Multivariate extreme value identification using depth functions

Fateh Chebana* and Taha B.M.J. Ouarda

Canada Research Chair on the Estimation of Hydrometeorological Variables,
INRS-ETE, 490 rue de la Couronne, Quebec (QC),
Canada G1K 9A9

***Corresponding author:** Tel: +1 (418) 654-2542
Fax: +1 (418) 654-2600
Email: fateh.chebana@ete.inrs.ca

October 13th 2010

Revised version submitted to Environmetrics

Abstract

Extreme value theory (EVT) is commonly applied in several fields such as finance, hydrology and environmental modeling. It is extensively developed in the univariate setting. A number of studies have focused on the extension of EVT to the multivariate context. However, most of these studies are based on a direct extension of univariate extremes. In the present paper, we present a procedure to identify the extremes in a multivariate sample. The present procedure is based on the statistical notion of depth function combined with the orientation of the observations. The extreme identification itself is important and it can also serve as basis for the modeling and the asymptotic studies. The proposed procedure is also employed to detect peaks-over-thresholds in the multivariate setting. This method is general and includes several special cases. Furthermore, it is flexible and can be applied to several situations depending on the degree of extreme event risk. The procedure is mainly motivated by application considerations. A simulation study is carried out to evaluate the performance of the procedure. An application, based on air quality data, is presented to show the various elements of the procedure. The procedure is also shown to be useful in other statistical areas.

1. Introduction

Extreme value theory (EVT) plays an important role in several fields such as finance, hydrology, insurance and Internet traffic, see e.g., de Haan and Ferreira [2006], Reiss and Thomas [2007]. EVT is extensively studied in the univariate setting when the extreme event is only described by one characteristic (e.g., Leadbetter et al. [1983] and Coles [2001]). In reality, extreme events are often described through a number of dependent variables. For instance, floods are described by their peak, volume and duration (e.g., Yue et al. [1999]) and air quality is monitored simultaneously through several variables such as the levels of ozone and nitrogen dioxide (e.g., Heffernan and Tawn [2004]). In the multivariate context, EVT is also developing increasingly to treat these events (e.g., Coles and Tawn [1991; 1994]; Mikosch [2005]; Heffernan and Tawn [2004]; Boldi and Davison [2007]; Li [2009] and the references therein).

In the univariate case, extremes are directly identified and the focus is then on modeling efforts. This scheme is commonly extended to the multivariate setting despite the fact that the extreme identification step is not similar and much more complex. In the multivariate literature, the focus is most often on the modeling of extremes, especially on describing the dependence of extreme observations, and also providing asymptotic results (e.g. Li [2009]). However, before developing these important issues (modeling and asymptotic behavior), it is important to correctly identify the notion of extreme in the multivariate context. That is, it is important to specify with respect to which characteristic an observation is extreme, and to identify and quantify the impacts (social, economic) of such extremes. In some fields, such as hydrology and car manufacturing, it is not appropriate to employ asymptotic results since sample sizes are generally small.

In the univariate case, the limiting distribution of the block-maxima is shown to belong to the class of the generalized extreme value (GEV) distributions. For practical considerations, even in the univariate setting, where the extremes are simply and clearly identified, the GEV distributions do not represent a systematic choice to fit extremes as shown in El Adlouni et al. [2010]. Several other distributions are appropriate for extreme modeling and should be considered as well, such as the Halphen family, Pearson, log-Pearson and Gamma (see e.g., Hosking and Wallis [1997]).

The usual extension of EVT to the multivariate setting is based merely on the *component-wise* maxima or minima of the vector sample. This extension is not appropriate since the obtained point does not necessarily belong to the sample and this extension is based on mathematical and theoretical justifications. In the recent literature, the component-wise approach was criticized from both the theoretical and the practical points of view (see e.g., Smith [2004] and Salvadori et al. [2007]). The multivariate EVT approaches neglected the identification step and focused on the modeling and asymptotic aspects.

One of the drawbacks of the multivariate EVT approaches, especially the component-wise extension, is the absence of a convenient notion of ordering in the multivariate context. The notion of order can be statistically extended to the multivariate context using depth functions, see e.g. Zuo and Serfling [2000]. Multivariate EVT does not take advantage of the potential of depth functions. In addition, in the multivariate setting the notion of the median is not employed to define extremes. Usually, in the univariate framework, an extreme observation is the one for which the deviation from the median is the highest. A drawback of the block-maxima approach identification is that it could select extremes in blocks even if all values in a particular block are low and could identify only one extreme in blocks where several high values should be identified in a particular block.

The aim of the present paper is to propose a procedure to identify extreme values in a multivariate sample. Then, modeling and studying the asymptotic properties will be based on the appropriate extremes. The present study has an exploratory objective, rather than modeling or inferential. As pointed out in Liu et al. [1999], depth values do not provide enough information from the multivariate sample to define extremes in the present context. The proposed procedure is based on a combination of depth functions with orientations of the observations with respect to the median. This combination (depth and orientation) is also employed by Serfling [2002] to introduce multivariate median-oriented quantiles. The proposed extreme identification procedure is motivated and defined by practical considerations.

In the univariate setting, extremes in a sample are selected as the minimum, the maximum or both. The choice is based on the underlying variable, the associated risk as well as the case study in hand. By extension, in the multivariate context, it is also of interest to focus on a *part* of the extreme observations by reducing the orientation space to a convenient part.

When dealing with extremes, an alternative of interest is also known as the peaks-over-threshold (POT) approach. In this situation we are interested in identifying all values "over" a given threshold. A more detailed description of the POT approach can be found for instance in Lang et al. [1999] and the references therein. Again as in the extremes, the multivariate POT theory focused on modeling and asymptotic results, see e.g. Reiss and Thomas [2007]. The present depth-based approach is generalized to identify POT observations. By analogy with extremes, the multivariate POT can also be focused on a part of the orientation space.

For reference and clarity of presentation, depth functions are presented briefly in section 2. In the following sections, we present a description and a general algorithm of the proposed method (section 3), we evaluate the consistency of the procedure in a simulation study (section 4), we illustrate it on a real-world air quality dataset (section 5), and we present a discussion (section 6). Conclusions and perspectives are presented in the last section.

2. Background: Depth functions

In the univariate setting, one of the most important notions related to extremes is the *ordering* of observations. In the multivariate setting several extensions of the order are developed and employed. Depth functions are introduced by Tukey [1975] to provide an outward ordering in a multivariate sample. A detailed description of the theoretical background of depth functions can be found in Zuo and Serfling [2000]. Depth functions are developed for several multivariate statistical applications, e.g. in Mizera and Müller, [2004] and Ghosh and Chaudhuri [2005] and are applied in several areas such as economic and social sciences by Caplin and Nalebuff [1991a; b], industrial quality control by Liu and Singh [1993] and in water sciences by Chebana and Ouarda [2008].

A statistical depth function $D(.;F)$, or simply $D(.)$, for a given cumulative distribution function F on R^d $(d \geq 1)$ is bounded and nonnegative which provides a F-based center-outward ordering of points x in R^d that satisfies the following properties:

i. *Affine invariance*

ii. *Maximality at center*

iii. *Monotonicity relative to the deepest point*

iv. *Vanishing at infinity*

For a formal definition of depth functions, the reader is referred to Zuo and Serfling [2000]. In the literature several kinds of depth functions are introduced and studied. Here we present a non exhaustive list of some of the key ones:

1. The *Mahalanobis depth* is given by:

$$MD(x;F) = \frac{1}{1 + d_A^2(x,\mu)} \tag{1}$$

where $d_A^2(x,y) = (x-y)' A^{-1} (x-y)$ is the Mahalanobis distance between two points $x, y \in R^d$ with respect to a positive definite matrix A, F is a given distribution and μ and A are any corresponding location and covariance measures, respectively.

2. The *Simplicial depth* whose expression is given by:

$$SD(x;P) = P\{x \in S[X_1,...,X_{d+1}]\} \tag{2}$$

where $S[X_1,...,X_{d+1}]$ is the random d-dimensional simplex with vertices $X_1,...,X_{d+1}$ which is a random sample from the distribution P.

3. The *Simplicial volume depth* is given through the expression :

$$SVD^{\alpha}(x;F) = \left(1 + E\left[\left(\frac{\Delta(S[x, X_1,...,X_d])}{\sqrt{\det(\Sigma)}}\right)^{\alpha}\right]\right)^{-1} \quad \text{for} \quad x \in R^d \tag{3}$$

where $\Delta(S[x, X_1,...,X_d])$ denotes the volume of the d-dimensional simplex $S[x, X_1,...,X_d]$, Σ is the covariance matrix of F and $\alpha > 0$.

4. The *Halfspace depth* is defined for $x \in R^d$ with respect to a probability P on R^d as:

$$HD(x;P) = \inf\{P(H) : H \text{ a closed halfspace that contains } x\} \tag{4}$$

A corresponding sample version of a statistical depth function $D(x;\ F)$ may be defined by replacing F with a suitable empirical function $\hat{F}_n$ and denoted by $D_n(x) = D(x;\hat{F}_n)$. The asymptotic properties of $D_n(x)$ are studied in several papers including Liu [1990], Massé [2002; 2004] and Lin and Chen [2006]. The evaluation of some depth functions is complex and requires approximations and specific algorithms. For instance, Miller et al. [2003] developed an algorithm for the computation of the halfspace depth. Recently, Massé and Plante [2009] provided a package in the R software to evaluate several depth functions.

3. Methodology

In this section we present a description of the proposed procedure followed by a general algorithm for practical implementation. Probabilistic formulations of the approach as well as a diagnostic of its use are presented. The main notations used throughout the paper are summarized in the notation list and are illustrated in the case study section.

3.1. Description of the methodology

Let $X_1,..., X_n$ be a R^d vector sample of size n, denoted $\Lambda_{n,d}$, where $X_i = (X_{i1},...,X_{id})$ and d is a positive integer ($d \geq 1$). Let M_n represent the multivariate median of the sample. It corresponds, in the present study, to the maximum depth value in the sample. It is natural to assume that the median is the "center" of the sample and an extreme is so with respect to the median. Hence, the median is considered as the origin of the multivariate space and then data are median-centered by translation. The orientation set of the observations is the unit hyper-sphere centered at the median M_n and denoted $\Omega^{d-1}(M_n)$. The space $\Omega^{d-1}(M_n)$ will be denoted Ω^{d-1} after translation of the data to be centered at M_n. In the bivariate case (d = 2), the unit sphere Ω^{d-1} reduces to the interval $[-1,\ 1]$, and it becomes $\{-1,+1\}$ in the univariate case.

For each observation i from the sample, we assign a depth value $D_i = D(X_i)$ and an orientation $u_i = u(X_i)$ from Ω^{d-1}. Note that in the bivariate case, the presentation is analogous to the polar coordinates but with depths instead of Euclidian distances. In higher dimensions, the expressions of an orientation u with respect to the Cartesian coordinates x are more complex and can be found, for instance, in Stanley [1990]. For a fixed orientation $u_0 \in \Omega^{d-1}$, we identify an extreme as the observation corresponding to the smallest depth value in that orientation.

Since the orientation space Ω^{d-1} is continuous, it is convenient to proceed to its discretization. To this end, let $\lambda \in (0,1]$ be a coefficient that defines a "partition" of Ω^{d-1}. The obtained partition of Ω^{d-1} is composed by the subsets $\Pi_{k,\lambda}$, $k = 1,...,n_e$ where n_e is related to λ as shown bellow. The coefficient λ represents the volume of each portion $\Pi_{k,\lambda}$. For each portion $\Pi_{k,\lambda}$, we select the observation corresponding to the smallest depth value in this portion as the extreme one. On the other hand, the coefficient λ indicates the size of the sub-sample composed by extreme observations, say, $n_e = \lfloor 1/\lambda^{d-1} \rfloor$ where $\lfloor . \rfloor$ represents the integer part of a real number. The condition $1 \le n_e \le n$ leads to the constraint on λ: $1/n \le \lambda^{d-1} \le 1$. We define the set of the identified extreme observations $\Sigma_n(\lambda, D)$ as:

$$\Sigma_n(\lambda, D) = \left\{ x_k \in \Lambda_{n,d} : D(x_k) \text{ is the samllest in } \Pi_{k,\lambda}, k = 1,...,n_e \right\} \tag{5}$$

As special cases for λ, we identify one observation ($n_e = 1$) as extreme over the whole sample for $\lambda = 1$. When λ is close to zero (which requires n to be large), the number of extreme observations is high. In the case where $\lambda^{d-1} = 1/n$, all the observations are identified as extremes, unless located on the same orientation.

The selection of the coefficient λ is important to define the extremes. In general, its selection depends on the context of the case study. At this stage, an acceptable general option could be related to the number of blocks (n_b) in the traditional block-maxima approach. Indeed, λ can be selected such that n_e is the same (or approximately) as n_b, that is $\lambda^{d-1} = 1/n_b$. For instance, blocks are generally related to the length of the temporal series, such as daily, weekly, monthly, seasonally or annually. This characteristic depends on the application field such as air quality modeling, hydrology or climatology. The coefficient λ should be small enough to lead to a reasonable number of extremes with which one can perform a statistical analysis and at the same time λ should be large enough to avoid obtaining a large number of extremes which would contradict the rarity principle of extremes. However, an optimal and automatic selection procedure of λ would be useful and should be developed in a future work.

By relating the extreme observations in $\Sigma_n(\lambda, D)$ we obtain the hyper-surface $\mathbb{C}_n(\lambda, D)$ which may be convex or not. The coefficient λ controls, in an inversely proportional way, the regularity (or smoothness) of $\mathbb{C}_n(\lambda, D)$. In terms of risk, a small value of λ indicates that the decision is hard and it should be taken with care whereas a large value of λ is associated to safer situations. Indeed, a small value of λ leads to several extreme combinations that should be considered to prevent the associated risk whereas large values of λ are representative for situations that require less attention. The coefficient λ can be interpreted as a confidence degree against the corresponding risk. Risk is often defined as the probability of occurrence of an extreme event (see Niwa [1989] and Ouarda and Labadie [2001]). Hence, a small value of λ indicates that a large number of extreme events have occurred and therefore the probability of occurrence of similar events is high. It is important to mention that, in contrast to the univariate case where we have one extreme observation (minimum or maximum), the existence of several extreme observations in the multivariate context is natural. It can be justified by the fact that several combinations of variable values lead to the same risk.

In the univariate case, extreme value refers to the maximum *or* the minimum of a sample. According to the problem to be treated, the focus is made on the minimum *or* the maximum *or* both. The above extreme identification procedure allows to generalize this aspect to the multivariate setting. Indeed, we consider the identification of extremes on a *part* T of Ω^{d-1} of orientations. Hence, the subdivision in portion of volume λ can be limited only to the part T instead of the whole set Ω^{d-1}. The corresponding set of extreme observations is given by:

$$\Sigma_n(\lambda, T, D) = \Sigma_n(\lambda, D) \cap T \qquad (6)$$

For instance, for $d = 2$ where $\Omega^1 = [-1,1]$, if the focus is on simultaneous non-exceedence events $\left(X \leq x,\ Y \leq y\right)$, it is convenient to choose $T =$ $[0, 0.5]$ which corresponds to the first quadrant. In the univariate case where $\Omega^0 = \left\{-1,1\right\}$, the maximum is associated to $T = \left\{1\right\}$ whereas the minimum is associated to $T = \left\{-1\right\}$. Note that the volume of the range T should be larger than λ. In the equality case λ = volume (T), we have $e_n = \Sigma_n(volume(T), T, D)$. By analogy with $\Sigma_n(\lambda, T, D)$, the hyper-surface $\mathbb{C}_n(\lambda, D)$ can be restricted to a given part T as:

$$\mathbb{C}_n(\lambda, T, D) = \mathbb{C}_n(\lambda, D) \cap T \qquad (7)$$

The present approach can be generalized for the identification of POTs for a given multivariate sample. In the univariate POT, one of the criteria used to define the threshold is based on a given percentile of the sample. Hence, in the multivariate framework we select in each λ-portion $\Pi_{k,\lambda}$ the observations for which the depth deviations from that of the median do not exceed a given proportion s ($0 \leq s \leq 1$) of the deviation of the minimum depth in $\Pi_{k,\lambda}$. That is the depth value is smaller than $D_{s,k} = D_{\max} - \left(D_{\max} - D_{\min,k}\right)s$ and, therefore, the set of POTs is given by:

$$\Sigma_n(\lambda, s, D) = \left\{x_j \in \Lambda_{n,d} \cap \Pi_{k,\lambda} : D(x_j) < D_{s,k},\ \ \Pi_{k,\lambda} \subset \Omega^{d-1}\right\} \qquad (8)$$

where $D_{\max}$ is the depth value of the median and $D_{\min,k}$ is the smallest depth value over $\Pi_{k,\lambda}$. Note that for a fixed value of the threshold s, the value $D_{s,k}$ is not necessarily the same for all $\Pi_{k,\lambda}$ since $D_{\min,k}$ depends on k. Clearly, the special case $s = 0$ leads to the selection of all data as

POTs whereas s = 1 identifies the extreme observations. As it is the case for extremes, the POT may also be of interest in a part $T \subset \Omega^{d-1}$. The corresponding set is:

$$\Sigma_n(\lambda, s, T, D) = \Sigma_n(\lambda, s, D) \cap T \tag{9}$$

The hyper-surfaces $\mathbb{C}_n(\lambda, s, D)$ and $\mathbb{C}_n(\lambda, s, T, D)$ can be defined, for instance, by connecting the observations with the largest depth value and smaller than $D_{s,k}$ in each portion $\Pi_{k,\lambda}$. Other options are presented in Sections 5 and 6.

In the depth-based approach, the case with $\lambda = 1$ corresponds to the usual POT approach in the whole data set. However, smaller values of λ are useful to adapt the approach in the presence of trend or seasonality in the data.

Now that the descriptive presentation of the depth-based identification approach is complete, we provide a brief probabilistic formulation of the approach. Assume that the original random vectors $X_1, X_2, \ldots$ have a multivariate distribution F on R^d, then the random vector of the depth-based extreme values is given by:

$$B_{k,\lambda} = \underset{X_i \in \Lambda_{n,d} \text{ with } u(X_i) \in \Pi_{k,\lambda}}{\arg\min} D(X_i) \tag{10}$$

The exact or asymptotic distribution of the random vector $B_{k,\lambda}$ is related to F as well as to the distribution of $D(X)$. Basically, the problem can be seen as a minimization of a special transformation $D(.)$ of the original random vectors X. In addition, in the present context, the study of $B_{k,\lambda}$ can be conducted by considering previous work such as Massé [2004; 2009], Arcones et al. [2006] and Zu and He [2006] where asymptotic results are obtained for $D(X)$. Further developments in this direction are outside the scope of the present study and are the subject of future work.

In a similar way the extremes on a range T can be defined as:

$$B_{k,\lambda,T} = \underset{X_i \in \Lambda_{n,d} \text{ with } u(X_i) \in \Pi_{k,\lambda} \cap T}{\arg\min} D(X_i) \tag{11}$$

The corresponding POTs on Ω^{d-1} and on a range T are defined respectively as follows:

$$K_{k,\lambda,s} = X_i * I\left\{D(X_i) \le D_{s,k}, u(X_i) \in \Pi_{k,\lambda}\right\} \tag{12}$$

$$K_{k,\lambda,s,T} = X_i * I\left\{D(X_i) \le D_{s,k}, u(X_i) \in \Pi_{k,\lambda} \cap T\right\} \tag{13}$$

where $I\{A\}$ stands for the indicator function of a set A, that is $I\{A\}=1$ if A holds and 0 if not. As indicated in the introduction, the block-maxima approach imposes a uniform repartition of the extremes over time (one extreme per time block). The proposed depth-based approach avoids this constraint since it is based on the magnitudes of the values and not on their time of occurrence. In situations where it is necessary to define time blocks, an intermediate option could be to combine both approaches by employing the depth-based approach in each large time block. A large time block is composed of a number of the usual blocks. For instance, large and usual blocks could be respectively season and month or year and season. An illustration is given in the case study where each season has four months and hence λ = ¼= 0.25 for each season.

Before presenting the procedure steps, we state a number of simple properties of the above concepts. For a given sample, using the same depth function D, we have:

$$\text{If } \lambda_1 < \lambda_2 \text{ and } \lambda_2/\lambda_1 \text{ is an integer, then } \Sigma_n(\lambda_2, D) \subset \Sigma_n(\lambda_1, D) \tag{14}$$

The condition λ_2/λ_1 is an integer insures that for each k, there exists k' such that $\Pi_{k,\lambda_2} \subset \Pi_{k',\lambda_1}$. A counter example is given in the case study section when λ_2/λ_1 is not an integer.

For a given sample, on the same part T, using the same depth function D, we have:

$$\text{If } s_1 < s_2, \quad \text{then} \quad \Sigma_n(\lambda, s_2, T, D) \subset \Sigma_n(\lambda, s_1, T, D) \text{ for a fixed } \lambda \tag{15}$$

$$\text{For } s < 1, \quad \text{we have} \quad \Sigma_n(\lambda, s, T, D) \subset \Sigma_n(\lambda, T, D) \tag{16}$$

3.2. Procedure steps

In the following we present the proposed procedure to identify extremes and POTs for a given multivariate sample. Identification of multivariate extremes requires a depth function D, a coefficient λ and, if necessary, a range $T \subset \Omega^{d-1}$. A threshold s in (0, 1] is also to be specified for POTs. On the basis of the description and notations introduced in Section 3.1, the POTs and the extremes are identified through the following steps:

1. Find the median M_n of the multivariate sample. In the present study, it corresponds to the largest depth value $D_{\max}$;
2. Standardize data, especially when variables are not of the same nature;
3. Evaluate, in the range $T \subseteq \Omega^{d-1}$, the depth D_i of the observation i using the selected depth function D, $i = 1,\ldots,n$;
4. Evaluate, in the range $T \subseteq \Omega^{d-1}$, the orientation u_i of the observation i, $i = 1,\ldots,n$;
5. Select, in the range $T \subseteq \Omega^{d-1}$, the observations for which the depth values are smaller than a threshold s of $D_{\min,k}$ in each λ-portion $\Pi_{k,\lambda}$, i.e., with depth smaller than $D_{s,k} = (1-s)D_{\max} + sD_{\min,k}$.

The above procedure is general and covers all possible scenarios depending on the special cases of T and s. Indeed, the range T is taken to be $T \subseteq \Omega^{d-1}$ and the threshold s is inclusively between 0 and 1. The case where T represents the whole space Ω^{d-1} is hence a special case. When $s = 1$, the identified observations are extremes whereas when $s < 1$, the identified observations are POTs. The observations corresponding to the depth values obtained in step 5 constitute, depending on s and T, one of the sets $\Sigma_n(\lambda, D)$, $\Sigma_n(\lambda, T, D)$, $\Sigma_n(\lambda, s, D)$ or $\Sigma_n(\lambda, s, T, D)$.

In step 1, various options are available in the literature to obtain the multivariate median, see e.g., Small [1990]; Liu et al. [1999] and Zuo and Serfling [2000]. The selection of a depth function, among the various options presented in the literature, depends on its convenience for the specific data in hand as well as the simplicity of its evaluation algorithm. Note that a depth function is more general than a simple transformed distance and it combines both geometry and statistics.

Generally depth functions are affine invariant, i.e., depth values remain the same after standardization of data. Therefore, step 2 is not required when the depth function is affine invariant. Note that some depth functions, such as the simplicial volume depth, meet this property only under some assumptions on the parent distribution. The reader is referred to Zuo and Serfling [2000] for more details.

The choice of a depth function may affect the identification of the extremes. As a first criterion to select a depth function, we propose to consider depth functions which are evaluated with respect to a given centre of the data (such as the median). This is the case for the Mahalanobis and the projection depth functions. Note that, in general, depth functions cannot be directly and analytically evaluated. To this end, numerical algorithms are required and a package in the R software is provided by Massé and Plante [2009] for several depth functions.

Since the distribution of the identified extremes is not developed in the present study, the following diagnostic strategy is proposed. First, if the data are highly correlated, the component-wise approach can be adopted to identify extremes and the corresponding models can be selected for further studies by considering the existing literature. This can be checked, for instance, from a scatter plot per block and an evaluation of different dependence association parameters such as the correlation coefficient, Spearman’s rho or Kendall’s tau (Joe [1997]). The scatter plot should have an elliptical shape in the first diagonal line and the values of the dependence parameters should be high to insure that the identified component-wise extremes are part or closely part of the data (a component-wise extreme is not necessarily an observation). Second, if the above situation does not occur, which can be the case for several multivariate data sets, we propose to consider the depth-based extreme identification. Even though this is an exploratory study and inferential concerns are a subject of future efforts, one can consider the identified sub-sample of extremes to select the appropriate distribution among those existing in the literature. This can be done on the basis of goodness-of-fit tests in the multivariate context. For instance, the univariate GEV and GPD distributions can be employed as marginal distributions and combined to a copula to obtain the whole multivariate distribution according to Sklar's theorem (Sklar [1959]). Among the available and convenient copulas, one can consider the extreme value copula or the Archimedean copula families.

Finally, even though the identified depth-based extremes are extreme observations by definition, it is advised to check them on the basis of the different types of plots such as scatter plots with the partitions $\Pi_{k,\lambda}$, depth-orientation plots and multiple chronological plots.

4. Procedure evaluation

In the present section, we evaluate the performance of the proposed procedure on the basis of simulations. To generate samples, we consider a bivariate distribution commonly used in EVT. The margins of this distribution are the Gumbel distribution given by:

$$F_X(x) = \exp\left\{-\exp\left(-\tfrac{x-\beta_X}{\alpha_X}\right)\right\},\ x \text{ real},\ \alpha_X > 0 \text{ and } \beta_X \text{ real} \qquad (17)$$

and the dependence structure is the Gumbel logistic copula expressed as:

$$C_\gamma(u,v) = \exp\left\{-\left[(-\log u)^\gamma + (-\log v)^\gamma\right]^{1/\gamma}\right\},\ \gamma \geq 1 \text{ and } 0 \leq u,v \leq 1 \qquad (18)$$

The Gumbel logistic copula C_γ is at the same time an Archimedean copula and an extreme value copula. The considered parameters of the marginal distributions are $\alpha_X = 300.22$, $\beta_X = 1239.80$

and $\alpha_Y = 15.85$, $\beta_Y = 51.85$. The parameter of the Gumbel logistic copula is considered to take each one of the values $\gamma = 1, 1.414, 3.162$ which correspond respectively to the correlation coefficient values $\rho = 0,\ 0.5, 0.9$. The parameters of the margins, with $\gamma = 1.414$, represent a real-world flood data set studied in Yue et al. [1999] corresponding to the Ashuapmushuan river basin in the province of Quebec, Canada. The sample generation is based on the algorithm developed by Ghoudi et al. [1998]. The number of the generated samples is taken to be $M = 2000$ samples (higher values lead to similar results).

For the evaluation of the procedure, we select values of $\lambda = 0.05$ and 0.10 and a value of $s = 0.90$ for the POTs. The procedure is judged consistent if for different samples of the same nature, all identified extremes are similar. Hence, we evaluate the consistency on the basis of the volume of the polygon composed by the identified extremes. For the k^{th} generated sample, let $V_e(k)$ be the volume of the polygon $\Pi_n^{(k)}(\lambda, D)$ composed by the set of the identified extreme observations $\Sigma_n^{(k)}(\lambda, D)$.

In a similar manner, the consistency of the POTs identification is evaluated by the volume of the area between the polygons composed by the extremes and the deepest POTs in each portion. Let $\tilde{\Sigma}_n^{(k)}(\lambda, s, D)$ be the subset of the identified POTs corresponding to the observations with the largest depth value among the POTs in each portion. Similarly, define $\tilde{\Pi}_n^{(k)}(\lambda, s, D)$ as the polygon composed by $\tilde{\Sigma}_n^{(k)}(\lambda, s, D)$ and let $V_{POT}(k)$ be the difference between the volumes of $\Pi_n^{(k)}(\lambda, D)$ and $\tilde{\Pi}_n^{(k)}(\lambda, s, D)$. Then, the evaluation is based on the mean and the standard-deviation over the M generated samples of $V_e(.)$ for the extremes and of $V_{POT}(.)$ for the POTs given respectively by :

$$M_e = \frac{1}{M}\sum_{k=1}^{M} V_e(k) \quad \text{and} \quad STD_e = \sqrt{\frac{1}{M-1}\sum_{k=1}^{M}\left(V_e(k) - M_e\right)^2} \tag{19}$$

$$M_{POT} = \frac{1}{M}\sum_{k=1}^{M} V_{POT}(k) \quad \text{and} \quad STD_{POT} = \sqrt{\frac{1}{M-1}\sum_{k=1}^{M}\left(V_{POT}(k) - M_{POT}\right)^2} \tag{20}$$

Table 1 presents the evaluation results obtained from the simulations. Results show the general consistency of the identified extremes and POTs for each one of the considered cases. Generally we observe that M_e values have a slight variation (e.g. between 0.85 and 0.97 for $\gamma = 1.414$ and $\lambda = 0.05$) with respect to n whereas STD_e decreases slightly (e.g. from 0.24 to 0.19 for the same case). On the other hand, M_{POT} increases with respect to n (e.g. from 0.16 to 0.59 for the same case) and STD_{POT} remains almost constant. Both values of M_{POT} and STD_{POT} are smaller than those corresponding to the extremes. This is mainly due to the definition of $V_e(.)$ and $V_{POT}(.)$ where $V_{POT} \leq V_e$. Hence, we always have $M_{POT} \leq M_e$ and we can also write $V_{POT}(k) \approx aV_e(k)$ for some constant $a < 1$ and independent of the index k since the values of s is constant (s = 0.9). Therefore, we have $STD_{POT} \approx aSTD_e \leq STD_e$. Furthermore, no significant differences are observed between the independence and moderate dependence cases ($\gamma = 1$ and $\gamma = 1.414$ respectively) whereas the case of higher dependence ($\gamma = 3.162$) produces clearly smaller values of both mean and standard-deviation. The reason could be related to the shape of the scatter plot of the last case which is more elliptical and concentrated as illustrated in Figure 1.

5. Case study

In this section we present a case study to illustrate the different aspects of the proposed methodology. We consider air quality monitoring data employed by Heffernan and Tawn [2004]. The data is represented by a series of summer daily maximum measurements (in parts per billion) of ground level ozone (O_3) and nitrogen dioxide (NO_2) in Leeds city centre, UK, during the years

1994-1998 inclusively with a sample size n = 578 measurements (with some missing data). The summer data set corresponds to observations during the months of April - July. The outliers mentioned in Heffernan and Tawn [2004] were excluded from these data sets.

The Mahalanobis depth (MD) function given in (1) is considered for its simplicity and convenient properties (Zuo and Serfling [2000]). The bivariate median is obtained as the observation that maximizes the MD function. For each observation, we obtain the corresponding MD value as well as the orientation u. Note that the MD function is affine invariant, i.e., depth values are the same for the original and the standardized data. The orientation space in the present case is the interval [-1, 1] since d = 2. Figure 2 illustrates the main employed notations for a selected value of λ (λ = 0.05). The corresponding sets of extremes $\Sigma_n(\lambda, D)$ and the associated curves $\mathbb{C}_n(\lambda, D)$ are presented in Figure 3 for each value of λ = 0.0625 and 0.05. The regularity of the curve $\mathbb{C}_n(\lambda, D)$ as well as the number of extreme observations depend on λ. As it can be seen, the number of extreme observations in the present case is n_e = 16 and 20 for λ = 0.0625 and 0.05 respectively. From Figures 3.a and 3.b, one can see that even though λ = 0.05 is smaller than λ = 0.0625, the set $\Sigma_n(0.0625, D)$ is not included in the set $\Sigma_n(0.05, D)$, since the ratio 0.0625/0.05 = 1.25 does not meet the condition of being an integer as specified in (15).

Since in air quality, high values of both variables O_3 and NO_2 are considered as extreme cases of air pollution, it is more realistic to restrict attention to the part T = [0, 0.5] representing the first quadrant as illustrated in Figure 2. The identified extremes and corresponding depth values are given in Table 2 for λ=0.05. Table 2 and Figure 4 indicate that depth values of the extremes vary between 0.0287 and 0.2590 which are associated to the most "outer" and "interior" observations.

For comparison purposes, we present in Table 3 and Figure 4 the component-wise extremes per month as well as those obtained by considering the depth-based approach on each season composed of four months. Therefore, to identify 20 extreme observations (four extremes in each season) we set λ=0.25 per season. Table 3 and Figure 4 indicate that out of 20 component-wise extremes only 4 correspond to observations. In addition, the component-wise extreme (68, 105), associated to the smallest depth value 0.0246, is very far from the observations. The two largest depth values 0.8198 and 0.3262 correspond respectively to the component-wise extremes (34, 41) and (45, 45). These two component-wise extremes are very close to the median and are unlikely to be true extremes. On the other hand, the depth-based approach per season identified 3 unusual extremes out of 20 which are relatively close to the median with large depth values 0.2836, 0.3550 and 0.4937.

In order to check the identified extremes by the different approaches, Figure 5 presents a multiple chronological plot of the series. Figure 5 shows that the component-wise extremes occur at different dates for each variable and in some situations these dates are very distant within the block such as in May 1994 and May 1997. On the one hand, the component-wise approach identified only one extreme in months with high O_3 and NO_2 (e.g. May 1995 and June 1996). On the other hand, it identified an extreme for ordinary months such as July 1998 where during the whole month the levels of O_3 and NO_2 are low. The extremes identified by the depth-based per season approach are generally clustered such as in June-July 1994 and April-May 1998.

It is important to point out that the usual scatter plot (O_3, NO_2) may be misleading since it is based on the Euclidian distance which represents more the geometric aspect of the data whereas the (u, D) plot is more appropriate since it exhibits the probabilistic aspects. For the entire data set, the (u, D) plot is presented in Figure 6 where the extremes are clearly shown. We observe that the range T = [0, 0.5] contains the observations with the lowest depth values. The above elements indicate that the depth-based approach seems more appropriate especially if we take into account the fact that the identified extremes are observations and are generally "far" from the median.

Figures 7a,b show the identified POTs for the studied data set where the threshold is taken to be s = 90% with λ = 1 and 0.05 on the whole data set. It is easier to visualize the threshold in the space (u, D) than in the space (O_3, NO_2) as it is shown in Figures 7c,d for the present case study. Again, we observe that almost all the identified POTs are found to be in the first quadrant $T = [0, 0.5]$.

As indicated in Section 3, it is of interest to consider other depth functions. In the following, we considered four depth functions (Mahalanobis, simplicial volume with $\alpha = 1$, halfspace and simplicial). Figure 8 illustrates the histograms for each one of the considered depth functions. It can be seen that the Mahalanobis depth is convenient for the current study. Indeed, the Mahalanobis histogram shows that the majority of data are in the centre (depth values between 0.1 and 0.9). However, a smaller portion of the observations is found to be very close to the median (depth values between 0.9 and 1) or is at the boarder (depth values approximately between 0 and 0.1). This distribution of depth values is natural and reflects the distribution of the data, especially the fact that extreme observations are rare. The simplicial volume depth (SVD) function, given in equation (3), could be a good choice as well.

6. Discussion

The identification of extremes treated in the present paper presents some similarities with a number of commonly used statistical techniques, either in their aims or in their concepts. The illustrations in this section are based on the above case study.

The hyper-surface $\mathbb{C}_n(\lambda, D)$ obtained by connecting the extreme observations can be employed to define the range of the multivariate sample. The theoretical version of $\mathbb{C}_n(\lambda, D)$ can be seen as the support of the corresponding multivariate distribution. In addition, the volume of the "inner set" with boundary $\mathbb{C}_n(\lambda, D)$ can be employed to measure and compare the spread of multivariate samples. This spread measure can be compared, for instance, to the measures proposed by Liu et al. [1999] which are also based on depth functions.

On the other hand, the hyper-surface $\mathbb{C}_n(\lambda, D)$ can also be viewed as the contour that includes the entire sample. In the bivariate case, it is possible to present this curve in the space (u, D) as shown in Figure 6. When presented in this manner, the contours can be associated to frontier estimation. Frontier estimation is a statistical technique useful and commonly employed in econometrics. The reader can refer, for instance, to Simar and Wilson [2000] for a review. Hence, the elements of the present procedure can be useful to frontier estimation problems. In addition, based on the presentation (u_i, D_i), the estimation of the curve $\mathbb{C}_n(\lambda, D)$ can be considered as in a regression analysis. However, in the current regression estimation we are dealing with minimum values of D whereas in the usual regression analysis the focus is on the mean values and on the global trend of the series. Furthermore, the presentation of the curve $\mathbb{C}_n(\lambda, D)$ in the space (u, D) as an open curve (function) is the opposite of the presentation used in time series analysis to illustrate seasonality trends (see e.g. Cunderlik et al., [2004] and Ouarda et al., [2006]).

An analogy can be established between the present procedure and the generalized additive model (GAM) estimation using spline functions [*Wood*, 2006]. Indeed, the coefficient λ in the present procedure has a similar role to the penalizing coefficient that controls the regularity of the estimated function in GAM inference. The estimated function using GAM is similar to the contour $\mathbb{C}_n(\lambda, D)$ shown in Figure 6. To be more general, and smoother, the contour $\mathbb{C}_n(\lambda, D)$ can be obtained by connecting the extreme observations by functions similar to splines instead of straight lines. An illustration is given in Figure 9 with $\lambda =$ 0.10 and 0.05.

Consequently, the developed tools in GAM inference can be adapted to the present context. One of these important tools is the generalized cross-validation technique which can be adapted to select the coefficient λ.

Note that one of the criteria to be imposed to the curve $\mathbb{C}_n(\lambda, D)$ is to include all data as well as to be the closest to the data. In other words, $\mathbb{C}_n(\lambda, D)$ should be the convex-hull on each portion $\Pi_{k,\lambda}$. The present procedure leads to a convex-hull that is not only geometrically-based but also statistically-based though depth functions.

7. Conclusions and future work

In the present paper a new procedure is proposed to identify extremes in a multivariate sample. The proposed procedure, as a natural extension of the univariate setting, is based on depth functions and the orientation of the observations toward the median. It can also be used to identify multivariate extremes in a POT framework. From a simulation study, the procedure is shown to be generally consistent. The procedure is applied to a case study representing environmental data. In addition, the component-wise maxima are shown not to represent realistic scenarios in several situations.

The identification of extremes is directly useful to build warning environmental or health systems. However, in numerous situations, the identification is not an end in itself. Indeed, the identification is an important step for the study of the asymptotic properties and the modeling of the identified extremes. It is also shown that the obtained extreme sets and curves are related to other statistical topics such as multivariate spread measures, frontier estimation and generalized additive modeling. The proposed procedure is general and offers a large degree of flexibility through the coefficient λ, the range T and the threshold s. It is useful to practitioners as well as to methodologists.

Even though a major part of the elements related to the procedure are treated in the present paper, others are worth developing in future work. An important issue to be developed is related to the inferential aspects of the approach including the modeling of the identified extremes. It is also of interest to optimize and automate the selection of the coefficient λ for a given data set. In addition, the coefficient λ does not need to be constant. This issue is analogue to the smoothing window in the kernel density nonparametric estimation. The impact of the choice of the depth function should also be studied thoroughly by considering different depth functions. More precisely, the consistency of the identified extremes according to the depth functions can be considered. Finally, it is of interest to associate the obtained extreme curve with a confidence band representing the identification errors.

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Canada Research Chair Program. The authors would like to thank Barbara Martel for her assistance. The authors wish to thank the Editor, Associate Editor, and two anonymous reviewers whose comments helped considerably improve the quality of the paper.

Main notation list:

d	Dimension of data vector
D	Depth function
$\lambda \in (0,1]$	Coefficient that defines a "grid" of the sphere Ω^{d-1} and represents the volume of a portion of the unit sphere Ω^{d-1}
MD	Mahalanobis depth function
M_n	Multivariate median of the sample
n	Sample size
$\Omega^{d-1}(M_n)$	Unit sphere centered at the median M_n. It represents the space of orientations
u	Orientation
$\Lambda_{n,d}$	d-dimension sample with size n
$\Pi_{k,\lambda}$	λ -portion from Ω^{d-1}
$\Sigma_n(\lambda, D)$	Set of extreme observations of the sample with coefficient λ using a depth function D
$B_{k,\lambda}$	The corresponding random vector of $\Sigma_n(\lambda, D)$
$\Sigma_n(\lambda, T, D)$	Restriction of the set $\Sigma_n(\lambda, D)$ on the range T
$B_{k,\lambda,T}$	The corresponding random vector of $\Sigma_n(\lambda, T, D)$
$\Sigma_n(\lambda, s, D)$	Set of POT observations of the sample over a threshold s with coefficient λ using a depth function D
$K_{k,\lambda,s}$	The corresponding random vector of $\Sigma_n(\lambda, s, D)$
$\Sigma_n(\lambda, s, T, D)$	Restriction of the set $\Sigma_n(\lambda, s, D)$ on the range T
$K_{k,\lambda,s,T}$	The corresponding random vector of $\Sigma_n(\lambda, s, T, D)$
e_n	Extreme observation when $\lambda = volume(T)$
$\mathbb{C}_n(\lambda, D)$ and $\mathbb{C}_n(\lambda, T, D)$	Extreme hyper-surfaces obtained by connecting the observations of respectively the sets $\Sigma_n(\lambda, D)$ and $\Sigma_n(\lambda, T, D)$
$\mathbb{C}_n(\lambda, s, D)$ and $\mathbb{C}_n(\lambda, s, T, D)$	POT hyper-surfaces obtained by connecting the observations of respectively the sets $\Sigma_n(\lambda, s, D)$ and $\Sigma_n(\lambda, s, T, D)$ in the space (u, D)

Bibliography

Arcones, M. A., H. Cui, and Y. Zuo (2006), Empirical depth processes, *Test*, *15*(1), 151-177.

Boldi, M. O., and A. C. Davison (2007), A mixture model for multivariate extremes, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *69*(2), 217-229.

Caplin, A., and B. Nalebuff (1991a), Aggregation and imperfect competition - on the existence of equilibrium, *Econometrica*, *59*(1), 25-59.

Caplin, A., and B. Nalebuff (1991b), Aggregation and social choice - a mean voter theorem, *Econometrica*, *59*(1), 1-23.

Chebana, F., and T. B. M. J. Ouarda (2008), Depth and homogeneity in regional flood frequency analysis, *Water Resources Research*, *44*(11).

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, xiv+208 pp., Springer-Verlag London Ltd., London.

Coles, S. G., and J. A. Tawn (1991), Modeling extreme multivariate events, *Journal of the Royal Statistical Society Series B-Methodological*, *53*(2), 377-392.

Coles, S. G., and J. A. Tawn (1994), Statistical-methods for multivariate extremes - an application to structural design, *Applied Statistics-Journal of the Royal Statistical Society Series C*, *43*(1), 1-48.

Cunderlik, J. M., T. B. M. J. Ouarda, and B. Bobée (2004), Determination of flood seasonality from hydrological records, *Hydrological Sciences Journal*, *49*(3), 511-526.

de Haan, L., and A. Ferreira (2006), *Extreme value theory, An introduction*, xviii+417 pp., Springer, New York.

El Adlouni, S., F. Chebana, and B. Bobee (2010), Generalized Extreme Value versus Halphen System: Exploratory Study, *Journal of Hydrologic Engineering*, *15*(2), 79-89.

Ghosh, A. K., and P. Chaudhuri (2005), On maximum depth and related classifiers, *Scandinavian Journal of Statistics*, *32*(2), 327-350.

Ghoudi, K., A. Khoudraji, and L. P. Rivest (1998), Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles, *Canadian Journal of Statistics*, *26*(1), 187-197.

Heffernan, J. E., and J. A. Tawn (2004), A conditional approach for multivariate extreme values, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *66*(3), 497-530.

Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, 240 pp., Cambridge University Press, Cambridge.

Joe, H. (1997), *Multivariate models and dependence concepts*, 1st ed., xviii, 399 p. pp., Chapman & Hall, London ; New York.

Lang, M., T. B. M. J. Ouarda, and B. Bobée (1999), Towards operational guidelines for over-threshold modeling, *Journal of Hydrology*, *225*(3-4), 103-117.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983), *Extremes and Related Properties of Random Sequences and Series.*, Springer, New York.

Li, H. (2009), Orthant tail dependence of multivariate extreme value distributions, *Journal of Multivariate Analysis*, *100*(1), 243-256.

Lin, L., and M. H. Chen (2006), Robust estimating equation based on statistical depth, *Statistical Papers*, *47*(2), 263-278.

Liu, R. Y. (1990), On a notion of data depth based on random simplices, *Annals of Statistics*, *18*(1), 405-414.

Liu, R. Y., and K. Singh (1993), A quality index based on data depth and multivariate rank tests, *J. Amer. Statist. Assoc.*, *8*, 252-260.

Liu, R. Y., J. M. Parelius, and K. Singh (1999), Multivariate analysis by data depth: Descriptive statistics, graphics and inference, *Annals of Statistics*, *27*(3), 783-858.

Massé, J.-C., and J.-F. Plante (2009), *Depth, an R package for depth functions in multivariate analysis (www.cran.r-project.org)*.

Massé, J. C. (2002), Asymptotics for the Tukey median, *Journal of Multivariate Analysis*, *81*(2), 286-300.

Massé, J. C. (2004), Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean, *Bernoulli*, *10*(3), 397-419.

Massé, J. C. (2009), Multivariate trimmed means based on the Tukey depth, *Journal of Statistical Planning and Inference*, *139*(2), 366-384.

Mikosch, T. (2005), How to model multivariate extremes if one must?, *Statistica Neerlandica*, *59*(3), 324-338.

Miller, K., S. Ramaswami, P. Rousseeuw, J. A. Sellares, D. Souvaine, I. Streinu, and A. Struyf (2003), Efficient computation of location depth contours by methods of computational geometry, *Statistics and Computing*, *13*(2), 153-162.

Mizera, I., and C. H. Müller (2004), Location-scale depth, *Journal of the American Statistical Association*, *99*(468), 949-966.

Niwa, K. (1989), *Knowledge-based risk management in engineering : a case study in human-computer cooperative systems*, xi, 132 p. pp., Wiley, New York.

Ouarda, T. B. M. J., and J. W. Labadie (2001), Chance-constrained optimal control for multireservoir system optimization and risk analysis, *Stochastic Environmental Research and Risk Assessment*, *15*(3), 185-204.

Ouarda, T. B. M. J., J. M. Cunderlik, A. St-Hilaire, M. Barbet, P. Bruneau, and B. Bobأ©e (2006), Data-based comparison of seasonality-based regional flood frequency methods, *Journal of Hydrology*, *330*(1-2), 329-339.

Reiss, R.-D., and M. Thomas (2007), *Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields.*, Third edition ed., Birkhäuser Verlag, Basel.

Salvadori, G., C. De Michele, N. T. Kottegoda, and R. Rosso (2007), *Extremes in Nature: An Approach Using Copulas*, Springer

Serfling, R. (2002), Quantile functions for multivariate analysis: Approaches and applications, *Statistica Neerlandica*, *56*(2), 214-232.

Simar, L., and P. W. Wilson (2000), Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis*, *13*(1), 49-78.

Sklar, M. (1959), Fonctions de r\'epartition \`a n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, *8*, 229--231.

Small, C. G. (1990), A survey of multidimensional medians, *International Statistical Review*, *58*(3), 263-277.

Smith, R. L. (2004), Discussion of "A conditional approach for multivariate extreme value" by J.E. Heffernan and J.A. Tawn., *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *66*(3), 530-532.

Stanley, C. R. (1990), Descriptive statistics for N-dimensional closed arrays: A spherical coordinate approach, *Mathematical Geology*, *22*(8), 933-956.

Tukey, J. W. (1975), Mathematics and picturing data, paper presented at Proceedings of the International Congress on Mathematics, Canadian Math. Congress.

Wood, S. N. (2006), *Generalized Additive Models, An introduction with R.*, Chapman & Hall/ CRC.

Yue, S., T. B. M. J. Ouarda, B. Bobée, P. Legendre, and P. Bruneau (1999), The Gumbel mixed model for flood frequency analysis, *Journal of Hydrology*, *226*(1-2), 88-100.
Zu, Y. J., and X. M. He (2006), On the limiting distributions of multivariate depth-based rank sum statistics and related tests, *Annals of Statistics*, *34*(6), 2879-2896.
Zuo, Y., and R. Serfling (2000), General notions of statistical depth function, *Annals of Statistics*, *28*(2), 461-482.

List of Tables and Figure Captions

Table 1: Evaluation of the consistency of the extreme and POT identification procedures based on the polygon volume.

Table 2: Original and standardized values of (O_3, NO_2) of the identified extreme observations corresponding to λ=0.05 as well as their *MD* depth values in the first quadrant.

Table 3: Values of (O_3, NO_2) of the identified extremes corresponding to $\lambda = 0.25$ within each season as well as their *MD* depth values in the first quadrant (left); similar values using the component-wise approach within monthly blocks (right).

Figure 1: Scatter plot illustration for samples generated with n = 300 and a) $\gamma = 1$ b) $\gamma = 1.414$ and c) $\gamma = 3.162$

Figure 2: Illustration of a λ-portion $\Pi_{k,\lambda}$, the orientation interval [-1, 1], the set of extreme observations $\Sigma_n(\lambda, D)$, the corresponding curve $\mathbb{C}_n(\lambda, D)$, and a part *T* as the first quadrant of (O_3, NO_2) for λ = 0.05 where *D* is *MD*

Figure 3: Identified extreme observations set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda, D)$ for a) λ = 0.0625 and b) λ = 0.05 where *D* is *MD*

Figure 4: Extremes identified as component-wise, depth-based (λ=0.05) and depth-based per season (λ=0.25) in the first quadrant

Figure 5: Chronological (O3, NO2) series and the extremes identified as component-wise, depth-based (λ=0.05) and depth based per season (λ=0.25) in the first quadrant. The vertical lines indicate month limits in each season

Figure 6: Identified extreme observation set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda, D)$ in the space (*u*, *D*) for a) λ = 0.10 and b) λ = 0.05 where *D* is *MD*

Figure 7: Identified POT observation set $\Sigma_n(\lambda, s, D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda, s, D)$ for a) λ = 1and b) λ = 0.05 where *D* is *MD* and *s* = 0.90; and for c) λ = 1and d) λ = 0.05 in the space (*u*, *D*)

Figure 8: Histograms of depth values of the data set (O_3, NO_2) for different depth functions

Figure 9: Identified extreme observation set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding smooth curve $\mathbb{C}_n(\lambda, D)$ in the space (*u*, *D*) for a) λ = 0.10 and b) λ = 0.05 where *D* is *MD*

Table 1: Evaluation of the consistency of the extreme and POT identification procedures based on the polygon volume.

		$\lambda = 0.05$				$\lambda = 0.1$			
		Extremes		POT with s = 0.9		Extremes		POT with s = 0.9	
		M_e	STD_e	M_{POT}	STD_{POT}	M_e	STD_e	M_{POT}	STD_{POT}
$\gamma = 1$									
	n = 100	0.87	0.23	0.14	0.07	1.09	0.26	0.29	0.11
	n = 300	0.95	0.21	0.36	0.09	1.08	0.23	0.52	0.12
	n = 500	0.96	0.20	0.45	0.09	1.06	0.20	0.57	0.11
	n= 1000	0.96	0.18	0.54	0.10	1.03	0.18	0.63	0.11
$\gamma = 1.414$									
	n = 100	0.85	0.24	0.16	0.07	1.05	0.28	0.31	0.11
	n = 300	0.95	0.22	0.39	0.09	1.05	0.23	0.54	0.11
	n = 500	0.95	0.20	0.48	0.09	1.05	0.20	0.61	0.11
	n= 1000	0.97	0.19	0.59	0.10	1.04	0.19	0.68	0.12
$\gamma = 3.162$									
	n = 100	0.44	0.16	0.11	0.05	0.45	0.17	0.13	0.07
	n = 300	0.46	0.13	0.19	0.05	0.45	0.14	0.21	0.07
	n = 500	0.46	0.12	0.22	0.05	0.46	0.13	0.24	0.07
	n= 1000	0.46	0.11	0.27	0.06	0.45	0.11	0.28	0.07

Table 2: Original and standardized values of (O_3, NO_2) of the identified extreme observations corresponding to λ=0.05 as well as their *MD* depth values in the first quadrant.

O3	NO2	Standardized O3	Standardized NO2	Depth
74	37	0.7167	0.0286	0.0528
80	40	0.8167	0.0714	0.0418
64	44	0.5500	0.1286	0.0894
84	53	0.8833	0.2571	0.0365
71	52	0.6667	0.2429	0.0615
53	46	0.3667	0.1571	0.1759
71	61	0.6667	0.3714	0.0565
65	60	0.5667	0.3571	0.0733
58	59	0.4500	0.3429	0.1018
64	70	0.5500	0.5000	0.0617
69	86	0.6333	0.7286	0.0372
63	79	0.5333	0.6286	0.0505
58	85	0.4500	0.7143	0.0471
40	**55**	**0.1500**	**0.2857**	**0.2590**
42	61	0.1833	0.3714	0.1712
38	60	0.1167	0.3571	0.1938
46	**105**	**0.2500**	**1.0000**	**0.0287**
36	62	0.0833	0.3857	0.1722
37	82	0.1000	0.6714	0.0621
32	58	0.0167	0.3286	0.2206

Bold character indicates the (O3, NO2) corresponding to the largest and smallest depth values.

Table 3: Values of (O_3, NO_2) of the identified extremes corresponding to $\lambda = 0.25$ within each season as well as their *MD* depth values in the first quadrant (left); similar values using the component-wise approach within monthly blocks (right).

Depth-based per season			Component-wise		
O3	**NO2**	**Depth**	**O3**	**NO2**	**Depth**
55	41	0.1329	45	58	0.1850
71	61	0.0456	46	62	0.1466
48	62	0.1339	53	78	0.0644
34	78	0.0806	71	71	0.0483
74	37	0.0640	46	60	0.1616
64	70	0.0945	71	86	0.0357
69	86	0.0566	*63*	*79*	*0.0505*
37	82	0.0678	74	60	0.0509
84	53	0.0453	52	62	0.1194
46	54	0.2595	41	55	0.2514
43	55	0.2836	**68**	**105**	**0.0246**
46	**105**	**0.0292**	*84*	*53*	*0.0365*
46	46	0.2100	39	60	0.1916
57	81	0.0439	58	66	0.0840
39	66	0.1138	53	51	0.1608
28	**47**	**0.4937**	*57*	*81*	*0.0542*
59	42	0.0694	47	63	0.1357
40	41	0.3550	59	62	0.0903
42	61	0.0989	45	45	0.3262
36	62	0.1114	***34***	***41***	***0.8198***

Bold character indicates the (O3, NO2) corresponding to the largest and smallest depth values.
Shaded character indicates component-wise extremes that coincide with observations.

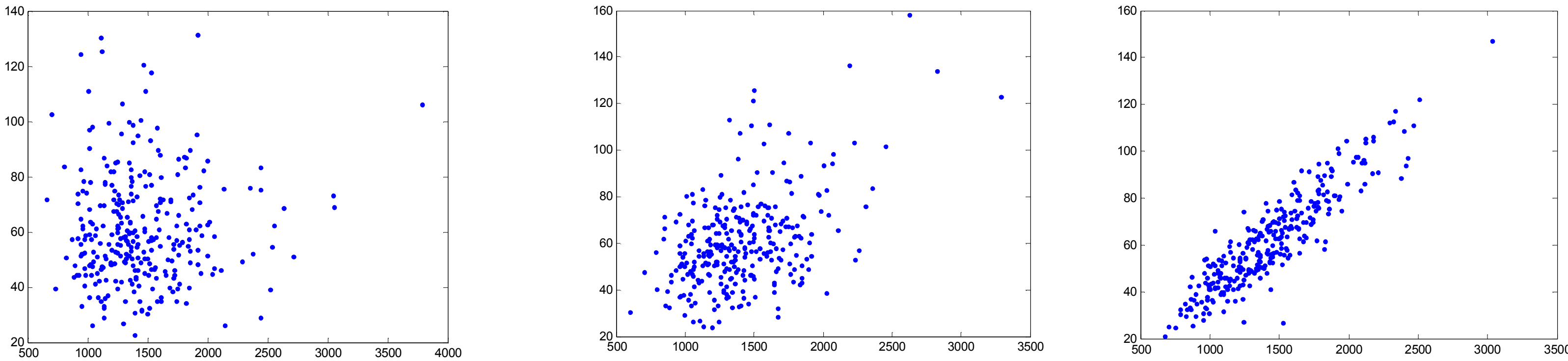


Figure 1: Scatter plot illustration for samples generated with $n = 300$ and a) $\gamma = 1$ b) $\gamma = 1.414$ and c) $\gamma = 3.162$

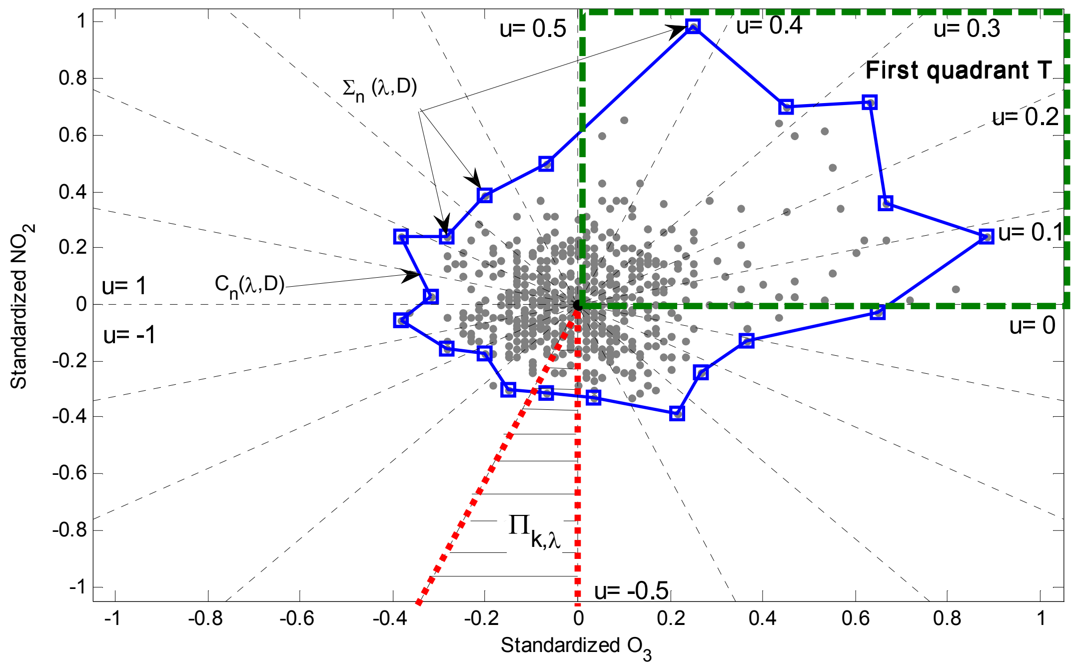
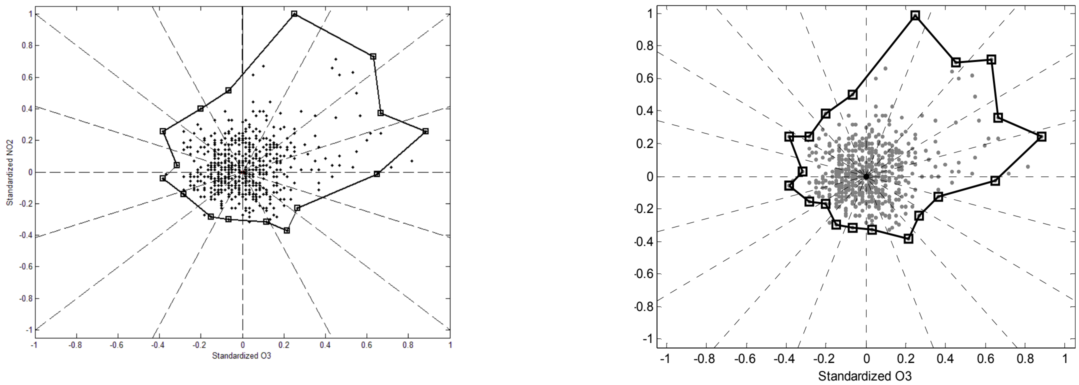


Figure 2: Illustration of a λ-portion $\Pi_{k,\lambda}$, the orientation interval [-1, 1], the set of extreme observations $\Sigma_n(\lambda, D)$, the corresponding curve $\mathbb{C}_n(\lambda, D)$, and a part T as the first quadrant of (O_3, NO_2) for $\lambda = 0.05$ where D is MD



a) b)

Figure 3: Identified extreme observations set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda, D)$ for a) $\lambda = 0.0625$ and b) $\lambda = 0.05$ where D is MD

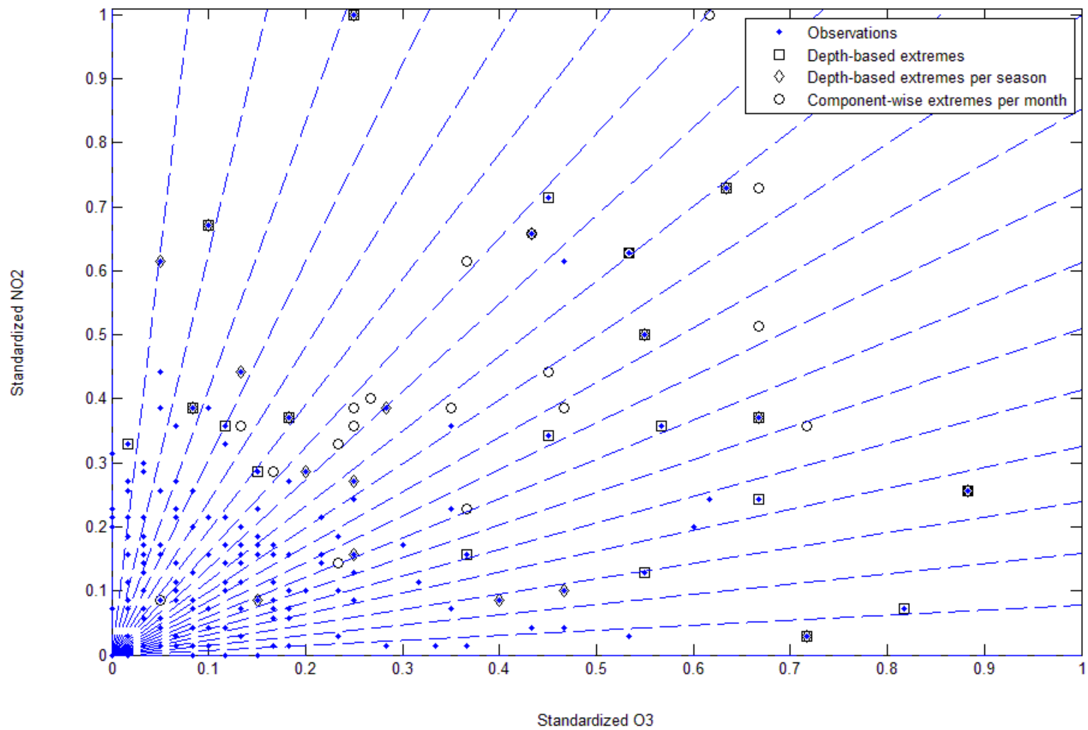


Figure 4: Extremes identified as component-wise, depth-based (λ=0.05) and depth-based per season (λ=0.25) in the first quadrant.

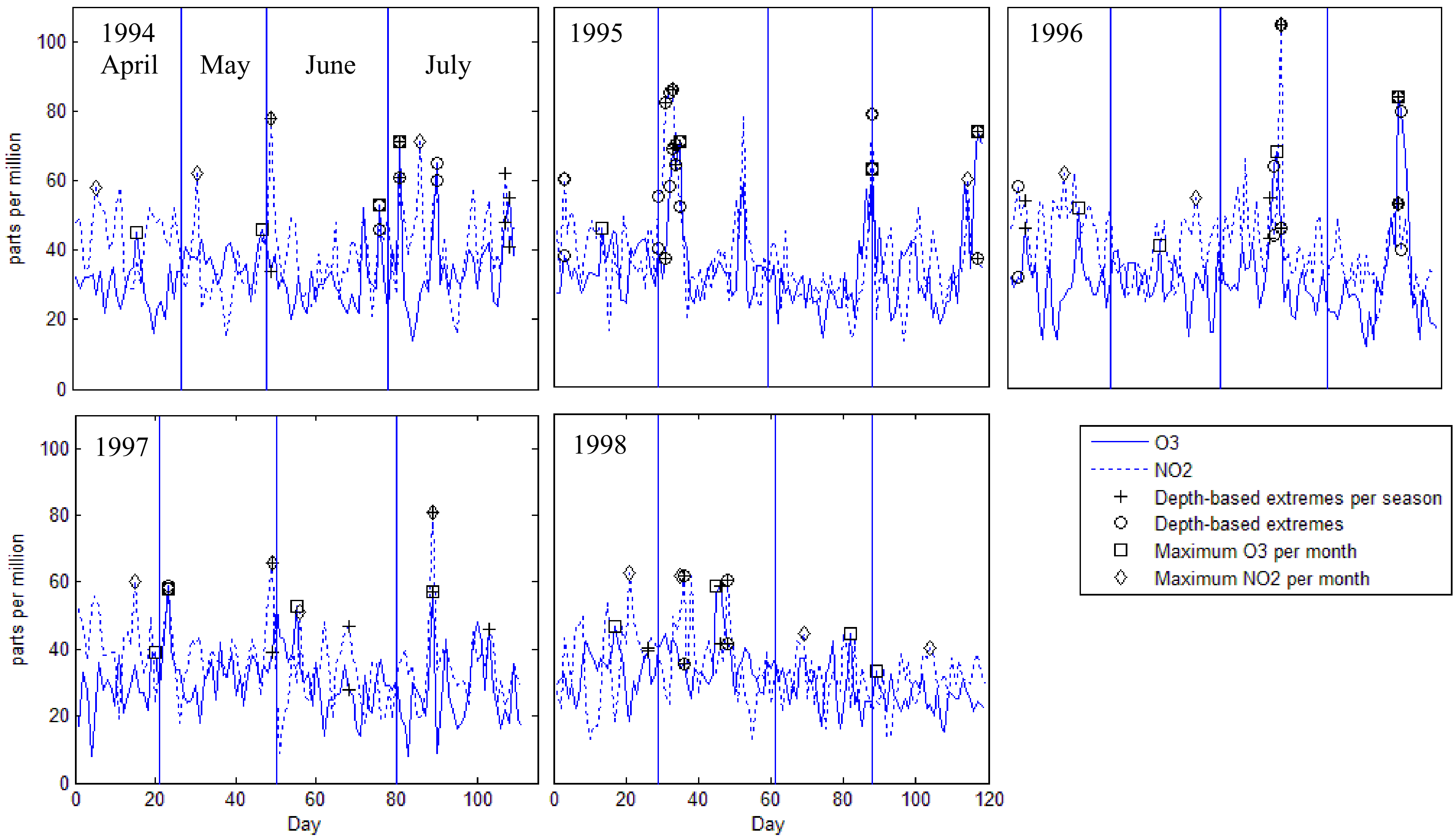
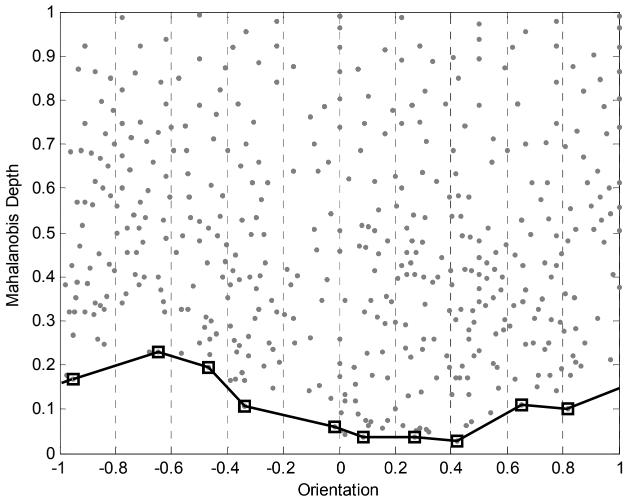
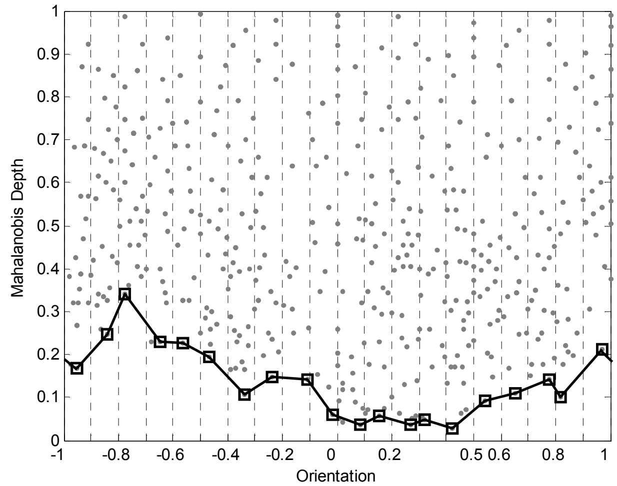


Figure 5: Chronological (O3, NO2) series and the extremes identified as component-wise, depth-based (λ=0.05) and depth based per season (λ=0.25) in the first quadrant. The vertical lines indicate month limits in each season

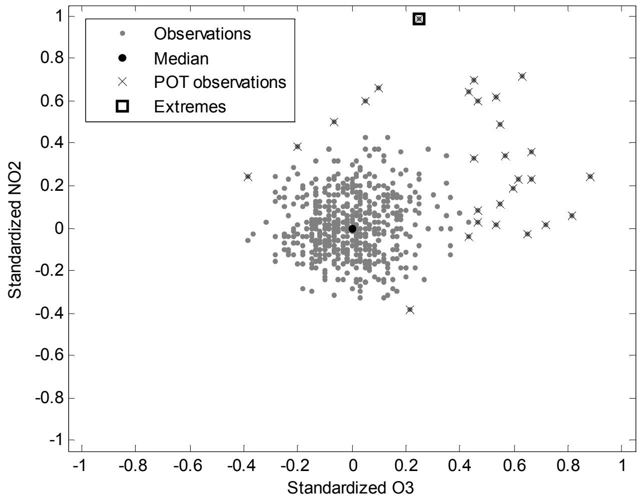




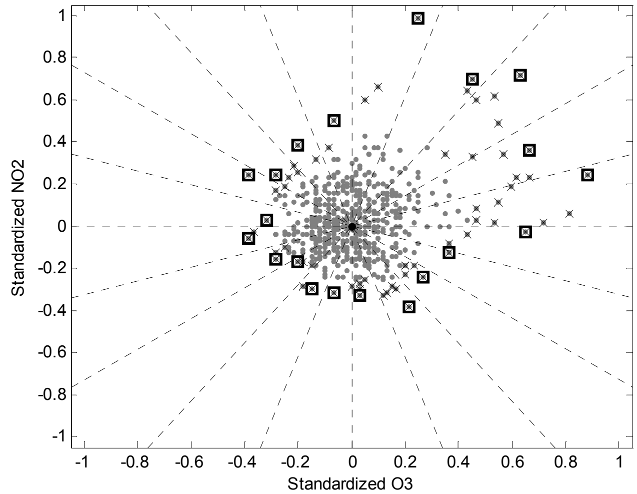
b)

a)

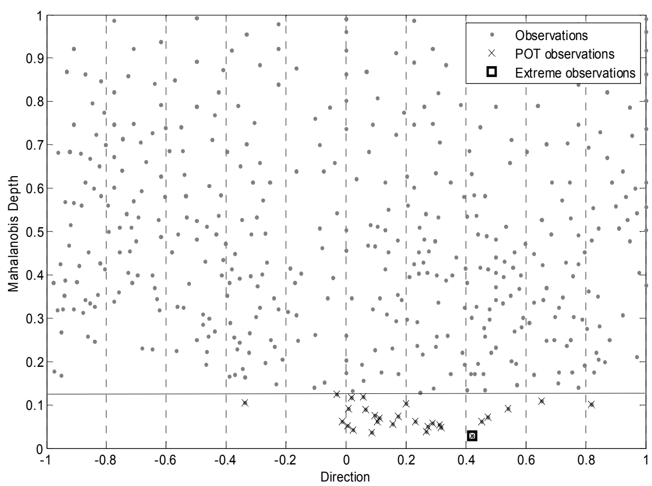
Figure 6: Identified extreme observation set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda, D)$ in the space (u, D) for a) $\lambda = 0.10$ and b) $\lambda = 0.05$ where D is MD



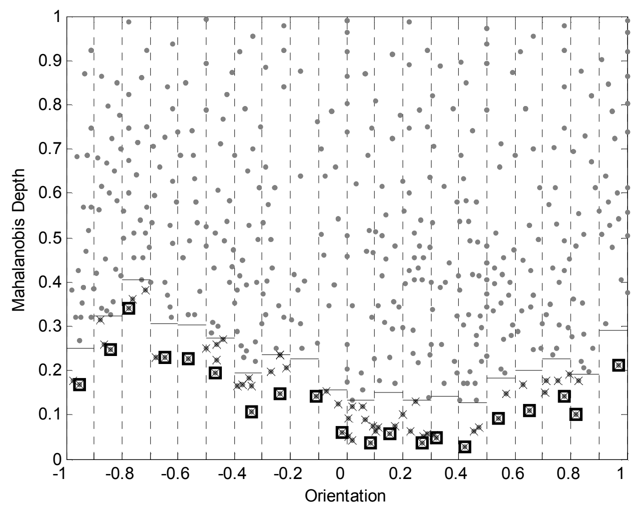
a)



b)

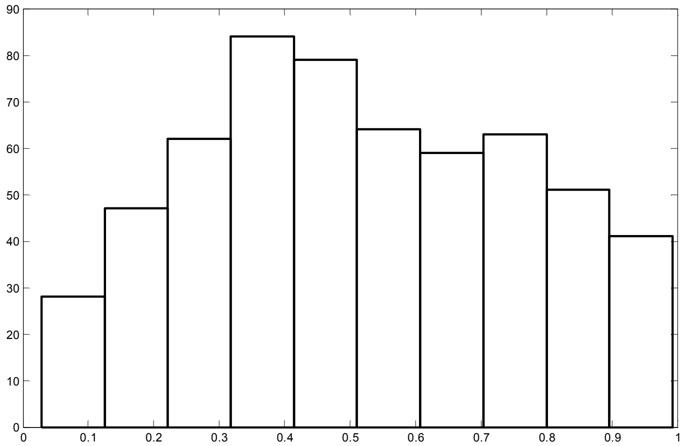


c)

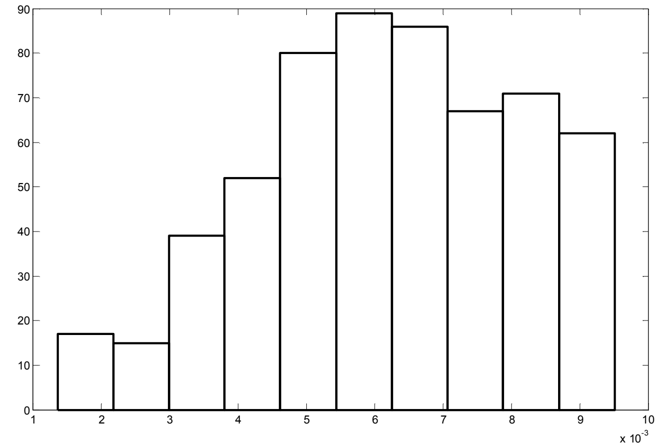


d)

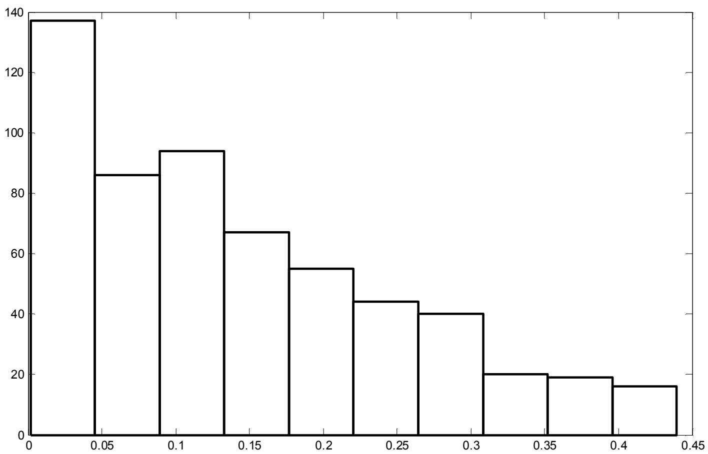
Figure 7: Identified POT observation set $\Sigma_n(\lambda,s,D)$ of (O_3, NO_2) and the corresponding curve $\mathbb{C}_n(\lambda,s,D)$ for a) λ = 1and b) λ = 0.05 where *D* is *MD* and *s* = 0.90; and for c) λ = 1and d) λ = 0.05 in the space (*u*, *D*)



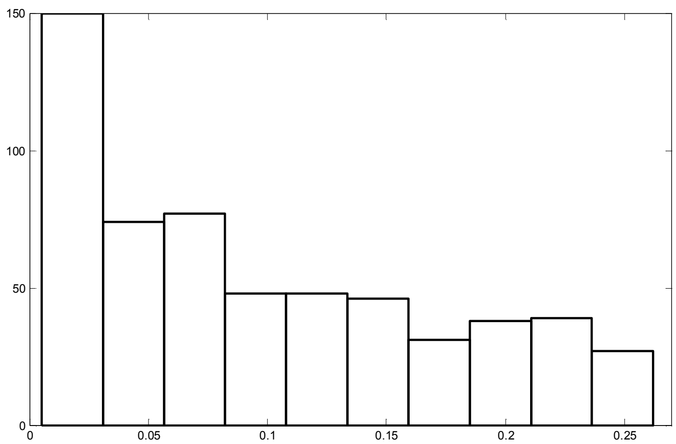
a) Mahalanobis depth



b) Simplicial volume depth

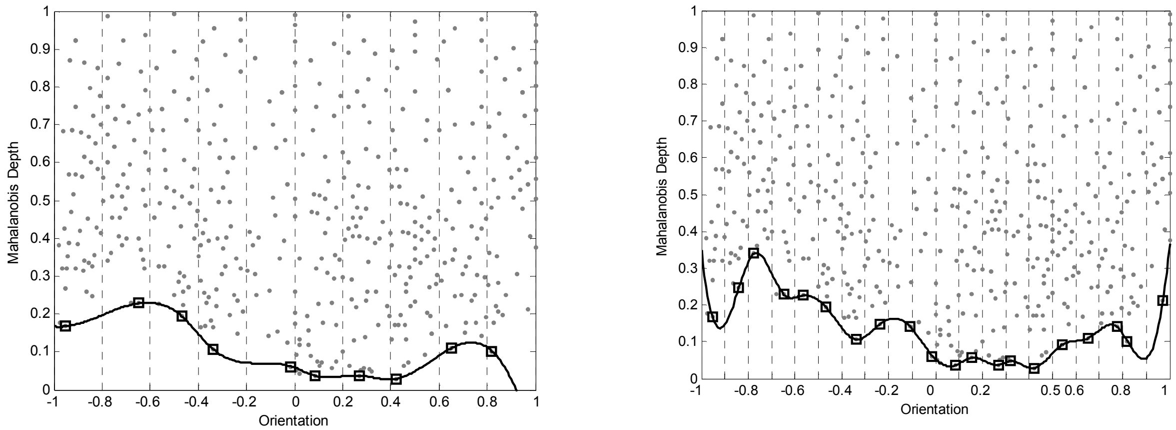


c) Halfspace depth



d) Simplicial depth

Figure 8: Histograms of depth values of the data set (O_3, NO_2) for different depth functions



a) b)

Figure 9: Identified extreme observation set $\Sigma_n(\lambda, D)$ of (O_3, NO_2) and the corresponding smooth curve $\mathbb{C}_n(\lambda, D)$ in the space (u, D) for a) $\lambda = 0.10$ and b) $\lambda = 0.05$ where D is MD