterative algorithm

According to Theorem 3.2 above, the asymptotic distribution of our estimator $\hat{\theta}$ does not depend on the nonparametric estimator for $\gamma_\theta(\theta^T X)$ (for a given θ). Nevertheless, for a finite sample size, the choice of the bandwidth has a significant impact on the procedure. This is the reason why we introduce an iterative algorithm to compute $\hat{\theta}$ and an appropriate h at the same time.

The procedure is the following: if at stage k of the algorithm, our current estimator is $\hat{\theta}^{(k)}$, we choose the bandwidth $h^{(k+1)}$ that maximizes the criterion $V_{n,k}(h) = M_n(\hat{\gamma}^{-i}_{\hat{\theta}^{(k)},h}, \hat{\sigma}; \hat{\theta}^{(k)})$ (where $\hat{\gamma}^{-i}_{\theta,h}$ denotes the cross-validated version of the nonparametric estimator (2.8), see Beirlant and Goegebeur (2004)), Next, using this selected bandwidth, we update θ by maximizing the criterion $M_n(\hat{\gamma}_{\theta,h^{(k+1)}}, \hat{\sigma}; \theta)$ with respect to θ.

We can summarize the algorithm in the following way:

Step 0: Initialization by some estimator $\hat{\theta}^{(0)}$ (see Section 4.2).

Step k.1: Consider a finite grid of values of $h : \mathcal{H}$. Compute

$$h^{(k)} = \arg\max_{h\in\mathcal{H}} V_{n,k-1}(h).$$

Step k.2: Compute

$$\hat{\theta}^{(k)} = \arg\max_{\theta\in\Theta} M_n(\hat{\gamma}_{\theta,h^{(k)}}, \hat{\sigma}; \theta).$$

Repeat steps k.1 and k.2 until convergence.

Remark 4.1.1 We do not discuss the choice of the bandwidth parameters that could be involved in $\hat{\sigma}(X)$ since they do not consist in the main estimation purpose of our methodology. Nevertheless, the algorithm described below can be easily adapted to also choose these bandwidths.

4.2 Initialization of the algorithm

We propose to take as an initial value $\hat{\theta}^{(0)}$ a preliminary consistent estimator (even with slow convergence rate). Based on a preliminary nonparametric estimator $\tilde{\gamma}(x)$ of $\gamma(x)$, compute

$$\hat{\theta}^{(0)} = \lambda \times \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_x \tilde{\gamma}(X_i) \right),$$

where λ is some normalizing constant that ensures that the absolute norm of $\hat{\theta}^{(0)}$ is equal to one (which is the identifiability condition used later in the simulations and the application). The idea behind this average derivative technique comes from the fact that $\nabla_x \gamma(x) = \theta_0 \gamma'(\theta_0^T x)$ if the model is true. Hence, the empirical mean in the definition of $\hat{\theta}^{(0)}$ is expected to be (almost) colinear with θ_0. The consistency of $\tilde{\gamma}$ should ensure the consistency of the technique. As an estimator $\tilde{\gamma}(x)$ (and $\nabla_x \tilde{\gamma}(x)$) of $\gamma(x)$ (and $\nabla_x \gamma(x)$), one may use an estimator based on the method of (conditional) moments, as for the estimator $\gamma_\theta^{(2)}$, or the local polynomial estimator of Beirlant and Goegebeur (2004) in a multivariate context. However, an initial estimator based on the moments should be favored in practice, since the local polynomial likelihood approach is extremely unstable when $d > 1$. In addition, it requires to estimate jointly $\gamma(x)$ and $\sigma(x)$, which is a pretty hard task. To avoid related numerical issues, we prefer to rely on the method of conditional moments, allowing us to estimate separately both parameters. This is this approach that we will use in the next sections to obtain initial estimations of θ_0 (and of $\sigma(x)$), both in the simulations and in the real data analysis.

5 Simulations and real data analysis

5.1 Simulations

In this section, we study the finite sample behavior of the proposed procedure. Especially, we want to know how our iterative procedure improves an initial estimation $\hat{\theta}^{(0)}$ of θ_0, based on the average derivative technique. We generate $B = 200$ samples of size $n \in \{1000, 1500, 2000\}$ from the GPD, following the single-index model described in the previous section. We specify two different functions $\gamma(x)$. The scale parameter (σ) is assumed to be known, constant and equal to 1. The covariates X are composed of $d \in \{3, 4, 5\}$ components, independently and uniformly distributed on [0, 1]. Analytically

$$Y \sim GPD(\gamma(x), \sigma), \tag{5.1}$$

$$X^{(p)} \sim U(0,1),\ p = 1, \cdots, d, \tag{5.2}$$

where the two different functions are:

Model 1: $\gamma(x) = (\sin(\sin(2\pi\theta_{0,(j)}^T x)) + 1)/7 + 0.1$,

Model 2: $\gamma(x) = (\sin(\sin(2\pi\theta_{0,(j)}^T x)) + 1)/3 + 0.3$,

with $\theta_{0,(1)} = [0.2\ 0.2\ 0.6]$, $\theta_{0,(2)} = [0.1\ 0.2\ 0.4\ 0.3]$, $\theta_{0,(3)} = [0.1\ 0.1\ 0.2\ 0.2\ 0.4]$ and $X^{(p)}$ denotes the p^{th} covariate. For an identification purpose, we impose to the l_1 norm of $\theta_{0,(j)}$ to be equal to 1 (this specific identification condition provided better numerical results). We use the average derivative technique based on a multivariate local polynomial regression of the moments (Ruppert and Wand, 1994) to obtain an initial estimation $\hat{\theta}_{(j)}^{(0)}$ of $\theta_{0,(j)}$ (we prefer to use the moment-based approach over the local polynomial likelihood approach of Beirlant and Goegebeur (2004), since we find the latter highly unstable when $d > 1$ and the source of numerical issues). We compare the quality of the estimation obtained with the proposed iterative procedure ($\hat{\theta}_{(j)}$) to this initial solution (in the latter, a subscript b is added to identify an estimator of θ_0 computed on sample b). In our iterative procedure, we use the local polynomial likelihood procedure of Beirlant and Goegebeur (2004), as well as their proposed leave-one-out cross-validation approach to

select the bandwidth. The estimator $\hat{\gamma}_{\theta,h}$ of γ_{θ_0} used here is defined by equation (2.8). To measure the quality of the estimation, we use the mean squared error ($MSE_{\theta_{(j)}}$) and the mean absolute error ($MAE_{\theta_{(j)}}$) criteria. These quantities are computed in the following way:

$$MSE_{\hat{\theta}_{(j)}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{p=1}^{d_j}(\hat{\theta}_{(j),b}(p) - \theta_{0,(j)}(p))^2, \tag{5.3}$$

$$MAE_{\hat{\theta}_{(j)}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{p=1}^{d_j}|\hat{\theta}_{(j),b}(p) - \theta_{0,(j)}(p)|, \tag{5.4}$$

for $j = 1, 2, 3$ and where $\theta_{0,(j)}(p)$ is the p^{th} element of the vector $\theta_{0,(j)}$.

To obtain indicators less sensitive to extreme values, we also compute the median squared error and the median absolute error, given by

$$mSE_{\hat{\theta}_{(j)}} = Md\left(\sum_{p=1}^{d_j}(\hat{\theta}_{(j),b}(p) - \theta_{0,(j)}(p))^2\right), b = 1, \cdots, B, \tag{5.5}$$

$$mAE_{\hat{\theta}_{(j)}} = Md\left(\sum_{p=1}^{d_j}|\hat{\theta}_{(j),b}(p) - \theta_{0,(j)}(p)|\right), b = 1, \cdots, B, \tag{5.6}$$

for $j = 1, 2, 3$ and where Md stands for empirical median.

We measure the quality of the final estimation $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ of $\gamma(x)$ and we compare it to the initial estimation performed with $\hat{\theta}^{(0)}$. To this end, we use the mean and median integrated squared error criteria, computed in the following way:

$$MISE_{\hat{\gamma}_\theta,(j)} = \frac{1}{B\cdot n}\sum_{b=1}^{B}\sum_{i=1}^{n}(\hat{\gamma}_{\theta,h}(\hat{\theta}_{(j),b}^T x_{j,b,i}) - \gamma(x_{j,b,i}))^2, \tag{5.7}$$

$$mISE_{\hat{\gamma}_\theta,(j)} = Md\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_{\theta,h}(\hat{\theta}_{(j),b}^T x_{j,b,i}) - \gamma(x_{j,b,i}))^2\right), b = 1, \cdots, B, \tag{5.8}$$

for $j = 1, 2, 3$ and $n = 1000, 1500, 2000$. Having at our disposal B random vectors $\{y_i, x_i\}_{1\leq i\leq n}$ following (5.2) and (5.1), $x_{j,b,i}$ denotes the i^{th} vector of covariates in the b^{th} simulated sample where $\theta_0 = \theta_{0,(j)}$; whereas $\hat{\gamma}_{\theta,h}(\hat{\theta}_{(j),b}^T x_{j,b,i})$ is the final estimation of $\gamma(x_{j,b,i})$ in sample b obtained with $\hat{\theta}_{(j),b}$. To measure the quality of the initial estimation, we obtain similar quantities simply by replacing $\hat{\theta}_{(j),b}$ by $\hat{\theta}_{(j),b}^{(0)}$ (the initial estimation of the single-index parameter in the b^{th} sample).

The bandwidths used to obtain the initial solution are selected with a leave-one-out cross-validation procedure over a grid between 0.1 and 0.5. We select the bandwidths $h_0^{(1)}$ and $h_0^{(2)}$ that minimize a least-square criterion:

$$h_0^{(1)} = arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} (y_i - \hat{m}_{h,-i}^{(1)}(x_i))^2, \tag{5.9}$$

$$h_0^{(2)} = arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} (y_i^2 - \hat{m}_{h,-i}^{(2)}(x_i))^2, \tag{5.10}$$

where $\hat{m}_{h,-i}^{(1)}(x_i)$ (respectively $\hat{m}_{h,-i}^{(2)}(x_i)$) is the cross-validated (multivariate) local polynomial estimation of the first (respectively second) conditional moment at point x_i, computed with a bandwidth $h \in \mathcal{H}$ and omitting the observation $\{y_i, x_i\}$.

At each iteration of the proposed procedure, we also choose the bandwidth of the local polynomial regression with the leave-one-out cross-validation procedure of Beirlant and Goegebeur (2004), using a grid of values depending on the range of $\hat{\theta}_{(j)}^{(k)T} x$. The kernel function is the bi-quadratic kernel function, as in Goegebeur et al. (2014) (this assumption does not influence much the final result) and is given by

$$K(x) = \frac{15}{16}(1 - x^2)^2 \mathbb{1}\{x \in [-1, 1]\}. \tag{5.11}$$

We perform a maximum of 20 iterations. Our final estimator $\hat{\theta}_{(j)}$ of $\theta_{0,(j)}$ is the one that maximizes the global likelihood function given by equation (2.4). To compare our procedure to the average derivative technique, we compute $RMSE_{\theta_{(j)}} = MSE_{\hat{\theta}_{(j)}}/MSE_{\theta_{(j)}^{(0)}}$, where $MSE_{\hat{\theta}_{(j)}}$ is the estimated MSE obtained with our procedure and $MSE_{\theta_{(j)}^{(0)}}$ is the MSE obtained with the average derivative technique. A ratio below 1 indicates that our procedure provides better results. Similarly, we compute $RmSE_{\theta_{(j)}}$, $RMAE_{\theta_{(j)}}$, $RmAE_{\theta_{(j)}}$, $RMISE_{\hat{\gamma}_{\hat{\theta}},(j)}$ and $RmISE_{\hat{\gamma}_{\hat{\theta}},(j)}$. The results are displayed in Tables 1 and 2.

For the first model (Table 1), we see that our iterative procedure is the best for all tested values of θ_0 and all considered criteria. The criteria based on the median emphasize the presence of several large errors (both with the average derivative method and our method) for $n = 1000$. It appears that when the average derivative method provides a very bad starting solution, we have some difficulties in improving a lot the final estimation of θ_0. Thus, controlling for these errors improves the results in favour of our procedure, as shown by the median criteria. When the sample size increases, the gap between the

average derivative technique and our procedure tends to increase. For a sample of size $n = 1000$, we observe a decrease of the MSE ranging from 5.4% to 16.7%. For $n = 1500$, the decrease varies between 19.4% and 27.3%. For $n = 2000$, the decrease ranges from 23.3% to 28.6%. The decrease is larger if we look at the mSE criterion. Similar figures and effects are observed for the MAE and mAE criteria (although the decreases related to these criteria are weaker). Looking at the criteria on $\gamma(x)$, we also perform better with our procedure, except for $n = 1000$ and $d = 5$ where there is no difference. When the sample size increases, we achieve a reduction of the MISE criterion up to 21%, whereas the mISE criterion decreases by up to 26%.

For the second model (Table 2), initial estimations provided by the average derivative technique are less good. Simultaneously, our final estimations display smaller error rates compared to the first model, with the consequence that the ratios are closer to zero. Considering the MSE criterion only, we observe a maximum improvement of 68.7%, 78.1% and 69.6% for the three different sets of covariates, respectively. The best improvement in term of MAE reaches an 85.6% decrease, compared to the estimation with the average derivative technique. Looking at the criteria on $\gamma(x)$, we are way better than the initial estimation. Our procedure enables a decrease of the MISE and mISE between 36% and 56% for $n = 1000$. When the sample size increases, the decreases are between 50% and 63% for $n = 1500$, and between 60% and 71% for $n = 2000$.

Hence, the proposed procedure improves well the initial estimation of θ_0. We notice that if the initial parameters are too far from the true parameters, we may have some difficulties in improving a lot this initial solution. Our procedure also improves the estimation of $\gamma(x)$, especially for the second DGP and when the sample size is larger than 1000. Eventually, notice that the good results for Model 2, compared to the initial estimation, are partly due to the fact that the initial solution based on local polynomial estimations of the second moment is biased when $\gamma(x) > 1/2$ (which is the case for some values of $\gamma(x)$ in the simulated samples). We could have used other local regression techniques that do not suffer from this drawback (e.g. local quantile regression or a multivariate version of Beirlant and Goegebeur (2004) estimator). However, as mentioned earlier, these approaches suffer from practical and theoretical issues (see, e.g. Zhang, 2010), that are beyond the scope of this paper and make them hard to implement.

Model 1

$n = 1000$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.204	0.245	0.833	0.154	0.174	0.885	0.175	0.185	0.946
$MAE_{\hat{\theta}_{(j)}}$	0.579	0.644	0.899	0.615	0.654	0.940	0.723	0.758	0.954
$mSE_{\hat{\theta}_{(j)}}$	0.110	0.151	0.729	0.104	0.130	0.800	0.120	0.156	0.769
$mAE_{\hat{\theta}_{(j)}}$	0.476	0.573	0.831	0.562	0.607	0.926	0.666	0.736	0.905
$MISE_{\hat{\gamma}_{\theta},(j)}$	$5.7e^{-3}$	$6.1e^{-3}$	0.94	$6.3e^{-3}$	$6.5e^{-3}$	0.97	$7.3e^{-3}$	$7.4e^{-3}$	0.99
$mISE_{\hat{\gamma}_{\theta},(j)}$	$5.1e^{-3}$	$5.6e^{-3}$	0.90	$5.4e^{-3}$	$5.7e^{-3}$	0.96	$6.8e^{-3}$	$6.8e^{-3}$	1.00
$n = 1500$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.149	0.205	0.727	0.121	0.161	0.751	0.124	0.153	0.806
$MAE_{\hat{\theta}_{(j)}}$	0.479	0.576	0.832	0.521	0.607	0.859	0.615	0.687	0.895
$mSE_{\hat{\theta}_{(j)}}$	0.079	0.125	0.632	0.080	0.116	0.685	0.100	0.132	0.757
$mAE_{\hat{\theta}_{(j)}}$	0.400	0.504	0.794	0.487	0.580	0.840	0.573	0.657	0.872
$MISE_{\hat{\gamma}_{\theta},(j)}$	$3.7e^{-3}$	$4.7e^{-3}$	0.79	$4.4e^{-3}$	$4.9e^{-3}$	0.91	$5.3e^{-3}$	$5.7e^{-3}$	0.94
$mISE_{\hat{\gamma}_{\theta},(j)}$	$3.3e^{-3}$	$4.1e^{-3}$	0.80	$3.6e^{-3}$	$4.4e^{-3}$	0.82	$4.7e^{-3}$	$5.4e^{-3}$	0.88
$n = 2000$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.140	0.196	0.714	0.091	0.127	0.720	0.099	0.129	0.767
$MAE_{\hat{\theta}_{(j)}}$	0.459	0.550	0.834	0.448	0.532	0.842	0.547	0.623	0.878
$mSE_{\hat{\theta}_{(j)}}$	0.064	0.099	0.647	0.053	0.075	0.668	0.077	0.110	0.700
$mAE_{\hat{\theta}_{(j)}}$	0.400	0.445	0.899	0.396	0.466	0.849	0.506	0.606	0.835
$MISE_{\hat{\gamma}_{\theta},(j)}$	$3.2e^{-3}$	$4.1e^{-3}$	0.80	$3.5e^{-3}$	$4.1e^{-3}$	0.87	$3.9e^{-3}$	$4.4e^{-3}$	0.89
$mISE_{\hat{\gamma}_{\theta},(j)}$	$2.7e^{-3}$	$3.4e^{-3}$	0.81	$2.9e^{-3}$	$3.9e^{-3}$	0.74	$3.4e^{-3}$	$4.1e^{-3}$	0.83

Table 1: Values of the various error rates and the ratio statistics, obtained with the average derivative technique and our iterative procedure, for Model 1.

Model 2									
$n = 1000$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.122	0.276	0.443	0.103	0.233	0.441	0.118	0.221	0.533
$MAE_{\hat{\theta}_{(j)}}$	0.411	0.713	0.576	0.473	0.758	0.623	0.533	0.829	0.692
$mSE_{\hat{\theta}_{(j)}}$	0.059	0.199	0.297	0.060	0.191	0.314	0.075	0.201	0.371
$mAE_{\hat{\theta}_{(j)}}$	0.369	0.685	0.539	0.413	0.725	0.569	0.496	0.804	0.618
$MISE_{\hat{\gamma}_\theta,(j)}$	$1.4e^{-2}$	$2.8e^{-2}$	0.50	$1.6e^{-2}$	$2.8e^{-2}$	0.58	$1.9e^{-2}$	$3e^{-2}$	0.64
$mISE_{\hat{\gamma}_\theta,(j)}$	$1.2e^{-2}$	$2.6e^{-2}$	0.44	$1.3e^{-2}$	$2.7e^{-2}$	0.49	$1.7e^{-2}$	$3e^{-2}$	0.56
$n = 1500$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.119	0.291	0.411	0.068	0.208	0.326	0.072	0.206	0.352
$MAE_{\hat{\theta}_{(j)}}$	0.381	0.717	0.531	0.372	0.709	0.525	0.459	0.804	0.570
$mSE_{\hat{\theta}_{(j)}}$	0.039	0.186	0.211	0.031	0.179	0.171	0.052	0.177	0.294
$mAE_{\hat{\theta}_{(j)}}$	0.315	0.649	0.486	0.302	0.699	0.432	0.423	0.788	0.537
$MISE_{\hat{\gamma}_\theta,(j)}$	$1.1e^{-2}$	$2.6e^{-2}$	0.41	$1.2e^{-2}$	$2.5e^{-2}$	0.47	$1.3e^{-2}$	$2.6e^{-2}$	0.50
$mISE_{\hat{\gamma}_\theta,(j)}$	$0.8e^{-2}$	$2.3e^{-2}$	0.37	$0.98e^{-2}$	$2.5e^{-2}$	0.39	$1.2e^{-2}$	$2.6e^{-2}$	0.45
$n = 2000$	$\hat{\theta}_{(1)}$	$\hat{\theta}^{(0)}_{(1)}$	Ratio(1)	$\hat{\theta}_{(2)}$	$\hat{\theta}^{(0)}_{(2)}$	Ratio(2)	$\hat{\theta}_{(3)}$	$\hat{\theta}^{(0)}_{(3)}$	Ratio(3)
$MSE_{\hat{\theta}_{(j)}}$	0.072	0.231	0.313	0.042	0.208	0.193	0.059	0.189	0.314
$MAE_{\hat{\theta}_{(j)}}$	0.299	0.664	0.451	0.281	0.705	0.399	0.401	0.759	0.528
$mSE_{\hat{\theta}_{(j)}}$	0.025	0.184	0.134	0.022	0.177	0.124	0.037	0.156	0.236
$mAE_{\hat{\theta}_{(j)}}$	0.245	0.600	0.408	0.251	0.697	0.360	0.355	0.714	0.497
$MISE_{\hat{\gamma}_\theta,(j)}$	$0.95e^{-2}$	$2.5e^{-2}$	0.38	$0.84e^{-2}$	$2.4e^{-2}$	0.35	$1.1e^{-2}$	$2.4e^{-2}$	0.44
$mISE_{\hat{\gamma}_\theta,(j)}$	$0.77e^{-2}$	$2.2e^{-2}$	0.35	$0.71e^{-2}$	$2.4e^{-2}$	0.29	$0.94e^{-2}$	$2.3e^{-2}$	0.40

Table 2: Values of the various error rates and the ratio statistics, obtained with the average derivative technique and our iterative procedure, for Model 2.

5.2 Real data analysis

In this section, we illustrate our methodology on a database of operational losses from the Italian bank UniCredit. These losses are defined as being *losses resulting from inadequate or failed internal processes, people and systems or from external events (...)* (Basel Committee on Banking Supervision, 2004). For regulatory purposes, banks are asked to

set aside a capital reserve to cover themselves against these losses. This capital reserve is a function of the 99.9^{th} order statistic of their severity distribution, usually modeled with a GPD in the Advanced Measurement Approach (AMA, Basel Committee on Banking Supervision, 2004). Traditionally, banks assume that this distribution is independent of the economic conditions but recent studies (Chernobai et al., 2011; Cope et al., 2012; Wang and Hsu, 2013; Chavez-Demoulin et al., 2015) suggest otherwise. Hence, the use of a GP-regression model in this context would be a natural extension of the traditional models and would be of interest to improve the adequacy of these capital reserves.

Our database consists of 3,862 operational losses recorded between 2005 and 2014. According to UniCredit internal classification system, these losses are related to the risk classes *Employment practices & Workplace safety* (EPWS), *Damages to Physical Assets* (DPA) and *Business Disruption & System Failures* (BDSF). We assume that (2.3), (2.9) and $\gamma(x) > 0$ hold so that the GPD can be considered as a valid approximation of the loss excess distribution $(Z - u_X(m))$ (see, e.g. Davison and Smith, 1990; Embrechts et al., 1997, for a similar approach). We use a value of 25,000€ as an estimate for the threshold parameter $u_X(m)$ (denoted $\hat{u}$ in the sequel) and assume that it does not depend on the covariates. We chose this value thanks to a mean excess plot that appears to be linear starting from 25,000€. See Scarrot and MacDonald (2012) for more considerations regarding the threshold selection. It gives us a final number of 585 losses above $\hat{u}$. Figure 1 shows the losses over time, whereas in Appendix 7.4, Table A1 gives descriptive statistics for this sample of losses.

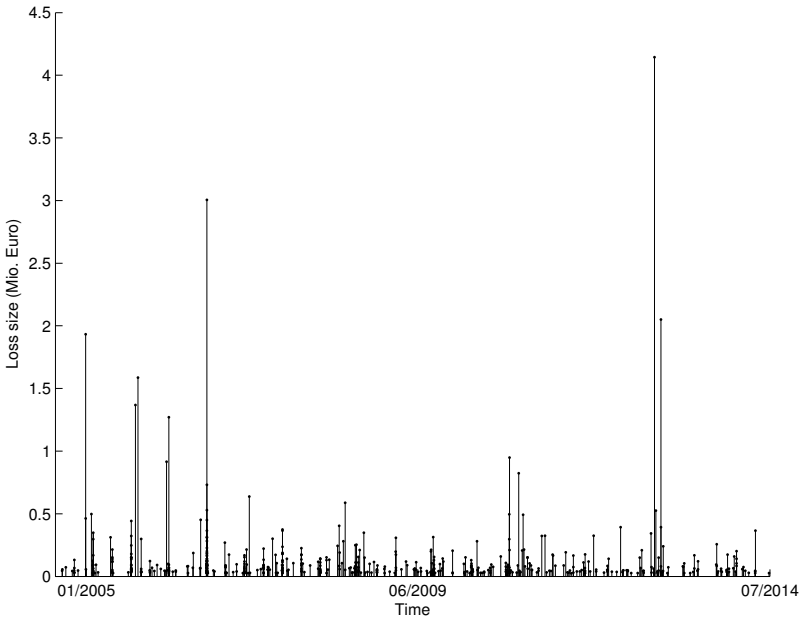

Figure 1: Size of the losses larger that 25,000€, for the period 01/2005-07/2014.

In addition to the sub-classification, we have at our disposal the following covariates: the percentage of the bank's revenue that comes from fees (PRF), the stock price of the bank, Thomson Reuters (TR) European stock index value, the long term (LT) bond rate and the Italian unemployment rate. These covariates are measured at a quarterly frequency. The PRF gives us the proportion of the revenue that does not come from taking a financial risk, but from executing an operation on behalf of its clients. It might be seen as a measure of the economic well-being of the bank: a high PRF indicates that the bank makes profit independently from the level of the interest rates. In the same idea, the stock price of the bank is another measure of its economic well-being (a high stock price indicates a good economic situation for the bank). Besides, Thomson Reuters index and the unemployment rate are measures of the overall performance of an economy. Similarly, high long term interest rates might indicate an increase in the perceived default risk and a defavourable economic situation. As noticed by Cope et al. (2012), high unemployment rates may have an impact on the quality of a bank's staff and may increase the overall crime rate of an economy, suggesting that the severity distribution of the considered operational losses may exhibit fatter tails in this situation. In addition, the same authors

notice that a booming economy may generate incentives for employees to commit frauds, thus increasing the probability of large operational losses.

In Appendix 7.4, Figure A1 shows the distributions of these covariates, whereas Table A2 displays the correlation matrix between covariates. We observe a strong positive correlation between the PRF and the stock price, as well as between the stock price and the TR index. On the other side, we observe that the LT bond rate is strongly negatively correlated with the PRF and the stock index. Lastly, the unemployment rate has a strong negative correlation with the PRF and the stock price. These high correlations between explanatory variables suggest that combining them into a smaller number of covariates might be a good way to perform a dimension reduction, not loosing too much of their explanatory powers in the process.

Similarly to what we do in Section 5.1, we obtain an initial estimation of θ_0 using the average derivative technique, with multivariate local polynomial estimations of the first and second conditional moments (Hristache et al., 2001). We choose the bandwidth parameter over a grid of values between 0.1 and 0.5, with a leave-one-out cross-validation procedure (to ease the selection, we scale all covariates to ensure that they vary between zero and one). More precisely, we select the bandwidth parameters that minimize the sum of squared differences between the cross-validated estimations of the first (respectively second) moment and the observed losses (respectively squared losses).

To obtain an initial estimation for the scale parameter $\sigma(x)$, we use the fact that (for $\gamma(x) < 1/2$)

$$\sigma(x) = \frac{m_1(x)}{2}\left(1 + \frac{m_1(x)^2}{m_2(x) - m_1(x)^2}\right), \tag{5.12}$$

where $m_1(x)$ and $m_2(x)$ are the conditional first and second moments of the excess loss $(Z - u_X(m))$. We estimate $m_1(x)$ and $m_2(x)$ with the same local polynomial estimators $\hat{m}_1$ and $\hat{m}_2$ used for the initial estimation of θ_0 (Hristache et al., 2001), and then we plug the estimates in equation 5.12 to obtain an estimated conditional scale parameter $\hat{\sigma}(x)$. Since we observe that the scale parameter has estimations close to each other in a small interval, we assume it constant across covariates and estimate it by

$$\hat{\sigma}^{(0)} = (1/m)\sum_{i=1}^{m}\hat{\sigma}(x_i). \tag{5.13}$$

This estimation is biased if $\gamma(x) > 1/2$.

To improve this initial guess, we re-estimate σ along the iterative procedure. We maximize, with respect to σ, the log-likelihood function given by equation (2.4) and where the tail indices are set to their estimated values at iteration k ($\hat{\gamma}_{\theta,h}(\hat{\theta}^{(k)T}x_i)$, for $i = 1, \cdots, 585$). Our final estimation is denoted $\hat{\sigma}^{FIN}$ and is obtained with the same maximization procedure, using $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ (our final estimation of $\gamma(x)$) in the likelihood function.

We perform a maximum of 20 iterations of the proposed procedure and constraint the l_1 norm of θ to be equal to 1. For the nonparametric estimation of the conditional tail index, we rely on the estimator of Beirlant and Goegebeur (2004). The bandwidth parameter is chosen with a leave-one-out cross-validation procedure, over a grid depending on the single-index range. The selected h is the one that minimizes the global likelihood function, where we replace the conditional tail index by its cross-validated estimate (as in Beirlant and Goegebeur, 2004).

We consider two combinations of the covariates. In the first combination, we pool all risk categories and we perform the estimation with respect to the economic covariates. Table 3 displays the initial ($\hat{\theta}^{(0)}$) and final ($\hat{\theta}$) estimations of θ_0, the estimated scale parameter and the selected bandwidth. We also compute the (Pearson) correlation coefficients between the single-index variable and the different covariates ($\hat{\rho}(\hat{\theta}^T x, x^{(p)}), p = 1, \cdots, 5$). This quantity indicates the intensity of the relationship between each covariate and the single-index variable. Figure 2 shows the estimated conditional tail index, as a function of the single-index variable.

In the first model (see Table 3), the PRF and the LT bond rate appear to be strongly correlated with the single-index variable. The PRF has the strongest correlation coefficient (-0.824). The stock price appears to be correlated with the single-index too, but the coefficient is smaller. The TR index and the unemployment rate exhibit very small correlation coefficients. Regarding the signs of $\hat{\theta}$ and $\hat{\rho}(\hat{\theta}^T x, x^{(p)})$ associated to the PRF and the stock price, we observe negative values (for both $\hat{\theta}$ and $\hat{\rho}(\hat{\theta}^T x, x^{(p)})$). Due to the shape of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ (Figure 2, left side), it indicates that an increase in these variables (all other things remaining equal) is associated with an increasing probability of very large losses.

Besides, we observe positive values of $\hat{\theta}$ and $\hat{\rho}(\hat{\theta}^T x, x^{(p)})$ for the LT bond rate, meaning that high values of the LT bond rate are associated with high values of the single-index variable, thus with low values of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$. Hence, high values of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ are associated with positive internal economic indicators (high PRF and high stock prices, low LT bond rates), whereas low values of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ are associated with negative indicators (high LT bond rates, low stock prices and low PRF). The considered macroeconomic covariates don't seem to have important explanatory powers.

In the second model, we consider simultaneously the effect of the economic covariates exhibiting the highest correlation coefficients (namely the PRF, the stock price and the LT bond rate) and the risk categories. We expect the risk categories to posses some explanatory powers, since regulators recommend to model separately the severity distribution of the losses from different categories (Basel Committee on Banking Supervision, 2004). We map the EPWS, DPA and BDSF categories into two binary variables (CAT1 and CAT2). CAT1 (resp. CAT2) takes value 1 when the loss belongs to the EPWS (resp. DPA) category, 0 otherwise. Table 3 displays the results of the regression analysis using these five covariates, whereas Figure 2 (right side) shows the estimated conditional tail index as a function of the single-index variable. The PRF and the LT bond rate variables appear to have the strongest correlation coefficients with the single-index (-0.426 and 0.497, respectively), whereas the CAT1 (i.e. EPWS) variable displays also an interesting correlation coefficient (-0.367). A negative sign indicates that high values of the considered variable are associated with negative values of the single-index variable, thus with high values of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ (as shown on Figure 2, right side). In other words, when the PRF is high, the LT bond rate is low and/or that a loss belongs to the EPWS category, the severity distribution has a fatter tail compared to low PRF, high LT bond rates and losses belonging to DPA or BDSF categories. These results are in line with those obtained with the first set of covariates. Besides, the stock price and the CAT2 (i.e DPA) variables display small correlation coefficients, suggesting that their explanatory power is small and that most of the available information are contained in the other variables.

For inference, we compute bootstrap confidence intervals for θ_0 and ρ. We use the estimated GP-regression model to generate $B = 2000$ resamples of size 585 (with fixed covariates). Then, we execute the iterative procedure to obtain bootstrap estimations $\hat{\theta}_b^*$,

$\hat{\sigma}_b^*$ and $\rho((\hat{\theta}_b^*)^T x, x^{(p)})$, with $b = 1, \cdots, B$ and $p = 1, \cdots, 5$. Applying first the Fisher z-transform to ensure that the bounds of the confidence intervals lie between -1 and 1, we are able to compute basic bootstrap confidence intervals. Table 3 displays the bounds of the 95% confidence intervals. We conclude that both the PRF and the LT bond rate have θ_0 and ρ parameters significantly different from zero. In the second model, we find in addition that the stock price has a single-index parameter significantly different from zero (but not its correlation coefficient). We cannot conclude anything for the categorical variables. We provide confidence intervals on σ in the same table.

To check the goodness-of-fit of the estimated models, we compare the empirical distribution of

$$e_i = \left(1/\hat{\gamma}_{\theta,h}(\hat{\theta}^T x_i)\right) \log\left(1 + \hat{\gamma}_{\theta,h}(\hat{\theta}^T x_i)(z_i - \hat{u})/\hat{\sigma}^{FIN}\right), i = 1, \cdots, 585, \tag{5.14}$$

with the standardized exponential distribution (see Chavez-Demoulin et al., 2015, for a similar approach). The QQ-plots displayed in Figure 3 show the good fits of the estimated models, even far in the tail (up to the quantile 99%). Model 2 seems a bit better, presumably because we take into account the category effect.

In summary, both GP-regression exercises indicate that low LT bond rates and high PRF are associated with high values of the tail index, whereas high long term bond rates and small PRF values correspond to smaller values of $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$. Also, the EPWS losses severity distribution has a fatter tail than losses from the DPA and BDSF categories (all other things remaining equal). It is in line with the higher kurtosis coefficient observed for the marginal distribution of EPWS excess losses (see Table A1). The relationship between the stock price and $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$ is more difficult to determine: at first look, an increase in stock price is associated with an increase in $\hat{\gamma}_{\theta,h}(\hat{\theta}^T x)$, but if we control for the risk category its influence mostly disappears. Macroeconomic variables do not display important explanatory powers.

Overall, these results suggest a positive link between economic well-being and the severity of operational losses, and are in line with the ones of Cope et al. (2012) (however, in our case, we observe such a relationship for the likelihood of large losses, and not for their expected size). They also highlight the need of flexible models, to investigate the nonlinear nature of the relationship between $\gamma(x)$ and the covariates. Nevertheless, this

application must be seen as an illustration of the potential of our statistical approach (e.g. as a preliminary step in the perspective of choosing a nonlinear parametric model), and not as a comprehensive empirical study, for which several issues should still be carefully considered in the semiparametric context (e.g. variable selection, model comparison and confidence interval issues).

Set of covariates 1

Stat.	PRF	Stock Price	TR index	LT bond	Unemp.
$\hat{\theta}^{(0)}$	-0.252	-0.135	0.268	-0.110	-0.235
$\hat{\theta}$	-0.534^{**}	-0.010	0.147	0.187^{*}	-0.123
	(-0.86, -0.40)	(-0.22, 0.29)	(-0.06, 0.52)	(-0.00, 0.61)	(-0.47,0.26)
$\rho(\hat{\theta}^T X, X^{(p)})$	-0.824^{**}	-0.354	0.042	0.692^{**}	0.035
	(-0.99,-0.53)	(-0.91, 0.60)	(-0.63, 0.73)	(0.43, 0.98)	(-0.75, 0.74)
	$h^{opt} = 0.123$		$\hat{\sigma}^{FIN} = 41283.3$ (34864, 46954.9)		

Set of covariates 2

Stat.	PRF	Stock Price	LT bond	CAT1	CAT2
$\hat{\theta}^{(0)}$	-0.451	0.120	0.348	-0.070	-0.011
$\hat{\theta}$	-0.356^{**}	0.131^{*}	0.331^{**}	-0.084	-0.098
	(-0.75, -0.26)	(-0.03, 0.37)	(0.12, 0.61)	(-0.22, 0.16)	(-0.44, 0.21)
$\rho(\hat{\theta}^T X, X^{(p)})$	-0.426^{*}	0.055	0.497^{**}	-0.366	-0.139
	(-0.89, 0.04)	(-0.69, 0.70)	(0.03, 0.9)	(-0.81, 0.65)	(-0.88, 0.54)
	$h^{opt} = 0.101$		$\hat{\sigma}^{FIN} = 42347.2$ (35994.7, 47993.6)		

Table 3: Initial and final estimations of θ_0, empirical correlation coefficients $\hat{\rho}(\hat{\theta}^T x, x^{(d)})$ and final estimate $\hat{\sigma}^{FIN}$ of σ, conditional on the first (top) and second (bottom) sets of covariates. h^{opt} is the cross-validated bandwidth used to perform the final regression. * (resp. **) indicates that the parameter is significantly different from 0 with a confidence level of 90% (resp. 95%). Bounds of the 95% confidence intervals are in parentheses.

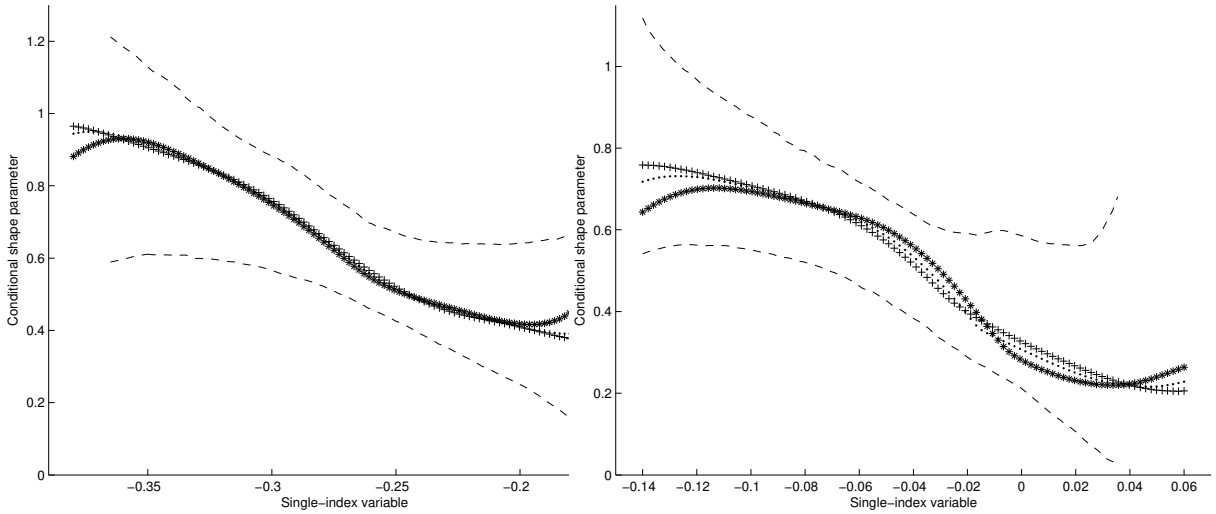

Figure 2: Estimated conditional tail index ($\hat{\gamma}(\hat{\theta}^T x)$) using the first (left) and second (right) set of covariates, as a function of the single-index variable. Dotted: estimation performed with h^{opt}. $*$ (respectively $\times$): estimations performed with $h = 0.9h^{opt}$ (respectively $h = 1.1h^{opt}$). Dashed: 95% bootstrap confidence bands obtained with $\hat{\theta}$ and h^{opt} ($B = 2000$).

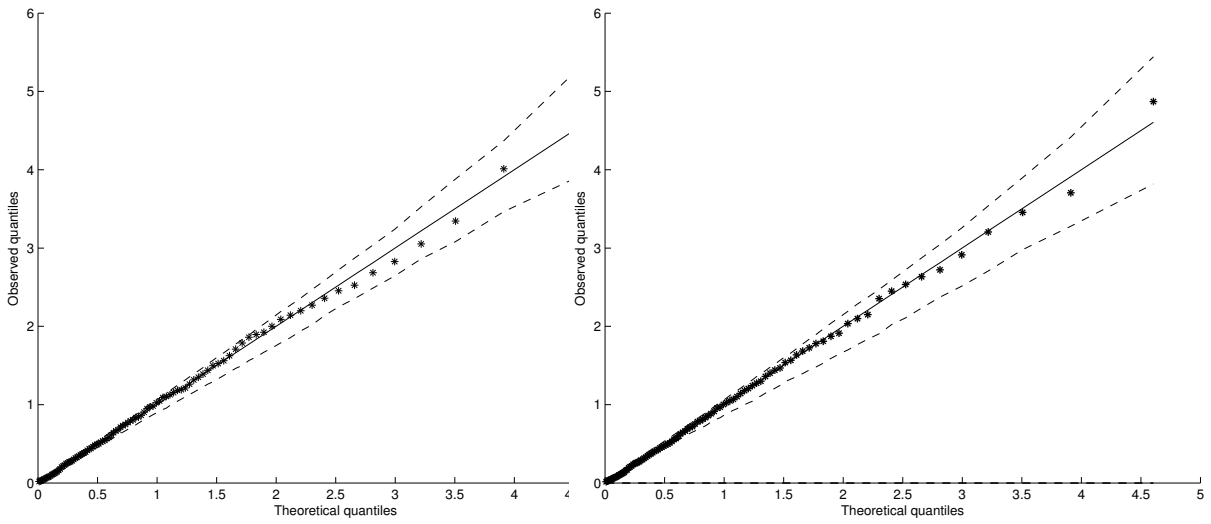

Figure 3: QQ-plot between the quantile of the standard exponential distribution (solid line) and the observed quantiles (*) of $e_i, i = 1, \cdots, 585$ (given by equation 5.14) for Model 1 (left) and Model 2 (right). Dashed: 95% bootstrap confidence intervals ($B = 2000$).

6 Conclusion

In this paper, we provide a new regression model for inference on the conditional tail index $\gamma(x)$ of a Generalized Pareto distribution. This model is based on a dimension reduction throughout a single-index assumption that makes it particularly valuable when the number of covariates is high. In this framework, purely nonparametric approaches usually have an erratic behavior, while purely parametric ones are often too rough to describe correctly the data. We propose a likelihood-based iterative procedure that makes the computation of the estimators easier. We provide asymptotic properties of the single-index parameter estimator under general assumptions and a simulation study investigates the finite sample behavior of this procedure. Lastly, we conduct a regression analysis with our methodology, on a novel database of financial operational losses from the bank UniCredit. Our results suggest that an improvement in the economic conditions increases the probability of large losses.

As already mentioned, several extensions of our approach may be proposed. First of all, one could also model the conditional parameter $\sigma(x)$ as a single-index model. Also $\gamma(x)$ itself could be modeled through a multiple index regression model, as it has been proposed in Chiou and Müller (2004) for mean-regression, the only limitation being, of course, that adding too many indices to the model will increase excessively the complexity of the problem. Lastly, an automatic variable selection procedure in this semiparametric context is obviously of prime interest so as to conduct more thorough empirical studies.

Acknowledgments

J. Hambuckers acknowledges the support of the Belgian National Fund for Scientific Research (FNRS) with a research fellow grant, and of the Research Training Group 1644 Scaling Problems in Statistics funded by the German Research Foundation (DFG).

C. Heuchenne acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

The authors warmly thank Dr. Fabio Piacenza (UniCredit ORM leader), for providing them the UniCredit data.

7 Appendix

In this section, we provide the proofs of Theorems 3.1 and 3.2 in Section 7.1 and 7.2. To simplify notations, we will use $Z_i^m = Z_i - u_{X_i}(m)$, $i = 1, \dots, m$.

7.1 Proof of Theorem 3.1

To prove consistency, it suffices to prove that $\sup_{\theta \in \theta} |M_n(\hat{\gamma}_\theta, \hat{\sigma}; \theta) - M(\gamma_\theta, \sigma; \theta)| = o_P(1)$, since θ_0 is defined as the unique maximizer of M from Assumption 1. This can be done

in two steps, showing that

$$\sup_{\theta\in\theta} |M_n(\hat{\gamma}_\theta, \hat{\sigma}; \theta) - M_{u_x}(\gamma_\theta, \sigma; \theta)| = o_P(1) \tag{7.1}$$

and

$$\sup_{\theta\in\theta} |M_{u_x}(\gamma_\theta, \sigma; \theta) - M(\gamma_\theta, \sigma; \theta)| = o_P(1). \tag{7.2}$$

The second equality is straightforward using expression (A.2) in Beirlant and Goegebeur (2004) and Assumption 4, and the first one is decomposed in

$$\sup_{\theta\in\theta} |M_n(\hat{\gamma}_\theta, \hat{\sigma}; \theta) - M_n(\gamma_\theta, \sigma; \theta)| = o_P(1) \tag{7.3}$$

and

$$\sup_{\theta\in\theta} |M_n(\gamma_\theta, \sigma; \theta) - M_{u_x}(\gamma_\theta, \sigma; \theta)| = o_P(1). \tag{7.4}$$

Equation (7.4) is a direct consequence of Assumption 2, while for (7.3), we write

$$\begin{aligned} M_n(\hat{\gamma}_\theta, \hat{\sigma}; \theta) - M_n(\gamma_\theta, \sigma; \theta) &= \frac{1}{n}\sum_{i=1}^{m} \mathbf{1}_{Z_i^m \geq 0} \left(\frac{1}{\gamma_\theta(\theta^T X_i)} + 1\right) \log\left(\frac{1 + \frac{Z_i^m \gamma_\theta(\theta^T X_i)}{\sigma(X_i)}}{1 + \frac{Z_i^m \hat{\gamma}_\theta(\theta^T X_i)}{\hat{\sigma}(X_i)}}\right) \\ &\quad - \frac{1}{n}\sum_{i=1}^{m} \mathbf{1}_{Z_i^m \geq 0} \left(\frac{1}{\hat{\gamma}_\theta(\theta^T X_i)} - \frac{1}{\gamma_\theta(\theta^T X_i)}\right) \\ &\quad \times \log\left(1 + \frac{Z_i^m \hat{\gamma}_\theta(\theta^T X_i)}{\hat{\sigma}(X_i)}\right) \end{aligned} \tag{7.5}$$

In decomposition (7.5), from Assumptions 3 and 4, the absolute value of the first sum on the right-hand side is bounded by some constant multiplied by

$$\frac{4C_\sigma}{c_\gamma}\left(\frac{1}{c_\gamma} + 1\right) \sup_{\theta,x} \left|\frac{\hat{\gamma}_\theta(\theta^T x)}{\hat{\sigma}(x)} - \frac{\gamma_\theta(\theta^T x)}{\sigma(x)}\right| \tag{7.6}$$

with probability tending to one (indeed, with probability tending to one, $c_\sigma/2 \leq \hat{\sigma}(x) \leq 2C_\sigma$ and $c_\gamma/2 \leq \hat{\gamma}_\theta(\theta^T x) \leq 2C_\gamma$ from the uniform consistency property of these two estimators). The supremum in (7.6) tends to zero from Assumptions 3 and 4, showing that the first sum in (7.5) tends to zero uniformly in θ. Similarly, the second sum in (7.5) can be bounded by

$$\sup_{\theta,x} \left|\frac{1}{\hat{\gamma}_\theta(\theta^T x)} - \frac{1}{\gamma_\theta(\theta^T x)}\right| \frac{1}{n}\sum_{i=1}^{m} \log\left(1 + \frac{2Z_i^m C_\gamma}{c_\sigma/2}\right) \mathbf{1}_{Z_i^m \geq 0},$$

with the supremum tending to zero by Assumptions 3 and 4.

7.2 Proof of Theorem 3.2

By definition of $\hat{\theta}$, we have $\nabla_\theta M_n(\hat{\gamma}_{\hat{\theta}}, \hat{\sigma}; \hat{\theta}) = 0$. Therefore, from a first order Taylor expansion,

$$\nabla_\theta M_n(\hat{\gamma}_{\theta_0}, \hat{\sigma}; \theta_0) = (\theta_0 - \hat{\theta})^T \nabla_\theta^2 M_n(\hat{\gamma}_{\tilde{\theta}}, \hat{\sigma}; \tilde{\theta}),$$

where $\tilde{\theta}$ tends to θ_0 due to the consistency of $\hat{\theta}$. Therefore, $\nabla_\theta^2 M_n(\hat{\gamma}_{\tilde{\theta}}, \hat{\sigma}; \tilde{\theta}) = \Sigma + o_P(1)$ using Assumptions 3, 8 and 12 ii).

Hence, the result of Theorem 3.2 follows if we show that $n^{1/2}\nabla_\theta M_n(\hat{\gamma}_{\theta_0}, \hat{\sigma}; \theta_0) \Longrightarrow \mathcal{N}(0, V)$. To show this convergence, decompose

$$\begin{aligned}
\nabla_\theta M_n(\hat{\gamma}_{\theta_0}, \hat{\sigma}; \theta_0) \;=\; & \frac{1}{n}\sum_{i=1}^m \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i)\partial_\gamma l(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\hat{\sigma}(X_i)})\mathbf{1}_{Z_i^m \geq 0} \\
& + \frac{1}{n}\sum_{i=1}^m \{\nabla_\theta \hat{\gamma}_{\theta_0}(\theta_0, X_i) - \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i)\}\partial_\gamma l(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\hat{\sigma}(X_i)})\mathbf{1}_{Z_i^m \geq 0} \\
& + \frac{1}{n}\sum_{i=1}^m \{\nabla_\theta \hat{\gamma}_{\theta_0}(\theta_0, X_i) - \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i)\}\{\hat{\gamma}_{\theta_0}(\theta_0^T X_i) - \gamma_{\theta_0}(\theta_0^T X_i)\} \\
& \times \partial_\gamma^2 l(\tilde{\gamma}_i; \frac{Z_i^m}{\hat{\sigma}(X_i)})\mathbf{1}_{Z_i^m \geq 0} \\
& + \frac{1}{n}\sum_{i=1}^m \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i)\{\hat{\gamma}_{\theta_0}(\theta_0^T X_i) - \gamma_{\theta_0}(\theta_0^T X_i)\} \\
& \times \partial_\gamma^2 l(\tilde{\gamma}_i; \frac{Z_i^m}{\hat{\sigma}(X_i)})\mathbf{1}_{Z_i^m \geq 0} \\
& =: A_{1n} + A_{2n} + A_{3n} + A_{4n}, \qquad (7.7)
\end{aligned}$$

where $\tilde{\gamma}_i$ tends to $\gamma_{\theta_0}(\theta_0^T X_i)$ due to the uniform consistency of $\hat{\gamma}_{\theta_0}$. From the uniform consistency rates of $\nabla_\theta \hat{\gamma}_{\theta_0}$ and $\hat{\gamma}_{\theta_0}$, we get that $A_{3n} = o_P(n^{-1/2})$.

First, let's treat the term A_{4n}. It can be written

$$\begin{aligned}
A_{4n} &= \frac{1}{n}\sum_{i=1}^{m}\nabla_\theta\gamma_{\theta_0}(\theta_0,X_i)\{\hat{\gamma}_{\theta_0}(\theta_0^TX_i)-\gamma_{\theta_0}(\theta_0^TX_i)\}\\
&\quad\times\partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^TX_i);\frac{Z_i^m}{\sigma(X_i)})\mathbf{1}_{Z_i^m\geq 0}+o_P(n^{-1/2})\\
&= \frac{1}{n}\sum_{i=1}^{m}\nabla_\theta\gamma_{\theta_0}(\theta_0,X_i)\{\hat{\gamma}_{\theta_0}(\theta_0^TX_i)-\gamma_{\theta_0}(\theta_0^TX_i)\}\\
&\quad\times\partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^TX_i);\frac{Z_i^m}{\sigma(X_i)})\mathbf{1}_{Z_i^m\geq 0}\mathbf{1}_{\log(1+Z_i^m)>(mp_m)^\beta}\\
&\quad+\frac{1}{n}\sum_{i=1}^{m}\nabla_\theta\gamma_{\theta_0}(\theta_0,X_i)\{\hat{\gamma}_{\theta_0}(\theta_0^TX_i)-\gamma_{\theta_0}(\theta_0^TX_i)\}\\
&\quad\times\partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^TX_i);\frac{Z_i^m}{\sigma(X_i)})\mathbf{1}_{Z_i^m\geq 0}\mathbf{1}_{\log(1+Z_i^m)\leq(mp_m)^\beta}+o_P(n^{-1/2})\\
&= A_{41n}+A_{42n}+o_P(n^{-1/2})
\end{aligned}$$

for some $\beta>0$ (to be fixed further). Since $E[\log(1+Y)]<+\infty$ (to treat the derivative of order 3 of the log-likelihood function with respect to its first variable) and $\partial^{21}_{\gamma\sigma}l(\tilde{\gamma}_i;\frac{Z_i^m}{\tilde{\sigma}_i})$ ($\tilde{\sigma}_i$ lies between $\sigma(X_i)$ and $\hat{\sigma}(X_i)$) is uniformly bounded for n sufficiently large, the last term $o_P(n^{-1/2})$ is uniform by the uniform consistencies of $\hat{\gamma}_{\theta_0}$ and $\hat{\sigma}(X_i)$. Next, for all $\varepsilon>0$,

$$\begin{aligned}
P((mp_m)^{-1/2}n|A_{41n}|>\varepsilon) &\leq C\frac{\eta_n(mp_m)^{1/2}E[\log(1+Z^m)\mathbf{1}_{\log(1+Z^m)\geq(mp_m)^\beta}|Z^m\geq 0]}{\varepsilon}\\
&\leq C'\frac{\eta_n(mp_m)^{1/2}(\mathbb{P}(\log(1+Z^m)>(mp_m)^\beta|Z^m\geq 0)^{(K_m-1/K_m)}}{\varepsilon}\\
&\leq \frac{C'c_n(mp_m)^{1/4-\beta(K_m-1)}}{\varepsilon},
\end{aligned}$$

for some $C,C'>0$, where $c_n\to 0$ and K_m is a constant that can be chosen as large as needed ($E[(\log(1+Z^m))^{K_m}|Z^m\geq 0]<\infty$ for any K_m). As a consequence, $(mp_m)^{-1/2}nA_{41n}=o_P(1)$ and $A_{41n}=o_P(n^{-1/2})$ since $(m/n)p_m=O_P(1)$.

We then treat the term A_{42n}. The function $x\to\hat{\gamma}_{\theta_0}(\theta_0^Tx)-\gamma_{\theta_0}(\theta_0^Tx)$ belongs to a Donsker class with probability tending to one by Assumption 9 i) and is multiplied by the function $(x,z)\to\partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^Tx);(z-u_x(m))/\sigma(x))\mathbf{1}_{0\leq z-u_x(m)\leq\exp((mp_m)^\beta)-1}$, $m\in\mathbb{Z}_0^+$ (the set of positive integers). The first factor of this function $\partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^Tx);(z-u_x(m))/\sigma(x))$ is Donsker by Assumption 7 ii). For the second factor, we restrict to the class of functions

$(x,z) \to \mathbf{1}_{z \leq u_x(m)+\exp((mp_m)^\beta)-1}$, $m \in \mathbb{Z}_0^+$, for which $\mathbf{1}_{0 \leq z-u_x(m)}$ is a particular case. Since $u_x(m)+\exp((mp_m)^\beta)-1$ is an increasing function of m, we denote $m_1 = 1, m_2, \ldots, m_k = \infty$ and divide $P(Z \leq u_X(m) + \exp((mp_m)^\beta) - 1)$ into $k-1$ intervals of length $O(\varepsilon^2)$ ($P(Z \leq u_X(m_i) + \exp((m_i p_{m_i})^\beta) - 1) - P(Z \leq u_X(m_{i-1}) + \exp((m_{i-1} p_{m_{i-1}})^\beta) - 1) = O(\varepsilon^2)$, $i = 2, \ldots, k$). The above class of functions is therefore Donsker with a bracketing number equal to $O(\varepsilon^{-2})$ (using the $L_2(P)-$norm). As a consequence, the resulting class $\mathcal{H} = \{(x,z) \to \nabla_\theta \gamma_{\theta_0}(\theta_0, x)(f_1^1(x) - f_1^2(x)) f_2(x,z) : f_1^1, f_1^2 \in \mathcal{F}_\gamma, f_2 \in \mathcal{F}_{\partial^2}\}$ is Donsker. Applying Lemma 19.36 of van der Vaart (1998) for the functions h_m, $h_m \in \mathcal{H}$, we have

$$\sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m h_m(X_i, Z_i) - \int h_m(x,z) d\mathbb{P}^u_{X,Z}(x,z)\right] = o_P((mp_m)^{v/8-1/4} p_m^{1/2-v/4})$$

uniformly in $h_m \in \mathcal{H}$ and where $\mathbb{P}^u(x,z)$ is the joint distribution of (X,Z). Indeed, for $\delta^2 = o(m^{-1/2} p_m^{1/2}) = E[h_m^2(X,Z)]$, we have

$$\begin{aligned} & J(\delta, \mathcal{H}, L_2(\mathbb{P}^u_{X,Z}(x,z)))(1 + \frac{J(\delta, \mathcal{H}, L_2(\mathbb{P}^u_{X,Z}(x,z)))\ \|h_m\|_\infty}{\delta^2 \sqrt{m}}) \\ &= o((mp_m)^{v/8-1/4} p_m^{1/2-v/4}) + O(m^{\beta-3/4+v/4} p_m^{\beta-1/4-v/4}), \end{aligned}$$

where $\|\cdot\|_\infty$ denotes the infinite norm. Defining

$$\begin{aligned} A^*_{421n}(x,z) &= \nabla_\theta \gamma_{\theta_0}(\theta_0, x)\{\hat{\gamma}_{\theta_0}(\theta_0^T x) - \gamma_{\theta_0}(\theta_0^T x)\} \\ &\quad \times \partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^T x); \frac{z - u_x(m)}{\sigma(x)}) \mathbf{1}_{z \geq u_x(m)} \mathbf{1}_{\log(1+z-u_x(m)) \leq (mp_m)^\beta} \end{aligned}$$

and

$$A_{421n} = \frac{m}{n}\left\{\frac{1}{m}\sum_{i=1}^m A^*_{421n}(X_i, Z_i) - \int A^*_{421n}(x,z) d\mathbb{P}^u_{X,Z}(x,z)\right\},$$

we easily obtain

$$n^{1/2} A_{421n} = \frac{\sqrt{m}}{\sqrt{n}} o_P((mp_m)^{v/8-1/4} p_m^{1/2-v/4}) = o_P((mp_m)^{v/8-1/4} p_m^{-v/4}) = o_P(1).$$

To reduce $n^{1/2} A_{421n}$ to the single term above, we used for example, $K_m = 2$ and $\beta = 1/4$ but other choices are possible. It then remains for $A_{42n} = A_{421n} + A_{422n}$,

$$\begin{aligned} A_{422n} &= \frac{mp_m}{n} \int \nabla_\theta \gamma_{\theta_0}(\theta_0, x)\{\hat{\gamma}_{\theta_0}(\theta_0^T x) - \gamma_{\theta_0}(\theta_0^T x)\} \partial^2_\gamma l(\gamma_{\theta_0}(\theta_0^T x); w) d\mathbb{P}_{X,W}(x,w) \\ &\quad + o_P(n^{-1/2}), \end{aligned}$$

where $\mathbb{P}_{X,W}(x,w)$ corresponds to the joint distribution of $(X, Z^m/\sigma(X))$ given $Z^m \geq 0$. By first integrating with respect to the distribution of $Z^m/\sigma(X)$ given X and $Z^m \geq 0$, the conditional mean (given X and $Z^m \geq 0$) of $\partial_\gamma^2 l(\gamma_{\theta_0}(\theta_0^T X); \frac{Z^m}{\sigma(X)})$ tends to a function which only depends on $\theta_0^T X$ such that $E[\partial_\gamma^2 l(\gamma_{\theta_0}(\theta_0^T X); \frac{Z^m}{\sigma(X)})|X, Z^m \geq 0] = Q(\theta_0^T X) + o(n^{-1/2})$ (see expression (A.6) in Beirlant and Goegebeur (2004)). As a consequence, the first term of A_{422n} is

$$\begin{aligned}
&\frac{mp_m}{n} \int \nabla_\theta \gamma_{\theta_0}(\theta_0, x)\{\hat{\gamma}_{\theta_0}(\theta_0^T x) - \gamma_{\theta_0}(\theta_0^T x)\} Q(\theta_0^T x) d\mathbb{P}_X(x) + o_P(n^{-1/2}) \\
&= \frac{mp_m}{n} \int \{\hat{\gamma}_{\theta_0}(z) - \gamma_{\theta_0}(z)\} Q(z) \int \nabla_\theta \gamma_{\theta_0}(\theta_0, x) d\mathbb{P}_{X|\theta_0^T X}(x|z) d\mathbb{P}_{\theta_0^T X}(z) + o_P(n^{-1/2}),
\end{aligned}$$

where $\mathbb{P}_{X|\theta_0^T X}(\cdot|\cdot)$ and $\mathbb{P}_{\theta_0^T X}(\cdot)$ denote the distributions of X given $\theta_0^T X$ and $\theta_0^T X$ respectively. Next, using Assumption 10, $A_{4n} = o_P(n^{-1/2})$.

To study A_{1n}, first replace $\hat{\sigma}$ by σ, leading to

$$\begin{aligned}
A_{1n} &= \frac{1}{n} \sum_{i=1}^m \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i) \partial_\gamma l(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\sigma(X_i)}) \mathbf{1}_{Z_i^m \geq 0} \\
&\quad + \frac{1}{n} \sum_{i=1}^m [\hat{\sigma}(X_i) - \sigma(X_i)] \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i) \partial_{\gamma\sigma}^{11} l(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\sigma(X_i)}) \mathbf{1}_{Z_i^m \geq 0} \\
&\quad + \frac{1}{n} \sum_{i=1}^m \frac{[\hat{\sigma}(X_i) - \sigma(X_i)]^2}{2} \nabla_\theta \gamma_{\theta_0}(\theta_0, X_i) \partial_{\gamma\sigma}^{12} l(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\tilde{\sigma}_i}) \mathbf{1}_{Z_i^m \geq 0} \\
&=: A_{11n} + A_{12n} + A_{13n}, \qquad (7.8)
\end{aligned}$$

where $\tilde{\sigma}_i$ tends to $\sigma(X_i)$ due to the uniform consistency of $\hat{\sigma}$. We easily see that $A_{13n} = o_P(n^{-1/2})$ due to the uniform convergence rate of $\hat{\sigma}$.

Next, we treat the term A_{12n}. By Assumptions 7 i) and 11 (and similarly to the term A_{42n}),

$$g_m : (x,z) \to (\hat{\sigma}(x) - \sigma(x)) \nabla_\theta \gamma_{\theta_0}(\theta_0, x) \partial_{\gamma\sigma}^{11} l(\gamma_{\theta_0}(\theta_0^T x); (z - u_x(m))/\sigma(x)) \mathbf{1}_{z \geq u_x(m)}, m \in \mathbb{Z}_0^+$$

belongs to a Donsker class $\mathcal{G}$ with probability tending to one. Since this function is bounded by a constant times $\mathbf{1}_{z \geq u_x(m)} \sup_x |\hat{\sigma}(x) - \sigma(x)|$, its second moment is $O_P(\eta_n^2 p_m)$.

Then,

$$\begin{aligned} A_{12n} &= \frac{m}{n}\left[\frac{1}{m}\sum_{i=1}^{m} g_m(X_i,Z_i) - \int g_m(x,z)d\mathbb{P}^u_{X,Z}(x,z)\right] \\ &\quad + \frac{m}{n}p_m \int g_m(x,z)d\mathbb{P}_{X,Z}(x,z) \\ &= A_{121n} + \frac{m}{n}\mathbb{P}(Z^m \geq 0)A_{122n} \end{aligned}$$

and A_{121n} is treated similarly to A_{421n} above. For A_{122n}, we have

$$\begin{aligned} A_{122n} &= \frac{1}{n}\sum_{j=1}^{m}\int \nu_n(x,X_j,Z_j^m)\mathbf{1}_{Z_j^m\geq 0}\nabla_\theta\gamma_{\theta_0}(\theta_0,x)\partial^{11}_{\gamma\sigma}l(\gamma_{\theta_0}(\theta_0^T x);\frac{z-u_x(m)}{\sigma(x)})d\mathbb{P}_{X,Z}(x,z) \\ &\quad + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{j=1}^{m}\int \nu_n(x,X_j,Z_j^m)\mathbf{1}_{Z_j^m\geq 0}\nabla_\theta\gamma_{\theta_0}(\theta_0,x) \\ &\qquad \times \int_0^{+\infty} \partial^{11}_{\gamma\sigma}l(\gamma_{\theta_0}(\theta_0^T x);w)d\mathbb{P}_{W|X}(w|x)d\mathbb{P}_X(x) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{j=1}^{m}\nu(X_j,Z_j^m)\mathbf{1}_{Z_j^m\geq 0}\nabla_\theta\gamma_{\theta_0}(\theta_0,X_j) \\ &\qquad \times \int_0^{+\infty} \partial^{11}_{\gamma\sigma}l(\gamma_{\theta_0}(\theta_0^T X_j);w)d\mathbb{P}_{W|X}(w|X_j)f_{X|Z\geq u_X(m)}(X_j) + o_P(n^{-1/2}), \end{aligned}$$

where $\mathbb{P}_X(\cdot)$ and $\mathbb{P}_{W|X}(\cdot|\cdot)$ denote the distributions X and $Z^m/\sigma(X)$ given $X=x$ (these distributions are defined given $Z^m \geq 0$) respectively and the last equality follows from Assumptions 11 and 12 i) and iii). Finally, using expression (A7) in Beirlant and Goegebeur (2004)

$$A_{122n} = \frac{1}{n}\sum_{j=1}^{m}\frac{-\nu(X_j,Z_j^m)\mathbf{1}_{Z_j^m\geq 0}\nabla_\theta\gamma_{\theta_0}(\theta_0,X_j)f_{X|Z\geq u_X(m)}(X_j)}{\sigma(X_j)(1+\gamma_{\theta_0}(\theta_0^T X_j))(1+2\gamma_{\theta_0}(\theta_0^T X_j))} + o_P(n^{-1/2}).$$

Since $E[\nu^2(X,Z^m)|Z^m \geq 0]$ is bounded as well as the other factors in the main term of the above expression, the Chebichev inequality leads to $A_{122n} = O_P(m^{1/2}p_m^{1/2}/n)$ and the factor $(m/n)p_m$ before A_{122n} in A_{12n} can be replaced by one so that $A_{12n} = A_{122n} + o_P(n^{-1/2})$.

Next, we compute the mean of each term of the sum in A_{11n}. We obtain

$$p_m E\left[\nabla_\theta \gamma_{\theta_0}(\theta_0, X_i) E\left[\partial_\gamma l\left(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\sigma(X_i)}\right) | X_i, Z_i^m \geq 0\right] | Z_i^m \geq 0\right],$$

where

$$\begin{aligned} & E\left[\partial_\gamma l\left(\gamma_{\theta_0}(\theta_0^T X_i); \frac{Z_i^m}{\sigma(X_i)}\right) | X_i, Z_i^m \geq 0\right] \\ &= \frac{c(X_i)\phi(u_{X_i}(m)|X_i)}{\gamma_{\theta_0}(\theta_0^T X_i)(\frac{1}{\gamma_{\theta_0}(\theta_0^T X_i)} - \rho(X_i))(1 + \frac{1}{\gamma_{\theta_0}(\theta_0^T X_i))} - \rho(X_i))} + o(n^{-1/2}), \end{aligned} \tag{7.9}$$

from (A.4) in Beirlant and Goegebeur (2004). The sum of means of the terms in A_{11n} is then $mp_m o(n^{-1/2})$. Finally, an application of the Lyapounov Theorem ($E[(\log(1+Y))^3] < +\infty$) leads to $n^{1/2}(A_{11n} + A_{122n}) \Longrightarrow \mathcal{N}(0, V)$. More precisely, the Lyapounov ratio is of order $(mp_m)^{-1/2}$ and tends to 0 since $mp_m \to \infty$. From this, $\frac{n}{\sqrt{m}}(A_{11n} + A_{12n})$ has a variance

$$\begin{aligned} & E[\eta_{\theta_0}(X, Z^m)\eta_{\theta_0}^T(X, Z^m)] \\ = & \int\int \tilde{\eta}_{\theta_0}(x, y)\tilde{\eta}_{\theta_0}^T(x, y) dF_{Y|X}(y|x) d\tilde{F}_X(x) p_m + o_P(p_m) \\ = & E[\tilde{\eta}_{\theta_0}(X, Y)\tilde{\eta}_{\theta_0}^T(X, Y)]p_m + o_P(p_m), \end{aligned}$$

using Assumption 13 and where $F_{Y|X}(y|x)$ is the Generalized Pareto cumulative distribution function of Y given X. It results that $\frac{n}{\sqrt{p_m m}}(A_{11n} + A_{12n}) \Longrightarrow \mathcal{N}(0, V)$ and by Slutsky Theorem, $n^{1/2}(A_{11n} + A_{122n}) \Longrightarrow \mathcal{N}(0, V)$.

The term A_{2n} is treated similarly to A_{12n} and A_{42n}. Using Assumptions 7 *ii*) and 9 *ii*), lemma 19.36 of van der Vaart (1998) leads to

$$\begin{aligned} A_{2n} &= \frac{mp_m}{n}\int \{\nabla_\theta \hat{\gamma}_{\theta_0}(\theta_0, x) - \nabla_\theta \gamma_{\theta_0}(\theta_0, x)\}\partial_\gamma l(\gamma_{\theta_0}(\theta_0^T x); w)\mathbf{1}_{w \geq 0} d\mathbb{P}_{X,W}(x, w) + o_P(n^{-1/2}), \\ &= \frac{mp_m}{n}\int \{\nabla_\theta \hat{\gamma}_{\theta_0}(\theta_0, x) - \nabla_\theta \gamma_{\theta_0}(\theta_0, x)\}\int_0^{+\infty} \partial_\gamma l(\gamma_{\theta_0}(\theta_0^T x); w) d\mathbb{P}_{W|X}(w|x) d\mathbb{P}_X(x) \\ &\quad + o_P(n^{-1/2}). \end{aligned}$$

By expression 7.9 above, $A_{2n} = o_P(n^{-1/2})$ and this finishes the proof.

Lemma 7.1 *Assume that $u \to m_{\theta_0}(u)$ has a continuous derivative m'_{θ_0}. Then*

$$\nabla_\theta m_{\theta_0}(\theta_0, x) = (x - E[X|\theta_0^T X])m'_{\theta_0}(\theta_0^T x).$$

Proof. We have

$$\begin{aligned} m_\theta(\theta^T X) &= E\left[Y\sigma(X)^{-1}|\theta^T X\right] \\ &= E\left[E\left[Y\sigma(X)^{-1}|X\right]|\theta^T X\right] \\ &= E\left[E\left[Y\sigma(X)^{-1}|\theta_0^T X\right]|\theta^T X\right] \\ &= E\left[m_{\theta_0}(\theta_0^T X)|\theta^T X\right]. \end{aligned}$$

Let $\alpha(X,\theta) = \theta_0^T X - \theta^T X$. Define

$$\Gamma_X(\theta_1,\theta_2) = E\left[m_{\theta_0}(\alpha(X,\theta_1) + \theta_2^T X)|\theta_2^T X\right].$$

We have $m_\theta(\theta^T X) = \Gamma(\theta,\theta)$. Hence,

$$\nabla_\theta m_{\theta_0}(\theta_0, X) = \nabla_{\theta_1}\Gamma_X(\theta_0,\theta_0) + \nabla_{\theta_2}\Gamma_X(\theta_0,\theta_0).$$

Moreover,

$$\nabla_{\theta_1}\Gamma_X(\theta_0,\theta_0) = -m'_{\theta_0}(\theta_0^T X)E\left[X|\theta_0^T X\right] \tag{7.10}$$

$$\nabla_{\theta_2}\Gamma_X(\theta_0,\theta_0) = m'_{\theta_0}(\theta_0^T X)X. \tag{7.11}$$

■

7.3 Discussion on the assumptions on the nonparametric estimators

The consistency and asymptotic normality of $\hat{\theta}$ rely on consistency properties of the nonparametric estimators $\hat{\gamma}_\theta$ (Assumptions 3 and 8) and specific conditions on γ_θ (Assumptions 9 and 10). We briefly show conditions under which they hold, in the case of the estimator $\hat{\gamma}_\theta^{(2)}$ but this can also be achieved similarly for other estimators (as suggested and explained in Sections 2 and 3). For the purpose of simplicity, we consider the case where Y_i given X_i, $i = 1,\ldots,n$, exactly follows a GPD. However, for the Z_i that follow a distribution of the type (2.9) in the present context with the assumptions described in Section 3, the Y_i can be replaced by the nonnegative $Z_i - u_{X_i}(m)$ and the resulting sums and integrals can be treated as in the proofs developed for Theorems 3.1 and 3.2.

Let $\eta_n = h^2 + n^{-1/2}h^{-1/2}$, where h is the smoothing parameter involved in $\hat{\gamma}_\theta^{(2)}$. Since we assumed $\sup_x |\hat{\sigma}(x) - \sigma(x)| = O_P(\eta_n)$, we have $\hat{\gamma}_\theta^{(2)}(u) = 1 - m_\theta^*(u)^{-1} + O_P(\eta_n)$, where the O_P-rate holds uniformly in θ and u (with $\sup_{\theta,u} \gamma_\theta^{(2)}(u) < 1$), and where

$$m_\theta^*(u) = \sum_{i=1}^n \frac{K\left(\frac{\theta^T X_i - u}{h}\right)}{\sum_{j=1}^n K\left(\frac{\theta^T X_j - u}{h}\right)} \frac{Y_i}{\sigma(X_i)}.$$

Assume that K is twice continuously differentiable with compact support and bounded derivatives up to order 2. If we assume that $\inf_{\theta,u} f_{\theta^T X}(u) > 0$, where $f_{\theta^T X}$ denotes the density of $\theta^T X$ (alternatively, we can relax this assumption by using some trimming strategy as in Härdle et al. (1993) or Delecroix et al. (2006)), $\sup_{\theta,u} |m_\theta^*(u) - m_\theta(u)| = O_P(\eta_n)$ provided that $E[Y_i^2/\sigma^2(X_i)] < \infty$ and that $m_\theta(u)$ is twice continuously differentiable with bounded derivatives (uniformly in θ and u). A proof of this assertion can be found in Lopez et al. (2013) in a more general framework. This shows that Assumption 3 holds in this case.

A proof for the convergence rate $\eta_n' = h^2 + n^{-1/2}h^{-3/2}$ of $\nabla_\theta m_{\theta_0}^*(\theta_0, x)$ can be found in Lopez et al. (2013) (uniformly in x); the uniform (in x and θ) convergence of $\nabla_\theta^j m_\theta^*(\theta, x)$, $j = 1, 2$, is achieved provided that $\eta_n'' = h^2 + n^{-1/2}h^{-5/2}$ tends to zero (also proved in this paper). This shows that the conditions of Assumptions 8 hold provided that $nh^5 \to \infty$ and $nh^8 \to 0$.

Let us now discuss Assumption 9. Let $\mathcal{C}^1(M) = \{f : \mathbb{R} \to \mathbb{R}; \|f\|_\infty + \|f'\|_\infty \leq M\}$: $\mathcal{C}^1(M)$ is a Donsker class of functions from Corollary 2.7.2 in van der Vaart and Wellner (1996). Using the notations of Assumption 9, if we take, for point i), $\mathcal{F}_\gamma = \mathcal{C}^1(M)$, then, due to the uniform convergence of $\hat{\gamma}_{\theta_0}^{(2)}(\theta_0^T \cdot)$ and of its first order derivative (see Lopez et al. (2013)), $\hat{\gamma}_{\theta_0}^{(2)}(\theta_0^T \cdot)$ belongs to $\mathcal{F}_\gamma$ with probability tending to one provided that $\gamma_{\theta_0}^{(2)}(\theta_0^T \cdot)$ does (with bounded m_{θ_0}') and that M is taken large enough.

Moreover, it follows from Lemma 7.1 that

$$\nabla_\theta m_{\theta_0}(\theta_0, x) = (x - E[X|\theta_0^T x]) m_{\theta_0}'(\theta_0^T x).$$

Recall that $\nabla_\theta \gamma_{\theta_0}(\theta_0, x) = m_{\theta_0}(\theta_0^T x)^{-2} \nabla_\theta m_{\theta_0}(\theta_0, x)$ and that $m_{\theta_0}(\theta_0^T x)$ is bounded away from 0. If we assume that m_{θ_0}' and $f_{\theta_0^T X}$ are C^1, we see that $\nabla_\theta m_{\theta_0}(\theta_0, \cdot)$ belongs to the class of functions

$$\mathcal{F}_\nabla := x\mathcal{C}^1(M) + \mathcal{C}^1(M).$$

for M large enough. Using standard kernel estimators arguments and under some additional assumptions, we can show the uniform consistency of $\nabla_\theta m^*_{\theta_0}(\theta_0, x)$ and its first derivative. This together with the uniform consistency of $m^*_{\theta_0}(\theta_0^T x)$ (and its first derivative) enables $\nabla_\theta \hat{\gamma}^{(2)}_{\theta_0}(\theta_0, x)$ to belong to $\mathcal{F}_\nabla$ with probability tending to one. Next, observe that $\mathcal{F}_\nabla$ is a Donsker class of functions, since $\mathcal{C}^1(M)$ is, and using Examples 2.10.7 and 2.10.10 in van der Vaart and Wellner (1996).

7.4 Additional information on the data

Risk category	Mean	Median	Std	Kurtosis	Max	γ^{UNC}	#
EPWS	159,190	61,270	739,600	345.41	15,034,330	0.662	479
DPA	149,630	62,215	310,130	29.61	1,933,100	0.776	39
BDSF	97,650	62,196	105,030	9.39	525,620	0.452	67
All	151,510	61,742	674,990	408.31	15,034,330	0.651	585

Table A1: Descriptive statistics of the empirical severity distribution. Losses are split between risks sub-classifications (lines 1 to 3) and pooled together (line 4). The column γ^{UNC} gives the unconditional maximum likelihood estimator of the GPD tail index.

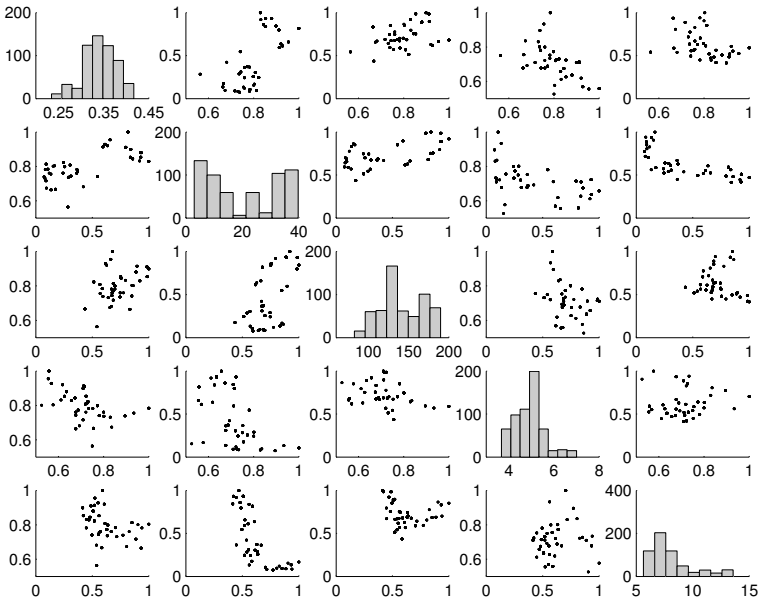

Figure A1: On the diagonal: empirical distribution of the considered covariates (from top left to bottom right: PRF (in %), stock price (in €), Thomson Reuters index, long term bond rate (in %) and unemployment rate (in %)). In the other cells: scatter plot of one covariate on another, normalized between zero and one.

ρ_{ij}	PRF	Stock Price	TR index	LT bond	Unemp.
PRF	1	0.692	0.383	-0.54	-0.433
Stock Price	0.692	1	0.665	-0.557	-0.73
TR index	0.383	0.665	1	-0.253	-0.299
LT bond	-0.54	-0.557	-0.253	1	0.156
Unemp.	-0.433	-0.73	-0.299	0.156	1
Loss size	0.128	0.078	0.037	-0.048	-0.072

Table A2: Correlation matrix between covariates. The last row displays the Spearman's correlation coefficient with the loss size.

References

Basel Committee on Banking Supervision (2004). Basel II: international convergence of capital measurement and capital standards. A revised framework. Report, Bank of International Settlements.

Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200.

Beirlant, J. and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Comput. Statist. Data Anal.*, 42(4):595–619.

Beirlant, J. and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions. *J. Multivariate Anal.*, 89(1):97–118.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and applications, With contributions from Daniel De Waal and Chris Ferro.

Bouaziz, O. and Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2):514–542.

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (in press, 2015). An extreme value approach for modeling Operational Risk losses depending on covariates. *J. Risk. Insur.*

Chernobai, A., Jorion, P., and Yu, F. (2011). The derminants of Operational Risk in U.S. financial institutions. *J. Financ. Quant. Anal.*, 46(8):1683–1725.

Chiou, J.-M. and Müller, H.-G. (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *Scand. J. Stat*, 31:367–386.

Cope, E., Piche, M., and Walter, J. (2012). Macroenvironmental determinants of operational loss severity. *J. Bank. Financ.*, 36(5):1362–1380.

Csörgő, S. and Viharos, L. (1998). Estimating the tail index. In *Asymptotic methods in probability and statistics (Ottawa, ON, 1997)*, pages 833–881. North-Holland, Amsterdam.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Stat. Soc. B Met.*, 52(3):393–442.

Delecroix, M., Härdle, W., and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, 86(2):213–226.

Delecroix, M., Hristache, M., and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, 136(3):730–769.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*, volume 648. Springer - Verlag, Berlin.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. of Math. (2)*, 44:423–453.

Goegebeur, Y., Guillou, A., and Schorgen, A. (2014). Nonparametric regression estimation of conditional tails - the random covariate case. *Statistics: A Journal of Theoretical and Applied Statistics*, 48(4):732–755.

Goldie, C. and Smith, R. (1987). Slow variation with remainder: a survey of the theory and its applications. *Quart. J. Math. Oxford. Ser.*, 38(2):45–71.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, 58(1-2):71–120.

Lopez, O., Patilea, V., and Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli*, 19(3):721–747.

Pickands, III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3:119–131.

Ruppert, D. and Wand, M. (1994). Multivariate Locally Weighted Least Squares Regression. *Ann. Statist.*, 22(3):1346–1370.

Scarrot, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT*, 10(1):33–60.

Smith, R. L. (1987). Estimating tails of probability distributions. *Ann. Statist.*, 15(3):1174–1207.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes with applications to statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Wang, T. and Hsu, C. (2013). Board composition and operational risk events of financial institutions. *J. Bank. Financ.*, 37(6):2042–2051.

Wu, T. Z., Yu, K., and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.*, 101(7):1607–1621.

Xia, Y., Härdle, W. K., and Linton, O. (2011). Optimal smoothing for a computationally and statistically efficient single index estimator. In *Exploring research frontiers in contemporary statistics and econometrics*, pages 229–261. Physica-Verlag/Springer, Heidelberg.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.*, 94(448):1275–1285.

Xia, Y., Tong, H., and Li, W. K. (1999). On extended partially linear single-index models. *Biometrika*, 86(4):831–842.

Zhang, J. (2010). Improving on estimation for the generalized pareto distribution. *Technometrics*, 52(3):335–339.