



HAL
open science

Bernstein adaptive nonparametric conditional sampling: a new method for rare event probability estimation

Elias Fekhari, Vincent Chabridon, Bertrand Iooss, Joseph Muré

► To cite this version:

Elias Fekhari, Vincent Chabridon, Bertrand Iooss, Joseph Muré. Bernstein adaptive nonparametric conditional sampling: a new method for rare event probability estimation. International Conference on Application of Statistics and Probability in Civil Engineering, Jul 2023, Dublin, Ireland. hal-04052861v2

HAL Id: hal-04052861

<https://hal.science/hal-04052861v2>

Submitted on 20 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bernstein adaptive nonparametric conditional sampling: a new method for rare event probability estimation

Elias Fekhari^{a,b}, Vincent Chabridon^a, Bertrand Iooss^{a,b}, Joseph Muré^a

^a*EDF R&D, 6 Quai Watier, Chatou, 78401, France*

^b*Université Côte d'Azur, 28 Avenue de Valrose, Nice, 06103, France*

Abstract

In the context of reliability assessment, estimating a failure probability associated to a rare event is a common task. To do so, various techniques have been proposed to overcome traditional crude Monte Carlo which becomes intractable in such a context. Among others, Subset Simulation is a widely used technique which relies on “splitting” the rare event probability into a sequence (i.e., a product) of less rare conditional probabilities associated to nested failure events, easier to estimate. However, this technique relies on simulating samples conditionally to the failure event by means of Monte Carlo Markov chain algorithms. These algorithms enable, at convergence, to simulate according to the target density. However, in practice, it often produces non-independent and identically distributed (i.i.d.) samples due to the correlation between Markov chains. In the present work, we propose another way to sample conditionally to the nested failure events in order to get i.i.d. samples which can be required (e.g., to perform dedicated sensitivity analysis). The proposed algorithm relies on a nonparametric fit of the conditional joint distribution using a combined kernel density estimation for marginals fitting and the Empirical Bernstein Copula (EBC). Thus, this new method presents some similarities with “Nonparametric Adaptive Importance Sampling” but addresses the problem of copula fitting by means of EBC. The proposed algorithm is tested on three toy-cases and its performances are compared with those obtained from Subset Sampling.

Keywords: Reliability analysis, Uncertainty propagation, Copulas, Subset Sampling

Email address: elias.fekhari@edf.fr (Elias Fekhari)

6 **1. INTRODUCTION**

7 Reliability analysis of a system is often associated with rare event probability estimation.
 8 Considering that the system’s performance is modeled by a deterministic scalar function g :
 9 $\mathcal{D}_{\mathbf{x}} \subseteq \mathbb{R}^d \rightarrow \mathbb{R}$, called *limit-state function* and a critical threshold on the system’s output
 10 $y_{\text{th}} \in \mathbb{R}$, one can define the *failure domain* as $\mathcal{F}_{\mathbf{x}} := \{\mathbf{x} \in \mathcal{D}_{\mathbf{x}} | g(\mathbf{x}) \leq y_{\text{th}}\}$. Uncertain inputs
 11 are represented by a continuous random vector $\mathbf{X} \in \mathcal{D}_{\mathbf{x}}$ assumed to be distributed according
 12 to its joint probability density function (PDF) $f_{\mathbf{X}}$. In this context, uncertainty propagation
 13 consists in composing the random vector \mathbf{X} by the function g to get an output variable of
 14 interest $Y = g(\mathbf{X}) \in \mathbb{R}$. A usual risk measure in reliability analysis is the *failure probability*,
 15 denoted by p_f , and defined as the probability that the system exceeds the threshold y_{th} :

$$p_f := \mathbb{P}(g(\mathbf{X}) \leq y_{\text{th}}) = \int_{\mathcal{D}_{\mathbf{x}}} \mathbb{1}_{\mathcal{F}_{\mathbf{x}}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} \quad (1)$$

16 where $\mathbb{1}_{\mathcal{F}_{\mathbf{x}}}(\cdot)$ is the indicator function of the failure domain such that $\mathbb{1}_{\mathcal{F}_{\mathbf{x}}}(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathcal{F}_{\mathbf{x}}$ and
 17 $\mathbb{1}_{\mathcal{F}_{\mathbf{x}}}(\mathbf{x}) = 0$ otherwise. Rare event problems are usually solved in the so-called *standard normal*
 18 *space* after applying an “iso-probabilistic transformation” which can be either the Rosenblatt
 19 or the generalized Nataf one (Lebrun, 2013). Additionally, the limit-state function g can be
 20 viewed as an input-output “black-box” model which can be costly to evaluate (e.g., a complex
 21 numerical model), making the failure probability estimation nontrivial. When the limit-state
 22 function is a costly computer model, one can build a surrogate model and use specific active
 23 learning methods (see, e.g., Moustapha et al. (2022)). However, using surrogate models is not
 24 always possible for practical engineering applications as they might introduce another level
 25 of approximation, which can be prohibitive from safety auditing. Moreover, their validation
 26 as well as their behavior with respect to large input dimension case make also their use quite
 27 complex (see, e.g., (Marrel et al., 2022)).

28 Going back to the rare event estimation literature, one can consider two major types
 29 of techniques for failure probability calculation (Morio and Balesdent, 2015): (i) Geomet-
 30 ric approaches, such as the *first-/second-order reliability method* (FORM/SORM) whose aim
 31 is to approximate the limit-state function by a first-/second-order Taylor expansion at the
 32 most probable failure point; (ii) Simulation-based techniques such as the *crude Monte Carlo*
 33 method. Unfortunately, FORM/SORM methods do not provide a lot of statistical information

34 as they are purely geometric approaches. Meanwhile, estimating a rare event probability by
 35 crude Monte Carlo becomes rapidly intractable. To overcome this limit, advanced simulation
 36 techniques have been developed: among others, one can mention several “variance reduction
 37 methods” such as the non-adaptive and adaptive versions of the *Importance Sampling* (Ru-
 38 binstein and Kroese, 2008) (either parametric, using the Cross-Entropy method Kurtz and
 39 Song (2013), or nonparametric Morio (2011)) and splitting techniques (C erou et al., 2012)
 40 such as the *Subset Simulation* (SS) Au and Beck (2001). In these techniques, the idea is to
 41 write the rare event p_f as a product of larger conditional probabilities, each one of them being
 42 easier to estimate. To generate intermediary conditional samples, this method uses Markov
 43 chain Monte Carlo (MCMC) sampling, which presents numerous versions (Papaioannou et al.,
 44 2015). However, MCMC algorithms are known to be highly tunable algorithms which produce
 45 non-i.i.d. samples, which consequently, cannot be used for direct statistical estimation (e.g.,
 46 failure probability or sensitivity indices (Da Veiga et al., 2021)).

47 The present work proposes a new rare event estimation method, adopting the same sequen-
 48 tial structure as SS while using a strictly different sampling mechanism to generate conditional
 49 samples. This method intends to fit the intermediary conditional distributions with a nonpara-
 50 metric tool called the *Empirical Bernstein Copula*. Contrarily to SS, the proposed method
 51 named “Bernstein adaptive nonparametric conditional sampling” (BANCS), generates i.i.d.
 52 samples of the intermediary conditional distributions. For instance, a practical use of such
 53 i.i.d. samples can be to estimate dedicated reliability-oriented sensitivity indices (see, e.g.,
 54 Chabridon et al. (2021); Marrel and Chabridon (2021)).

55 In this paper, Section 2 will recall the methodology of subset sampling and probabilistic
 56 modeling. Then, Section 3 will introduce the BANCS method for rare event estimation.
 57 Section 4 will apply this method to three toy-cases and analyze the results with respect to SS
 58 performances. Then, the last section present some conclusions and research perspectives.

59 2. BACKGROUND

60 2.1. Subset sampling

61 Subset sampling splits the failure event \mathcal{F}_x into an intersection of $k_{\#}$ intermediary events
 62 $\mathcal{F}_x = \cap_{k=1}^{k_{\#}} \mathcal{F}_{[k]}$. Each are nested such that $\mathcal{F}_{[1]} \supset \dots \supset \mathcal{F}_{[k_{\#}]} = \mathcal{F}_x$. The failure probability is

63 then expressed as a product of conditional probabilities:

$$p_f = \mathbb{P}(\mathcal{F}_{\mathbf{x}}) = \mathbb{P}(\cap_{k=1}^{k_{\#}} \mathcal{F}_{[k]}) = \prod_{k=1}^{k_{\#}} \mathbb{P}(\mathcal{F}_{[k]} | \mathcal{F}_{[k-1]}). \quad (2)$$

64 From a practical point of view, the analyst tunes the algorithm by setting the intermediary
65 probabilities $\mathbb{P}(\mathcal{F}_{[k]} | \mathcal{F}_{[k-1]}) = p_0, \forall k \in \{1, \dots, k_{\#}\}$. Then, the corresponding quantiles $q_{[1]}^{p_0} >$
66 $\dots > q_{[k_{\#}]}^{p_0}$ are estimated for each conditional subset samples $\mathbf{X}_{[k],N}$ of size N . Note that the
67 initial quantile is estimated by crude Monte Carlo sampling on the input PDF $f_{\mathbf{x}}$. Following
68 conditional subset samples are generated by MCMC sampling of $f_{\mathbf{x}}(\mathbf{x} | \mathcal{F}_{[k-1]})$, using as seeds
69 initialisation points the $n = Np_0$ samples given by $\mathbf{A}_{[k],n} = \{\mathbf{X}_{[k-1]}^{(j)} \subset \mathbf{X}_{[k-1],N} | g(\mathbf{X}_{[k-1]}^{(j)}) >$
70 $\hat{q}_{[k-1]}^{\alpha}\}_{j=1}^n$. This process is repeated until an intermediary quantile exceeds the threshold:
71 $\hat{q}_{[k_{\#}]}^{p_0} < y_{\text{th}}$. Finally, the failure probability is estimated by:

$$p_f \approx \hat{p}_f^{\text{SS}} = p_0^{k_{\#}-1} \frac{1}{N} \sum_{j=1}^N \mathbb{1}_{\{g(\mathbf{x}) \leq y_{\text{th}}\}}(\mathbf{X}_{[k_{\#}],N}^{(j)}). \quad (3)$$

72 In practice, the subset sample size should be large enough to properly estimate intermediary
73 quantiles, which leads Au and Beck (2001) to recommend setting $p_0 = 0.1$. SS efficiency
74 depends on the proper choice and tuning of the MCMC algorithm (Papaioannou et al., 2015).
75 Our work uses the SS implementation from `OpenTURNS`¹ (Baudin et al., 2017) which integrates
76 a component-wise Metropolis-Hastings algorithm. As an alternative to generating samples on
77 a conditional distribution by MCMC, one could try to fit this conditional distribution.

78 2.2. Multivariate modeling using copulas

79 The Sklar theorem (Joe, 1997) affirms that the multivariate distribution of any random
80 vector $\mathbf{X} \in \mathbb{R}^d$ can be broken down into two objects:

- 81 1. A set of univariate marginal distributions to describe the behavior of the individual
82 variables;
- 83 2. A function describing the dependence structure between all variables, called a copula.

84 This theorem states that considering a random vector $\mathbf{X} \in \mathbb{R}^d$, with its distribution F and its
85 marginals $\{F_i\}_{i=1}^d$, there exists a copula $C : [0, 1]^d \rightarrow [0, 1]$, such that:

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \quad (4)$$

¹<https://openturns.github.io/www/index.html>

86 It allows us to divide the problem of fitting a joint distribution into two independent
87 problems: fitting the marginals and fitting the copula. Note that when the joint distribution
88 is continuous, this copula is unique. Provided a dataset, this framework allows to combine a
89 parametric (or nonparametric) fit of marginals with a parametric (or nonparametric) fit of the
90 copula. When the distribution’s dimension is higher than two, one can perform a parametric
91 fit using vine copulas (Joe and Kurowicka, 2011), implying the choice of multiple types of
92 parametric copulas. Otherwise, nonparametric fit by multivariate kernel density estimation
93 (KDE) presents a computational burden as soon as the dimension increases (Chabridon et al.,
94 2021). Since univariate marginals are usually well-fitted with nonparametric tools (e.g., KDE),
95 let us introduce an effective nonparametric method for copula fitting.

96 3. A NEW COPULA-BASED CONDITIONAL SAMPLING METHOD

97 3.1. Empirical Bernstein copula

98 Copulas are continuous and bounded functions defined on a compact set (the unit hyper-
99 cube). Bernstein polynomials allow to uniformly approximate as closely as desired any contin-
100 uous and real-valued function defined on a compact set (Weierstrass approximation theorem).
101 Therefore, they are good candidates to approximate unknown copulas. This concept was intro-
102 duced as *empirical Bernstein copula* (EBC) by Sancetta and Satchell (2004) for applications in
103 economics and risk management. Later on, Segers et al. (2017) offered further asymptotic stud-
104 ies. Formally, the multivariate Bernstein polynomial for a function $C : [0, 1]^d \rightarrow \mathbb{R}$ on a grid
105 over the unit hypercube $G := \left\{ \frac{0}{m_1}, \dots, \frac{m_1}{m_1} \right\} \times \dots \times \left\{ \frac{0}{m_d}, \dots, \frac{m_d}{m_d} \right\}$, $\mathbf{m} = (m_1, \dots, m_d) \in \mathbb{N}^d$,
106 writes:

$$B_{\mathbf{m}}(C)(\mathbf{u}) := \sum_{t_1=0}^{m_1} \dots \sum_{t_d=0}^{m_d} C \left(\frac{t_1}{m_1}, \dots, \frac{t_d}{m_d} \right) \prod_{j=1}^d P_{m_j, t_j}(u_j), \quad (5)$$

107 with $\mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d$, and the Bernstein polynomial $P_{m,t}(u) := \frac{t!}{m!(t-m)!} u^m (1-u)^{t-m}$.
108 Notice how the grid definition implies the polynomial’s order. When C is a copula, then $B_{\mathbf{m}}(C)$
109 is called “Bernstein copula”. Therefore, the empirical Bernstein copula is an application of the
110 Bernstein polynomial in Eq. (5) to the so-called “empirical copula”.

111 In practice, considering a sample $\mathbf{X}_n = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \in \mathbb{R}^{np}$ and the associated ranked

112 sample $\mathbf{R}_n = \{\mathbf{r}^{(1)}, \dots, \mathbf{r}^{(n)}\}$, the corresponding empirical copula writes:

$$C_n(\mathbf{u}) := \frac{1}{n} \sum_{i=0}^n \prod_{j=1}^p \mathbb{1} \left\{ \frac{r_j^{(i)}}{n} \leq u_j \right\}, \quad (6)$$

113 with $\mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d$. In the following, the polynomial order is set as equal in each
 114 dimension: $\{m_i = m\}_{j=1}^d$. Theoretically, the tuning parameter can be optimized to minimize
 115 an ‘‘Mean Integrated Squared Error’’ (MISE), leading to a bias-variance tradeoff. Formally,
 116 the MISE of the empirical Bernstein copula $B_{\mathbf{m}}(C_n)$ is defined as follows:

$$\mathbb{E}[\|B_{\mathbf{m}}(C_n) - C\|_2^2] = \mathbb{E} \left[\int_{\mathbb{R}^d} (B_{\mathbf{m}}(C_n)(\mathbf{u}) - C(\mathbf{u}))^2 d\mathbf{u} \right]. \quad (7)$$

117 Then, Sancetta and Satchell (2004) prove in their Theorem 3 that:

- 118 • $B_{\mathbf{m}}(C_n)(\mathbf{u}) \rightarrow C(\mathbf{u})$ for any $u_j \in]0, 1[$ if $\frac{m^{d/2}}{n} \rightarrow 0$, when $m, n \rightarrow \infty$.
- 119 • The optimal order of the polynomial in terms of MISE is: $m \lesssim m_{\text{IMSE}} = n^{2/(d+4)}$, $\forall u_j \in$
 120 $]0, 1[$. The sign \lesssim means ‘‘less than or approximately’’.

121 Let us remark that in the special case $m = n$, also called the ‘‘Beta copula’’ in Segers
 122 et al. (2017), the bias is very small while the variance gets large. To illustrate the previous
 123 theorem, Lasserre (2022) represents the evolution of the m_{IMSE} for different dimensions and
 124 sample sizes (see Fig. 1). In high dimension, the values of m_{IMSE} tend towards one, which is
 125 equivalent to the independent copula. Therefore, high-dimensional problems should be divided
 126 into a product of smaller problems on which the EBC is tractable. Provided a large enough
 127 learning set \mathbf{X}_n , KDE fitting of marginals combined with EBC fitting of the copula delivers
 128 good results even on complex dependence structures. Moreover, EBC provides an explicit
 129 expression, making a Monte Carlo generation of i.i.d. samples simple. In the following, this
 130 nonparametric tool is used to fit the intermediary conditional distributions present in subset
 131 sampling.

132 3.2. Bernstein adaptive nonparametric conditional sampling (BANCS) method

133 This new method reuses the main idea from SS while employing a different approach to
 134 generate conditional samples. Instead of using MCMC sampling, the conditional distribution is
 135 firstly fitted by a nonparametric procedure, before sampling on this nonparametric model. As

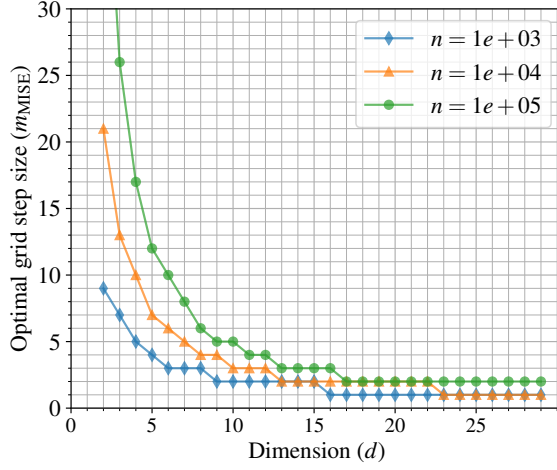


Figure 1: Evolution of m_{MISE} for different dimensions and sample sizes.

136 described in Algorithm 1, conditional sampling is done on a distribution composed by merging
 137 marginals $\{\widehat{F}_i\}_{i=1}^d$ fitted by KDE, with a copula $B_{\mathbf{m}}(C_n)$ fitted by EBC. Fig. 2 illustrate the
 138 nonparametric fit and conditional sampling in BANCS method on a two-dimensional reliability
 139 problem (later introduced as “toy-case #1”). At iteration k , after estimating the intermediary
 140 quantile $\widehat{q}_{[k]}^{p_0}$, a nonparametric model is fitted on $\mathbf{A}_{[k+1],n}$ and used to generate the next N -
 141 sized subset sample $\mathbf{X}_{[k+1],N}$. Note that the BANCS method does not require iso-probabilistic
 142 transform.

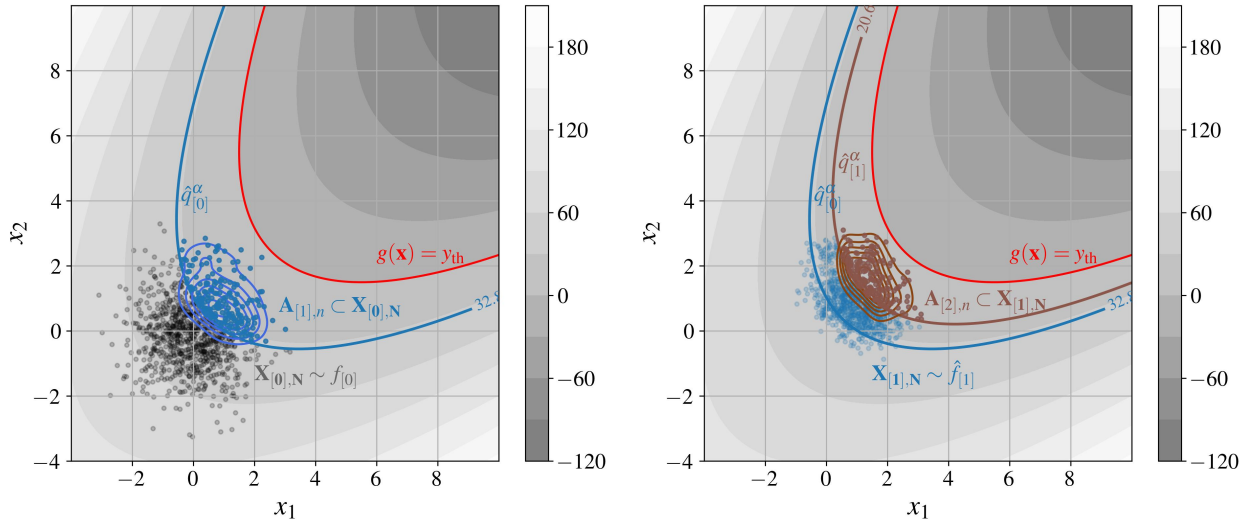


Figure 2: BANCS on toy-case #1: illustration of nonparametric fit at the first iteration.

Algorithm 1 Bernstein adaptive nonparametric conditional sampling (BANCS).

▷ *Inputs:* ◁

$f_{\mathbf{X}}$, joint PDF of the inputs

$g(\cdot)$, limit-state function

$y_{\text{th}} \in \mathbb{R}$, threshold defining the failure event

N , number of samples per iteration

$m \in \mathbb{N}$, parameter of the EBC fitting

$p_0 \in]0, 1[$, empirical quantile order (rarity parameter)

▷ *Algorithm:* ◁

Set $k = 0$ and $f_{[0]} = f_{\mathbf{X}}$

Sample $\mathbf{X}_{[0],N} = \{\mathbf{X}_{[0]}^{(j)}\}_{j=1}^N \stackrel{\text{i.i.d.}}{\sim} f_{[0]}$

Evaluate $G_{[0],N} = \{g(\mathbf{X}_{[0]}^{(j)})\}_{j=1}^N$

Estimate the empirical p_0 -quantile $\hat{q}_{[0]}^{p_0}$ of the set $G_{[0],N}$

while $\hat{q}_{[k]}^{p_0} > y_{\text{th}}$ **do**

 Subsample $\mathbf{A}_{[k+1],n} = \{\mathbf{X}_{[k]}^{(j)} \subset \mathbf{X}_{[k],N} | g(\mathbf{X}_{[k]}^{(j)}) > \hat{q}_{[k]}^{p_0}\}_{j=1}^n$

 Fit marginals of the subset $\mathbf{A}_{[k+1],n}$ by KDE $\{\hat{F}_i\}_{i=1}^d$

 Fit the copula of the subset $\mathbf{A}_{[k+1],n}$ by EBC $B_{\mathbf{m}}(C_n)$

 Build a CDF $\hat{F}_{[k+1]}(\mathbf{x}) = B_{\mathbf{m}}(C_n)(\hat{F}_1(x_1), \dots, \hat{F}_d(x_d))$

 Sample $\mathbf{X}_{[k+1],N} = \{\mathbf{X}_{[k+1]}^{(j)}\}_{j=1}^N \stackrel{\text{i.i.d.}}{\sim} \hat{f}_{[k+1]}$

 Evaluate $G_{[k+1],N} = \{g(\mathbf{X}_{[k+1]}^{(j)})\}_{j=1}^N$

 Estimate the empirical p_0 -quantile $\hat{q}_{[k+1]}^{p_0}$ of $G_{[k+1],N}$

 Set $k = k + 1$

Set total iteration number $k_{\#} = k - 1$

Estimate $\hat{p}_{\text{f}} = (1 - p_0)^{k_{\#}} \cdot \frac{1}{N} \sum_{j=1}^N \mathbb{1}_{\{g(\mathbf{X}_{[k_{\#}] }^{(j)}) \geq y_{\text{th}}\}}(\mathbf{X}_{[k_{\#}] }^{(j)})$

▷ *Outputs:* ◁

\hat{p}_{f} , estimate of p_{f}

143 As discussed in the previous section, EBC fitting is tuned by the Bernstein polynomial
144 of order m , implying a bias-variance trade off. In Fig. 2, conditional distributions fitted by
145 EBC (blue and brown isolines) seem to present a slight bias since they overlay the quantiles.
146 However, reducing this bias implies decreasing the tuning parameter m , until $m = 1$, which
147 is equivalent to an independent copula. Tools to control the goodness of fit of nonparametric
148 conditional distributions are also available. As an example, let us consider the fitted condi-
149 tional distribution at the first iteration (visible in Fig. 2). Its quantile-quantile plot in Fig. 3
150 (left) shows a good fit of the two marginals by KDE. Then, the goodness of fit of copulas can

151 be evaluated by Kendall’s plot, represented in Fig. 3 (right). This fit is also good, even if a
 152 slight bias is again visible.

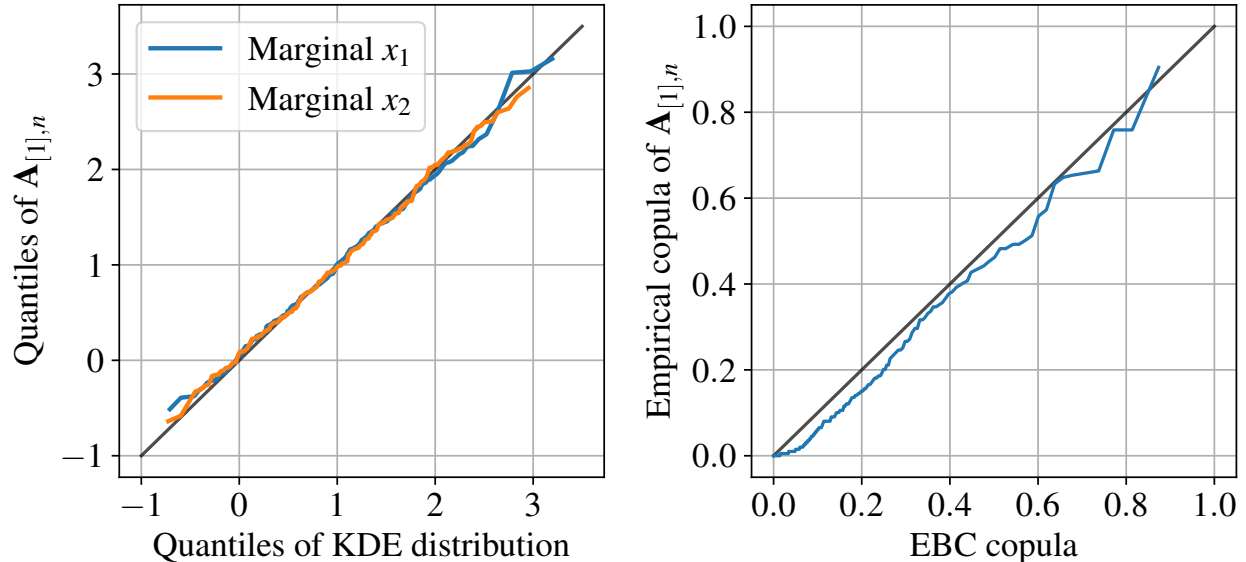


Figure 3: QQ-plot (left) for KDE on marginals and Kendall plot (right) for EBC on the copula of the conditional distribution from Fig. 2.

153 4. NUMERICAL EXPERIMENTS

154 In the following analytical numerical experiments, the intermediary probabilities were set
 155 to $p_0 = 0.1$, allowing a fair comparison with subset sampling. Then, the subset sample size
 156 is set to $N = 10^4$, in order to get a reasonable sample size $n = Np_0 = 10^3$ to perform the
 157 nonparametric fitting. EBC tuning is setup to minimize the MISE in Eq. (7): $m = 1 + n^{\frac{2}{d+4}}$. In
 158 order to take into account the variability of the method’s results, each experiment is repeated
 159 100 times, allowing the computation of a coefficient of variation $\hat{\delta} = \frac{\sigma_{\hat{p}_i}}{\mu_{\hat{p}_i}}$. Note that an
 160 implementation of the BANCS method and the following numerical experiments are available
 161 in a Git repository².

162 4.1. Toy-case #1: Parabolic reliability problem

163 Let us define the parabolic reliability problem, considering the function $g_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$:

$$g_1(\mathbf{x}) = (x_1 - x_2)^2 - 8(x_1 + x_2 - 5), \quad (8)$$

²<https://github.com/efekhari27/icasp14>

164 with the input random vector $\mathbf{X} = (X_1, X_2)$ following a standard 2-dimensional normal dis-
 165 tribution. The reliability problem consists in evaluating: $p_{f,1} = \mathbb{P}(g_1(\mathbf{X}) \leq 0) = 1.31 \times 10^{-4}$.

166 *4.2. Toy-case #2: Four-branch reliability problem*

Let us define the four-branch reliability problem (originally proposed by Waarts (2000)), considering the following function $g_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$:

$$g_2(\mathbf{x}) = \min \begin{pmatrix} 5 + 0.1(x_1 - x_2)^2 - \frac{(x_1+x_2)}{\sqrt{2}} \\ 5 + 0.1(x_1 - x_2)^2 + \frac{(x_1+x_2)}{\sqrt{2}} \\ (x_1 - x_2) + \frac{9}{\sqrt{2}} \\ (x_2 - x_1) + \frac{9}{\sqrt{2}} \end{pmatrix}, \quad (9)$$

167 with the input random vector $\mathbf{X} = (X_1, X_2)$ following a standard 2-dimensional normal dis-
 168 tribution. The reliability problem consists in evaluating: $p_{f,2} = \mathbb{P}(g_2(\mathbf{X}) \leq 0) = 2.21 \times 10^{-4}$.

169 *4.3. Toy-case #3: higher-dimensional reliability problem*

170 Let us define the higher-dimensional reliability problem (proposed by Yun et al. (2018)),
 171 considering the following function $g_3 : \mathbb{R}^7 \rightarrow \mathbb{R}$:

$$g_3(\mathbf{x}) = 15.59 \times 10^4 - \frac{x_1 x_3^2 x_2^4 - 4x_5 x_6 x_7^2 + x_4(x_6 + 4x_5 + 2x_6 x_7)}{2x_3^2 x_4 x_5 (x_4 + x_6 + 2x_6 x_7)}, \quad (10)$$

172 with the input random vector $\mathbf{X} = (X_1, \dots, X_7)$, following a product of normal distributions
 173 defined in Yun et al. (2018). The reliability problem consists in evaluating: $p_{f,3} = \mathbb{P}(g_3(\mathbf{X}) \leq$
 174 $0) = 8.10 \times 10^{-3}$.

175 *4.4. Results analysis*

176 Results of our numerical experiments are presented graphically (for 2-dimensional prob-
 177 lems) in Figure 4, and numerically in Table 1. In the same fashion as the previous illustrations,
 178 the figures represent the intermediary quantiles $\widehat{q}_{[k]}^{p_0}$ estimated over conditional samples of size
 179 $N = 10^4$. Moreover, samples $\mathbf{A}_{[k+1],n}$ exceeding these quantiles are also represented in the
 180 same color. Notice how the last estimated quantile is set to the problem threshold $y_{\text{th}} = 0$. To
 181 capture the dispersion of BANCS estimation, 100 repetitions were realized. Let us notice that
 182 for each toy-case, BANCS well estimates the failure probabilities' orders of magnitude. Yet the
 183 numerical values in Table 1 consistently present a positive bias, leading to an overestimated

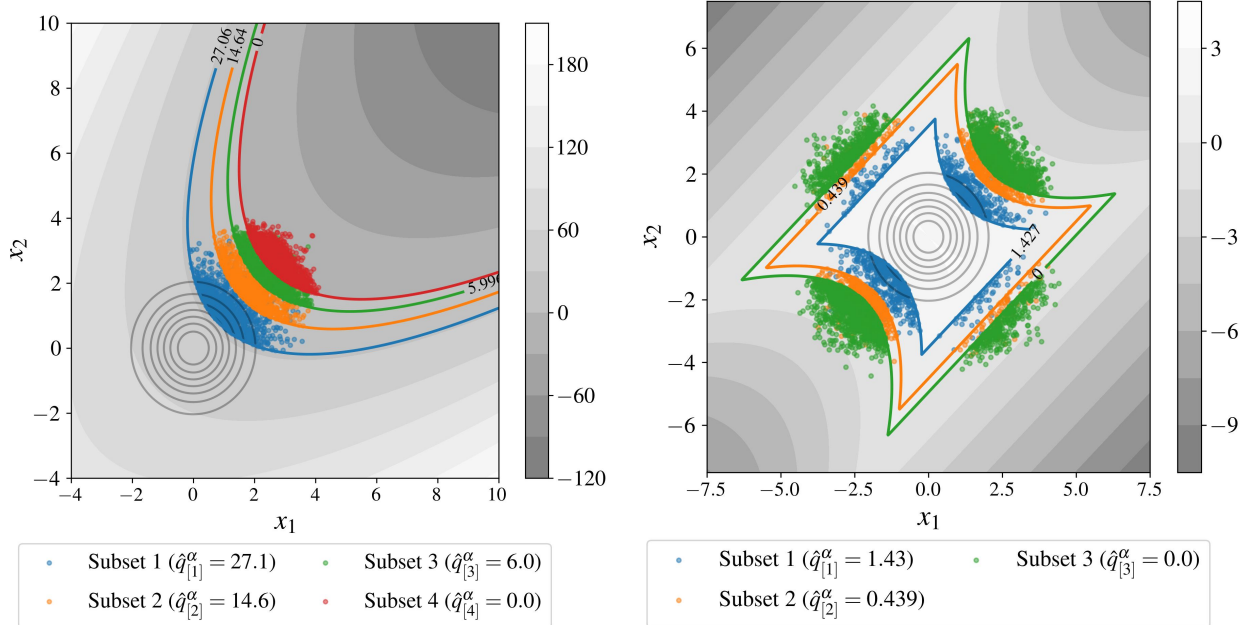


Figure 4: BANCS sampling steps on toy-case #1 (left) and #2 (right).

184 failure probability. This bias is partially explained by the EBC tuning choice and could be
 185 reduced at the expense of a slightly higher variance.

186 The variance obtained with the repetitions is quite large. Although, part of it is due to
 187 the fact that the algorithm might compute a different total number of subsets (e.g., toy-case
 188 #1 is either solved in four or five subsets). Overall, considering the EBC tuning from Eq. (7),
 189 BANCS performs worst than SS on toy-cases #1 and #2 but performs as well as SS on the
 190 toy-case #3. This might be due to the fact that toy-case #3 has a higher input dimension.
 191 However, one can note that SS coefficient of variation is computed by an approximation,
 192 tending to underestimate the true coefficient of variation (see e.g., Papaioannou et al. (2015)).

Table 1: Results of the numerical experiments (subset sample size $N = 10^4$, $p_0 = 0.1$).

	d	p_f^{ref}	\hat{p}_f^{BANCS}	$\hat{\delta}^{\text{BANCS}}$	\hat{p}_f^{SS}	$\hat{\delta}^{\text{SS}}$
Toy-case #1	2	1.31×10^{-4}	2.67×10^{-4}	24%	1.30×10^{-4}	9%
Toy-case #2	2	2.21×10^{-4}	4.23×10^{-4}	7%	2.24×10^{-4}	6%
Toy-case #3	7	8.10×10^{-3}	9.32×10^{-3}	15%	8.92×10^{-3}	6%

193 5. CONCLUSION

194 Subset Simulation uses MCMC sampling to generate its intermediary conditional samples.
195 However, MCMC algorithms tends to be complex to tune and does not generate i.i.d. condi-
196 tional samples. In this work, a new method is proposed, replacing MCMC sampling with a
197 simpler procedure. An intermediary conditional distribution is first fitted by a nonparametric
198 approach, mixing kernel density estimation for fitting the marginals and Empirical Bernstein
199 Copula (EBC) for fitting the copula. Then, the resulting allows to perform direct Monte Carlo
200 sampling. This method is named “Bernstein adaptive nonparametric conditional sampling”
201 (BANCS) and is applied to three toy-cases (two 2-dimensional and one 7-dimensional) and
202 compared with SS.

203 The method shows promising results, even though a small positive bias consistently ap-
204 pears. This issue results from EBC tuning, creating a bias-variance tradeoff in the copula
205 fit. Theoretical works offer optimal tuning, allowing us to find the optimal compromise. In
206 our numerical experiments, an empirical estimation of BANCS variance is computed over a
207 set of repetitions. BANCS estimated coefficient of variation is higher than SS approximated
208 coefficient of variation. This work can be further explored by building an approximation of
209 BANCS variance and confidence interval. One major advantage remains that the samples
210 generated at each iteration are i.i.d. leading to a possible use of these samples to perform
211 global reliability-oriented sensitivity analysis (Marrel and Chabridon, 2021) in order to detect
212 and analyze the most influential input variables leading to failure.

213 **Funding Statement.** This study is part of HIPERWIND project which has received funding
214 from the European Union’s Horizon 2020 Research and Innovation Programme under Grant
215 Agreement No. 101006689.

216 References

- 217 Au, S.-K. and Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions
218 by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277.
- 219 Baudin, M., Dutfoy, A., Iooss, B., and Popelin, A.-L. (2017). Openturns: an industrial software

220 for uncertainty quantification in simulation. In *Handbook of uncertainty quantification*, pages
221 2001–2038. Springer.

222 Cérou, F., Del Moral, P., Furon, T., and Guyader, A. (2012). Sequential monte carlo for rare
223 event estimation. *Statistics and computing*, 22:795–808.

224 Chabridon, V., Balesdent, M., Perrin, G., Morio, J., Bourinet, J.-M., and Gayton, N. (2021).
225 Global reliability-oriented sensitivity analysis under distribution parameter uncertainty. *Me-*
226 *chanical Engineering under Uncertainties: From Classical Approaches to Some Recent De-*
227 *velopments*, pages 237–277.

228 Da Veiga, S., Gamboa, F., Iooss, B., and Prieur, C. (2021). *Basics and Trends in Sensitivity*
229 *Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

230 Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and
231 Hall.

232 Joe, H. and Kurowicka, D. (2011). *Dependence modeling: vine copula handbook*. World
233 Scientific.

234 Kurtz, N. and Song, J. (2013). Cross-entropy-based adaptive importance sampling using
235 Gaussian mixture. *Structural Safety*, 42:35–44.

236 Lasserre, M. (2022). *Apprentissages dans les réseaux bayésiens à base de copules non-*
237 *paramétriques*. PhD thesis, Sorbonne Université.

238 Lebrun, R. (2013). *Contributions à la modélisation de la dépendance stochastique*. PhD thesis,
239 Université Paris-Diderot – Paris VII. (in English).

240 Marrel, A. and Chabridon, V. (2021). Statistical developments for target and conditional
241 sensitivity analysis: Application on safety studies for nuclear reactor. *Reliability Engineering*
242 *& System Safety*, 214:107711.

243 Marrel, A., Iooss, B., and Chabridon, V. (2022). The ICSCREAM methodology: Identification
244 of penalizing configurations in computer experiments using screening and metamodel –
245 Applications in thermal-hydraulics. *Nuclear Science and Engineering*, 196:301–321.

- 246 Morio, J. (2011). Non-parametric adaptive importance sampling for the probability estimation
247 of a launcher impact position. *Reliability Engineering and System Safety*, 96(1):178–183.
- 248 Morio, J. and Balesdent, M. (2015). *Estimation of Rare Event Probabilities in Complex*
249 *Aerospace and Other Systems: A Practical Approach*. Woodhead Publishing, Elsevier.
- 250 Moustapha, M., Marelli, S., and Sudret, B. (2022). Active learning for structural reliability:
251 Survey, general framework and benchmark. *Structural Safety*, 96:102174.
- 252 Papaioannou, I., Betz, W., Zwirgmaier, K., and Straub, D. (2015). MCMC algorithms for
253 Subset Simulation. *Probabilistic Engineering Mechanics*, 41:89–103.
- 254 Rubinstein, R. Y. and Kroese, D. P. (2008). *Simulation and the Monte Carlo Method*. Wiley,
255 Second ed. edition.
- 256 Sancetta, A. and Satchell, S. (2004). The Bernstein copula and its applications to modeling
257 and approximations of multivariate distributions. *Econometric Theory*, 20(3):535–562.
- 258 Segers, J., Sibuya, M., and Tsukahara, H. (2017). The empirical beta copula. *Journal of*
259 *Multivariate Analysis*, 155:35–51.
- 260 Waarts, P. (2000). *Structural reliability using finite element methods: an appraisal of direc-*
261 *tional adaptive response surface sampling (DARS)*. PhD thesis, Technical University of
262 Delft, The Netherlands.
- 263 Yun, W., Lu, Z., Zhang, Y., and Jiang, X. (2018). An efficient global reliability sensitivity
264 analysis algorithm based on classification of model output and subset simulation. *Structural*
265 *Safety*, 74:49–57.