Nonparametric estimation of distributions of order statistics with application to nuclear engineering

C. Butucea
LAMA UMR 8050 (UPE-MLV)
Université Paris-Est, Marne-la-Vallée, France

J-F. Delmas
CERMICS (ENPC)
Université Paris-Est, Marne-la-Vallée, France

A. Dutfoy & R. Fischer
Department of Industrial Risk Management
EDF Research & Development, Clamart, France

ABSTRACT:
In this paper we focus on the modelling of d-dimensional random vectors with fixed marginals and components which verify an ordering constraint almost surely. We aim to calculate the distribution which maximizes the Shannon-entropy under these constraints. We provide the solution with explicit formulas to the maximum entropy problem, which has a particular form. Namely, the density of the optimal joint distribution becomes a product of univariate functions on the support of the random vector.
To exploit the special structure of such distributions, we propose a nonparametric approach to estimate the density of the joint distribution based on an available sample. We propose an exponential model based on series of quasi-orthogonal polynomials specially designed to suit this particular structure. We show that by exploiting the special features of the model, we achieve a fast convergence rate which depends only linearly on the dimension of our random vector, thus does not suffer from the curse of dimensionality which arises when dealing with high-dimensional density estimation problems.
We apply the proposed method to an industrial application case involving the estimation of mechanical flaw dimensions in a component of a power plant, with experimental data available. We compare the results obtained with the maximum entropy approach to previously considered modelling schemes.

1 INTRODUCTION

In various cases of modelling the propagation of uncertainty in complex numerical simulation schemes, the engineer needs to create a probabilistic model for the joint distribution of a vector of uncertain quantities, that takes into consideration the available statistical information (or expert judgement). This could include detailed knowledge of the marginal distribution of each uncertain quantity, certain aspects of the dependence structure between the components of the vector, or it can also be a physical constraint between these quantities, for example an order relationship. The construction of a model compatible with all the information is crucial in order to correctly evaluate a probabilistic reliability criterion by Monte Carlo methods.

In this paper we focus on the joint distribution of ordered random vectors. That is, we consider random vectors $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ that verify $\mathbb{P}(X \in S) = 1$, with $S = \{(x_1, \dots, x_d) \in \mathbb{R}^d, x_1 \leq \cdots \leq x_d\}$. We will refer to these vectors as vectors of order statistics without reference to a parent distribution. We assume that the one-dimensional marginal distributions $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$ are given, where $\mathbf{F}_i$ is the cumulative distribution function of X_i. In an industrial context, information on the individual behaviour of the components of the random vector is often available, but less attention is given to the modelling of the dependence structure between the components. Given the marginals, the joint distribution of the ordered random vector can be characterized by its connecting copula, which contains all information on the dependence structure. Copulas of order statistics derived from an underlying distribution by sorting its components in ascending order were considered in (Avérous, Genest, & C Kochar 2005) in the i.i.d. case and in (Navarro & Spizzichino 2010) with a general set-up. (Lebrun & Dutfoy 2014) characterizes the possible copula functions of ordered random vectors with given marginals by describing their supports, and proposes an admissible construction of such a copula called sub-square copula.

Among the possible joint distributions (if there exists any), our aim is to find the model which has minimum information content in addition to our constraints. We measure the uncertainty of our joint model by the Shannon entropy, which has been widely used in uncertainty and risk measurement in the literature. Under some mild constraints imposed on the marginals, we calculate the density f^* of the distribution which maximizes the entropy in Section 2, which admits a product form on its support S, see Remark 2.1. We use the optimization techniques for infinite dimensional constraints developed by (Borwein, Lewis, & Nussbaum 1994) to obtain this result.

In the second part of the paper, we propose a nonparametric exponential series estimator for densities with product form on S. We propose a model in which we approximate the logarithm of the density by series of quasi-orthogonal polynomials specially designed to suit this particular structure. Approximation of log-densities by polynomials appear in (Good 1963) as an application of the maximum entropy principle, while (Crain 1977) show existence and consistency of the maximum likelihood estimation. We measure the quality of the density estimator by the Kullback-Leibler divergence. We show that Kullback-Leibler divergence of the estimator and the true underlying density converges fast in probability to zero. Convergence rates in probability for the exponential series estimator was given in (Barron & Sheu 1991) for $d = 1$ and (Wu 2010) for $d \geq 2$ without the reduced support.

We apply the estimation method to an industrial case involving the estimation of flaw dimensions in a passive component in a power plant. The correct modelling of these quantities is quintessential for assessing the failure probability of these components, as the propagation of such flaws can lead to rupture, damaging the integrity of the component. We compare our methodology to the previous work of (Remy, Popelin, & Feng 2012) where a parametric copula approach was considered. This model does not take into consideration the ordering constraint between the dimensions. We evaluate the failure probability for the component with each method and compare the results. We also include a simulated case study, with data generated from a maximum entropy distribution of order statistics to demonstrate the efficiency of the nonparametric approach.

The rest of the paper is organised as follows. In Section 2, we give the explicit solution to the problem of maximum entropy distribution of order statistics with given marginals, which has a product form on S. In Section 3 we introduce the exponential model for estimating densities of product form on S, and we show that this estimator converges rapidly to the true density in Kullback-Leibler divergence. Section 4 contains the definition and some properties of the basis functions used throughout the estimation process. Finally in Section 5, we apply the nonparametric method with a real dataset (Section 5.4.1) and with simulated data (Section 5.4.2) to assess its performance compared to other approaches.

2 MAXIMUM ENTROPY DISTRIBUTION OF ORDER STATISTICS WITH GIVEN MARGINALS

In this section, we give the joint distribution of an ordered random vector with fixed marginals which maximizes the differential entropy H defined as, for a random variable Z with density f_Z:

$$H(Z) = -\int f_Z(z) \log f_Z(z) dz,$$

and $H(Z) = -\infty$ if Z does not have a density. If F_Z denotes the distribution function of Z, we use the convention $H(F_Z) = H(Z)$. In an information-theoretic interpretation, the maximum entropy distribution is the least informative among the distributions which verify the constraints. For a d-dimensional random vector $X = (X_1, \ldots, X_d)$ with distribution function F and copula function C_F, the entropy of F can be decomposed into the sum of the entropy of its one-dimensional marginals $\mathbf{F}_i$, $1 \leq i \leq d$, plus the entropy of C_F (see (Zhao & Lin 2011)):

$$H(F) = \sum_{i=1}^{d} H(\mathbf{F}_i) + H(C_F).$$

In our case, since the marginals $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$ are fixed, maximizing the entropy of the joint distribution F of the ordered random vector equivalent to maximizing the entropy of its copula C_F. A similar problem was considered in (Butucea et al. 2015a)

We give the solution to this problem under the condition that $\mathbf{F}_i > \mathbf{F}_{i+1}$ on the common part of their supports $\{t \in \mathbb{R}; 1 > \mathbf{F}_i(t), \mathbf{F}_{i+1}(t) > 0\}$ for $1 \leq i \leq d-1$. If

$$\sum_{i=2}^{d} \int_{\mathbb{R}} \mathbf{f}_i(s) \left| \log \left(\mathbf{F}_{(i-1)}(s) - \mathbf{F}_{(i)}(s) \right) \right| ds < +\infty,$$

and $H(\mathbf{F}_i) > -\infty$ for all $1 \leq i \leq d$, then there is a unique distribution $F_{\mathbf{F}}$ which maximizes the entropy, which has density function $f_{\mathbf{F}}$ given by, for

$x = (x_1, \dots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^{d} \left(\frac{\mathbf{f}_i(x_i)}{\mathbf{F}_{i-1}(x_i) - \mathbf{F}_i(x_i)} \times \exp\left(- \int_{x_{i-1}}^{x_i} \frac{\mathbf{f}_i(s)}{\mathbf{F}_{i-1}(s) - \mathbf{F}_i(s)} \, ds \right) \right) \mathbf{1}_S(x), \tag{2}$$

where $\mathbf{f}_i$ is the marginal density function of X_i, $1 \leq i \leq d$. See (Butucea et al. 2015b) for the detailed proofs.

Remark 2.1. We remark that the density has a product form on the domain S:

$$f_{\mathbf{F}}(x) = \prod_{i=1}^{d} p_i(x_i) \mathbf{1}_S(x)$$

for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. This is also the unique distribution of with fixed marginals whose density is a product of univariate functions on S.

3 ESTIMATION OF $F_{\mathbf{F}}$ BY AN EXPONENTIAL MODEL

In this section, we concentrate on the estimation of densities with product form on S by a nonparametric approach given a sample of n independent observations $X^1, \dots, X^n$. We restrict ourselves to densities supported on $\triangle = [0,1]^d \cap S$. Such a density, say f^0, can be written in the form, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$:

$$f^0(x) = \exp\left(\sum_{i=1}^{d} \ell_i^0(x_i) - \mathrm{a}_0 \right) \mathbf{1}_\triangle(x), \tag{1}$$

with ℓ_i^0 bounded,measurable functions on I for all $1 \leq i \leq d$, and normalizing constant a_0. For $r \in \mathbb{N}^*$, let W_r^2 denote the Sobolev space of functions on $[0,1]$, such that the $(r-1)$-th derivative is absolutely continuous and the L^2 norm of the r-th derivative is finite. Let ℓ_i^0 belong to the Sobolev space $W_{r_i}^2$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. We present an exponential model specifically designed to estimate such densities. This exponential model is a multivariate version of the family considered in (Barron & Sheu 1991) in a univariate setting. Essentially, we approximate the functions ℓ_i^0 by a family of polynomials $(\varphi_{i,k}, k \in \mathbb{N})$, which are orthonormal for each $1 \leq i \leq d$ with respect to q, the uniform weight function on the support $\triangle$: $q(x) = \mathbf{1}_\triangle(x)$. The estimator takes the form, for $x = (x_1, \dots, x_d) \in \triangle$:

$$f_\theta(x) = \exp\left(\sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) - \psi(\theta) \right),$$

with

$$\psi(\theta) = \log\left(\int_\triangle \exp\left(\sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) \right) dx \right).$$

This model is a reduced version of the multidimensional exponential series estimator introduced in (Wu 2010), as we have only kept the univariate terms of the basis since the logarithm of the target density is the sum of univariate functions. Furthermore, we have restricted our model to $\triangle$ instead of the hyper-cube I^d, and we have chosen the basis functions $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ which are suited for this support. The choice for the polynomials $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ will be given in Section 4 where we discuss some of their key properties. In particular, the polynomials $(\varphi_{i,k}, k \in \mathbb{N})$ are orthonormal for each $1 \leq i \leq d$ with respect to q, but for $i \neq j$, the families $(\varphi_{i,k}, k \in \mathbb{N})$ and $(\varphi_{j,k}, k \in \mathbb{N})$ are not orthogonal with respect to q, see Lemma 4.3. We estimate the parameters $(\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$ by solving the maximum likelihood equations:

$$\int_\triangle \varphi_{i,k}(x_i) f_{\hat{\theta}}(x) \, dx = \frac{1}{n} \sum_{j=1}^{n} \varphi_{i,k}(X^j). \tag{3}$$

For $m = (m_1, \dots, m_d)$ basis functions used, we estimate f by $\hat{f}_{m,n} = f_{\hat{\theta}}$. If such $\hat{\theta}_n$ does not exist, let us take $\hat{f}_{m,n} = d! \mathbf{1}_\triangle$.

We measure the quality of the estimation $\hat{f}_{m,n}$ of f^0 by the Kullback-Leibler divergence $D(f^0 || \hat{f}_{m,n})$ defined as:

$$D(f^0 || \hat{f}_{m,n}) = \int_\triangle f^0 \log\left(\frac{f^0}{\hat{f}_{m,n}} \right).$$

Convergence rates for nonparametric density estimators have been given by (Hall 1987) for kernel density estimators, (Barron & Sheu 1991) and (Wu 2010) for the exponential series estimators, (Barron et al. 1992) for histogram-based estimators, and (Koo & Kim 1996) for wavelet-based log-density estimators.

We show that if we take $m_i = m_i(n)$ members of the families $(\varphi_{i,k}, k \in \mathbb{N})$, $1 \leq i \leq d$, and let m_i grow with n in an appropriate way, then the maximum likelihood estimator $f_{\hat{\theta}}$ exists with probability tending to 1, and converges rapidly. The main result is given by the following Theorem whose proof is detailed in (Butucea et al. 2015b).

Theorem 3.1. *Let $f^0 \in \mathcal{P}(\triangle)$ be a probability density with a product form given by (1). Assume the functions ℓ_i^0 belongs to the Sobolev space $W_{r_i}^2$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $(X^n, n \in \mathbb{N}^*)$ be i.i.d. random variables with density distribution f^0. We consider sequences $m_i = m_i(n) \to \infty$ as $n \to \infty$, such that*

$$\left(\sum_{i=1}^{d} m_i \right)^{2d} \left(\sum_{i=1}^{d} m_i^{-2r_i} \right) \to 0, \tag{4}$$

$$\frac{\left(\sum_{i=1}^d m_i\right)^{2d+1}}{n} \to 0. \tag{5}$$

The maximum likelihood estimator $\hat{f}_{m,n} = f_{\hat{\theta}_n}$, with $\hat{\theta}_n$ obtained by solving (3), exists with probability tending to 1 as $n \to \infty$. The Kullback-Leibler distance $D\left(f^0 \| \hat{f}_{m,n}\right)$ of $\hat{f}_{m,n}$ to f^0 converges in probability to 0 with the convergence rate:

$$D\left(f^0 \| \hat{f}_{m,n}\right) = O_p\left(\sum_{i=1}^d \left(m_i^{-2r_i} + \frac{m_i}{n}\right)\right). \tag{6}$$

The proof of this Theorem relies on the classic bias-variance decomposition of the estimation error.

Remark 3.2. Notice that this is the sum of the same univariate convergence rates as in (Barron & Sheu 1991). Let us take $m_i = O(n^{1/(2r_i+1)})$. Then the conditions (4) and (5) are satisfied, and we obtain that $D\left(f \| \hat{f}_{m,n}\right) = O_p(\sum_{i=1}^d n^{-2r_i/(2r_i+1)})$. This is further of the order of $n^{-2\min(r)/(2\min(r)+1)}$ corresponding to the least smooth ℓ_i^0. The obtained convergence rate is optimal in the minimax sense for one-dimensional nonparametric density estimation as shown in (Yang & Barron 1999)

4 ORTHONORMAL SERIES OF POLYNOMIALS ON $\triangle$

In this section we precise the choice for the basis functions $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$. Essentially, they are shifted versions of certain Jacobi polynomials. We first recall the definition and some properties of Jacobi polynomials, then we define the functions $((\varphi_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ and give some of their key properties. They are easy to implement as the Jacobi polynomials are readily available in most mathematical computational platforms.

4.1 *Jacobi polynomials*

The following results can be found in (Abramowitz & Stegun 1970) p. 774. The Jacobi polynomials $(P_k^{(\alpha,\beta)}, k \in \mathbb{N})$ for $\alpha, \beta \in (-1, +\infty)$ are series of orthogonal polynomials with respect to the measure $w_{\alpha,\beta}(t)\mathbf{1}_{[-1,1]}(t)\,dt$, with

$$w_{\alpha,\beta}(t) = (1-t)^\alpha (1+t)^\beta \quad \text{for } t \in [-1,1].$$

They are given by Rodrigues' formula, for $t \in [-1,1]$, $k \in \mathbb{N}$:

$$P_k^{(\alpha,\beta)}(t) = \frac{(-1)^k}{2^k k! w_{\alpha,\beta}(t)} \frac{d^k}{dt^k}\left[w_{\alpha,\beta}(t)(1-t^2)^k\right].$$

The normalizing constants are given by:

$$\int_{-1}^1 P_k^{(\alpha,\beta)}(t) P_\ell^{(\alpha,\beta)}(t) w_{\alpha,\beta}(t)\,dt = \tag{7}$$

$$\mathbf{1}_{\{k=\ell\}} \frac{2^{\alpha+\beta+1}}{2k+\alpha+\beta+1} \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+\alpha+\beta+1)k!}.$$

In what follows, we will be interested in Jacobi polynomials with $\alpha = d - i$ and $\beta = i - 1$, which are orthogonal to the weight function $w_{d-i,i-1}(t) = \mathbf{1}_{[-1,1]}(t)(1-t)^{d-i}(1+t)^{i-1}$.

4.2 *Definition of the basis functions*

Based on the Jacobi polynomials, we define a shifted version, normalized with respect to the measure q and adapted to the interval $I = [0,1]$.

Definition 4.1. *For $1 \leq i \leq d$, $k \in \mathbb{N}$, we define for $t \in I$:*

$$\varphi_{i,k}(t) = \rho_{i,k}\sqrt{(d-i)!(i-1)!}\, P_k^{(d-i,i-1)}(2t-1),$$

with

$$\rho_{i,k} = \sqrt{\frac{(2k+d)k!(k+d-1)!}{((k+d-i)!(k+i-1)!)}}. \tag{8}$$

Let q_i, $1 \leq i \leq d$ be the one-dimensional marginals of the measure q:

$$q_i(t) = \frac{(1-t)^{d-i} t^{i-1}}{(d-i)!(i-1)!} \mathbf{1}_I(t). \tag{9}$$

According to the following Lemma, the polynomials $(\varphi_{i,k}, k \in \mathbb{N})$ form an orthonormal basis of $L^2(q_i)$ for all $1 \leq i \leq d$. The proof is elementary and we leave it to the Reader.

Lemma 4.2. *For $1 \leq i \leq d$, $k, \ell \in \mathbb{N}$, we have:*

$$\int_I \varphi_{i,k}\varphi_{i,\ell}\, q_i = \mathbf{1}_{\{k=\ell\}}.$$

4.3 *Mixed scalar products*

We give the mixed scalar products of $(\varphi_{[i],k}, k \in \mathbb{N})$ and $(\varphi_{[j],\ell}, \ell \in \mathbb{N})$, $1 \leq i < j \leq d$ with respect q.

Lemma 4.3. *For $1 \leq i < j \leq d$ and $k, \ell \in \mathbb{N}$, we have:*

$$\int \varphi_{i,k}(x_i)\, \varphi_{j,\ell}(x_j) q(x) =$$

$$\mathbf{1}_{\{k=\ell\}} \sqrt{\frac{(j-1)!(d-i)!}{(i-1)!(d-j)!}} \sqrt{\frac{(k+d-j)!(k+i-1)!}{(k+d-i)!(k+j-1)!}}.$$

We also have $\left|\int \varphi_{i,k}(x_i)\, \varphi_{j,\ell}(x_j) q(x)\right| \leq 1$ for all $k, \ell \in \mathbb{N}$.

We refer to (Butucea et al. 2015b) for the proof of this Lemma. This shows that the family of functions $(\varphi_{i,k}, 1 \leq i \leq d, k \in \mathbb{N})$ is not orthogonal with respect to the Lebesgue measure on $\triangle$. This result also allows us to bound the $L^2(q)$ norm of the function $\sum_{i=1}^{d}\sum_{k=1}^{m_i}\theta_{i,k}\varphi_{i,k}(x_i)$ from either sides by the Euclidean norm of the vector of parameters $\theta = (\theta_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ given by $\|\theta\| = \sum_{i=1}^{d}\sum_{k=1}^{m_i}\theta_{i,k}^2$. These bounds were used to control the bias error of the estimator.

Lemma 4.4. *For all $\theta = (\theta_{i,k}, k \in \mathbb{N}), 1 \leq i \leq d)$ we have:*

$$\frac{\|\theta\|}{\sqrt{d}} \leq \int \left(\sum_{i=1}^{d}\sum_{k=1}^{m_i}\theta_{i,k}\varphi_{i,k}(x_i)\right)^2 q(x) \leq \sqrt{d}\,\|\theta\|.$$

Once again, the proof can be found in (Butucea et al. 2015b).

5 APPLICATION TO NUCLEAR ENGINEERING DATA

5.1 *Industrial context and the dataset*

In this section, we apply the proposed methodology to estimate the joint distribution of the dimensions of flaws of a passive component in an EDF electric power plant. These flaws may lead to a crack under the severe stress to which the material is exposed, endangering the integrity of the component. The model predicting the propagation of the flaws requires its size (given by Length × Depth) as an input parameter, therefore the joint modelling of the distribution of these two quantities is crucial. Since higher values of the size of the flaws are more penalizing for the occurrence of a crack, we prefer a model which is not only adequate for the dataset, but assigns relatively great probability to higher values of these dimensions to obtain a conservative estimation of the failure probability of the component.

EDF possesses a database of joint measurements of these quantities which contains $n = 198$ measurements obtained by supervised experimentations along with 341 observations registered during regular inspections of the components in operation. We will only consider the database coming from the experimentations as these can be considered statistically perfect, whereas the inspection data is subject to measurement uncertainty and detection threshold.

Both sets of data suggest that the dimensions verify the ordering constraint, since for every pair of dimensions we have that the length of the flaw is greater than the depth. The currently applied modelling schemes does not take into consideration this aspect of the dataset. Figure 1 presents the experimentation dataset after applying a strictly monotone transformation on both dimensions to obtain values on $[0, 1]$.

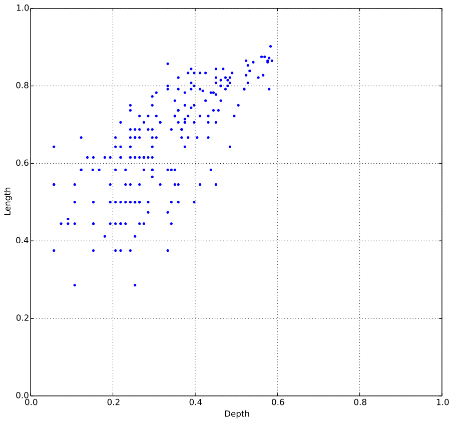


Figure 1: Scatter-plot of the transformed data set

5.2 *Available modelling schemes*

We will compare our approach to the currently approved modelling scheme as well as a method proposed by the former conference paper (Remy et al. 2012). In what follows, we note by L the random variable of the length of the flaw and by D the random variable of the depth.

5.2.1 *Reference model*

The first method used in current statistical studies of this problem at EDF consists of modelling the joint distribution of the pair (D, R), where $R = D/L$ is the random variable of the ratio between the two dimensions. The model takes the assumption that D and R are independent, and propose the following distributions for these variables (we omit the parameters of the distributions for confidentiality reasons):

- F_D: Weibull with two parameters,
- F_R: Log-normal.

5.2.2 *Parametric model*

(Remy et al. 2012) propose a parametric copula-based approach for modelling of the dependence structure between D and L. Copula theory allows to separate the modelling of the marginals and the dependence structure. The joint distribution function $F_{(D,L)}$ of the pair (D, L) can be expressed by Sklar's Theorem as:

$$F_{(D,L)}(d,l) = C_{(D,L)}(F_D(d), F_L(l)),$$

where F_D,F_L are the marginal distribution functions of D and L, and $C_{(D,L)}$ is the connecting copula containing all information on the dependence. We refer to (Nelsen 2006) for an overview of copula theory. In this setting, both dimensions D and L are modelled by Weibull distributions with two parameters. For the connecting copula $C_{(D,L)}$ the Authors consider multiple parametric families such as Gaussian, Frank or Gumbel copulas. They estimate, based on the dataset, the parameters in each family by various

Table 1: Estimated parameters for $m = 1, 2, 3, 4$.

m	$\hat{\theta}_{1,k}$	$\hat{\theta}_{2,k}$
1	$\hat{\theta}_{1,1} = -0.000307772$	$\hat{\theta}_{2,1} = -0.0277476$
2	$\hat{\theta}_{1,1} = -0.523519$	$\hat{\theta}_{2,1} = 0.295835$
	$\hat{\theta}_{1,2} = -1.06206$	$\hat{\theta}_{2,2} = -0.814702$
3	$\hat{\theta}_{1,1} = -0.545568$	$\hat{\theta}_{2,1} = -0.0445993$
	$\hat{\theta}_{1,2} = -1.10107$	$\hat{\theta}_{2,2} = -0.603401$
	$\hat{\theta}_{1,3} = -0.00991902$	$\hat{\theta}_{2,3} = -0.310838$
4	$\hat{\theta}_{1,1} = -1.82941$	$\hat{\theta}_{2,1} = 0.759716$
	$\hat{\theta}_{1,2} = -2.73921$	$\hat{\theta}_{2,2} = -2.43278$
	$\hat{\theta}_{1,3} = -1.1029$	$\hat{\theta}_{2,3} = 0.626079$
	$\hat{\theta}_{1,4} = -0.631885$	$\hat{\theta}_{2,4} = -1.03101$

methods, and compare the resulting joint distributions in order to determine the most relevant model. In conclusion, the Gumbel copula proved to give the most satisfactory results according to the graphical criterion of the Kendall plots and the Cramér-von-Mises goodness-of-fit test .

5.3 *Estimation of the nonparametric model*

For the nonparametric model, we have first transformed the dataset by using the monotone transformation T given by, for $x \in \mathbb{R}^+$:

$$T(x) = \frac{cx}{cx+1},$$

with c a constant. This is necessary since the estimation procedure requires a sample distributed on $\triangle$. The impact of the choice of the transformation function T as well as the constant c on the estimation quality has not been addressed in this paper. We choose an equal number of parameters $m = m_1 = m_2$ for both dimensions. We estimate the parameters $\theta = (\theta_{i,k}; 1 \leq i \leq 2, 1 \leq k \leq m)$ by maximizing the function G given by:

$$G(\theta) = \sum_{i=1}^{2}\sum_{k=1}^{m} \theta_{i,k}\hat{\mu}_{i,k} - \psi(\theta)$$

with $\hat{\mu}_{1,k} = (1/n)\sum_{j=1}^{n} \varphi_{1,k}(D^j)$, $\hat{\mu}_{2,k} = (1/n)\sum_{j=1}^{n} \varphi_{2,k}(L^j)$, and $\psi(\theta)$ is given by (2). This is equivalent to solving equation (3). We estimate our model for increasing values of m, using the result of the previous estimation with fewer parameters as described in (Wu 2003). We use the TNC algorithm of the OpenTURNS library for Python to numerically maximize G. The estimated parameters for $m = 1, 2, 3, 4$ can be found in Table 1.

5.4 *Comparison of the competing models*

5.4.1 *Fitting to the empirical data*

Here we compare the three different approaches in terms of goodness-of-fit to the underlying dataset and the resulting failure probability. For the reference and parametric model, we utilize the parameters obtained in the previous studies. For the nonparametric model, we take $m_1 = m_2 = 4$. In Figure 2, the densities obtained from each model can be seen along with the dataset. One can observe that the support of the nonparametric model is indeed the half plane S, whereas the other two models allow the variables to take values such that $L < D$. In Table 2 we calculated the log-likelihood of each model along with the BIC value. According to these values, the parametric model seems the most adapted for the sample followed by the reference model and the nonparametric model. The results suggest that distribution of the sample may not belong to the family of maximum entropy distributions of order statistics, and there may exist a hidden constraint that needs to be taken into consideration.

Table 2: Log-likelihood and BIC of the competing models with empirical data.

Model	Copula	Log-likelihood	BIC
Reference	-	-957.399	1930.663
Parametric	Gumbel	-927.196	1880.833
Nonparametric	MaxEntropy	-998.516	2039.338

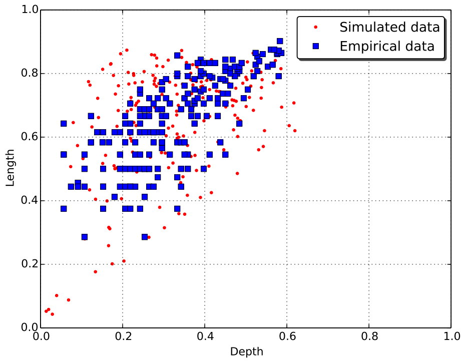
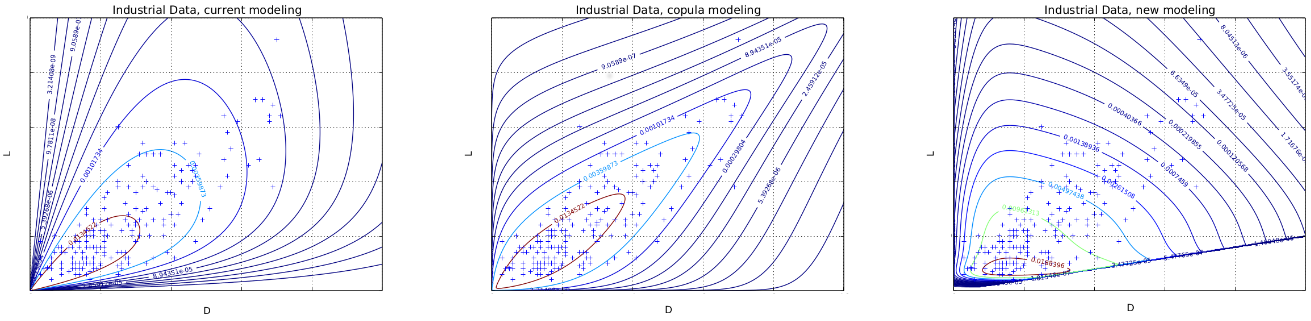


Figure 3: Scatter-plot of the two transformed data sets.

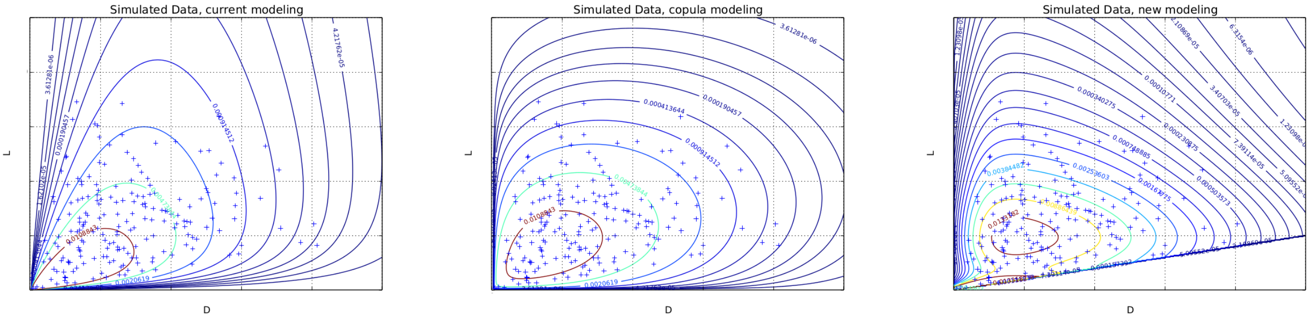
5.4.2 *Fitting to simulated data*

In order to show that effectiveness of the nonparametric model when the underlying distribution belongs to the family of maximum entropy distributions of order statistics, we simulate a dataset with 198 entries from the maximum entropy distribution with the same Weibull marginals which were used to construct the parametric model in 5.2.2. Figure 3 shows the difference between the two sets of data. We re-estimated all the parameters of the three competing models, and Table 3 shows the log-likelihood and BIC values for each model. For the parametric model, we made estimations using the Frank, Gumbel and Normal (Gaussian) family of copulas. The results confirm that if the underlying distribution belongs to the family of maximum entropy distributions of order statistics, then the nonparametric model outperforms the reference and parametric models.



(a) Reference model (b) Parametric model (Gumbel) (c) Nonparametric model

Figure 2: Isodensities for the competing models with empirical data.



(a) Reference model (b) Parametric model (Normal) (c) Nonparametric model

Figure 4: Isodensities for the competing models with simulated data.

Table 3: Log-likelihood and BIC of the competing models with simulated data.

Model	Copula	Log-likelihood	BIC
Reference	-	-1050.075	2116.016
Parametric	Frank	-1031.315	2089.072
Parametric	Gumbel	-1030.492	2087.425
Parametric	Normal	-1021.243	2068.928
Nonparametric	MaxEntropy	-995.058	2032.423

Table 4: Failure probabilities calculated with the competing models using an importance sampling method with 10^{-4} simulations.

Model	$\hat{R}^f_{model}$	c_{model}
Reference	1	2.09%
Parametric	1.022	1.99%
Nonparametric	0.148	4.14%

5.4.3 *Failure probability*

We use the joint distribution of the pair (D, L) estimated in three different ways in 5.4.1 to carry out a Monte Carlo study to determine the impact of the modelling on the component failure probability. The failure probability P^f is the probability that one of the output factors of the fracture mechanics model stays below a certain threshold. To estimate this probability, we couple the fracture mechanics model with the OpenTURNS platform. The fracture mechanics model takes 15 input variables, we assume that the pair (L, D) is independent of the rest of the variables whose values are fixed at an average level for this study. We evaluate the failure probabilities by Monte-Carlo simulations with importance sampling using $N = 10^4$ simulations. The simulations provide the estimators $\hat{P}^f_{model}$ for the three models. The results are summarized in Table 4, where we give the estimated failure probabilities relative to the failure probability of the reference model, that is the ratio:

$$\hat{R}^f_{model} = \frac{\hat{P}^f_{model}}{\hat{P}^f_{ref.model}},$$

and the coefficient of variation c_{model} given by:

$$c_{model} = \sqrt{\frac{1 - \hat{P}^f_{model}}{\hat{P}^f_{model} N}}.$$

We observe that the nonparametric model estimates the failure probability to be much lower than the other two models. This is due to the fact that a failure usually occurs when both D and L assume high values. The Gumbel copula ensures a high positive tail dependence, leading to more frequent common high values, whereas the nonparametric model, as Figure 2 suggests, gives more probabilistic mass to the upper-left zones with greater L values but smaller D values.

6 CONCLUSIONS

In this paper we draw attention to the importance of modelling the dependence structure of random variables appearing in uncertainty quantification studies. The modelling should take into consideration all the available statistical data, but ensure a maximum of freedom besides this knowledge. We presented the family of maximum entropy distribution of ordered random variables as well as a nonparametric estimation procedure to efficiently estimate such distributions. We examined its statistical performance in an uncertainty quantification study compared to some other approaches. We have seen that when the underlying data set comes from a distribution which belongs to the family of maximum entropy distributions of order statistics, the nonparametric density estimation approach proposed in Section 3 performs well. When applied to the industrial case study, we observe a decline in the performance of the nonparametric estimator, suggesting that there are some hidden constraints in addition to the ordering which was not taken into consideration by this approach (for example the high upper tail dependence). The failure probability calculations shows that the dependence modelling have a significant impact on the estimation of failure risks.

In following studies we would like to determine, via extensive simulation studies, the cases where such distributions may give more favourable results compared to other approaches. We would like to give a testing procedure to determine whether the underlying data set comes from a maximum entropy distribution or there are other hidden constraints which needs to be taken into consideration. An aggregation method is also under development to give an adaptive nonparametric estimator of the maximum entropy distribution which performs as well as the nonparametric model with an optimal number of parameters chosen based on the density's unknown regularity.

7 ACKNOWLEDGEMENTS

The authors of this paper would like to thank Emilie Dautrème and Roman Sueur for providing the code for the fracture mechanics model, and Régis Lebrun for his technical aid with OpenTURNS. We would also like to thank the anonymous Reviewers for their useful comments to improve the paper.

REFERENCES

Abramowitz, M. & I. A. Stegun (1970). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Dover Publications.

Avérous, J., C. Genest, & S. C Kochar (2005). On the dependence structure of order statistics. *Journal of multivariate analysis 94*(1), 159–171.

Barron, A. R., L. Gyorfi, & E. C. van der Meulen (1992). Distribution estimation consistent in total variation and in two types of information divergence. *Information Theory, IEEE Transactions on 38*(5), 1437–1454.

Barron, A. R. & C.-H. Sheu (1991). Approximation of density functions by sequences of exponential families. *The Annals of Statistics 19*(3), 1347–1369.

Borwein, J., A. Lewis, & R. Nussbaum (1994). Entropy minimization, DAD problems, and doubly stochastic kernels. *Journal of Functional Analysis 123*(2), 264 – 307.

Butucea, C., J.-F. Delmas, A. Dutfoy, & R. Fischer (2015a). Maximum entropy copula with given diagonal section. *Journal of Multivariate Analysis 137*(0), 61 – 81.

Butucea, C., J.-F. Delmas, A. Dutfoy, & R. Fischer (2015b). Maximum entropy distribution of order statistics with given marginals. *Working paper*.

Crain, B. R. (1977). An information theoretic approach to approximating a probability distribution. *SIAM Journal on Applied Mathematics 32*(2), 339–346.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 911–934.

Hall, P. (1987). On kullback-leibler loss and density estimation. *The Annals of Statistics*, 1491–1519.

Koo, J.-Y. & W.-C. Kim (1996). Wavelet density estimation by approximation of log-densities. *Statistics & probability letters 26*(3), 271–278.

Lebrun, R. & A. Dutfoy (2014). Copulas for order statistics with prescribed margins. *Journal of Multivariate Analysis 128*, 120–133.

Navarro, J. & F. Spizzichino (2010). On the relationships between copulas of order statistics and marginal distributions. *Statist. Probab. Lett. 80*(5-6), 473–479.

Nelsen, R. B. (2006). *An introduction to copulas* (Second ed.). Springer Series in Statistics. New York: Springer.

Remy, E., A.-L. Popelin, & A. Feng (2012). Modelling dependence using copulas - an implementation in the field of structural reliability. Paper presented at PSAM 11 and ESREL 2012 Conference on Probabilistic Safety Assessment.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics 115*(2), 347–354.

Wu, X. (2010). Exponential series estimator of multivariate densities. *Journal of Econometrics 156*(2), 354–366.

Yang, Y. & A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*(5), 1564–1599.

Zhao, N. & W. T. Lin (2011). A copula entropy approach to correlation measurement at the country level. *Applied Mathematics and Computation 218*(2), 628 – 642.