RKHS classification for multivariate extreme-value analysis

Miguel Piera-Martinez, Emmanuel Vazquez, Eric Walter, Gilles Fleury

RKHS classification for multivariate extreme-value analysis

Miguel Piera Martínez[1], Emmanuel Vazquez[2,*], Éric Walter[3] and Gilles Fleury[2]

1. European Space Agency, OPS-GSS section, 64293 Darmstadt, Germany
2. SUPELEC, 91192 Gif-sur-Yvette, France
3. Laboratoire des Signaux et Systèmes, CNRS, SUPELEC, UNIV. PARIS-SUD, 91192 Gif-sur-Yvette, France

* Corresponding author: emmanuel.vazquez@supelec.fr

Abstract — In many engineering applications, data samples are expensive to get and limited in number. In such a difficult context, this paper shows how classification based on Reproducing Kernel Hilbert Space (RKHS) can be used in conjunction with Extreme Value Theory (EVT) to estimate *extreme multivariate quantiles* and *small probabilities of failure*. For estimating extreme multivariate quantiles, RKHS one-class classification makes it possible to map vector-valued data onto $\mathbb{R}$, so as to estimate a high quantile of a univariate distribution by means of EVT. In order to estimate small probabilities of failure we basically apply multivariate EVT, however EVT is hampered by the fact that many samples may be needed before observing a single tail event. By means of a new method again based on RKHS classification, we can partially solve this problem and increase the proportion of tail events in the samples collected.

keywords Extreme value theory; RKHS classification; Multivariate quantiles; Failure probability

1 Introduction

This paper advocates the combination of Reproducing Kernel Hilbert Space (RKHS)-based classification with extreme value theory (EVT) to estimate two quantities of fundamental importance in engineering, namely *extreme multivariate quantiles* and *small probabilities of failure*.

A standard approach to build multivariate quantiles of a probability distribution is to estimate the level sets of its density (Molchanov, 1990). It is well

known that RKHS classification can be applied for the estimation of the level sets of a density (Vert and Vert, 2006). However, an accurate estimation of tail probabilities, which corresponds to the estimation of large level sets, should in general be based on large data sets. Extreme value theory (EVT), by proposing a generic parametric form for univariate tail distributions, makes it possible to obtain estimates with smaller variance than RKHS classification. To extend its applicability to a multivariate setting, a possible approach, explored in this paper, is to map the vector-valued data onto $\mathbb{R}$, so as to consider univariate distributions again. The first contribution of this paper is to explain how such a mapping can be carried out in the framework of RKHS classification.

Next, we shall show that classification methods can also be used to improve extreme value analysis in the case of failure probability estimation. Assume that $h : \mathbb{X} \subseteq \mathbb{R}^d \to \mathbb{Y} \subseteq \mathbb{R}^q$ is a multivariate vector-valued continuous function, and let X be a random vector on $\mathbb{X}$, with probability measure $\mathsf{P}_{\mathbb{X}}$. Our objective is to estimate the probability of failure

$$\mathsf{P}_f = \mathsf{P}_{\mathbb{X}}\{x\,;\; h(x) \in \Upsilon\}\,, \tag{1}$$

where Υ is some predefined failure domain, such that P_f is small. Drawing samples in $\mathbb{X}$ according to $\mathsf{P}_{\mathbb{X}}$, turns out to be a very inefficient approach to the evaluation of P_f, so many estimation methods use some type of importance sampling (e.g., Au and Beck, 2001 ; Homem-de Mello and Rubinstein, 2002 ; Nie and Ellingwood, 2004). A direct analysis of the tail of $h(X)$ using EVT compares poorly against such methods based on importance sampling. Indeed, EVT can be used to extrapolate the far-tail behaviour of a distribution from samples in the near-tail, but the method is hampered by the fact that many samples may be needed before a single tail event is observed. We propose to bypass this limitation by using a two-step method; in the first step, samples are drawn according to $\mathsf{P}_{\mathbb{X}}$ and a classification method is applied to build a subset $\Gamma \subset \mathbb{X}$ that contains the failure set $\{x; h(x) \in \Upsilon\}$; in the second step, samples are drawn in Γ by rejection sampling and EVT estimation is conducted from these samples.

Although EVT and RKHS classification have both been extensively studied, surprisingly enough, they do not seem to have been combined for treating multivariate extreme-value analysis problems, as suggested in this paper. Sections 2 and 3 briefly recall these two basic frameworks, Section 4 shows how EVT and one-class classification can be used to estimate extreme multidimensional quantiles, and Section 5 presents the estimation of failure probabilities via EVT and two-class classification. Simple illustrative examples, inspired by actual problems of robust system design in engineering, are also presented.

2 RKHS-based classification

Let $\{\mathbb{X}_j\}_{j\in\mathcal{J}}$ be a finite partition of $\mathbb{X}$ and $\mathbb{T} = \{t_j\}_{j\in\mathcal{J}}$ be a corresponding set of labels. Assume that there exists a function $\Phi^* : \mathbb{X} \to \mathbb{T}$ such that $\forall j \in \mathcal{J}$, $x \in \mathbb{X}_j \implies \Phi^*(x) = t_j$. From a set of training data $\{(x_i, \Phi^*(x_i)), i = 1, \ldots, n\}$,

a classification algorithm attempts to build a decision function $\Phi : \mathbb{X} \to \mathbb{T}$, which is an estimate of Φ^*, such that $\forall j \in \mathcal{J}$, $x \in \mathbb{X}_j \implies \Phi(x) = t_j$.

One- and two-class classifications both correspond to partitions with two subsets and we shall use the standard notations, $\mathbb{X} = \mathbb{X}_{-1} \cup \mathbb{X}_1$, with $t_{-1} = -1$ and $t_1 = 1$. The most standard type of classification has two classes. One-class classification is employed when data points are not labeled, but are instead supposed to belong with high frequency to one particular subset, $\mathbb{X}_1$ say.

With obvious notations, an important family of decision functions can be written as

$$\Phi(x) = \mathbf{1}_{f(x) \geq u} - \mathbf{1}_{f(x) < u},$$

where $f : \mathbb{X} \to \mathbb{R}$ is a smooth function such that, for some given u, and for all $i = 1, \ldots, n$, $f(x_i) < u$ if $x_i \in \mathbb{X}_{-1}$ and $f(x_i) \geq u$ if $x_i \in \mathbb{X}_1$. Functions of this family receive different names depending on how f is built. In this paper, we focus on RKHS-based methods (see, e.g., Wendland, 2005).

Let $\mathcal{F}$ be a Hilbert space of real-valued functions defined on $\mathbb{X}$, with scalar product denoted by $(\cdot, \cdot)_{\mathcal{F}}$. If there exists a function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, called a *reproducing kernel*, such that

$$\forall f \in \mathcal{F}, \quad \forall x \in \mathbb{X} \quad f(x) = (f, k(x, \cdot))_{\mathcal{F}}\,,$$

then $\mathcal{F}$ is an RKHS (Aronszajn, 1950).

A one-class RKHS classifier is obtained by solving the program

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{C}\sum_{i=1}^{n} l(f(x_i))\,, \tag{2}$$

where C is a tuning parameter, $\|f\|_{\mathcal{F}}^2$ is a regularization term and l is a loss term that penalizes functions f such that $f(x_i) < u$. For instance, a one-class *support vector machine* (SVM) is obtained with the hinge loss function $l(v) = \max(0, u - v)$ (see, e.g., Schölkopf et al., 2001).

A two-class RKHS classifier is obtained by taking $u = 0$ and solving the program

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{C}\sum_{i=1}^{n} l(f(x_i), \Phi^*(x_i))\,, \tag{3}$$

where l is a loss function that penalizes functions f such that $f(x_i) \geq 0$ when $\Phi^*(x_i) = -1$, and $f(x_i) < 0$ when $\Phi^*(x_i) = 1$. A standard choice is $l(f(x_i), \Phi^*(x_i)) = (f(x_i) - \Phi^*(x_i))^2$. Another example is the hinge loss function $l(f(x_i), \Phi^*(x_i)) = \max(0, \alpha - f(x_i)\Phi^*(x_i))$, which leads to standard two-class SVM with α-margins (see, e.g., Schölkopf and Smola, 2002).

Each of the programs (2) and (3) admits a unique solution f^*, which can be written as

$$f^* = \sum_{i=1}^{n} a_i k(x_i, \cdot)\,. \tag{4}$$

As a consequence, building one-class and two-class SVM boils down to solving quadratic finite-dimensional optimization problems (Schölkopf and Smola, 2002).

3 Extreme value theory

Assume, for the time being, that X is scalar. Under some technical conditions, the Pickands-Balkema-de Haan Theorem (Embrechts et al., 1997, page 354) suggests the following semi-parametric model (called the *generalized Pareto distribution* (GPD) model)

$$\begin{aligned} F(x) := \mathsf{P}\{X \leq x\} \approx 1 - \mathsf{P}\{X > u\} \left(1 + \xi \frac{x-u}{\beta}\right)^{-1/\xi}, \\ \forall\, x \text{ such that } x - u > 0 \text{ and } 1 + \xi \frac{x-u}{\beta} > 0\,, \end{aligned} \tag{5}$$

for the tail of the distribution of X above a given threshold u near the upper bound x_0 of its support. The validity of this model is asymptotic: the higher u, the more accurate the model becomes (see, e.g., (Coles, 2001) for a comprehensive discussion).

Once the threshold u has been fixed, the parameters ξ and β may be estimated by maximum likelihood or by the method of moments, among others (Embrechts et al., 1997, pages 327-348). Only the samples of X above the threshold are used for the estimation of ξ and β, as the model is only considered valid for $X > u$. $\mathsf{P}\{X > u\}$ is evaluated empirically by the Monte-Carlo method. Asymptotic validity suggests choosing a large u, but the number of data points available for the estimation of ξ and β then decreases, which leads to a higher variance (a bias-variance trade-off). Various numerical methods can be used in order to determine a suitable compromise (Dupuis, 1998).

A quantile at level $1 - \alpha$ may be estimated by inverting (5),

$$Q_{1-\alpha} \approx u + \frac{\beta}{\xi} \left[\left(\frac{\mathsf{P}\{X > u\}}{1-\alpha} \right)^{\xi} - 1 \right]. \tag{6}$$

For the estimation of failure probabilities in Section 5, we shall need a multidimensional extension of univariate EVT. By analogy with threshold methods for univariate extremes, in which the approximate generalized Pareto distribution is treated as exact for sufficiently high thresholds, Smith (1993) proposes, under some technical conditions, the following approximation[1] of the c.d.f of a d-dimensional vector, based on (Resnick, 1987, proposition 5.15),

$$\begin{aligned} F(x) \approx 1 - \upsilon \left(\frac{1}{\mathsf{P}\{X_1 > u_1\}} \Big(1 + \xi_1 \frac{x_1 - u_1}{\beta_1}\Big)^{1/\xi_1}, \cdots, \right. \\ \left. \frac{1}{\mathsf{P}\{X_d > u_d\}} \Big(1 + \xi_d \frac{x_d - u_d}{\beta_d}\Big)^{1/\xi_d} \right), \end{aligned} \tag{7}$$

valid for $x_1 \geq u_1, \cdots, x_d \geq u_d$. The function υ describes the dependence between the variables. A number of possible parameterizations for υ have been

[1] This model is only valid under asymptotic dependence. See (Ledford and Tawn, 1996) for its extension under asymptotic independence.

proposed (see Kotz and Nadarajah (2000) for an overview). For $d = 2$, one of the most popular model is the logistic model

$$\upsilon_\alpha(x_1, x_2) = \left(x_1^{-1/\alpha} + x_2^{-1/\alpha}\right)^\alpha . \tag{8}$$

A presentation of the various parameterizations and their relevance is out of the scope of this paper.

The estimation procedure consists of two steps (Coles, 2001). First, a univariate analysis of the marginal distributions is performed to estimate their tail behaviours. The thresholds $u_1, \cdots, u_d$ are chosen separately to make a trade-off between bias and variance for each marginal variable. The parameters $\xi_1, \cdots, \xi_d$ and $\beta_1, \cdots, \beta_d$ are estimated for each component using one of the methods cited above. Next, the parameters defined in the dependence structure (α in case of the logistic model) are estimated using censored likelihood (Smith, 1993). Only the samples for which at least one component is extreme ($x_i > u_i$) are used for this estimation.

4 Estimating extreme multidimensional quantiles with one-class SVM

Let $\mathcal{Q}$ be a class of measurable subsets of $\mathbb{X}$ and λ be a real-valued function defined on $\mathcal{Q}$. A *multivariate quantile* $Q_{1-\alpha}$ with respect to $(\mathbb{X}, \mathsf{P}, \mathcal{Q}, \lambda)$ is defined as a set $Q \in \mathcal{Q}$ that reaches the infimum

$$c(\alpha) = \inf\{\lambda(Q) : Q \in \mathcal{Q}\,,\ \mathsf{P}(Q) \geq 1 - \alpha\}\,, \quad 0 < \alpha \leq 1\,,$$

where $c(\alpha)$ is called *generalized quantile function* (Einmahl and Mason, 1992). Note that $Q_{1-\alpha}$ is not necessarily unique. If $\mathcal{Q}$ is the family of closed sets in $\mathbb{R}^d$ and λ the Lebesgue measure, then $Q_{1-\alpha}$ is a minimum-volume set that contains at least a $(1-\alpha)$-fraction of the probability mass.

Minimum-volume set estimation has been extensively studied. Sager (1979) and Hartigan (1987) address the particular case where $\mathcal{Q}$ is the class of convex closed sets in $\mathbb{R}^2$. Nolan (1991) works with ellipsoidal sets. Tsybakov (1997) uses piecewise-polynomial estimators. Whatever the class considered, the quality of the estimation decreases when the number of data points decreases and when the probability of the set of interest becomes closer to one (or zero).

Nunez-Garcia et al. (2003) show that density level sets correspond to minimum-volume sets. The reciprocal is not true: a minimum-volume set is not necessarily a density level set. Let $\widehat{f}_n(x)$ be an estimator of the density of X based on an n-sample of X. There exists c_α such that $\widehat{Q}_{1-\alpha} = \{x \in \mathbb{X} : \widehat{f}_n(x) \geq c_\alpha\}$ is a minimum-volume set estimator. Such estimators are called *plug-in estimators* (Molchanov, 1990).

To estimate extreme multidimensional quantiles, that is $Q_{1-\alpha}$ with $0 < \alpha \leq 1/n$, we need to extrapolate the behaviour of the available data. Classical methods are not suited to this case, hence the interest of the method proposed in this paper.

We consider a parameterized class of subsets $\mathcal{Q} = \{B_\rho,\, \rho \in \mathbb{R}^+\}$ such that

$$\rho_1 < \rho_2 \implies B_{\rho_1} \supset B_{\rho_2}\,. \tag{9}$$

λ is defined by $\lambda(B_\rho) := \rho$, for every $B_\rho \in \mathcal{Q}$.

In order to build such a parameterized class of subsets, we use the one-class SVM classifier obtained by solving the following program (Schölkopf et al., 2001)

$$\min_{f\in\mathcal{F},\xi\in\mathbb{R}^n,\rho\in\mathbb{R}} \quad \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \frac{1}{n\nu}\sum_{i=1}^{n}(\xi_i - \rho)\;, \tag{10}$$

$$\text{subject to} \begin{cases} f(x_i) = (f, k(x_i,\cdot))_{\mathcal{F}} \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{11}$$

where the parameter $\nu \in (0,1]$ controls the trade-off between the regularization term $\|f\|_{\mathcal{F}}^2$ and the constraints. The ξ_is are called *slack variables*. Denote by $f^*_{\nu,n}$, ξ^* and ρ^* the solutions of (10, 11). This program is a particular case of (2). The convergence of one-class SVM to minimum-volume sets has been proved recently (Vert and Vert, 2006), which motivates our choice of one-class SVM over other classification methods.

The fraction of the data points such that $(f^*_{\nu,n}, k(x,\cdot))_{\mathcal{F}} - \rho^* = f^*_{\nu,n}(x) - \rho^* < 0$ tends to ν when $n \to \infty$ (Schölkopf et al., 2001). Take $B^{\nu,n}_\rho := \{x \in \mathbb{X} : f^*_{\nu,n}(x) > \rho\}$ and let $\mathcal{Q}_n$ be the family of subsets

$$\{B^{\nu,n}_\rho; \rho \geq 0\}\,. \tag{12}$$

$\mathcal{Q}_n$ satisfies (9) and

$$\mathsf{P}(B^{\nu=\alpha,n}_{\rho^*}) \to 1-\alpha\,, \tag{13}$$

when n tends to infinity. Vert and Vert (2006) show that $f^*_{\nu,n}(x)$ is a density estimator of X truncated at ρ^* if k is a Gaussian kernel. $B^{\nu=\alpha,n}_{\rho^*}$ therefore tends to a $(1-\alpha)$-quantile of minimal volume (Nunez-Garcia et al., 2003).

However, when $\alpha \leq 1/n$, estimating $Q_{1-\alpha}$ by $B^{\nu=\alpha,n}_{\rho^*}$ is not a viable option, as the convergence of (13) is very slow. We use EVT instead, in order to estimate the tail of $f_{\nu,n}(X)$.

Our objective is to accelerate the convergence of $\mathsf{P}(B^{\nu,n}_\rho) = \mathsf{P}\{x \in \mathbb{X} : f_{\nu,n}(x) > \rho\}$ to $1-\alpha$ when n goes to infinity. For this purpose, we look for a more suitable tuning of ν and ρ than the choice $\rho = \rho^*$ and $\nu = \alpha$ suggested by (13). In order to improve the convergence rate, and thus to reduce the number of data points required, we need to extrapolate the behaviour of the data. The main idea is to transform the multi-dimensional problem into a one-dimensional one using the fact that

$$\mathsf{P}\{x \in \mathbb{X} : f_{\nu,n}(x) > \rho\} = \mathsf{P}_{f_{\nu,n}}(]\rho, +\infty[)\,,$$

where $\mathsf{P}_{f_{\nu,n}}$ is a probability on $\mathbb{R}$ defined as the probability image of P by $f_{\nu,n}$. Our algorithm consists of two steps.

(a) Program (10, 11) is solved for a Gaussian kernel and a value of ν chosen so that there are enough support vectors (typically $\nu = 0.5$). One thus obtains a member of the family $\mathcal{Q}_n$ defined by (12).

(b) EVT is used to estimate ρ^{**} from $\{f_{\nu,n}(x_i)\,;\ i = 1, \ldots, n\}$ such that

$$\mathsf{P}_{f_{\nu,n}}(]\rho^{**}, +\infty[) \approx 1 - \alpha\,.$$

Example — In order to illustrate the convergence acceleration achieved by using the method presented in the previous section, we take $X = (X_1, X_2)^{\mathsf{T}}$ where

$$X_1 = 5\sin(0.125\pi + \theta) + V \ \text{ and } \ X_2 = 5\cos(0.125\pi + \theta) + V\,.$$

The random variable θ follows a uniform distribution $U(0, 1.25\pi)$ and V follows a normal distribution $N(0, 1)$.

Figure 1 shows 200 samples of X and $\widehat{Q}_{0.999} = B_{\rho^{**}}^{\nu=0.5,n}$ estimated by the proposed method[2].

Table 1 compares the means and standard deviations of $P(\widehat{Q}_{0.999})$, when $\widehat{Q}_{0.999}$ is estimated using the SVM only and when our methodology combining SVM and extreme value theory is used.

	$\mathsf{P}(X \in \widehat{Q}_{0.999,\,\mathrm{SVM}})$	$\mathsf{P}(X \in \widehat{Q}_{0.999,\,\mathrm{SVM+EVT}})$
$n = 50$	0.7419 (0.0769)	0.9816 (0.02816)
$n = 100$	0.8437 (0.0486)	0.9923 (0.0145)
$n = 200$	0.9186 (0.0149)	0.9964 (0.0042)
$n = 400$	0.9561 (0.0072)	0.9979 (0.0025)

Table 1: Comparison, based on 100 trials, of the mean (and standard deviation) of $P(\widehat{Q}_{0.999})$ when $\widehat{Q}_{0.999}$ is computed using the SVM only and with combining SVM and EVT.

This example illustrates the superior performance of the new method. The speed of convergence of the estimator remains to be studied, as well as its sensitivity to the dimension of $\mathbb{X}$, the tail of the underlying distribution and the choice of the kernel.

5 RKHS classification for the estimation of failure probabilities

The estimation of failure probabilities could, at least in principle, be carried out directly by using EVT, since the choice of a probability distribution $\mathsf{P}_{\mathbb{X}}$ induces

[2]The choice of σ is of course an important problem in practice, not considered here for the sake of brevity. Choosing σ, or more generally the kernel, is very similar to choosing the kernel in the Parzen-Rozenblatt density estimator, for which numerous procedures have been developed (see for instance Duong (2004) and Scott (1992)). Here, we used $\sigma = ((\mathrm{var}(X_1) + \mathrm{var}(X_2))/2)^{1/2}$.

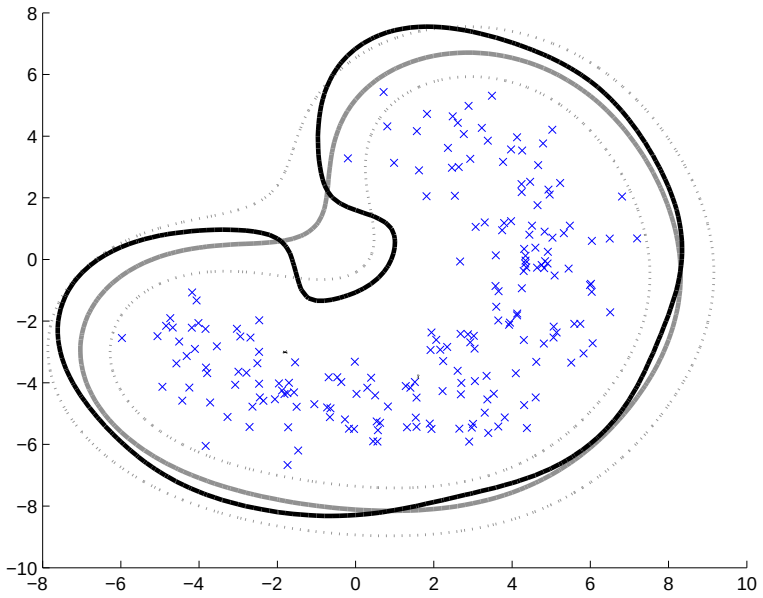

Figure 1: 0.999 -quantile as estimated from 200 sample values (solid line in gray), 90% confidence regions are also drawn (dotted lines). The Gaussian kernel with $\sigma = 3.2$ has been used. The actual minimum-volume quantile (as estimated from 10^5 sample values) is in black.

a probability distribution on $\mathbb{Y}$, defined by

$$\forall A \subset \mathbb{Y}\,, \quad \mathsf{P}_{\mathbb{Y}}(A) := \mathsf{P}_{\mathbb{X}}(h^{-1}(A))$$

A key point for the applicability of EVT, however, is the availability of a sufficiently large number of samples of $Y = h(X)$ above u_i $(i = 1, \cdots, q)$ in (7). This number increases at the rate $n\mathsf{P}\{Y_i > u_i\}$ for each marginal $i = 1, \cdots, q$, where n is the sample size. If $\mathsf{P}\{Y_i > u_i\}$ is small for at least one marginal, which is usually the case because u_i has to be large in order to reduce bias, many samples are needed to get enough data above each u_i to ensure an acceptable variance. Unfortunately, when estimating failure probabilities (1), the number of available samples if often small (the collection of samples may involve complex and time-consuming simulations or the realization of prototypes), so EVT can seldom be applied directly in practice.

To bypass this difficulty, we propose to use the following two-step approach, based on RKHS classification.

(a) Choose a set $S \subset \mathbb{Y}$ that contains Υ. Draw n_1 samples x_i in $\mathbb{X}$ according

to $\mathsf{P}_{\mathbb{X}}$. Build the decision function Φ of an RKHS classifier, such that $\forall x \in \mathbb{X}$, $\Phi(x) = 1$ when $h(x) \in S$, and $\Phi(x) = -1$ otherwise. Estimate $\mathsf{P}(h(X) \in S)$ by $\mathrm{card}\{x_i\,;\; h(x_i) \in S\}/n_1$.

(b) By rejection sampling driven by the RKHS classifier, draw n_2 samples x_i in $\mathbb{X}$ according to $\mathsf{P}_{\mathbb{X}}$ and satisfying $\Phi(x_i) = 1$. Perform an EVT analysis based on these samples, which means (i) selecting the thresholds $u_1, \cdots, u_q$ such that the EVT model (7) is valid, (ii) estimating the probabilities $\mathsf{P}\{Y_i > u_i\}$, $i = 1, \cdots, q$, with the correction needed to take into account the effect of rejection sampling, (iii) estimating (ξ_i, β_i), $i = 1, \cdots, q$.

Denote by $\widehat{F}_{\mathrm{EVT}}$ the model of the tail distribution of Y obtained by using (7). The failure probability is estimated by

$$\widehat{P}_f = \int_{\Upsilon} \frac{\partial \widehat{F}_{\mathrm{EVT}}(y)}{\partial y^{\mathsf{T}}} dy\,. \tag{14}$$

The total number of evaluations of h is then equal to $n = n_1 + n_2$. In many engineering applications, the evaluation of $h(x)$ is expensive and the budget for such evaluations is limited. Hence the interest of the approach proposed, which leads to a smaller variance of the estimators than would have been obtained if the RKHS classifier had not been used to select suitable values of x. Figure 2 illustrates the approach.

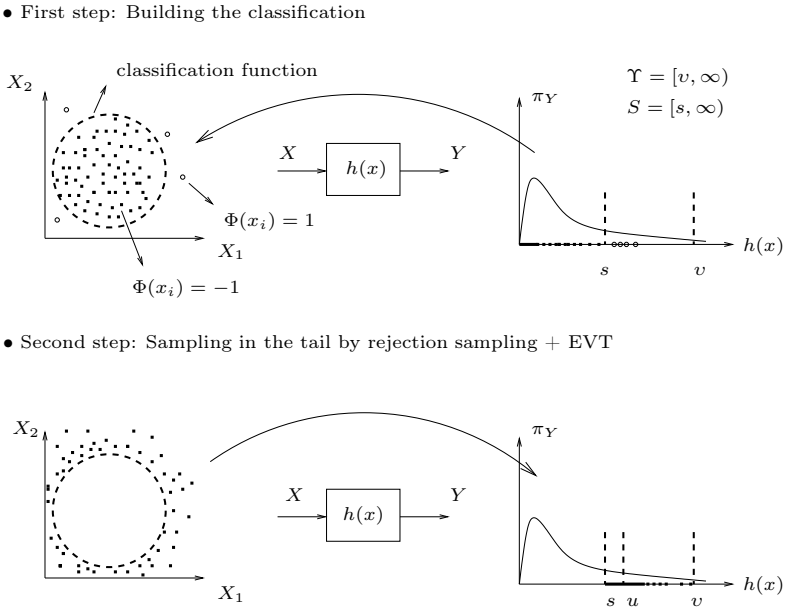

Figure 2: Two-step estimation of failure probabilities based on RKHS classification ($d = 2$ and $q = 1$).

Example — Extreme values play an important role in engineering because they may correspond to abnormal or dangerous operating conditions. The following example is an academic version of an aeronautical problem. A rough approximation of the deflexion of the tip of the wing of an airplane is given by a beam-deflection formula

$$d = h(f, l, E, I) = -\frac{fl^3}{6EI}\,,$$

where f is a force acting on the wing, l is the wing span, E is the modulus of elasticity, and I is the area moment of inertia. In this example, we shall assume that $f \sim \mathcal{N}(1,1)$, $l \sim \mathcal{N}(10,1)$, $E \sim \mathcal{N}(50,1)$, $I \sim \mathcal{N}(5,1)$ and that the failure domain Υ is $(-\infty, -4]$. During actual airplane conception, more complex and more realistic models are of course employed, but it should be noted that the methods advocated here are especially interesting for such complex models, for which simulation budget is severely limited.

Table 2 shows failure probabilities as estimated by means of classical EVT and the method proposed. We choose $S = (-\infty, 1]$. A two-class SVM is used for classification. The threshold u, above which the GPD model is valid, has been set to 1.5, based on the *mean excess plot* method (Coles, 2001). The actual value of the failure probability is considered to be the result of a Monte Carlo experiment with 10^6 samples. Thus, we get $P_f = 0.0028$.

	$\widehat{P}_f^{\text{SVM+EVT}}$	$\widehat{P}^{\text{EVT}}$
$n_1 = 50, n_2 = 50 \implies n = 100$	0.0034 (0.0031)	0.0033 (0.0034)
$n_1 = 50, n_2 = 150 \implies n = 200$	0.0028 (0.0016)	0.0032 (0.0030)
$n_1 = 50, n_2 = 450 \implies n = 500$	0.0028 (0.0010)	0.0028 (0.0019)
$n_1 = 50, n_2 = 950 \implies n = 1000$	0.0028 ($8.1\,10^{-4}$)	0.0028 (0.0015)
$n_1 = 50, n_2 = 4950 \implies n = 5000$	0.0028 ($3.7\,10^{-4}$)	0.0028 ($8.4\,10^{-4}$)

Table 2: Comparison, based on 200 trials, of the mean (and standard deviation) of $\widehat{P}_f^{\text{SVM+EVT}}$ and $\widehat{P}_f^{\text{EVT}}$.

This example shows the superior performance of the method proposed. However, the performance depends on the choice of S, n_1 and n_2. The optimal choice for these parameters remains to be studied.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68: 337–404, 1950.

S.K. Au and J.L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engrg. Mech.*, 16:263–277, 2001.

S.G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

T. Duong. *Bandwidth Selectors for Multivariate Kernel Density Estimation*. PhD thesis, Univ. Western Australia, 2004.

D.J. Dupuis. Exceedances over high thresholds: a guide to threshold selection. *Extremes*, 1(3):251–261, 1998.

J.H.J. Einmahl and D.M. Mason. Generalized quantile processes. *Ann. Statist.*, 20(2):1062–1078, 1992.

P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997.

J.A. Hartigan. Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82:267–270, 1987.

T. Homem-de Mello and R.Y. Rubinstein. Rare event probability estimation using cross-entropy. In E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 310–319. Winter Simulation Conferenc, 2002.

S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London, 2000.

A.W. Ledford and J.A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187, 1996.

I.S. Molchanov. Empirical estimation of distribution quantiles of random closed sets. *Theory Probab. Appl.*, 35:594–600, 1990.

J. Nie and B.R. Ellingwood. A new directional simulation method for system reliability I: Application of deterministic point sets. *Probabilistic Engrg. Mech.*, 19(4):425–436, 2004.

D. Nolan. The excess mass ellipsoid. *J. Multivariate Anal.*, 39:348–371, 1991.

J. Nunez-Garcia, Z. Kutalik, K.H. Cho, and O. Wolkenhauer. Level sets and minimum volume sets of probability distribution functions. *Approximate Reasoning*, 34:25–47, 2003.

S.I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Verlag, 1987.

T.W. Sager. An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.*, 74:329–339, 1979.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

D.W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. Wiley series in Probability, 1992.

R.L. Smith. Multivariate threshold methods. Technical report, National Institute of Statistical Sciences, 1993. www.niss.org/technicalreports/tr7.pdf.

A.B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997.

R. Vert and J.P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. *J. Mach. Learn. Res.*, 7:817–854, 2006.

H. Wendland. *Scattered Data Approximation*. Monographs on Applied and Computational Mathematics. Cambridge Univ. Press, Cambridge, 2005.