Probability in the Engineering and Informational Sciences, **34**, 2020, 14–26.
doi:10.1017/S026996481800030X

MODELING EXPERIMENTAL DATA WITH POLYNOMIALS CHAOS

Emeline Gayrard
Laboratoire de Mathématiques Blaise Pascal (LMBP), CNRS UMR 6620
Université Clermont Auvergne, Campus Universitaire des Cézeaux, 3 place Vasarely,
TSA 60026/CS 60026, 63 178 Aubière Cedex, France
E-mail: Emeline.Gayrard@uca.fr

Cédric Chauvière
Laboratoire de Mathématiques Blaise Pascal (LMBP), CNRS UMR 6620
Université Clermont Auvergne, Campus Universitaire des Cézeaux, 3 place Vasarely,
TSA 60026/CS 60026, 63 178 Aubière Cedex, France
E-mail: Cedric.Chauviere@uca.fr

Hacène Djellout
Laboratoire de Mathématiques Blaise Pascal (LMBP), CNRS UMR 6620
Université Clermont Auvergne, Campus Universitaire des Cézeaux, 3 place Vasarely,
TSA 60026/CS 60026, 63 178 Aubière Cedex, France
E-mail: Hacene.Djellout@uca.fr

Pierre Bonnet
Institut Pascal, CNRS UMR 6602
Université Clermont Auvergne, Campus Universitaire des Cézeaux, 4 Avenue Blaise Pascal,
TSA 60026/CS 60026, 63178 Aubière Cedex,
France
E-mail: pierre.bonnet@univ-bpclermont.fr

Given a raw data sample, the purpose of this paper is to design a numerical procedure to model this sample under the form of polynomial chaos expansion. The coefficients of the polynomial are computed as the solution to a constrained optimization problem. The procedure is first validated on samples coming from a known distribution and it is then applied to raw experimental data of unknown distribution. Numerical experiments show that only five coefficients of the Chaos expansions are required to get an accurate representation of a sample.

Keywords: chaos expansions, experimental data modeling, numerical methods

1. INTRODUCTION

Since the work of Wiener in 1938 [13] and the seminal book of Ghanem and Spanos in the 90s [5], polynomial chaos (PC) have been very successful at solving a wide variety of

© Cambridge University Press 2018 0269-9648/18 $25.00

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

random problems coming from different fields: mechanics [8], electromagnetism [4],... and so on to name a few. In most, if not all problems, PC expansion (PCE) is an efficient way to construct a model response surface in order to study the propagation of uncertainty through a complex system. Typically, some uncertain input parameters are represented by a random variable (say X) and a random output quantity of interest may be written as

$$Y = \sum_{i=0}^{n} \alpha_i \Psi_i(X),$$

where $\{\Psi_i\}_{0\leq i\leq n}$ form an orthonormal polynomial basis of degree n. Then, macroscopic quantities of the output Y, such as the mean, the variance or even Sobol's indices [11] can be computed solely from the knowledge of the coefficients α_i. Although this procedure looks simple and appealing, some care must be taken. Indeed, the choice of the orthonormal basis Ψ_i is critical for an accurate representation of the random output Y (see [14]). This problem can be solved by constructing an adapted PC basis from the knowledge of a sample of the random variable Y (see [1,10]). Such a procedure usually referred to as adaptive PC (aPC), has been meeting increasing interest over the past few years since it does not require a-priori knowledge of the form of the random input X to build the basis Ψ_i. However, for some applications, such as the one presented in this paper, the knowledge of the random variable (r.v.) X is required and therefore the aPC cannot be used.

Surprisingly, the problem of modeling a sample through a PC representation has received little attention among the scientific community. However, such a problem is neither trivial nor an irrelevant matter. Indeed, when one has to model a sample data, an a-priori distribution is chosen and its parameters may be estimated using a maximum likelihood estimator or the method of moments, for example [12]. Although distributions of different shapes are plentiful, it is still possible that none of them will be able to represent faithfully the sample. In this paper, we propose an alternative procedure that is, representing a general sample of a r.v. Y through a PC expansion. For this problem, none of the standard technique available in the literature (should it be intrusive or non-intrusive) can be used to compute the coefficients α_i. The key point why these techniques, or some of its refinements described above fail, is that they all rely on a functional relationship of the form $Y = f(X)$, that is, for a given value of X, one should be able to compute the corresponding value of Y. When the only information at hand is raw data, such a relationship is not available and new schemes should be thought of. Some experiments have been carried out using a Bayesian approach with PC. However, maximizing the likelihood function is a major issue and may require several approximations [2]. In this paper, we adopt a simpler and more didactic approach that does not require prior knowledge of Bayesian statistics. We build an optimization problem based on the inverse cumulative distribution of a sample of X and Y for computing the coefficients α_i. We found that adding constraints to the problem increased the robustness of the procedure. More specifically, we impose that the empirical mean and the variance of the sample be the same as the one of its chaos representation.

The proposed procedure is numerically tested by first generating a sample coming from a polynomial function $Y = P(X)$, where X is a r.v. of known distribution. Then, we show that the numerical method we propose converges to the right polynomial coefficients. As a second numerical experiment, the sample is generated from a function of the form $Y = f(X)$, where f is chosen to be an exponential function. The coefficients are then compared with the one obtained by a Galerkin projection of f on the PC basis. Finally, we test our procedure on raw data coming from an experiment, and we show that PC expansion significantly outperforms the representation of the same sample under the form of a Weibull distribution.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

The paper is organized as follows: in the next section, we restate some basic results about PC and we explain why, for our problem, the standard procedure cannot be used to compute PC coefficients. Then, by developing a constrained optimization problem, we show how to compute PC coefficients that model a given sample. Section 4 is devoted to numerical experiments: we first test our procedure on academic examples of known distribution and then we extend it to an experimental sample of an unknown distribution. Finally, Section 5 draws some conclusions.

2. PC EXPANSION

Let denote by $(\Omega, \mathcal{A}, \mathcal{P})$ the probability space, where as usual Ω is the set of all possible outcomes, $\mathcal{A}$ is a σ-algebra over Ω, and $\mathcal{P}$ is a function $\mathcal{A} \to [0,1]$ that gives a probability measure on $\mathcal{A}$. Consider an $\mathbb{R}$-valued random variable X that describes input uncertainties. The probability law of X may be defined by the probability density function $p(x)$ defined on $(\Omega, \mathcal{A}, \mathcal{P})$. Let us now denote $L_2(\Omega, \mathcal{A}, \mathcal{P})$ the space of real random variables with finite second-order moments. $L_2(\Omega, \mathcal{A}, \mathcal{P})$ is a Hilbert space which can be provided with a set of complete orthogonal polynomial basis $\{\Psi_j(x)\}_{j\geq 0}$ that are consistent with the density of X. For example, Legendre polynomials are associated with uniform distributions whereas Hermite polynomials are associated with Gaussian distributions. A given output function $Y = Y(X)$ which is assumed to belong to $L_2(\Omega, \mathcal{A}, \mathcal{P})$ can be represented as [5]

$$Y(X) = \sum_{j=0}^{\infty} \alpha_j \Psi_j(X). \tag{2.1}$$

In numerical simulations, this series is truncated by keeping terms up to a degree n

$$Y(X) \approx \widetilde{Y}(X) = \sum_{j=0}^{n} \alpha_j \Psi_j(X). \tag{2.2}$$

Once an approximation of the response system under the form of a PC expansion has been found, it is easy to compute various macroscopic quantities. For example, the expectation of $\widetilde{Y}$ is simply the first coefficient of the PC expansion, that is, $E(\widetilde{Y}) = \alpha_0$. Similarly, the variance of $\widetilde{Y}$ is given by

$$\mathrm{var}(\widetilde{Y}) = \sum_{j=1}^{n} {\alpha_j}^2.$$

In the next section, we briefly restate two ways of computing the coefficients α_j when a functional relationship of the form $Y = f(X)$ is known: Galerkin projection and linear regression. Then, we explain why none of the two procedures can be applied within the framework of this paper and we propose a new method.

3. COMPUTATION OF PC COEFFICIENTS

3.1. Functional Relationship $Y = f(X)$ Available

3.1.1. Galerkin projection Assume a surface response $\widetilde{Y}(X)$ of the form (2.2) that model the system $Y = f(X)$, where the expression of f does not need to be known explicitly. We can think of f as being a black box representing our model such that for any given input

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

value x_i the response of the system $y_i = f(x_i)$ is available. To compute the coefficients α_j of (2.2), we multiply both sides by $\Psi_i(X)$, take the expectation and use the linearity of the expectation to end up with

$$E(f(X)\Psi_i(X)) = \sum_{j=0}^{n} \alpha_j E(\Psi_i(X)\Psi_j(X)). \tag{3.1}$$

Using the orthonormality of the polynomials $\Psi_j(X)$, that is, $E(\Psi_i(X)\Psi_j(X)) = \delta_{ij}$ leads to

$$\alpha_i = E(f(X)\Psi_i(X)) = \int_I f(x)\Psi_i(x)p(x)dx, \tag{3.2}$$

The above integral can be evaluated with Gaussian quadrature rules based on n_q points, leading to

$$\alpha_i \approx \sum_{k=1}^{n_q} \omega_k f(x_k)\Psi_i(x_k), \tag{3.3}$$

where x_k are the quadrature points and ω_k are the quadrature weight. Also the procedure sketched above has been well described in the literature, it was important to restate its main elements in order to emphasize that the computation of α_i requires the evaluation of $f(x_k)$ at *specified* points x_k.

3.1.2. Linear regression The second most popular way of computing the coefficients α_j of (2.2) consists in solving this equation in the least square sense. For that, a collection of N response values $f(x_k)$ is computed for the input values x_k. The PC coefficients are sought as the minimum of the function $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$g(\alpha_0, \ldots, \alpha_n) = \sum_{k=1}^{N} \left(f(x_k) - \sum_{j=0}^{n} \alpha_j \Psi_j(x_k) \right)^2.$$

It can be shown that the solution $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_n)^T$ to such problem is

$$\boldsymbol{\alpha} = (M^T M)^{-1} M^T \mathbf{f}, \tag{3.4}$$

where $\mathbf{f} = (f(x_1), \ldots, f(x_N))^T$ and M is a $N \times (n+1)$ matrix defined by

$$M = \begin{pmatrix} \Psi_0(x_1) & \cdots & \Psi_n(x_1) \\ \vdots & \ddots & \vdots \\ \Psi_0(x_N) & \cdots & \Psi_n(x_N) \end{pmatrix}.$$

Here again, we wish to emphasize that the regression procedure to identify the coefficients α_j rely on the computation of the system response $f(x_k)$ for a specified input value x_k. To the best of our knowledge, all the available method in the literature to compute PC coefficients, rely on such input/output relationship. However, the problem we want to solve in this paper does not meet that condition and a new procedure, described in the next section, is proposed.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

3.2. Functional Relationship $Y = f(X)$ Not Available

Assume the problem at hand is now to model a sample $(y_{1,\dots,}y_{N_Y})$ of N_Y independent and identically distributed r.v. as follows:

$$Y = \sum_{j=0}^{n} \alpha_j \Psi_j(X), \tag{3.5}$$

where X is a r.v. of known distribution. Such a representation of a sample contains the same kind of information that the pdf of Y would do, and in both cases one has to fit parameters. For a probability density, one assumes a possible shape and then fit its parameters using a maximum likelihood estimator, for example. When the PC expansion is used, none of the previous techniques can be used to compute the PC coefficients, and an optimization problem has to be devised instead. In both cases, macroscopic quantities (expectation, variance ...) are then easy to compute. However, we believe that a representation under the form (3.5) can offer significant benefits in determining a possible pdf. First, it gives a greater freedom of choice when no standard probability density functions at hand present suitable shapes. Then, it is easy to generate a new sample for the r.v. Y from a sample of X with the formula (3.5). On the other hand, generating samples from non standard distributions may be a challenging task.

Let $(x_{1,\dots,}x_{N_X})$ be a sample of N_X independent and identically distributed r.v. having the same pdf as X. Such sample can easily be generated numerically with standard pseudo-random number generator. Let $(z_{1,\dots,}z_{N_X})$ be a sample of size N_X such that for $k = 1, \dots, N_X$, we have

$$z_k = \sum_{j=0}^{n} \alpha_j \Psi_j(x_k).$$

We denote by $\widetilde{F_Y}$ the empirical cumulative density function (cdf) built from the sample $(y_{1,\dots,}y_{N_Y})$ and by $\widetilde{F}_{(\alpha_0,\dots,\alpha_n)}$ the empirical cdf built from the sample $(z_{1,\dots,}z_{N_X})$ for a given set of coefficients $(\alpha_0, \dots, \alpha_n)$. Let $0 = t_0 < t_1 < \cdots < t_r = 1$ be a uniform subdivision of the interval $[0, 1]$ such that $t_{l+1} - t_l = 1/n$. Finally, we define the function c (referred to as 'cost function" hereafter) as follows:

$$\begin{aligned} c \quad &: \quad \mathbb{R}^{n+1} \to \mathbb{R} \\ (\alpha_0, \dots, \alpha_n) \quad &\mapsto \quad \sum_{l=0}^{r} \left| \widetilde{F}_Y^{-1}(t_l) - \widetilde{F}^{-1}_{(\alpha_0,\dots,\alpha_n)}(t_l) \right|, \end{aligned} \tag{3.6}$$

where $\widetilde{F}^{-1}$ stands for the inverse (reciprocal) empirical cdf of $\widetilde{F}$. For a sample of size N, it may be defined as a function $[0, 1] \to \mathbb{R}^+$ such that for all $y \in [0, 1]$,

$$\begin{aligned} \widetilde{F}^{-1}(y) &= \inf \left\{ x \in \mathbb{R} \ / \ \widetilde{F}(x) \geq y \right\} \\ &= \inf \left\{ x \in \mathbb{R} \ / \ \left(\frac{\text{number of elements} \leq x}{N} \right) \geq y \right\} \\ &= \inf \left\{ x \in \mathbb{R} \ / \ (\text{number of elements} \leq x) \geq Ny \right\} \end{aligned} \tag{3.7}$$

The inf function in the definition of the inverse empirical cdf, ensures that in case of multiple identical observations in the sample, $x = \widetilde{F}^{-1}(y)$ is defined as the smallest value of x such that $\widetilde{F}(x) \geq y$.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

Then, the coefficients of the PC expansion (3.5) are sought as the solution to the following constrained optimization problem:

$$\begin{cases} \alpha = (\alpha_0, \ldots, \alpha_n) = \underset{\alpha \in \mathbb{R}^{n+1}}{\arg\min} \left(c(\alpha)\right) \\ \alpha_0 = \widetilde{E}(Y) \\ \sum_{j=1}^{n} (\alpha_j)^2 = \widetilde{\mathrm{var}}(Y) \end{cases},$$

where $\widetilde{E}(Y)$ and $\widetilde{\mathrm{var}}(Y)$ are the empirical expectation and variance computed from the sample $(y_{1,\ldots,}y_{N_Y})$ as follows:

$$\widetilde{E}(Y) = \frac{1}{N_Y} \sum_{k=1}^{N_Y} y_k \quad \text{and} \quad \widetilde{\mathrm{var}}(Y) = \frac{1}{N_Y} \sum_{k=1}^{N_Y} \left(y_k - \widetilde{E}(Y)\right)^2$$

The constraints are directly incorporated into the cost function such that the constrained optimization problem can be solved as an unconstrained one. Then, we use the Nelder–Mead simplex algorithm as described in Lagarias et al. [7] to find the minimum of the cost function.

Remarks:

- The optimization problem is built to ensure that at $(r+1)$ points $0 = t_0 < t_1 < \cdots < t_r = 1$ of the interval $[0, 1]$, the inverse cdf of the r.v. Y matches the inverse cdf of the PC representation $\sum_{j=0}^{n} \alpha_j \Psi_j(X)$ of Y. The two constraints were found to greatly improve the robustness of the optimization problem. They impose that the empirical mean and the empirical variance of the sample $(y_{1,\ldots,}y_{N_Y})$ be exactly the mean and the variance of the PC expansion.
- We have chosen to match the inverse cdf of the two samples instead of the cdf, mainly for implementation and efficiency reasons. Numerical experiments using the cdf instead of the inverse cdf was found to give similar results.
- The cost function (3.6) to be minimized rely on a L^1 norm. Other norms could be used (such as the L^2 norm, for example), however, we found that the L^1 norm gave consistently more robust results for the optimization problem.

4. NUMERICAL RESULTS

4.1. The Kolmogorov–Smirnov Test

Before moving on to some numerical examples, we briefly describe the Kolmogorov–Smirnov (KS) test which is chosen to assess the quality of the PC representation of the sample. This test measures the discrepancy between the empirical cdf coming from the experimental sample (denoted by $\widetilde{F}_Y$) and the cdf of the PC expansion (denoted by $\widetilde{F}_{(\alpha_0,\ldots,\alpha_n)}$). Typically, the KS criterion is based on a L^∞ norm (it looks at the worst point of discrepancy between $\widetilde{F}_Y$ and $\widetilde{F}_{(\alpha_0,\ldots,\alpha_n)}$) and take the form (see [3] for details)

$$\mathrm{KS} = \sup_{x \in I} \left(\widetilde{F}_Y(x) - \widetilde{F}_{(\alpha_0,\ldots,\alpha_n)}(x) \right),$$

where I is the support of the distributions to be compared. This test can also be used to compare two samples, and in this case, the exact cdf has to be replaced by the empirical cdf.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

TABLE 1. Confidence intervals of PC coefficients for $n = 3$ (polynomial example).

$\alpha \setminus N_Y$	$N_Y = 500$	$N_Y = 2,000$	$N_Y = 8,000$	Exact coefficients
α_0	[0.9954; 1.0048]	[0.9967; 1.0009]	[0.9989; 1.0018]	1
α_1	[0.5008; 0.5241]	[0.5106; 0.5185]	[0.5158; 0.5212]	0.5
α_2	[0.2061; 0.2308]	[0.2036; 0.2092]	[0.2049; 0.2082]	0.25
α_3	[0.1275; 0.1500]	[0.1308; 0.1384]	[0.1380; 0.1432]	0.125

4.2. Data Coming From a Polynomial Function

As a first numerical experiment, we consider an orthogonal Hermite polynomial basis $\{\Psi_0, \ldots, \Psi_4\}$ of degree 4 and we define the r.v Y as

$$Y = \Psi_0(X) + \frac{1}{2}\Psi_1(X) + \frac{1}{4}\Psi_2(X) + \frac{1}{8}\Psi_3(X) + \frac{1}{16}\Psi_4(X), \tag{4.1}$$

where X is a Gaussian r.v. with mean 0 and variance 1.

We generate what will be considered as an 'experimental' sample for Y, which, in this case, is generated numerically with pseudo-random numbers for X and using formula (4.1). Our numerical procedure is tested for three different sizes of the experimental sample, namely $N_Y = 500$, 2,000, 8,000. The numerical method described in Section 3.2 for representing the r.v. Y under the form of a PC expansion requires the generation of a sample of size N_X for X. The robustness of the method should be tested with respect to the generation of this sample and therefore results for the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$ are given under the form of 95% confidence intervals:

$$\left[\overline{\alpha}_i - 1.96\frac{\sigma_{\alpha_i}}{\sqrt{p}},\ \overline{\alpha}_i + 1.96\frac{\sigma_{\alpha_i}}{\sqrt{p}}\right] \quad \text{for } i = 0, \ldots, 4$$

where $\overline{\alpha}_i$ and σ_{α_i} are the empirical mean and standard deviation of α_i computed from samples $(\alpha_0^k, \cdots, \alpha_4^k)_{1 \le k \le p}$ of size $p = 100$. Here we take $N_X = 10^5$ and $r = 3,000$, where $(r + 1)$ is the number of points where the inverse cdf of Y matches the inverse cdf of the PC expansion (see Eq. (3.6)). Tables 1 and 2 show results with increasing values of N_Y together with the exact coefficients for $n = 3$ and $n = 4$, respectively. We see that as the size of the experimental sample of Y increases, the confidence intervals tighten around the exact values of the PC expansion for $n = 4$. This is confirmed by the computation of the KS test whose average values are found to be 0.0263, 0.0133, and 0.0080 as the size of the sample for Y increases from $N_Y = 500$ to $N_Y = 8,000$. Such results show that as we increase the size of the experimental sample, its cdf $\widetilde{F}_Y(x)$ converges to the cdf of the PC expansion. For $n = 3$, the confidence intervals do not generally contain the value of the exact coefficient since a degree 3 polynomial is not able to properly represent a degree 4 polynomial. Accordingly, the KS test does not go below 0.0280 for $N_Y = 8,000$.

Furthermore, on Figures 1 and 2, we have plotted the pdf and the cdf of the experimental sample when $N_Y = 2,000$, together with the exact pdf/cdf of the r.v. given by Eq. (4.1) and the pdf/cdf of its PC representation. For this last case, the coefficients of the PC expansion are taken equal to their empirical mean value, which is also the central value of the confidence interval. We can see that the three curves match very well.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

TABLE **2.** Confidence intervals of PC coefficients for $n = 4$ (polynomial example).

$\alpha \setminus N_Y$	$N_Y = 500$	$N_Y = 2,000$	$N_Y = 8,000$	Exact coefficients
α_0	[0.9973; 1.0076]	[0.9984; 1.0030]	[0.9980; 1.0002]	1
α_1	[0.4810; 0.5052]	[0.4953; 0.5019]	[0.4976; 0.5009]	0.5
α_2	[0.2428; 0.2690]	[0.2426; 0.2520]	[0.2474; 0.2522]	0.25
α_3	[0.1140; 0.1304]	[0.1180; 0.1250]	[0.1227; 0.1265]	0.125
α_4	[0.0580; 0.0745]	[0.0563; 0.0636]	[0.0606; 0.0647]	0.0625

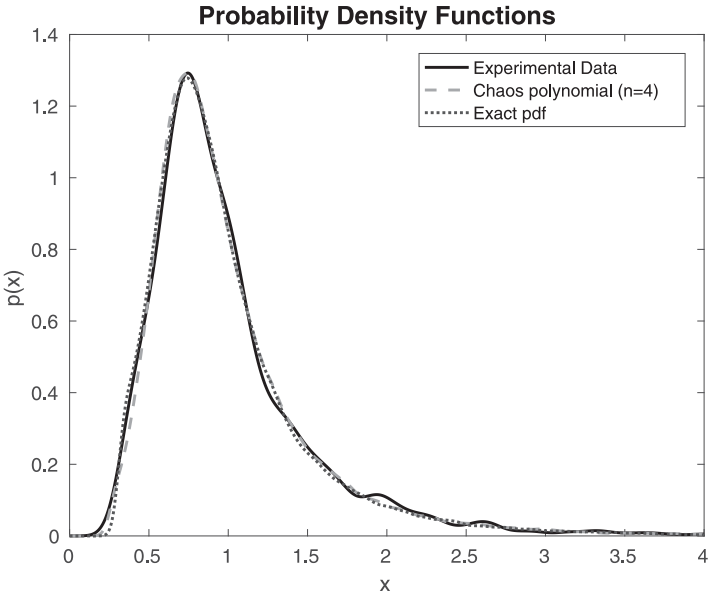


FIGURE **1.** PDF: Polynomial example.

4.3. Data Coming From a Non-Polynomial Function

As a second example, we consider a non-polynomial function of the form

$$Y = e^{0.5X}, \tag{4.2}$$

where X is a Gaussian r.v. with mean 0 and variance 1. In that case, it is not possible to have an exact representation of the r.v. Y under the form of a polynomial. Instead, an approximate of Y under the form of a PC expansion might be given by its Galerkin projection onto a Hermite basis as explained in Section 3.1. The PC coefficients obtained with this projection (see Eq. (3.2)) will be referred as 'Galerkin coefficients'. What will be considered as the 'experimental sample' of size N_Y is obtained by generating a sample of size N_Y for the r.v. Y and using formula (4.2). Note that the r.v. Y defined by the relation (4.2) follows a log-normal distribution. All the parameters N_X, N_Y, and r are set to the same values as in the previous example. For this example, nothing can ensure that the Galerkin coefficients are in the confidence interval. Indeed, a linear regression, for example, would have given a different set of reference PC coefficients. Tables 3 and 4 show results with increasing values of N_Y for $n = 3$ and $n = 4$, respectively.

We notice that for $n = 4$, as the size of the sample increases from $N_Y = 500$ to $N_Y = 8,000$, we get closer to the Galerkin coefficients, which is not the case for $n = 3$.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

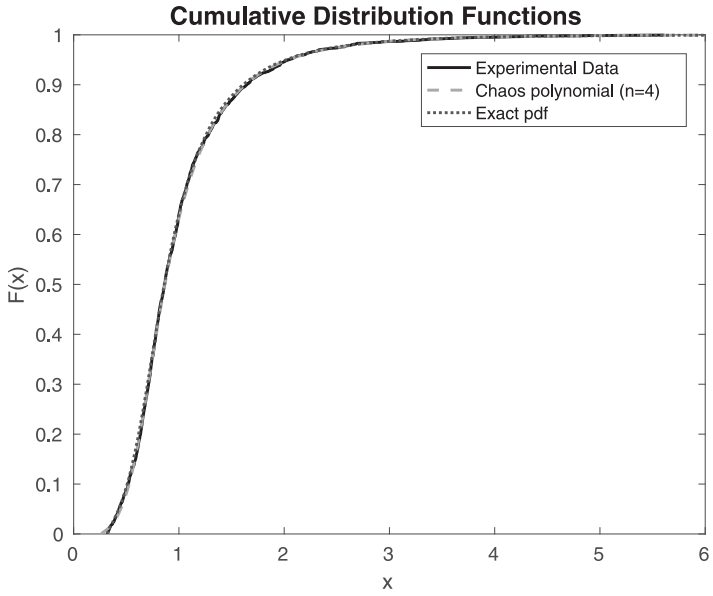


FIGURE **2.** CDF: Polynomial example.

TABLE 3. Confidence intervals of PC coefficients for $n = 3$ (exponential example).

$\alpha \setminus N_Y$	$N_Y = 500$	$N_Y = 2,000$	$N_Y = 8,000$	Galerkin coefficients
α_0	[1.1270; 1.1371]	[1.1300; 1.1352]	[1.1313; 1.1342]	1.1331
α_1	[0.5589; 0.5690]	[0.5670; 0.5730]	[0.5670; 0.5702]	0.5665
α_2	[0.1834; 0.1921]	[0.1915; 0.1964]	[0.1898; 0.1922]	0.1999
α_3	[0.0447; 0.0542]	[0.0534; 0.0583]	[0.0533; 0.0560]	0.0550

TABLE 4. Confidence intervals of PC coefficients for $n = 4$ (exponential example).

$\alpha \setminus N_Y$	$N_Y = 500$	$N_Y = 2,000$	$N_Y = 8,000$	Galerkin coefficients
α_0	[1.1210; 1.1318]	[1.1297; 1.1348]	[1.1315; 1.1346]	1.1331
α_1	[0.5539; 0.5651]	[0.5646; 0.5701]	[0.5655; 0.5686]	0.5666
α_2	[0.1868; 0.1987]	[0.1962; 0.2033]	[0.1983; 0.2015]	0.2003
α_3	[0.0446; 0.0545]	[0.0519; 0.0574]	[0.0539; 0.0564]	0.0577
α_4	[0.0030; 0.0125]	[0.0101; 0.0160]	[0.0114; 0.0140]	0.0139

Figures 3 and 4, show the pdf and the cdf of the experimental sample, the pdf/cdf of the PC representation of the sample and the exact pdf/cdf of the r.v. Y when $N_Y = 2,000$. In that case, we note that the pdf of the r.v. $Y = e^{0.5X}$ does not match exactly its PC expansion, due to the truncature error coming from the approximation of an exponential function by a polynomial of degree four.

For this example, the average values of the KS test for $n = 4$ are 0.0233, 0.0119, and 0.0070 for $N_Y = 500$, 2,000, and 8,000, respectively. Here again, those figures show the good convergence properties of our PC representation of the sample.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

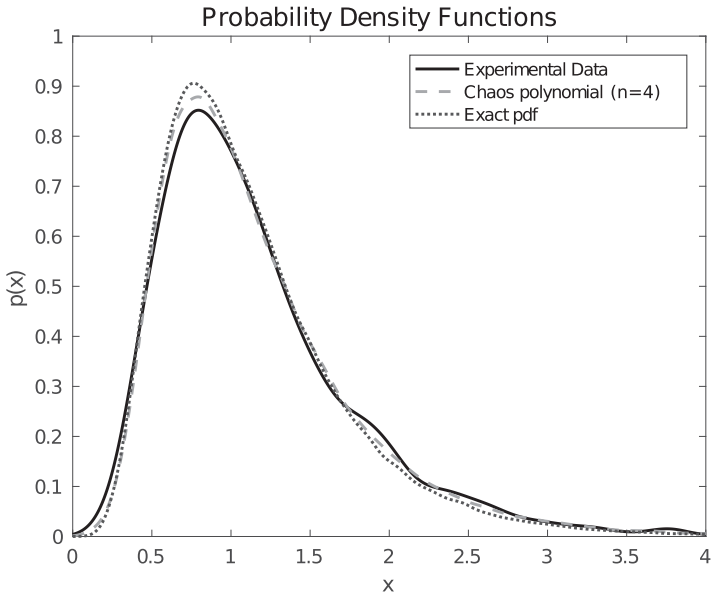


FIGURE **3.** PDF: Exponential example.

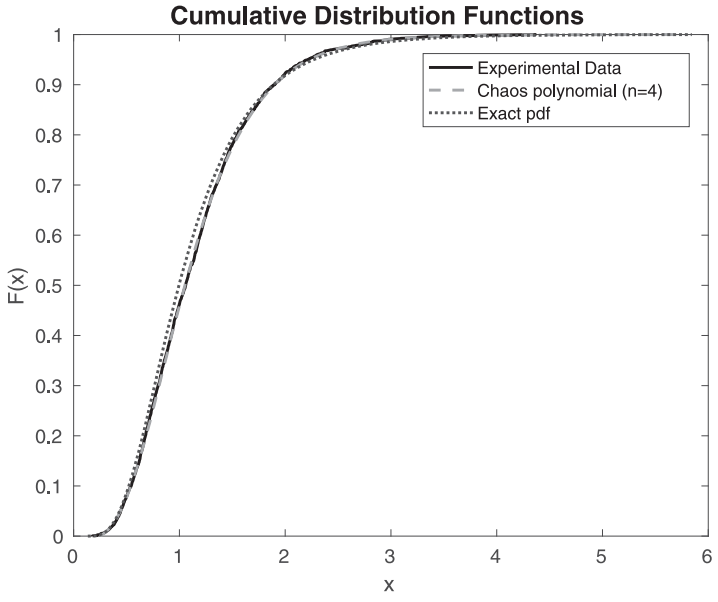


FIGURE **4.** CDF: Exponential example.

4.4. Experimental Data

Having tested and validated our numerical procedure for samples of known distribution, we now move on to the case of a true experimental sample. The sample of interest comes from the explosion of a tube under the effect of high internal pressure. After collecting the fragments, they are weighted individually and we are interested in modeling the weight of the fragments as a r.v. Y of the form (3.5). Since the exact form of the distribution of Y

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

TABLE 5. Confidence intervals of PC coefficients for $n = 3$ (experimental example).

α	
α_0	1.1423
α_1	$[1.8379; 1.8409]$
α_2	$[0.1388; 0.1540]$
α_3	$[-0.5672; -0.5534]$

TABLE 6. Confidence intervals of PC coefficients for $n = 4$ (experimental example).

α	
α_0	1.1423
α_1	$[1.8739; 1.8743]$
α_2	$[0.2504; 0.2542]$
α_3	$[-0.3446; -0.3347]$
α_4	$[0.1674; 0.1749]$

is not known, we compare our results with what is usually done in the literature (see [6,9], for example), that is, modeling the sample by a two-parameters Weibull distribution

$$p(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \ x > 0 \tag{4.3}$$

The two parameters (k, λ) are computed using a maximum likelihood estimation technique. Note that for $k = 1$, the Weibull reduces to an exponential distribution. This observation drives us to choose Laguerre polynomials as PC basis since they are orthogonal with respect to an exponential distribution.

Then we apply the numerical procedure described in Section 3 with $N_Y = 778$, $N_X = 10^5$ and $r = 400$ in order to compute the coefficients of the PC representation of our sample. Results are given in the Tables 5 and 6 for $n = 3$ and $n = 4$, respectively. The first coefficient α_0 is no longer given under the form of a confidence interval since it is equal to the empirical mean of the experimental sample, as a result of the constraint to the optimization problem.

In that case, the PC coefficients cannot be compared against the exact solution, however the confidence intervals are sufficiently narrow to demonstrate the robustness and the convergence of the proposed method. Using the maximum likelihood estimation technique, the parameters of the Weibull distribution (4.3) that best fit the sample are found to be $(\lambda, k) = (0.6483, 0.5372)$.

Figure 5 shows the pdf of the experimental data together with the pdf of this Weibull representation and its PC expansion. We clearly see that the Weibull fail to catch some important features of the experimental data. In particular, the maximum of the pdf cannot be properly modeled since, for the optimal parameters of the Weibull, it is a strictly decreasing function. On the other hand, the PC representation of the sample does catch this feature. We also represent the cdf for the three cases on Figure 6. Although the cdf somewhat averages the differences, it is clear that the cdf coming from PC expansion matches the experimental sample better than the cdf of the Weibull does. For this last example, the value of the KS test is about 50% higher for the Weibull (0.0930) than for the PC expansion with $n = 4$ (0.0615). For $n = 3$, the KS test gives results (KS = 0.0917) similar to the Weibull representation of the sample.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

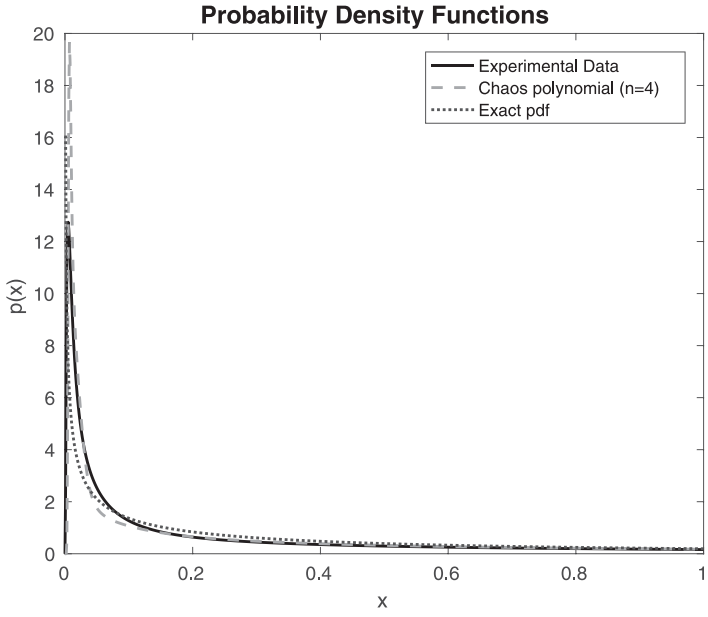


FIGURE **5.** PDF: experimental example.

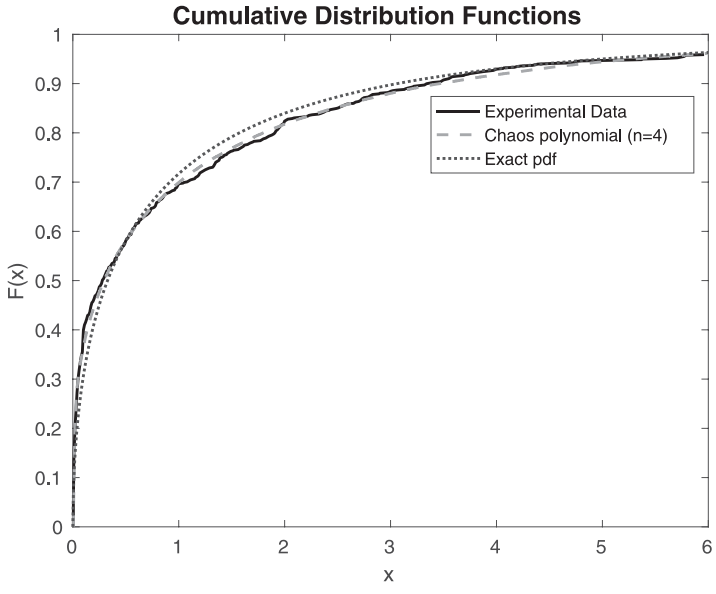


FIGURE **6.** CDF: experimental example.

5. CONCLUSION

In this paper, we have proposed a numerical procedure to model an experimental sample under the form of a PC expansion. It was shown that in this context, a standard technique such as Galerkin projection or linear regression could not be used to compute the PC coefficients. An alternative technique, based on a constrained optimization problem was provided and tested. Two academic examples of known solution have shown the robustness

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press

and the accuracy of the proposed method. Then, we have modeled an experimental masses sample coming from the explosion of a tube under the effect of high internal pressure using this procedure. It was shown that representing the sample under the form of a PC expansion of degree $n = 4$ was more accurate than using a standard Weibull distribution. This example has shown that using PC expansion offers an alternative choice when no distribution can adequately model a sample.

Acknowledgments

The authors would like to acknowledge the CEA/ GRAMAT for supporting this research (Contract Réf. CEA: C27358 Réf. UCA : DRV/VALO 2017-093).

References

1. Ahlfeld, R., Belkouchi, B., & Montomoli, F. (2016). SAMBA: sparse approximation of moment-based arbitrary polynomial chaos. *Journal of Computational Physics* 320: 1–16.
2. Arnst, M., Ghanem, R., & Soize, C. (2010). Identification of Bayesian posteriors for coefficients for chaos expansions. *Journal of Computational Physics* 229(9): 3134–3154.
3. Bonnini, S., Corain, L., Marozzi, M., & Salmaso, L (2014). *Nonparametric hypothesis testing*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd., Rank and permutation methods with applications in *R*, Presentation of the book by Fortunato Pesarin.
4. Chauvière, C., Hesthaven, J.S., & Lurati, L. (2006). Computational modeling of uncertainty in time-domain electromagnetics. *SIAM Journal on Scientific Computing* 28(2): 751–775.
5. Ghanem, R.G. & Spanos, P.D. (1991). *Stochastic finite elements: a spectral approach*. New York: Springer-Verlag.
6. Grady, D. (2006). *Fragmentation of rings and shells*. Shock wave and high pressure phenomena. New York, Berlin: Springer, The Legacy of N.F. Mott.
7. Lagarias, J.C., Reeds, J.A., Wright, M.H., & Wright, P.E. (1999). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization* 9(1): 112–147.
8. Le Maître, O.P., Reagan, M.T., Najm, H.N., Ghanem, R.G., & Knio, O.M. (2002). A stochastic projection method for fluid flow. II. Random process. *Journal of Computational Physics* 181(1): 9–44.
9. Levy, S., Molinari, J.F., Vicari, I., & Davison, A. (2010). Dynamic fragmentation of a ring: Predictable fragment mass distributions. *Physics Review* 82(6): 066105-1–066105-6.
10. Oladyshkin, S. & Nowak, W. (2012). Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliability Engineering & System Safety* 106: 179–190.
11. Sudret, M. (2008). Global sensitivity analysis using polynomial Chaos expansion. *Reliability Engineering & System Safety* 93(7): 964–979.
12. Wasserman, L. (2004). *All of statistics. Springer texts in statistics*. New York: Springer-Verlag, A concise course in statistical inference.
13. Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics* 60(4): 897–936.
14. Xiu, D. & Karniadakis, G.E. (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *Journal on Scientific Computing* 24(2): 619–644.

https://doi.org/10.1017/S026996481800030X Published online by Cambridge University Press