HAL
open science

# Identification of polynomial chaos representations in high dimension from a set of realizations

Guillaume Perrin, Christian Soize, Denis Duhamel, Christine Fünfschilling

## To cite this version:

Guillaume Perrin, Christian Soize, Denis Duhamel, Christine Fünfschilling. Identification of polynomial chaos representations in high dimension from a set of realizations. SIAM Journal on Scientific Computing, 2012, 34 (6), pp.A2917-A2945. 10.1137/11084950X . hal-00770006

**HAL Id: hal-00770006**

**https://hal.science/hal-00770006v1**

Submitted on 4 Jan 2013

**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.

# IDENTIFICATION OF POLYNOMIAL CHAOS REPRESENTATIONS IN HIGH DIMENSION FROM A SET OF REALIZATIONS

G. PERRIN*†‡, C. SOIZE*, D. DUHAMEL†, AND C. FUNFSCHILLING‡

**Abstract.**
This paper deals with the identification in high dimension of polynomial chaos expansion of random vectors from a set of realizations. Due to numerical and memory constraints, the usual polynomial chaos identification methods are based on a series of truncations that induces a numerical bias. This bias becomes very detrimental to the convergence analysis of polynomial chaos identification in high dimension. This paper therefore proposes a new formulation of the usual polynomial chaos identification algorithms to avoid this numerical bias. After a review of the polynomial chaos identification method, the influence of the numerical bias on the identification accuracy is quantified. The new formulation is then described in details, and illustrated on two examples.

**Key words.** polynomial chaos expansion, high dimension, computation.

**AMS subject classifications.** 60H35, 60H15, 60H25, 60H40, 65C50

**1. Introduction.** In spite of always more accurate numerical solvers, deterministic models are not able to represent most of the experimental data, which are variable and often uncertain by nature. Hence, the application fields of non deterministic modeling, which can take into account the model parameters variability as well as the model error uncertainties, has kept increasing. Uncertainties are therefore introduced in computational mechanical models with more and more degrees of freedom. In this context, the characterization of the probability distribution $P_{\boldsymbol{\eta}}(d\boldsymbol{x})$ of $N_\eta$-dimension random vector $\boldsymbol{\eta}$ from sets of experimental measurements is bound to play a key role, in particular, in high dimension, that is to say for a large value of $N_\eta$. In this work, it is assumed that $P_{\boldsymbol{\eta}}(d\boldsymbol{x}) = p_{\boldsymbol{\eta}}(\boldsymbol{x})d\boldsymbol{x}$ in which the probability density function (PDF) $p_{\boldsymbol{\eta}}$ is a function in the set $\mathcal{F}(\mathcal{D}, \mathbb{R}^+)$ of all the positive-valued functions defined on any part $\mathcal{D}$ of $\mathbb{R}^{N_\eta}$ and for which integral over $\mathcal{D}$ is 1.

Two kinds of methods can be used to build such a PDF: the direct and the indirect methods. Among the direct methods, the Prior Algebraic Stochastic Modeling (PASM) methods postulate an algebraic representation $\boldsymbol{\eta} \approx t^{\text{alg}}(\boldsymbol{\Xi}, \boldsymbol{w})$, with $t^{\text{alg}}$ a prior transformation, $\boldsymbol{\Xi}$ a given random vector and $\boldsymbol{w}$ a vector of parameters to identify. In the same category, the methods based on the Information Theory and the Maximum Entropy Principle (MEP) have been developped (see [13] and [27]) to compute $p_{\boldsymbol{\eta}}$ from the only available information of random vector $\boldsymbol{\eta}$. This information can be seen as the admissible set $\mathcal{C}^{\text{ad}}$ for $p_{\boldsymbol{\eta}}$:

$$\begin{aligned}\mathcal{C}^{\text{ad}} = \Big\{ p_{\boldsymbol{\eta}} \in \mathcal{F}(\mathcal{D}, \mathbb{R}^+) \mid & \int_{\mathcal{D}} p_{\boldsymbol{\eta}}(\boldsymbol{x})d\boldsymbol{x} = 1, \\ & \forall 1 \leq m \leq M, \quad \int_{\mathcal{D}} \boldsymbol{g}_m(\boldsymbol{x}) p_{\boldsymbol{\eta}}(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{f}_m \Big\},\end{aligned} \tag{1.1}$$

*Université Paris-Est, Modélisation et Simulation Multi-Échelle (MSME UMR 8208 CNRS), 5 Bd. Descartes, 77454 Marne-la-Vallée, France (christian.soize@univ-paris-est.fr).

†Université Paris-Est, Navier (ENPC-IFSTTAR-CNRS UMR 8205), Ecole Nationale des Ponts et Chaussées, 6 et 8 Avenue Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne-la-Vallée, Cedex 2, France (denis.duhamel@enpc.fr)

‡SNCF, Innovation and Research Department, Immeuble Lumière, 40 avenue des Terroirs de France, 75611, Paris, Cedex 12, France (guillaume.perrin@sncf.fr,christine.funfschilling@sncf.fr).

where $\{\boldsymbol{f}_m,\ 1 \leq m \leq M\}$ gathers $M$ given vectors which are respectively associated with a given vector-valued functions $\{\boldsymbol{g}_m,\ 1 \leq m \leq M\}$. Hence, the MPE allows building $p_{\boldsymbol{\eta}}$ as the solution of the optimization problem:

$$p_{\boldsymbol{\eta}} = \arg \max_{p_{\boldsymbol{\eta}} \in \mathcal{C}^{\mathrm{ad}}} \left\{ - \int_{\mathcal{D}} p_{\boldsymbol{\eta}}(\boldsymbol{x}) \log \left( p_{\boldsymbol{\eta}}(\boldsymbol{x}) \right) d\boldsymbol{x} \right\}. \tag{1.2}$$

On the other hand, the indirect methods allow the construction of the PDF $p_{\boldsymbol{\eta}}$ of the considered random vector $\boldsymbol{\eta}$ from a transformation $\boldsymbol{t}$ of a known random vector $\boldsymbol{\xi} = \left(\xi_1, ..., \xi_{N_g}\right)$ of given dimension $N_g \leq N_\eta$:

$$\boldsymbol{\eta} = \boldsymbol{t}\left(\boldsymbol{\xi}\right), \tag{1.3}$$

defining a transformation $\mathbb{T}$ between $p_{\boldsymbol{\eta}}$ and the PDF $p_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$:

$$p_{\boldsymbol{\eta}} = \mathbb{T}\left(p_{\boldsymbol{\xi}}\right). \tag{1.4}$$

The construction of the transformation $\boldsymbol{t}$ is thus the key point of these indirect methods. In this context, the isoprobabilistic transformations such as the Nataf transformation (see [20]) or the Rosenblatt transformation (see [23]) have allowed the development of interesting results in the second part of the twentieth century but are still limited to very small dimension cases and not to the high dimension case considered in this work. Nowadays, the most popular indirect methods are the polynomial chaos expansion (PCE) methods, which have been first introduced by Wiener [33] for stochastic processes, and pioneered by Ghanem and Spanos [10] [11] for the use of it in computational sciences. In the last decade, this very promising method has thus been applied in many works (see, for instance [1], [2], [3], [4], [5], [7], [8], [9], [12], [14], [15], [16], [19], [18], [17], [21], [22], [24], [26], [28], [31], [32], [25], [34]). The PCE is based on a direct projection of the random vector $\boldsymbol{\eta}$ on a chosen hilbertian basis $\mathcal{B}_{orth} = \left\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \boldsymbol{\alpha} \in \mathbb{N}^{N_g}\right\}$ of all the second-order random vectors with values in $\mathbb{R}^{N_\eta}$:

$$\boldsymbol{\eta} = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^{N_g}} \boldsymbol{y}^{(\boldsymbol{\alpha})} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{1.5}$$

$$\boldsymbol{\xi} \mapsto \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = X_{\alpha_1}(\xi_1) \otimes ... \otimes X_{\alpha_{N_g}}(\xi_{N_g}), \tag{1.6}$$

where $x \mapsto X_{\alpha_\ell}(x)$ is the normalized polynomial basis of degree $\alpha_\ell$ associated with the PDF $p_{\xi_\ell}$ of the random variable $\xi_\ell$, and $\boldsymbol{\alpha}$ is the multi-index of the multidimensional polynomial basis element $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$. Building the transformation $\boldsymbol{t}$ requires therefore the construction of the projection vectors $\left\{\boldsymbol{y}^{(\boldsymbol{\alpha})},\ \boldsymbol{\alpha} \in \mathbb{N}^{N_g}\right\}$.

The present work is devoted to the identification in high dimension of the PCE coefficients $\left\{\boldsymbol{y}^{(\boldsymbol{\alpha})},\ \boldsymbol{\alpha} \in \mathbb{N}^{N_g}\right\}$, when the only available information on the random vector $\boldsymbol{\eta}$ is a set of $\nu^{\mathrm{exp}}$ independent realizations $\left\{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(\nu^{\mathrm{exp}})}\right\}$.

In practice, the PCE of $\boldsymbol{\eta}$ has first to be truncated:

$$\boldsymbol{\eta} \approx \boldsymbol{\eta}^{\text{chaos}}(N) = \sum_{\boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p} \boldsymbol{y}^{(\boldsymbol{\alpha})} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{1.7}$$

$$\boldsymbol{\mathcal{A}}_p = \left\{ \boldsymbol{\alpha} = \left(\alpha_1, ..., \alpha_{N_g}\right) \mid |\boldsymbol{\alpha}| = \sum_{\ell=1}^{N_g} \alpha_\ell \leq p \right\} = \left\{ \boldsymbol{\alpha}^{(1)}, \cdots, \boldsymbol{\alpha}^{(N)} \right\}, \tag{1.8}$$

where $\boldsymbol{\eta}^{\text{chaos}}(N)$ is the projection of $\boldsymbol{\eta}$ on the $N$-dimension subspace spanned by $\{\ \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}),\ \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\ \} \subset \mathcal{B}_{orth}$. It can be noticed that $N$ increases very quickly with respect to the dimension $N_g$ of $\boldsymbol{\xi}$ and the maximum degree $p$ of the truncated basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}),\ \boldsymbol{\alpha} \in \mathcal{A}_p\}$, as:

$$N = (N_g + p)! / \left(N_g!\ p!\right). \tag{1.9}$$

Methods to perform the convergence analysis in high dimension with respect to a given error threshold on the PCE residue $\boldsymbol{\eta} - \boldsymbol{\eta}^{\text{chaos}}(N)$ are therefore of great concern to justify the truncation parameters $N_g$ and $p$.

In this prospect, the article [29] provides advanced algorithms to compute the PCE coefficients from the $\nu^{\text{exp}}$ independent realizations $\left\{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(\nu^{\text{exp}})}\right\}$ by focusing on the maximization of the likelihood. In particular, one of the key point of these algorithms is the calculation of $\left(N \times \nu^{\text{chaos}}\right)$ real matrix $[\Psi]$ of independent realizations of the truncated PCE basis $\{\psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\xi}\right),\ \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\}$:

$$[\Psi] = \left[\boldsymbol{\Psi}\left(\boldsymbol{\xi}\left(\theta_1\right), p\right) \quad \cdots \quad \boldsymbol{\Psi}\left(\boldsymbol{\xi}\left(\theta_{\nu^{\text{chaos}}}\right), p\right)\right], \tag{1.10}$$

$$\boldsymbol{\Psi}\left(\boldsymbol{\xi}, p\right) = \left(\psi_{\boldsymbol{\alpha}^{(1)}}\left(\xi_1, \cdots, \xi_{N_g}\right), \cdots, \psi_{\boldsymbol{\alpha}^{(N)}}\left(\xi_1, \cdots, \xi_{N_g}\right)\right), \tag{1.11}$$

where the set $\{\boldsymbol{\xi}\left(\theta_1\right), \cdots, \boldsymbol{\xi}\left(\theta_{\nu^{\text{chaos}}}\right)\}$ gathers $\nu^{\text{chaos}}$ independent realizations of the random vector $\boldsymbol{\xi}$.

Recurrence formula or algebraic explicit representations are generally used to compute such matrix $[\Psi]$, which are supposed to verify the asymptotical property:

$$\lim_{\nu^{\text{chaos}} \to +\infty} \frac{1}{\nu^{\text{chaos}}} [\Psi][\Psi]^T = [I_N], \tag{1.12}$$

as a direct consequence of the orthonormality of the PCE basis $\{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\}$, where $[I_N]$ is the $N$-dimension identity matrix.

However, for numerically admissible values of $\nu^{\text{chaos}}$ (between 1000 and 10000), it has been shown in [30] that the difference $\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N]$ can be very significant when high values of the maximum degree $p$ can be encountered with simultaneously significant values of $N_g$. This difference induces a detrimental bias in the PCE identification, which makes the convergence of classical PCE in high dimension very difficult.

In [30], it is therefore proposed a method using singular matrix decomposition to numerically adapt classical generations of $[\Psi]$, and make this difference be zero for any values of $p$ and $N_g$. Nevertheless, this conditionning on $[\Psi]$ modifies the initial structure of $[\Psi]$, and makes the identified PCE coefficients $\left\{\boldsymbol{y}^{(\boldsymbol{\alpha})}, \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\right\}$ impossible to be reused on an other matrix $[\Psi^*]$ of $\nu^{\text{chaos},*}$ new realizations of $\boldsymbol{\Psi}(\boldsymbol{\xi}, p)$.

As an extension of the works described in [29] and [30], this article proposes an original decomposition of the PCE coefficients $\left\{\boldsymbol{y}^{(\boldsymbol{\alpha})}, \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\right\}$, that reduces the numerical bias introduced during the identification by the finite dimension of $[\Psi]$ and for large values of degree $p$. This new formulation is particulary adapted to the high dimension, and allows the identified coefficients to be reused for other matrix of realizations $[\Psi^*]$.

In Section 2, the PCE identification from a set of experimental data with an arbitrary measure is described. In particular, the role played by the matrix of independent realizations $[\Psi]$ is emphasized. Section 3 focuses on the convergence properties of this matrix $[\Psi]$ with respect to three statistical measures, and describes an innovative method to generate this matrix without using computational recurrence formula nor algebraic explicit representation. In Section 4, the new formulation of the PCE identification problem is given. Finally, are presented in Section 5 two applications of the former method with a Gaussian measure.

**2. PCE identification of random vectors from a set of independent realizations.** In this section, a description of the PCE identification with respect to an arbitrary measure is given. The objective is to summarize the different key steps of the PCE identification method and the way they are practically implemented.

After having defined the theoretical frame of the PCE identification, the cost-function that leads to the computation of the PCE coefficients $\left\{\boldsymbol{y}^{(\boldsymbol{\alpha})},\ \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\right\}$ is presented, for given truncation parameters $N_g$ and $p$. At last, to justify the choice of these truncation parameters, a method to perform the convergence analysis is introduced.

**2.1. Theoretical frame.** Let $(\Theta, \mathcal{T}, \mathcal{P})$ be a probability space. Let $L^2_{\mathcal{P}}\left(\Theta, \mathbb{R}^{N_\eta}\right)$ be the space of all the second-order $N_\eta$-dimension random vectors defined on $(\Theta, \mathcal{T}, \mathcal{P})$ with values in $\mathbb{R}^{N_\eta}$, equipped with the inner product $\langle ., . \rangle$:

$$\langle \boldsymbol{U}, \boldsymbol{V} \rangle = \int_\Theta \boldsymbol{U}^T(\theta)\boldsymbol{V}(\theta)dP(\theta) = E\left(\boldsymbol{U}^T\boldsymbol{V}\right), \quad \forall\, \boldsymbol{U}, \boldsymbol{V} \in L^2_{\mathcal{P}}\left(\Theta, \mathbb{R}^{N_\eta}\right), \tag{2.1}$$

where $E\left(.\right)$ is the mathematical expectation.

Let $\boldsymbol{\eta} = \left(\eta_1, \cdots, \eta_{N_\eta}\right)$ be an element of $L^2_{\mathcal{P}}\left(\Theta, \mathbb{R}^{N_\eta}\right)$. It is assumed that $\nu^{\text{exp}}$ independent realizations $\left\{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(\nu^{\text{exp}})}\right\}$ of $\boldsymbol{\eta}$ are known and gathered in the $(N_\eta \times \nu^{\text{exp}})$ real matrix $[\eta^{\text{exp}}]$:

$$[\eta^{\text{exp}}] = \left[\boldsymbol{\eta}^{(1)} \quad \cdots \quad \boldsymbol{\eta}^{(\nu^{\text{exp}})}\right]. \tag{2.2}$$

Equation (1.7) can be rewritten as:

$$\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p), \tag{2.3}$$

$$[y] = \left[ \boldsymbol{y}^{(\boldsymbol{\alpha}^{(1)})} \ \cdots \ \boldsymbol{y}^{(\boldsymbol{\alpha}^{(N)})} \right]. \tag{2.4}$$

The orthonormality property of the projection basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}),\ \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\}$ yields the condition:

$$E\left(\boldsymbol{\Psi}(\boldsymbol{\xi},p)\boldsymbol{\Psi}(\boldsymbol{\xi},p)^T\right) = [I_N]. \tag{2.5}$$

Since $\psi_{\boldsymbol{\alpha}^{(1)}}(\boldsymbol{\xi}) = 1$, it can be seen that:

$$E\left(\boldsymbol{\eta}^{\text{chaos}}(N)\right) = \boldsymbol{y}^{(\boldsymbol{\alpha}^{(1)})}. \tag{2.6}$$

Let $[R_\eta]$ and $[R_\eta^{\text{chaos}}(N)]$ be the autocorrelation matrix of the random vectors $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^{\text{chaos}}(N)$:

$$[R_\eta] = E\left(\boldsymbol{\eta}\boldsymbol{\eta}^T\right), \tag{2.7}$$

$$\left[R_\eta^{\text{chaos}}(N)\right] = E\left(\boldsymbol{\eta}^{\text{chaos}}(N)\left(\boldsymbol{\eta}^{\text{chaos}}(N)\right)^T\right) = [y]E\left(\boldsymbol{\Psi}(\boldsymbol{\xi},p)\boldsymbol{\Psi}(\boldsymbol{\xi},p)^T\right)[y]^T = [y][y]^T. \tag{2.8}$$

**2.2. Identification of the polynomial chaos expansion coefficients.** In this section, particular values of the truncation parameters $N_g$ and $p$ are considered. Let $\mathcal{M}_{N_\eta N}$ be the space of all the $(N_\eta \times N)$ real matrices. For a given value of $[y^*]$ in $\mathcal{M}_{N_\eta N}$, the random vector $\boldsymbol{U}([y^*]) = [y^*]\boldsymbol{\Psi}(\boldsymbol{\xi},p)$ is a $N_\eta$-dimension random vector, for which the autocorrelation is equal to $[y^*][y^*]^T$. Let $p_{\boldsymbol{U}([y^*])}$ be its multidimensional PDF.

When the only available information on $\boldsymbol{\eta}$ is a set of $\nu^{\text{exp}}$ independent realizations, the optimal coefficients matrix $[y]$ of its truncated PCE, $\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi},p)$, can be seen as the argument which maximizes the log-likelihood $\mathcal{L}_{\boldsymbol{U}([y^*])}([\eta^{\text{exp}}])$ of $\boldsymbol{U}([y^*])$:

$$[y] = \arg \max_{[y^*]\in\mathcal{M}_{N_\eta N}} \mathcal{L}_{\boldsymbol{U}([y^*])}([\eta^{\text{exp}}]), \tag{2.9}$$

$$\mathcal{L}_{\boldsymbol{U}([y^*])}([\eta^{\text{exp}}]) = \sum_{i=1}^{\nu^{\text{exp}}} \ln p_{\boldsymbol{U}([y^*])}\left(\boldsymbol{\eta}^{(i)}\right). \tag{2.10}$$

**2.3. Practical solving of the log-likelihood maximization.**

**2.3.1. The need for statistical algorithms to maximize the log-likelihood.** The log-likelihood $\mathcal{L}_{\boldsymbol{U}([y^*])}([\eta^{\text{exp}}])$ being non-convex, deterministic algorithms such as gradient algorithms cannot be applied to solve Eq. (2.9), and random search algorithms have to be used. Hence, the precision of the PCE has to be correlated to a numerical cost $M$, which corresponds to a number of independent trials of $[y^*]$ in $\mathcal{M}_{N_\eta N}$. Let $\mathcal{Y} = \left\{[y^*]^{(r)},\ 1 \le r \le M\right\}$ be a set of $M$ elements, which have been

chosen randomly in $\mathcal{M}_{N_\eta N}$. For a given numerical cost $M$, the most accurate PCE coefficients matrix $[y]$ is approximated by:

$$[y] \approx [y_{\mathcal{Y}}] = \arg \max_{[y^*]\in\mathcal{Y}} \mathcal{L}_{\boldsymbol{U}([y^*])}\left([\eta^{\text{exp}}]\right). \tag{2.11}$$

**2.3.2. Restriction of the maximization domain.** From the $\nu^{\text{exp}}$ independent realizations $\left\{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(\nu^{\text{exp}})}\right\}$, the mean value $E(\boldsymbol{\eta})$ and the autocorrelation matrix $[R_\eta]$ of $\boldsymbol{\eta}$ can be estimated by:

$$E\left(\boldsymbol{\eta}\right) \approx \widehat{\boldsymbol{\eta}}(\nu^{\text{exp}}) = \frac{1}{\nu^{\text{exp}}} \sum_{i=1}^{\nu^{\text{exp}}} \boldsymbol{\eta}^{(i)}, \tag{2.12}$$

$$[R_\eta] \approx [\widehat{R}_\eta(\nu^{\text{exp}})] = \frac{1}{\nu^{\text{exp}}} \sum_{i=1}^{\nu^{\text{exp}}} \boldsymbol{\eta}^{(i)} \left(\boldsymbol{\eta}^{(i)}\right)^T = \frac{1}{\nu^{\text{exp}}} [\eta^{\text{exp}}][\eta^{\text{exp}}]^T. \tag{2.13}$$

A good way to improve the efficiency of the numerical identification of $[y]$ is then to restrict the research set to $\mathcal{O}_\eta \subset \mathcal{M}_{N_\eta N}$, with:

$$\begin{aligned} \mathcal{O}_\eta = \Big\{ [y] = \left[\boldsymbol{y}^{(\boldsymbol{\alpha}^{(1)})}, \cdots, \boldsymbol{y}^{(\boldsymbol{\alpha}^{(N)})}\right] \in \mathcal{M}_{N_\eta N} \mid \\ \boldsymbol{y}^{(\boldsymbol{\alpha}^{(1)})} = \widehat{\boldsymbol{\eta}}(\nu^{\text{exp}}),\ [y][y]^T = [\widehat{R}_\eta(\nu^{\text{exp}})] \Big\}, \end{aligned} \tag{2.14}$$

which, taking into account Eqs. (2.6) and (2.8), guarantees by construction that:

$$\left\{ \begin{array}{l} [R_\eta^{\text{chaos}}(N)] = [\widehat{R}_\eta(\nu^{\text{exp}})], \\ E\left(\boldsymbol{\eta}^{\text{chaos}}(N)\right) = \widehat{\boldsymbol{\eta}}(\nu^{\text{exp}}). \end{array} \right. \tag{2.15}$$

Hence, the PCE coefficients matrix $[y]$ can be approximated as the argument in $\mathcal{O}_\eta$ that maximizes the log-likelihood $\mathcal{L}_{\boldsymbol{U}([y^*])}\left([\eta^{\text{exp}}]\right)$. By defining $\mathcal{W}$ the set that gathers $M$ randomly raised elements of $\mathcal{O}_\eta$, $[y]$ can then be assessed as the solution of the new optimization problem:

$$[y] \approx [y_{\mathcal{W}}] = \arg \max_{[y^*]\in\mathcal{W}} \mathcal{L}_{\boldsymbol{U}([y^*])}\left([\eta^{\text{exp}}]\right). \tag{2.16}$$

**2.3.3. Approximation of the log-likelihood function.** From a particular matrix of realizations $[\Psi]$ (which is defined in Eq. (1.10)), if $[y^*]$ is an element of $\mathcal{O}_\eta$, $\nu^{\text{chaos}}$ independent realizations $\left\{\boldsymbol{U}\left([y^*], \theta_n\right) = [y^*]\boldsymbol{\Psi}\left(\boldsymbol{\xi}(\theta_n), p\right), 1 \leq n \leq \nu^{\text{chaos}}\right\}$ of the random vector $\boldsymbol{U}([y^*])$ can be computed and gathered in the matrix $[U]$:

$$[U] = [\boldsymbol{U}\left([y^*], \theta_1\right) \quad \cdots \quad \boldsymbol{U}\left([y^*], \theta_{\nu^{\text{chaos}}}\right)] = [y^*][\Psi]. \tag{2.17}$$

Hence, using Gaussian Kernels, the PDF $p_{\boldsymbol{U}([y^*])}$ of $\boldsymbol{U}\left([y^*]\right)$ can be directly estimated by its non parametric estimator $\widehat{p}_{\boldsymbol{U}}$:

$$\forall \boldsymbol{x} \in \mathbb{R}^{N_\eta}, \quad p_{\boldsymbol{U}([y^*])}(\boldsymbol{x}) \approx$$
$$\widehat{p}_{\boldsymbol{U}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{N_\eta/2}\,\nu^{\text{chaos}}\prod_{k=1}^{N_\eta} h_k} \sum_{n=1}^{\nu^{\text{chaos}}} \exp\left(-\frac{1}{2}\sum_{k=1}^{N_\eta}\left(\frac{x_k - U_k([y^*],\theta_n)}{h_k}\right)^2\right), \tag{2.18}$$

where $\boldsymbol{h} = \left(h_1, \cdots, h_{N_\eta}\right)$ is the multidimensionnal optimal Silverman bandwith vector (see [6]) of the Kernel smoothing estimation of $p_{\boldsymbol{U}([y^*])}$:

$$\forall\ 1 \leq k \leq N_\eta,\ h_k = \widehat{\sigma}_{U_k}\left(\frac{4}{(2+N_\eta)\nu^{\text{exp}}}\right)^{1/(N_\eta+4)}, \tag{2.19}$$

where $\widehat{\sigma}_{U_k}$ is the empirical estimation of the standard deviation of each component $U_k$ of $\boldsymbol{U}$. It has to be noticed that $\widehat{p}_{\boldsymbol{U}}$ only depends on the bandwidth vector $\boldsymbol{h}$, and the two matrices $[y^*]$ and $[\Psi]$. Hence, according to the Eqs. (2.10), (2.17) and (2.18), for a given value of $\nu^{\text{chaos}}$, the maximization of the log-likelihood function $\mathcal{L}_{\boldsymbol{U}([y^*])}$ can be replaced by the maximization of the cost-function $\mathcal{C}([\eta^{\text{exp}}],[y^*],[\Psi])$ such that:

$$[y] \approx [y_{\mathcal{O}_\eta}] = \arg\max_{[y^*]\in\mathcal{O}_\eta} \mathcal{C}([\eta^{\text{exp}}],[y^*],[\Psi]), \tag{2.20}$$

where:

$$\mathcal{C}([\eta^{\text{exp}}],[y^*],[\Psi]) = \mathcal{C}_C + \mathcal{C}_V([\eta^{\text{exp}}],[y^*],[\Psi]), \tag{2.21}$$

$$\mathcal{C}_C = -\nu^{\text{exp}} \ln\left((2\pi)^{N_\eta/2}\,\nu^{\text{chaos}}\prod_{k=1}^{N_\eta} h_k\right), \tag{2.22}$$

$$\mathcal{C}_V([\eta^{\text{exp}}],[y^*],[\Psi]) = \sum_{i=1}^{\nu^{\text{exp}}} \ln\left(\sum_{n=1}^{\nu^{\text{chaos}}} \exp\left(-\frac{1}{2}\sum_{k=1}^{N_\eta}\left(\frac{\eta_k^{(i)} - U_k([y^*],\theta_n)}{h_k}\right)^2\right)\right). \tag{2.23}$$

Hence, the optimization problem defined by Eq. (2.16) can finally be estimated by:

$$[y] \approx [y^M_{\mathcal{O}_\eta}] = \arg\max_{[y^*]\in\mathcal{W}} \mathcal{C}\left([\nu^{\text{exp}}],[y^*],[\Psi]\right). \tag{2.24}$$

The optimization problem defined by Eq. (2.24) is now supposed to be solved with the advanced algorithms described in [29] to optimize the trials of the elements of $\mathcal{W}$ for a given computation cost $M$. The higher the value of $M$ is, the better the PCE identification should be. Therefore, this value has to be chosen as high as possible while respecting the computational ressource limitation.

**2.3.4. Accuracy of the PCE identification.** For a given computation cost $M$, let $[y^M_{\mathcal{O}_\eta}]$ be an optimal solution of Eq. (2.24). $[y^M_{\mathcal{O}_\eta}]$ is a numerical estimation of the PCE coefficients matrix $[y]$. For a new $\left(N \times \nu^{\text{chaos},*}\right)$ real matrix $[\Psi^*]$ of independent realizations ($\nu^{\text{chaos},*}$ can be higher than $\nu^{\text{chaos}}$), the robustness of $[y^M_{\mathcal{O}_\eta}]$ regarding the choice of $[\Psi]$ can then be estimated by comparing $\mathcal{C}\left([\eta^{\text{exp}}], [y^M_{\mathcal{O}_\eta}], [\Psi]\right)$ and $\mathcal{C}\left([\eta^{\text{exp}}], [y^M_{\mathcal{O}_\eta}], [\Psi^*]\right)$. In addition, if $\nu^{\text{exp}}$ new independent realizations of $\boldsymbol{\eta}$ were available and gathered in the matrix $[\eta^{\text{exp,new}}]$, the over-learning of the method could be measured by comparing $\mathcal{C}\left([\eta^{\text{exp}}], [y^M_{\mathcal{O}_\eta}], [\Psi]\right)$ and $\mathcal{C}\left([\eta^{\text{exp,new}}], [y^M_{\mathcal{O}_\eta}], [\Psi]\right)$. At last, for the same computation cost $M$, if $[y^{M,\text{new}}_{\mathcal{O}_\eta}]$ is a new optimal solution of Eq. (2.24), the global accuracy of the identification stems from the comparison between $\mathcal{C}\left([\eta^{\text{exp,new}}], [y^M_{\mathcal{O}_\eta}], [\Psi^*]\right)$ and $\mathcal{C}\left([\eta^{\text{exp,new}}], [y^{M,\text{new}}_{\mathcal{O}_\eta}], [\Psi^*]\right)$.

**2.4. Identification of the PCE truncation parameters.** As shown in Introduction, two truncation parameters, $N_g$ and $p$, appear in the truncated PCE, $\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p)$, of $\boldsymbol{\eta}$. The values of these parameters have to be determined from a convergence analysis. The objective of this section is thus to give the fundamental elements to perform such a convergence analysis.

**2.4.1. Definition of a log error function.** For each component $\eta_k^{\text{chaos}}(N)$ of the truncated PCE, $\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p)$, of $\boldsymbol{\eta}$, the $L^1$-log error function $err_k$ is introduced as described in [29]:

$$\forall\ 1 \leq k \leq N_\eta, \quad err_k(N_g, p) = \int_{\text{BI}_k} \left| log_{10}\left(p_{\eta_k}(x_k)\right) - log_{10}\left(p_{\eta_k^{\text{chaos}}}(x_k)\right) \right| dx_k, \quad (2.25)$$

where:

- $\text{BI}_k$ is the support of $\eta_k^{\text{exp}}$;
- $p_{\eta_k}$ and $p_{\eta_k^{\text{chaos}}}$ are the PDF of $\eta_k$ and $\eta_k^{\text{chaos}}$ respectively.

The multidimensional error function $err(N_g, p)$ is then deduced from the unidimensional $L^1$-log error function as:

$$err(N_g, p) = \sum_{k=1}^{N_\eta} err_k(N_g, p). \quad (2.26)$$

The parameters $N_g$ and $p$ have thus to be determined to minimize the multidimensional $L^1$-log error function $err(N_g, p)$.

For given values of truncation parameters $N_g$ and $p$, it is reminded that PCE coefficients matrix $[y]$ is searched in order to maximize the multidimensional log-likelihood function, which allows us to consider a priori strongly correlated problems. Once this matrix $[y]$ is identified, it is possible to generate as many independent realizations of truncated PCE $\boldsymbol{\eta}^{\text{chaos}}(N)$ as needed to estimate as precisely as possible the non parametric estimator $\widehat{p}_{\boldsymbol{U}}$ of its multidimensional PDF. The number $\nu^{\text{exp}}$ of available experimental realizations of $\boldsymbol{\eta}$ is however limited. This number is generally too small for the non parametric estimator of multidimensional PDF $p_{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ to be relevant, whereas it is most of the time large enough to define the estimators of the marginals of $p_{\boldsymbol{\eta}}$. Therefore, the log-error functions defined by Eqs. (2.25) and (2.26)

only consider the marginals of the PDF of $p_{\boldsymbol{\eta}}$ and $p_{\boldsymbol{\eta}}^{\text{chaos}}$. In addition, the logarithm function has been introduced in order to measure the errors of the very small values of the probability density function (the tails of the probability density function).

**2.4.2. Definition of an admissible set for the truncation parameters.** As it exists an isoprobabilistic transformation between $\boldsymbol{\eta}$ and $\left(\Xi_1, \cdots, \Xi_{N_\eta}\right)$, where $\{\Xi_k,\ 1 \leq k \leq N_\eta\}$ is a set of $N_\eta$ independent centered normalized Gaussian random variables, the convergence analysis can be restricted to the values of $N_g$ which verify:

$$N_g \leq N_\eta. \tag{2.27}$$

Moreover, imposing the $(N_\eta \times N)$ real matrix $[y]$ to be in $\mathcal{O}_\eta$ amounts to imposing $\frac{N_\eta(N_\eta+3)}{2}$ constraints on $[y]$, which implies:

$$N_\eta N \geq \frac{N_\eta\left(N_\eta+3\right)}{2} \Leftrightarrow N \geq \frac{N_\eta+3}{2}. \tag{2.28}$$

However, the algorithms developed in [29], on which the solving of the optimization problem, defined by Eq. (2.24), is based, need the more restrictive condition:

$$N \geq N_\eta + 1. \tag{2.29}$$

We will therefore consider $\mathcal{Q}(N_\eta)$ the set of the admissible values for $p$ and $N_g$ with:

$$\mathcal{Q}(N_\eta) = \left\{(p, N_g) \in \mathbb{N}^2,\ \mid N_g \leq N_\eta,\ N = (N_g+p)!/(N_g!\ p!) \geq N_\eta + 1\right\}. \tag{2.30}$$

Theoretically, increasing $p$ and $N_g$ adds terms in the PCE of the considered random vector, and therefore should induce the decrease of the error function:

$$\begin{aligned} \forall\ p^* \geq p,\ N_g^* \geq N_g,\ \ err(N_g, p) \geq \max\left\{err(N_g^*, p), err(N_g, p^*)\right\} \geq \\ \min\left\{err(N_g^*, p), err(N_g, p^*)\right\} \geq err(N_g^*, p^*). \end{aligned} \tag{2.31}$$

However, the higher the values of $p$ and $N_g$ are, the bigger the PCE coefficients matrix is, the harder the numerical identification is. Hence, introducing $\varepsilon$ as an error threshold, which has to be adapted to the problem, let $\mathcal{P}(\varepsilon, N_\eta)$ be the set:

$$\mathcal{P}(\varepsilon, N_\eta) = \left\{(p, N_g) \in \mathcal{Q}(N_\eta) \mid err(N_g, p) \leq \varepsilon\right\}. \tag{2.32}$$

Finally, given the error threshold $\varepsilon$, rather than directly minimizing the $L^1$-log error function $err(N_g, p)$, it appears to be more accurate to look for the optimal values of $p$ and $N_g$ that minimize the size of the projection basis $N = (N_g+p)!/\left(N_g!\ p!\right)$:

$$(p, N_g) = \arg \min_{(p^*, N_g^*) \in \mathcal{P}(\varepsilon, N_\eta)} \left(N_g^* + p^*\right)!/\left(N_g^*!\ p^*!\right). \tag{2.33}$$

If the polynomial order (which is a priori unknown) of the non truncated PCE of $\boldsymbol{\eta}$ is infinite, it may not exist values of $p$ and $N_g$ in $\mathcal{P}(\varepsilon, N_\eta)$ for error function $err(N_g, p)$ to be inferior to small values of $\varepsilon$. In this case, the former algorithms can nevertheless be used to find the most accurate values of $p$ and $N_g$ with respect to an available computational cost.

**3. Adaptation of the PCE identification method in high dimension.** As it has been presented in the former sections, the $\left(N \times \nu^{\text{chaos}}\right)$ real matrix $[\Psi]$ gathers $\nu^{\text{chaos}}$ independent realizations of the $N$-dimension PCE basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \quad \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\}$. Eqs. (2.23) and (2.24) underline the fact that the numerical identification of the PCE coefficients $[y]$ can be seen as the minimization of a cost function involving the elements of the $\left(N_\eta \times \nu^{\text{chaos}}\right)$ real matrix of independent realizations $[U] = [y][\Psi]$ and the elements of the $(N_\eta \times \nu^{\text{exp}})$ real matrix $[\eta^{\text{exp}}] = [\boldsymbol{\eta}^{(1)} \quad \cdots \quad \boldsymbol{\eta}^{(\nu^{\text{exp}})}]$. In theoretical terms, this cost function should be minimum when the multidimensional PDF $p_{\boldsymbol{U}}$ of $\boldsymbol{U} = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p)$ is as near as possible to the multidimensional PDF $p_{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$. In practical terms, this cost function is however minimum when $\widehat{p}_{\boldsymbol{U}}$ is as near as possible to $\widehat{p}_{\boldsymbol{\eta}}$, where $\widehat{p}_{\boldsymbol{U}}$ and $\widehat{p}_{\boldsymbol{\eta}}$ are the multidimensional non parametric estimators of $p_{\boldsymbol{U}}$ and $p_{\boldsymbol{\eta}}$ defined by Eq. (2.18). With respect to $\nu^{\text{exp}}$ and $\nu^{\text{chaos}}$, three bias are then introduced in the PCE identification:

- a bias due to a lack of information on $\boldsymbol{\eta}$:

$$b^{(1)}(\nu^{\text{exp}}) = \int_{\mathbb{R}^{N_\eta}} |\widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x}) - p_{\boldsymbol{\eta}}(\boldsymbol{x})| d\boldsymbol{x}, \tag{3.1}$$

- a bias due to a lack of information on $\boldsymbol{U}$:

$$b^{(2)}(\nu^{\text{chaos}}) = \int_{\mathbb{R}^{N_\eta}} |\widehat{p}_{\boldsymbol{U}}(\boldsymbol{x}) - p_{\boldsymbol{U}}(\boldsymbol{x})| d\boldsymbol{x}, \tag{3.2}$$

- a bias due to the truncation and to the fact that the global maximum is not necessary reached:

$$b^{(3)}(\nu^{\text{exp}}, \nu^{\text{chaos}}) = \int_{\mathbb{R}^{N_\eta}} |\widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x}) - \widehat{p}_{\boldsymbol{U}}(\boldsymbol{x})| d\boldsymbol{x}. \tag{3.3}$$

These three bias could also be expressed with respect to the statistical moments of $\boldsymbol{\eta}$ and $\boldsymbol{U}$. For instance, when focusing on the autocorrelation matrix, let $err^1$, $err^2$ and $err^3$ be the autocorrelation errors corresponding respectively to the bias $b^{(1)}$, $b^{(2)}$ and $b^{(3)}$:

$$err^1(\nu^{\text{exp}}) = \left\| [R_\eta] - [\widehat{R}_\eta(\nu^{\text{exp}})] \right\|_F / \left\| [R_\eta] \right\|_F, \tag{3.4}$$

$$err^2(\nu^{\text{chaos}}) = \left\| [R_\eta^{\text{chaos}}(N)] - [\widehat{R}_U(\nu^{\text{chaos}})] \right\|_F / \left\| [R_\eta^{\text{chaos}}(N)] \right\|_F, \tag{3.5}$$

$$err^3(\nu^{\text{exp}}, \nu^{\text{chaos}}) = \left\| [\widehat{R}_U(\nu^{\text{chaos}})] - [\widehat{R}_\eta(\nu^{\text{exp}})] \right\|_F / \left\| [\widehat{R}_\eta(\nu^{\text{exp}})] \right\|_F, \tag{3.6}$$

where $\|.\|_F$ is the Frobenius norm of matrices, and where it is reminded from Eqs. (2.8) and (2.13) that:

$$\left\{ \begin{array}{c} \left[\widehat{R}_\eta(\nu^{\text{exp}})\right] = \frac{1}{\nu^{\text{exp}}}[\eta^{\text{exp}}][\eta^{\text{exp}}]^T, \\ \left[\widehat{R}_U(\nu^{\text{chaos}})\right] = \frac{1}{\nu^{\text{chaos}}}[U][U]^T = [y]\left(\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T\right)[y]^T, \\ \left[R_\eta^{\text{chaos}}(N)\right] = [y][y]^T. \end{array} \right. \tag{3.7}$$

Hence, the smaller these three errors are, the more precise the PCE identification is. The $\nu^{\text{exp}}$ independent realizations $\left\{\boldsymbol{\eta}^{(1)}, \cdots, \boldsymbol{\eta}^{(\nu^{\text{exp}})}\right\}$ being the maximum available information on $\boldsymbol{\eta}$, the bias $b^{(1)}$ and the autocorrelation error $err^1$ cannot be decreased, whereas the set $\mathcal{O}_\eta$, which was introduced to guarantee that $[R_\eta^{\text{chaos}}(N)] = [\widehat{R}_\eta(\nu^{\text{exp}})]$, aims at reducing $b^{(2)}$, $b^{(3)}$, $err^2$ and $err^3$. Therefore, imposing $[y]$ to be in $\mathcal{O}_\eta$ leads us to:

$$\begin{aligned} & err^2\left(\nu^{\text{chaos}}\right) = err^3\left(\nu^{\text{exp}}, \nu^{\text{chaos}}\right) \\ & = \left\| [y]\left(\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N]\right)[y]^T \right\|_F / \left\| [y][I_N][y]^T \right\|_F . \end{aligned} \tag{3.8}$$

The following asymptotical property can thus be deduced from Eq. (1.12):

$$\lim_{\nu^{\text{chaos}} \to +\infty} err^2\left(\nu^{\text{chaos}}\right) = \lim_{\nu^{\text{chaos}} \to +\infty} err^3\left(\nu^{\text{exp}}, \nu^{\text{chaos}}\right) = 0, \tag{3.9}$$

which is equivalent to say that the larger $\nu^{\text{chaos}}$ is, the more accurate the PCE identification should be. However, from a practical point of view, the value of $\nu^{\text{chaos}}$ is fixed by the available computation ressources. As an extension of the work presented in [30], this section aims at quantifying the divergence of the ratio:

$$r = \left\| \frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N] \right\|_F / \left\| [I_N] \right\|_F , \tag{3.10}$$

when the truncation parameters $N_g$ and $p$ increase for several statistical measures. From Eq. (3.8), $r$, defined by Eq. (3.10), can be seen as a general characterization of the autocorrelation errors $err^2$ and $err^3$. This divergence being very detrimental to the PCE identification in high dimension, a new decomposition of the PCE coefficient matrix $[y]$ will be then presented in this section to make $err^2$ and $err^3$ be zero for any value of $N_g$ and $p$.

**3.1. Decomposition of the matrix of independent realizations.** To better emphasize the influence of the truncation parameters on the ratio $r$, a rewriting of the matrix $[\Psi]$ is first presented.

**3.1.1. Theoretical basis of the decomposition.** From Eq. (1.10), matrix $[\Psi]$ gathers $\nu^{\text{chaos}}$ columns $\left\{\boldsymbol{\Psi}\left(\boldsymbol{\xi}(\theta_n), p\right),\ 1 \leq n \leq \nu^{\text{chaos}}\right\}$, which are independent realizations of the $N$-dimension PCE basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}),\ \boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}_p\}$. This basis being orthonormal leads us to the asymptotical condition on $[\Psi]$, defined by Eq. (1.12). Moreover, Eq. (1.6) implies that $[\Psi]$ can be expressed as:

$$[\Psi] = [A][M], \tag{3.11}$$

where $[A]$ is the $(N \times N)$ real matrix that gathers the coefficients of the orthonormal polynomials with respect to the probability measure of the $N_g$-dimension PCE germ, $\boldsymbol{\xi} = \left(\xi_1, \cdots, \xi_{N_g}\right)$, and $[M]$ is a $\left(N \times \nu^{\text{chaos}}\right)$ real matrix of $\nu^{\text{chaos}}$ independent realizations of the multi-index monomials $\mathcal{M}_{\boldsymbol{\alpha}}\left(\boldsymbol{\xi}\right) = \xi_1^{\alpha_1} \times \cdots \times \xi_{N_g}^{\alpha_{N_g}}$, for any value $\boldsymbol{\alpha}$ in $\boldsymbol{\mathcal{A}}_p$:

$$[M] = [\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}\left(\theta_1\right), p\right) \quad \cdots \quad \boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}\left(\theta_{\nu^{\text{chaos}}}\right), p\right)], \tag{3.12}$$

$$\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, p\right) = \left(\mathcal{M}_{\boldsymbol{\alpha}^{(1)}}\left(\boldsymbol{\xi}\right), \cdots, \mathcal{M}_{\boldsymbol{\alpha}^{(N)}}\left(\boldsymbol{\xi}\right)\right). \tag{3.13}$$

If $[A]$ is independent of $[M]$, Eq. (3.11) certifies that, if the columns of $[M]$ are independent, then the columns of $[\Psi]$ stay independent. Let $[R_{\boldsymbol{\mathcal{E}}}]$ be the autocorrelation matrix of the random vector $\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, p\right)$:

$$[R_{\boldsymbol{\mathcal{E}}}] = E\left(\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, p\right)\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, p\right)^T\right). \tag{3.14}$$

It can be deduced from Eqs. (1.12), (3.11), (3.12) and (3.14) that:

$$[R_{\boldsymbol{\mathcal{E}}}] = \lim_{\nu^{\text{chaos}} \to +\infty} \frac{1}{\nu^{\text{chaos}}}[M][M]^T = [A]^{-1}[A]^{-T}. \tag{3.15}$$

According to this decomposition, computing the classical Gram-Schmidt orthogonalization to identify the polynomial basis coefficients only requires the calculation of $[A]^{-T}$, which corresponds to the Cholesky decomposition matrix of the positive definite matrix $[R_{\boldsymbol{\mathcal{E}}}]$. Hence, by construction, the matrix $[\Psi]$ can be written as the product of a lower triangular matrix $[A]$ and a matrix $[M]$ of independent realizations of a multi-index random vector $\boldsymbol{\mathcal{E}}(\boldsymbol{\xi}, p)$.

**3.1.2. Practical computation of matrix $[\Psi]$.** Thanks to Eq. (3.11), matrix $[\Psi]$ can be numerically computed without requiring computational recurrence formula nor algebraic explicit representation. An illustration of the method is presented hereinafter for a PCE based on a Gaussian mesure. This development can be directly extented to any value of $p$ and $N_g$, as well as to other statistical measures. Let $\xi_1$ and $\xi_2$ be two independent normalized Gaussian random variables, such that $\boldsymbol{\xi} = (\xi_1, \xi_2)$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. Choosing $p = 2$ and $N_g = 2$, which corresponds to $N = 6$, leads to the following definition of $\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, p\right)$:

$$\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}, 2\right) = \left(1, \xi_1, \xi_2, \xi_1\xi_2, \xi_1^2, \xi_2^2\right). \tag{3.16}$$

According to this equation, matrix $[M]$ can thus be easily deduced from $\nu^{\text{chaos}}$ independent realizations of $\boldsymbol{\xi}$. Moreover, let $[\alpha]$ be the $(N_g \times N)$ real matrix which gathers the admissible values for $\boldsymbol{\alpha}$ in $\boldsymbol{\mathcal{A}}_p$:

$$[\alpha] = \begin{bmatrix} 0 & 1 & 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 0 & 2 \end{bmatrix} \leftrightarrow \boldsymbol{\mathcal{A}}_p = \{(0,0), (1,0), (0,1), (1,1), (2,0), (0,2)\}. \tag{3.17}$$

The random variables $\xi_1$ and $\xi_2$ being independent, normalized and Gaussian, the autocorelation matrix $[R_{\boldsymbol{\mathcal{E}}}]$ can thus be written as:

$$\begin{aligned} \forall i, j \in \{1, \cdots, N\}, \ [R_{\boldsymbol{\mathcal{E}}}]_{ij} &= E\left(\xi_1^{[\alpha]_{1i}+[\alpha]_{1j}} \times ... \times \xi_{N_g}^{[\alpha]_{N_g i}+[\alpha]_{N_g j}}\right) \\ &= E\left(\xi_1^{[\alpha]_{1i}+[\alpha]_{1j}}\right) \times ... \times E\left(\xi_{N_g}^{[\alpha]_{N_g i}+[\alpha]_{N_g j}}\right), \end{aligned} \tag{3.18}$$

where, for $1 \leq \ell \leq N_g$:

$$\begin{cases} E\left(\xi_\ell^q\right) = 0 \text{ if q is not even,} \\ E\left(\xi_\ell^q\right) = \frac{q!}{(q/2)!2^{q/2}} \text{ if q is even.} \end{cases} \tag{3.19}$$

Therefore, Eq. (3.15) allows us to numerically find back in $[A]$ the multidimensional Hermite polynomials $H_{\alpha_1} \times \cdots \times H_{\alpha_{N_g}}$:

$$\forall x \in \mathbb{R}, \begin{cases} H_0(x_1) \times H_0(x_2) = 1 \\ H_1(x_1) \times H_0(x_2) = x_1 \\ H_0(x_1) \times H_1(x_2) = x_2 \\ H_1(x_1) \times H_1(x_2) = x_1 x_2 \\ H_2(x_1) \times H_0(x_2) = \frac{x_1^2 - 1}{\sqrt{2}} \\ H_0(x_1) \times H_2(x_2) = \frac{x_2^2 - 1}{\sqrt{2}} \end{cases} \leftrightarrow [A] = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{-1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{2}} & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{3.20}$$

Noticing that:

- if $\xi$ is a random variable uniformly distributed on $[-1, 1]$:

$$\begin{cases} E\left(\xi^q\right) = 0 \text{ if q is not even,} \\ E\left(\xi^q\right) = \frac{1}{q+1} \text{ if q is even,} \end{cases} \tag{3.21}$$

- if the random variable $\xi$ is a random variable that is characterized by a normalized exponential distribution on $[0, +\infty[$:

$$E\left(\xi^q\right) = q!, \tag{3.22}$$

this method can directly be generalized to the uniform and exponential cases to compute the multidimensional Legendre and Laguerre polynomial coefficients, but also to an arbitrary probability measure for the germ $\boldsymbol{\xi}$.

**3.2. Influence of the truncation parameters and of the choice for the PCE probability measure.** The convergence properties of ratio $r$ when $\nu^{\text{chaos}}$ tends to infinity are strongly related to the statistical properties of germ $\boldsymbol{\xi}$. This section aims therefore to emphasize the dominant trends of this specific link, and to highlight the difficulties brought about by the divergence of ratio $r$, when trying to perform analysis of convergence in high dimension.

The definition of the Frobenius norm allows us to write that:

$$r = \left\| \frac{1}{\nu^{\text{chaos}}} [\Psi][\Psi]^T - [I_N] \right\|_F / \left\| [I_N] \right\|_F = \sqrt{N} \Sigma(\nu^{\text{chaos}}), \tag{3.23}$$

where $\Sigma(\nu^{\text{chaos}})$ is such that:

$$\left\{ \Sigma(\nu^{\text{chaos}}) \right\}^2 = \frac{1}{N^2} \sum_{1 \leq i,j \leq N} \left( \left( \frac{1}{\nu^{\text{chaos}}} [\Psi][\Psi]^T - [I_N] \right)_{ij} \right)^2. \tag{3.24}$$

By construction, $\left\{\Sigma(\nu^{\text{chaos}})\right\}^2$ is an assessment of the mean value of the squared difference between the elements of $\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T$ and the elements of the identity matrix $[I_N]$. Hence, if $\left\{\Sigma(\nu^{\text{chaos}})\right\}^2$ remains constant when the size $N$ of the polynomial basis increases, the ratio $r$ should increase as $\sqrt{N}$. Moreover, Eqs. (3.11) and (3.15) yield,

$$\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N] = [A]\left(\frac{1}{\nu^{\text{chaos}}}[M][M]^T - [R_{\boldsymbol{\mathcal{E}}}]\right)[A]^T. \tag{3.25}$$

For all $(i,j)$ in $\{1,\cdots,N\}^2$, $[R_{\boldsymbol{\mathcal{E}}}]_{ij}$ is such that:

$$[R_{\boldsymbol{\mathcal{E}}}]_{ij} = E\left(\xi_1^{\alpha_1^{(i)}+\alpha_1^{(j)}} \times \cdots \times \xi_{N_g}^{\alpha_{N_g}^{(i)}+\alpha_{N_g}^{(j)}}\right). \tag{3.26}$$

Let $[\widehat{R}_{\boldsymbol{\mathcal{E}}}]$ be the following estimator of $[R_{\boldsymbol{\mathcal{E}}}]$:

$$[\widehat{R}_{\boldsymbol{\mathcal{E}}}]_{ij} = \frac{1}{\nu^{\text{chaos}}}\sum_{n=1}^{\nu^{\text{chaos}}}\left(\Xi_1^{(n)}\right)^{\alpha_1^{(i)}+\alpha_1^{(j)}} \times \cdots \times \left(\Xi_{N_g}^{(n)}\right)^{\alpha_{N_g}^{(i)}+\alpha_{N_g}^{(j)}}, \tag{3.27}$$

where $\left\{\boldsymbol{\Xi}^{(n)} = \left(\Xi_1^{(n)},\cdots,\Xi_{N_g}^{(n)}\right),\ 1\le n\le \nu^{\text{chaos}}\right\}$ is a set of $\nu^{\text{chaos}}$ independent $N_g$-dimension random vectors, which have the same PDF than $\boldsymbol{\xi}$. The central limit theorem yields that, for all $(i,j)$ in $\{1,\cdots,N\}^2$, we have:

$$\sqrt{\frac{\nu^{\text{chaos}}}{\text{Var}\left(\xi_1^{\alpha_1^{(i)}+\alpha_1^{(j)}} \times \cdots \times \xi_{N_g}^{\alpha_{N_g}^{(i)}+\alpha_{N_g}^{(j)}}\right)}}\left([\widehat{R}_{\boldsymbol{\mathcal{E}}}]_{ij} - [R_{\boldsymbol{\mathcal{E}}}]_{ij}\right) \overset{\text{in law}}{\longrightarrow} \mathcal{N}(0,1), \tag{3.28}$$

where $\mathcal{N}(0,1)$ is the normalized Gaussian distribution, and Var(.) is the variance. Under this formalism, it can be noticed that $\frac{1}{\nu^{\text{chaos}}}[M][M]^T$ is one particular realization of $[\widehat{R}_{\boldsymbol{\mathcal{E}}}]$. Hence, from Eqs. (3.24), (3.25) and (3.28), we deduce that:

- if $\text{Var}\left(\xi_1^{\alpha_1^{(i)}+\alpha_1^{(j)}} \times \cdots \times \xi_{N_g}^{\alpha_{N_g}^{(i)}+\alpha_{N_g}^{(j)}}\right) \le \text{Var}\left(\xi_1^{\alpha_1^{(i)}+\alpha_1^{(j)}} \times \cdots \times \xi_{N_g+1}^{\alpha_{N_g+1}^{(i)}+\alpha_{N_g+1}^{(j)}}\right)$, then $\Sigma(\nu^{\text{chaos}})$ potentially increases with respect to $N_g$.
- if $\text{Var}\left(\xi_\ell^{\alpha_\ell^{(i)}}\right) \le \text{Var}\left(\xi_\ell^{\alpha_\ell^{(j)}}\right)$ for $\alpha_\ell^{(i)} \le \alpha_\ell^{(j)}$, then $\Sigma(\nu^{\text{chaos}})$ potentially increases with respect to $p$.

As an illustration, for each couple $(N_g,p)$ such that $1\le p\le 10$ and $1\le N_g\le 6$, three sets, $\{[\Psi_U^{(m)}(p,N_g)],\ 1\le m\le 1000\}$, $\{[\Psi_G^{(m)}(p,N_g)],\ 1\le m\le 1000\}$ and $\{[\Psi_E^{(m)}(p,N_g)],\ 1\le m\le 1000\}$, are computed, such that $[\Psi_U^{(m)}(p,N_g)]$, $[\Psi_G^{(m)}(p,N_g)]$ and $[\Psi_G^{(m)}(p,N_g)]$ refer to particular $\left(N\times\nu^{\text{chaos}}\right)$ real matrices of independent realisations of the basis $\{\psi_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha}(\boldsymbol{\xi})\in\boldsymbol{\mathcal{A}}_p\}$, in the uniform, the Gaussian and the exponential cases, respectively. Hence, defining:

$$\left\{\begin{array}{l} r_U^m(\nu^{\text{chaos}}) = \left\| \frac{1}{\nu^{\text{chaos}}}[\Psi_U^{(m)}(p,N_g)][\Psi_U^{(m)}(p,N_g)]^T - [I_N] \right\|_F / \|[I_N]\|_F \,, \\ r_G^m(\nu^{\text{chaos}}) = \left\| \frac{1}{\nu^{\text{chaos}}}[\Psi_G^{(m)}(p,N_g)][\Psi_G^{(m)}(p,N_g)]^T - [I_N] \right\|_F / \|[I_N]\|_F \,, \\ r_E^m(\nu^{\text{chaos}}) = \left\| \frac{1}{\nu^{\text{chaos}}}[\Psi_E^{(m)}(p,N_g)][\Psi_E^{(m)}(p,N_g)]^T - [I_N] \right\|_F / \|[I_N]\|_F \,, \end{array}\right. \tag{3.29}$$

allows us to compute, in each case, three approximations $err_U^{\text{ortho}}(p,N_g)$, $err_G^{\text{ortho}}(p,N_g)$ and $err_E^{\text{ortho}}(p,N_g)$ of the mean value of the ratio $r$, defined in Eq. (3.23), such that:

$$\left\{\begin{array}{l} err_U^{\text{ortho}}(p,N_g) = \frac{1}{1000}\sum_{m=1}^{1000} r_U^m(\nu^{\text{chaos}}), \\ err_G^{\text{ortho}}(p,N_g) = \frac{1}{1000}\sum_{m=1}^{1000} r_G^m(\nu^{\text{chaos}}), \\ err_E^{\text{ortho}}(p,N_g) = \frac{1}{1000}\sum_{m=1}^{1000} r_E^m(\nu^{\text{chaos}}). \end{array}\right. \tag{3.30}$$

For $\nu^{\text{chaos}} = 1000$, in figure 3.1, the two factors which make the ratio $r$ diverge with respect to $p$ and $N_g$ can therefore be emphasized. On the first hand, if increasing $p$ or $N_g$ does not increase the variance of the elements of $\boldsymbol{\mathcal{E}}(\boldsymbol{\xi},p)$, which is the case if the PCE germ $\boldsymbol{\xi}$ is characterized by an uniform distribution (see Eq. (3.21)), the ratio $r$ increases approximately as $\sqrt{N}$. On the other hand, if increasing $p$ or $N_g$ increases the variance of the element of $\boldsymbol{\mathcal{E}}(\boldsymbol{\xi},p)$, as it is the case if the PCE germ $\boldsymbol{\xi}$ is characterized by a Gaussian or exponential distribution (see Eqs. (3.19) and (3.22)), the ratio $r$ diverges very quickly with respect to the truncation parameters, and bias the PCE identification results.

As a conclusion, for a fixed value of $\nu^{\text{chaos}}$, the difference $\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N]$ increases when $p$ and $N_g$ increase. Therefore, imposing $[y]$ to be in $\mathcal{O}_\eta$ introduces a numerical bias in the PCE identification, which becomes very important when high values of $p$ and $N_g$ are needed. Such a phenomenon prevents thus to perform the analysis of convergence of the PCE in high dimension, especially when dealing with Gaussian and exponential PCE germs.

**3.3. Adaptation of the optimization problem.** In this section, fixed values for $\nu^{\text{chaos}}$, $p$ and $N_g$ are considered. According to the notations of Section 3.1, a $\left(N \times \nu^{\text{chaos}}\right)$ real matrix of independent realizations $[\Psi] = [A][M]$ can then be constructed. Under the condition $\nu^{\text{chaos}} \geq N$, $\frac{1}{\nu^{\text{chaos}}}[M][M]^T$ is positive definite by construction, which allows writing:

$$\frac{1}{\nu^{\text{chaos}}}[M][M]^T = [L][L]^T, \tag{3.31}$$

where $[L]$ is the Cholesky decomposition of $\frac{1}{\nu^{\text{chaos}}}[M][M]^T$, which yields:

$$\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T = [A][L][L]^T[A]^T = [B][B]^T, \tag{3.32}$$

$$[B] = [A][L]. \tag{3.33}$$

The matrix:

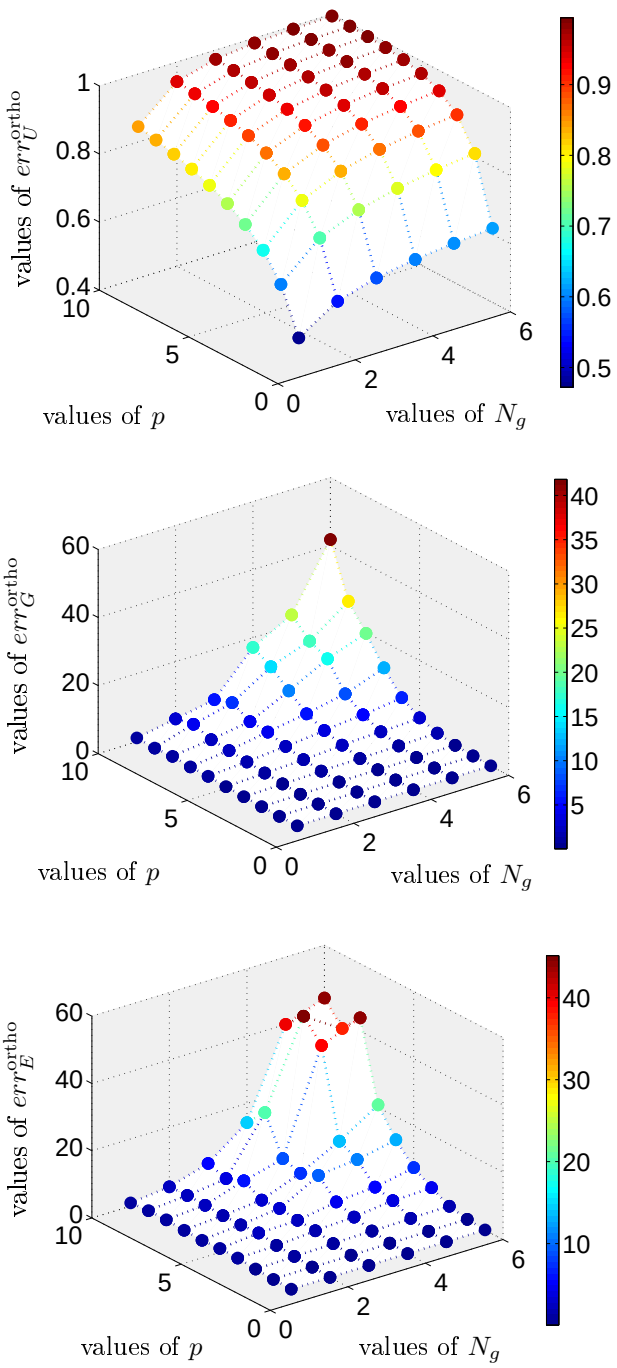



Figure 3.1: Graphs of the errors $err_U^{\text{ortho}}$, $err_G^{\text{ortho}}$ and $err_E^{\text{ortho}}$ with respect to the truncation parameters $N_g$ and $p$.

$$[\widetilde{\Psi}] = [B]^{-1}[\Psi], \tag{3.34}$$

is then introduced, such that by construction:

$$\frac{1}{\nu^{\text{chaos}}}[\widetilde{\Psi}][\widetilde{\Psi}]^T = [I_N]. \tag{3.35}$$

Using the notations of Section 2, let $[y^*]$ be a $(N_\eta \times N)$ real matrix such that the random vector $\boldsymbol{U}$ is defined as:

$$\boldsymbol{U} = [y^*]\boldsymbol{\Psi}(\boldsymbol{\xi}, p). \tag{3.36}$$

Hence, $\nu^{\text{chaos}}$ independent realizations of $\boldsymbol{U}$ can be directly deduced from the matrix $[\Psi]$ and gathered in the matrix $[U] = [y^*][\Psi]$. Defining $[z]$ such that:

$$[z] = [y^*][B], \tag{3.37}$$

therefore yields the equality:

$$[U] = [y^*][\Psi] = \left([z][B]^{-1}\right)\left([B][\widetilde{\Psi}]\right) = [z][\widetilde{\Psi}]. \tag{3.38}$$

If $[z]$ is in $\mathcal{O}_\eta$, $[z][z]^T = [\widehat{R}_\eta(\nu^{\text{exp}})]$, which implies that:

$$\left\{\begin{array}{l} \left[\widehat{R}_U(\nu^{\text{chaos}})\right] = \frac{1}{\nu^{\text{chaos}}}[U][U]^T = [z]\left(\frac{1}{\nu^{\text{chaos}}}[\widetilde{\Psi}][\widetilde{\Psi}]^T\right)[z]^T = [z][z]^T = [\widehat{R}_\eta(\nu^{\text{exp}})], \\ [R_U] = E\left(\boldsymbol{U}\boldsymbol{U}^T\right) = \lim_{\nu^{\text{chaos}}\to\infty}[\widehat{R}_U(\nu^{\text{chaos}})] = [\widehat{R}_\eta(\nu^{\text{exp}})]. \end{array}\right. \tag{3.39}$$

From Eqs. (3.5) and (3.6), it can thus be deduced that imposing $[z]$ to be an element of $\mathcal{O}_\eta$ guarantees that, for any $\nu^{\text{chaos}} \geq N$, we have $err^2(\nu^{\text{chaos}}) = err^3(\nu^{\text{exp}}, \nu^{\text{chaos}}) = 0$.

Hence, whereas the optimization problem defined by Eq. (2.24) is perturbed by autocorrelation errors, the new optimization problem:

$$\left\{\begin{array}{l} \left[y^M_{\mathcal{O}_\eta}\right] = \left[z^M_{\mathcal{O}_\eta}\right][B^{-1}], \\ \left[z^M_{\mathcal{O}_\eta}\right] = \arg\max_{[z^*]\in\mathcal{W}} \mathcal{C}\left([\eta^{\text{exp}}], [z^*], [\widetilde{\Psi}]\right), \end{array}\right. \tag{3.40}$$

is no more affected, which allows us to consider high values of the truncation parameters $N_g$ and $p$. Equation (3.38) underlines that the two former optimization problems are equivalent, as the independent realizations of $\boldsymbol{U}$ have just been rewritten. Only the research set, for the PCE coefficient matrix, has been modified, which allows the numerical bias due to the finite dimension of $[\Psi]$ to be reduced.

Finally, if $[y]$ is the coefficients matrix of the truncated PCE, $\boldsymbol{\eta}^{chaos}(N)$, of random vector $\boldsymbol{\eta}$, such that $\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p)$, a good estimation of $[y]$ in high dimension can be computed by solving the optimization problem defined by Eq. (3.40).

**3.4. Remarks on the new optimization problem.** It has to be noticed that $[\widetilde{\Psi}]$ is unique, and keeps exactly the same structure than $[\Psi]$. Indeed, let $[L^{\text{asym}}] = [A]^{-1}$ be the Cholesky decomposition matrix of the autocorrelation matrix $[R_{\boldsymbol{\varepsilon}}]$, which is defined by Eq. (3.14). Hence, from Eq. (3.11), $[\Psi] = [L^{\text{asym}}]^{-1}[M]$, which has to be compared to $[\widetilde{\Psi}] = [B]^{-1}[\Psi] = \left([L]^{-1}[A]^{-1}\right)([A][M]) = [L]^{-1}[M]$, where $[L]$ and $[L^{\text{asym}}]$ are two lower triangular matrices. Whereas $[L^{\text{asym}}]$ implies the asymptotical orthonormality, $[L]$ guarantees the numerical orthonormality. Moreover, from Eq. (3.40), the optimal PCE coefficients matrix $[y]$ is approximated as a product of two matrices:

$$[y] \approx \left[z^M_{\mathcal{O}_\eta}\right][B]^{-1}. \tag{3.41}$$

For a fixed value of $N$, $[B]$ is strongly dependent on $\nu^{\text{chaos}}$ and $[\Psi]$. From Eq. (1.12), it also verifies the asymptotical property:

$$\lim_{\nu^{\text{chaos}}\to\infty}[B] = [I_N], \tag{3.42}$$

which implies that $[z^M_{\mathcal{O}_\eta}]$ converges towards $[y]$ if sufficiently high values of $\nu^{\text{chaos}}$ is considered. Hence, the less dependent on $[\Psi]$ the matrix $[z^M_{\mathcal{O}_\eta}]$ is, the more accurate the choice of $\nu^{\text{chaos}}$ is, and the better the PCE identification is.

If another $\left(N \times \nu^{\text{chaos},*}\right)$ real matrix $[\Psi^*]$ of independent realizations is considered, the matrices $[B^*]$ and $[\widetilde{\Psi}^*] = [B^*]^{-1}[\Psi^*]$ can be computed according to Eqs. (3.33) and (3.34). As it has previously been seen, $[\Psi^*]$, $[\widetilde{\Psi}]$ and $[\widetilde{\Psi}^*]$ keep the same structure. The accuracy of $[z^M_{\mathcal{O}_\eta}]$ can thus be estimated by comparing $\mathcal{C}([\nu^{\text{exp}}], [z^M_{\mathcal{O}_\eta}], [\Psi][B]^{-1})$ and $\mathcal{C}([\nu^{\text{exp}}], [z^M_{\mathcal{O}_\eta}], [\Psi^*][B^*]^{-1})$.

In particular, $\nu^{\text{chaos},*}$ and $\nu^{\text{chaos}}$ can be different. Finally, once the coefficient matrix $[z^M_{\mathcal{O}_\eta}]$ has been computed, the higher $\nu^{\text{chaos},*}$ is, the more accurate and general the validation is.

**4. Application.** The method proposed in the two former sections is applied to two examples that both deal with the identification of the truncated PCE coefficients of a random vector $\boldsymbol{\eta}$ characterized by a multidimensional analytical distributions. The first one is built with $N_\eta = 3$, the second one with $N_\eta = 50$. According to the notations of the former sections, the set of the $\nu^{\text{exp}}$ independent realizations used in the PCE identification are gathered in $[\eta^{\text{exp}}]$. Another set of $\nu^{\text{ref}}$ independent realizations ($\nu^{\text{ref}} \gg \nu^{\text{exp}}$) is used as a reference to validate the different modelings. The idea of this section is thus to show to what extent the whole method described in the former sections allows computing convergence analysis and relevant PCE identification of the random vector $\boldsymbol{\eta}$ from a limited number of information, $[\eta^{\text{exp}}]$. Moreover, a distinction has to be made between the PDF modeling, achieved thanks to a PCE, and its estimation from PCE samples, computed thanks to non parametrical methods. In this context, let $\nu^{\text{chaos}}$ be the number of independent realizations used to carry out the PCE identification, and $\nu^{\text{chaos},*}$ the number of independent realizations of the identified PCE random vector, which will be used to draw graphical representations. For the two applications, the Gaussian measure is chosen for the PCE of $\boldsymbol{\eta}$, and $\nu^{\text{exp}} = 1000$.

It is reminded that, in this work, the term high dimension refers to the fact that the dimension $N_\eta$ of unknown random vector $\boldsymbol{\eta}$ is high.

**4.1. Application in low dimension.** The objective of this section is to apply the whole PCE method to a $N_\eta = 3$-dimension case. First, the statistical properties of the unknown random vector $\boldsymbol{\eta}$ are presented. Secondly, a convergence analysis is carried out in order to calculate the optimal truncation parameters $N_g$ and $p$ of the PCE, $\boldsymbol{\eta}^{\text{chaos}}(N)$, of $\boldsymbol{\eta}$. Then, the PCE coefficients are identified from the $\nu^{\text{exp}}$ independent realizations, $[\eta^{\text{exp}}]$, of $\boldsymbol{\eta}$. At last, the relevance of the PCE modeling is analysed.

**4.1.1. Generation of the random vector to identify.** Let $[X]$ be a $(3 \times 6)$ real-valued random matrix whose coefficients are uniformly and independently chosen between -1 and 1, such that $\boldsymbol{\eta}$ is defined according to the notations of Section 3 as:

$$\boldsymbol{\eta} = [X]\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}^{\text{exp}}, 2\right), \tag{4.1}$$

where $\boldsymbol{\xi}^{\text{exp}} = (\xi_1^{\text{exp}}, \xi_2^{\text{exp}})$ is a normalized Gaussian random vector which components are independent. The components of $\boldsymbol{\eta}$ are however strongly dependent, and the PCE truncation parameters to be found back by the convergence analysis are $p^{\text{exp}} = 2$ and $N_g^{\text{exp}} = 2$.

Let $\{\boldsymbol{\xi}^{\text{exp}}(\theta_1), \cdots, \boldsymbol{\xi}^{\text{exp}}(\theta_{\nu^{\text{exp}}})\}$ and $\{\boldsymbol{\xi}^{\text{exp}}(\theta_1), \cdots, \boldsymbol{\xi}^{\text{exp}}(\theta_{\nu^{\text{ref}}})\}$ be $\nu^{\text{exp}}$ and $\nu^{\text{ref}}$ independent realizations of the random vector $\boldsymbol{\xi}^{\text{exp}}$, such that the matrices of independent realizations $[\eta^{\text{exp}}]$ and $\left[\eta^{\text{ref}}\right]$ are given by:

$$[\eta^{\text{exp}}] = [X]\left[\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}^{\text{exp}}(\theta_1), 2\right) \quad \cdots \quad \boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}^{\text{exp}}(\theta_{\nu^{\text{exp}}}), 2\right)\right], \tag{4.2}$$

$$\left[\eta^{\text{ref}}\right] = [X]\left[\boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}^{\text{exp}}(\theta_1), 2\right) \quad \cdots \quad \boldsymbol{\mathcal{E}}\left(\boldsymbol{\xi}^{\text{exp}}(\theta_{\nu^{\text{ref}}}), 2\right)\right]. \tag{4.3}$$

Let $\left\{\widehat{p}_\eta^{\text{ref},k},\ 1 \leq k \leq 3\right\}$ be the Kernel smoothing estimations of the marginal PDFs of each component of $\boldsymbol{\eta}$, which are computed thanks to the $\nu^{\text{ref}}$ independent realizations of $\boldsymbol{\eta}$ gathered in $[\eta^{\text{ref}}]$. In this example, $\nu^{\text{ref}} = 2 \times 10^6 \gg \nu^{\text{exp}} = 1000$. It is reminded that the PCE identification of $\boldsymbol{\eta}$ is only achieved thanks to the matrix of independent realizations $[\eta^{\text{exp}}]$, which is considered as the only available information. The PDFs $\left\{\widehat{p}_\eta^{\text{ref},k},\ 1 \leq k \leq 3\right\}$ are moreover supposed to build the marginal PDFs of the reference $\boldsymbol{\eta}$.

**4.1.2. Identification of the PCE truncation parameters.** Using the notations of Section 2.4, the boundary intervals $\text{BI}_1$, $\text{BI}_2$ and $\text{BI}_3$ for which the convergence analysis is achieved, are chosen such that:

$$\forall\ 1 \leq k \leq 3,\ \text{BI}_k = \left\{x \in \mathbb{R} \mid \widehat{p}_\eta^{\text{ref},k}(x) \geq \frac{1}{\nu^{\text{exp}}}\right\}. \tag{4.4}$$

Figure 4.1 displays the reference marginal PDFs of $\boldsymbol{\eta}$, as well as the marginal PDFs $\left\{\widehat{p}_\eta^{\text{exp},k},\ 1 \leq k \leq 3\right\}$ estimated from the $\nu^{\text{exp}}$ independent realizations only. The $1/\nu^{\text{exp}}$ tolerance is also plotted so that the boundary intervals can therefore be deduced from these graphs.

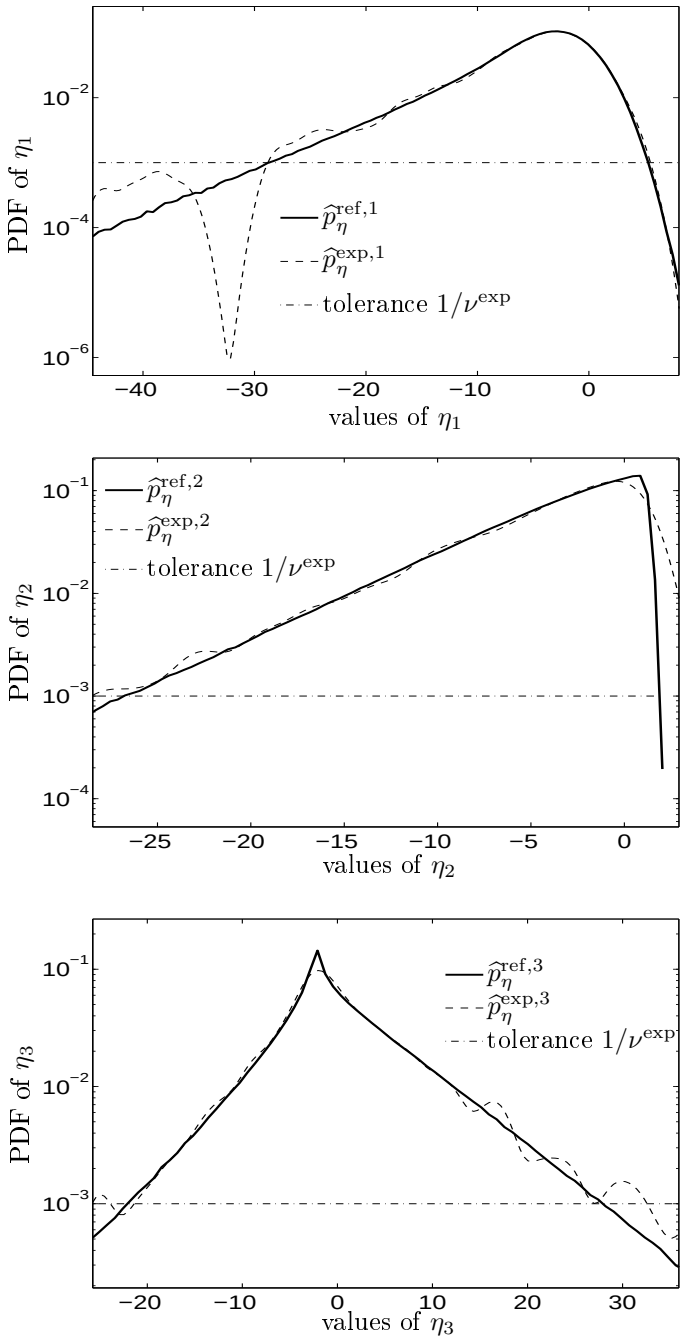



Figure 4.1: Graphs of the marginal PDFs of $\boldsymbol{\eta}$.

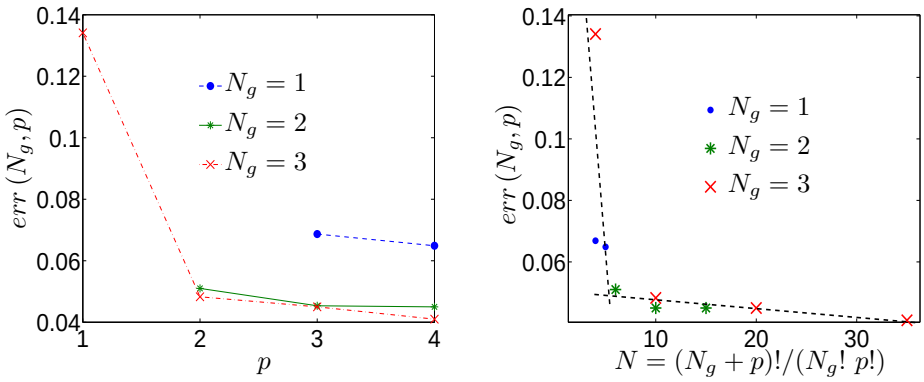



Figure 4.2: Convergence analysis of the PCE of $\boldsymbol{\eta}$.

Figure 4.2 shows the values of $err\,(N_g, p)$, for nine pairs $(N_g, p)$ in $\mathcal{Q}(3)$. On these graphs, the gradient break of $N \mapsto err\,(N)$ is observed at $N = 6$, which allows us to find back the initial solution $p^{\text{exp}} = 2$ and $N_g^{\text{exp}} = 2$. For this small dimension case, the optimal truncation parameters $p$ and $N_g$ given by the convergence analysis are equal to the parameters of the analytical reference PCE.

**4.1.3. PCE Identification.** The former convergence analysis leads us to the following PCE of $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} \approx \boldsymbol{\eta}^{\text{chaos}}(6) = \sum_{j=1}^{6} \boldsymbol{y}^j \Psi_j\,(\xi_1, \xi_2) = [y]\boldsymbol{\Psi}\,(\xi_1, \xi_2)\,, \tag{4.5}$$

where $\xi_1$ and $\xi_2$ are two independent normalized Gaussian random variables. We are now going to compare $[y^{\text{class}}]$ and $[y^{\text{new}}]$, where $[y^{\text{class}}]$ stems from the classical problem defined by Eq. (2.24), whereas $[y^{\text{new}}]$ comes from the maximization of the new formulation defined by Eq. (3.40). In this application, $\nu^{\text{chaos}} = 1000$, and the two PCE identifications have been computed thanks to the same numerical cost $M = 10^4$, which means that $M = 10^4$ independent random trials of $[y^{\text{class}}]$ and $[y^{\text{new}}]$ have been used to maximize the log-likelihood. The value of $M$ has been chosen sufficiently high for the PCE error function $err(N_g, p)$ to be independent of $M$. Hence, for a new matrix of independent realizations, $[\Psi^*]$, of size $\left(6 \times \nu^{\text{chaos},*}\right)$, independent realizations $[\eta^{\text{class}}(6)]$ and $[\eta^{\text{new}}(6)]$ of $\boldsymbol{\eta}^{\text{chaos}}(6)$ are deduced, with respect to the two optimization options:

$$[\eta^{\text{class}}(6)] = [y^{\text{class}}][\Psi^*], \tag{4.6}$$

$$[\eta^{\text{new}}(6)] = [y^{\text{new}}][\Psi^*]. \tag{4.7}$$

Let

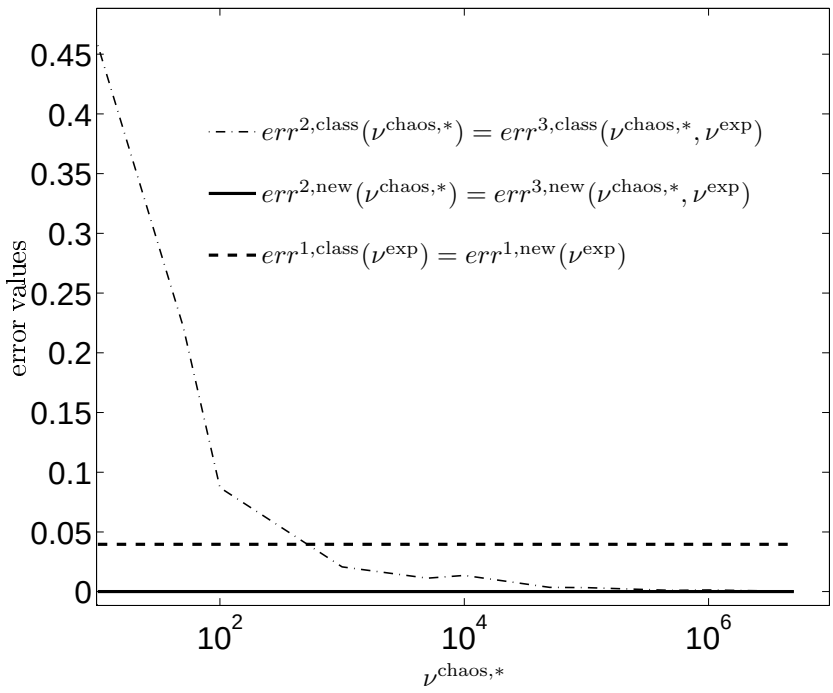



Figure 4.3: Convergence of the autocorrelation error functions with respect to $\nu^{\text{chaos},*}$.

$$[R_\eta^{\text{exp}}] = \frac{1}{\nu^{\text{exp}}}[\eta^{\text{exp}}][\eta^{\text{exp}}]^T, \tag{4.8}$$

$$[R_\eta^{\text{ref}}] = \frac{1}{\nu^{\text{ref}}}[\eta^{\text{ref}}][\eta^{\text{ref}}]^T, \tag{4.9}$$

$$[R_\eta^{\text{class}}] = \frac{1}{\nu^{\text{chaos},*}}[\eta^{\text{class}}(6)][\eta^{\text{class}}(6)]^T, \tag{4.10}$$

$$[R_\eta^{\text{new}}] = \frac{1}{\nu^{\text{chaos},*}}[\eta^{\text{new}}(6)][\eta^{\text{new}}(6)]^T \tag{4.11}$$

be four estimations of the autocorrelation matrix $[R_\eta]$ of $\boldsymbol{\eta}$. It is supposed that $[R_\eta^{\text{ref}}]$ is the best approximation of $[R_\eta]$ and will be considered as the reference. According to the Eqs. (3.4), (3.5) and (3.6), the autocorrelation errors $err^{1,\text{class}}$, $err^{2,\text{class}}$, $err^{3,\text{class}}$ and $err^{1,\text{new}}$, $err^{2,\text{new}}$, $err^{3,\text{new}}$ are then computed in each case. In figure 4.3, it can thus be verified that:

$$\forall\ \nu^{\text{chaos},*} \geq 6,\ err^{2,\text{new}}(\nu^{\text{chaos},*}) = err^{3,\text{new}}(\nu^{\text{chaos},*}, \nu^{\text{exp}}) = 0, \tag{4.12}$$

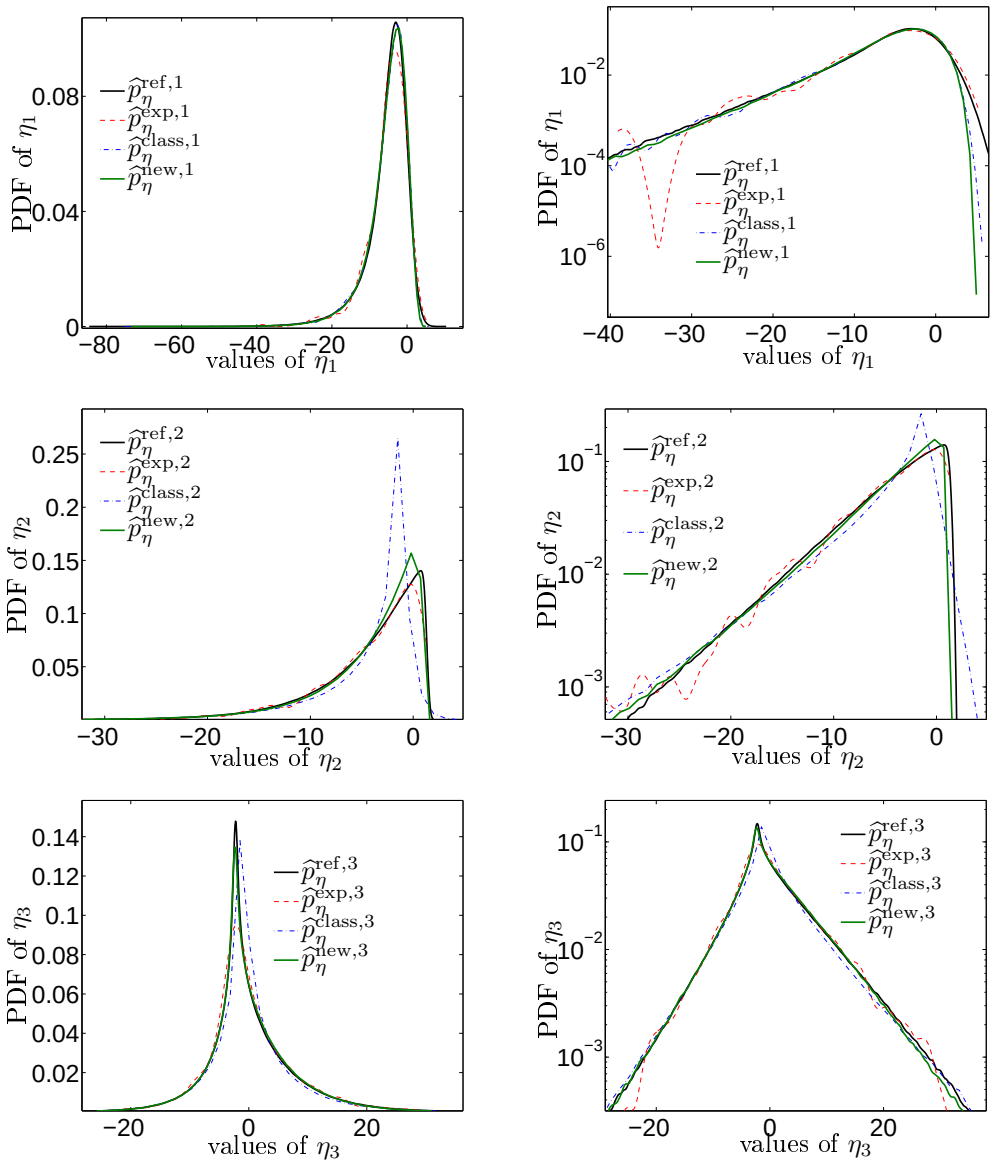



Figure 4.4: Comparison of the marginal PDFs of $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^{\mathrm{chaos}}(6)$.

$$\lim_{\nu^{\mathrm{chaos},*}\to+\infty} err^{2,\mathrm{class}}(\nu^{\mathrm{chaos},*}) = err^{3,\mathrm{class}}(\nu^{\mathrm{chaos},*},\nu^{\mathrm{exp}}) = 0. \tag{4.13}$$

In particular, for the value $\nu^{\mathrm{chaos},*} = \nu^{\mathrm{chaos}} = 1000$, it can be noticed that the values of $err^{2,\mathrm{class}}(\nu^{\mathrm{chaos},*})$ and $err^{3,\mathrm{class}}(\nu^{\mathrm{chaos},*},\nu^{\mathrm{exp}})$ are significant when compared to $err^{1,\mathrm{class}}(\nu^{\mathrm{exp}})$, which introduces an additive bias in the identification.

Figure 4.4 shows a comparison between the marginal PDFs $\widehat{p}_\eta^{\mathrm{ref},k}$, $\widehat{p}_\eta^{\mathrm{exp},k}$, $\widehat{p}_\eta^{\mathrm{class},k}$ and $\widehat{p}_\eta^{\mathrm{new},k}$, for $1 \leq k \leq 3$. These PDFs are estimated using Kernel smoothing on the independent realizations gathered in the matrices $[\eta^{\mathrm{ref}}]$, $[\eta^{\mathrm{exp}}]$, $[\eta^{\mathrm{class}}(6)]$ and $[\eta^{\mathrm{new}}(6)]$, respectively, with $\nu^{\mathrm{chaos},*} = 10^6 \gg \nu^{\mathrm{chaos}} = 1000$. First, from only $\nu^{\mathrm{exp}} = 1000$ independent realizations of $\boldsymbol{\eta}$, it can be seen that the marginal PDFs are well described by the PCE random vectors $\boldsymbol{\eta}^{\mathrm{new}}(6)$ and $\boldsymbol{\eta}^{\mathrm{class}}(6)$. In particular, the PDFs tails are very well characterized. The PCE method is therefore an extremely efficient tool to build arbitrary multidimensional PDFs. Secondly, it can be noticed that, for a same computational cost $M$, the new PCE identification formulation leads us to better results than the classical one. Finally, to still improve these PCE, more trials in $\mathcal{O}_\eta$ would be necessary to better characterize $[y^{\mathrm{class}}]$ and $[y^{\mathrm{new}}]$. In order to obtain a PCE that corresponds still more precisely to the reference random vector $\boldsymbol{\eta}$, an increase of $\nu^{\mathrm{exp}}$, that is to say, more information about $\boldsymbol{\eta}$, would have been required.

**4.1.4. Relevance of the PCE compared to Kernel Mixture and PASM.** From adequacy tests, likelihood estimations and graphical representations, the idea of this section is to show the assets of the new PCE formulation when dealing with the identification of multidimensional distributions from a limited knowledge on the random vector of interest $\boldsymbol{\eta}$ compared to Kernel Mixture (KM) and Prior Algebraic Stochastic Modeling (PASM). In this prospect, two PDFs $\widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x})$ and $\widehat{p}_{\boldsymbol{\eta}}^{\mathrm{PASM}}(\boldsymbol{x}, \boldsymbol{w})$ are built using a KM approach and a PASM method. The input data of these modelings are still the matrix of independent realizations $[\eta^{\mathrm{exp}}] = \left[\boldsymbol{\eta}^{(1)} \ \cdots \ \boldsymbol{\eta}^{(\nu^{\mathrm{exp}})}\right]$. Once the KM, the PASM and the two PCE projection matrices, $[y^{\mathrm{class}}]$ and $[y^{\mathrm{new}}]$, are constructed, $Q$ independent realizations are computed from the four distributions, from which comparisons to the reference solution are achieved. For this application, $Q = 10^6$.

**Construction of independent realisations.**

- **Kernel Mixture**.

Considering an independent Gaussian multidimensional Kernel, a non parametrical PDF $\widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x})$ is postulated as a sum of $\nu^{\mathrm{exp}}$ Gaussian PDFs $\{p_i,\ 1 \leq i \leq \nu^{\mathrm{exp}}\}$ to model $p_{\boldsymbol{\eta}}(\boldsymbol{x})$:

$$\widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x}) = \sum_{i=1}^{\nu^{\mathrm{exp}}} \frac{1}{\nu^{\mathrm{exp}}} p_i(\boldsymbol{x}), \tag{4.14}$$

$$p_i(\boldsymbol{x}) = \prod_{k=1}^{N_\eta} \frac{1}{\sqrt{2\pi} h_k} \exp\left(-\frac{1}{2}\left(\frac{x_k - \eta_k^i}{h_k}\right)^2\right), \tag{4.15}$$

$$\boldsymbol{h} = \widehat{\boldsymbol{\sigma}} \left(\frac{4}{(2 + N_\eta)\, \nu^{\mathrm{exp}}}\right)^{1/(N_\eta+4)}, \tag{4.16}$$

where $\boldsymbol{x} \mapsto p_i(\boldsymbol{x})$ is the $N_\eta$-dimension multivariate Gaussian PDF, with mean value $\boldsymbol{\eta}^{(i)}$ and covariance matrix $\begin{pmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & h_{N_\eta} \end{pmatrix}$, $\boldsymbol{h}$ is the multidimensionnal op-

timal Silverman bandwidth, and $\widehat{\sigma}_k$ is the empirical estimation of the standard deviation of each component $\eta_k$ of $\boldsymbol{\eta}$. Let $\boldsymbol{\eta}^{\text{ker}}$ be the Kernel Mixture characterized by the PDF $\boldsymbol{x} \mapsto \widehat{p}_{\boldsymbol{\eta}}(\boldsymbol{x})$. The $Q$ independent realizations $\left\{\boldsymbol{\eta}^{ker,1}, \cdots, \boldsymbol{\eta}^{ker,Q}\right\}$ of $\boldsymbol{\eta}^{\text{ker}}$ are then computed and gathered in the matrix $[\eta^{\text{ker}}]$.

- **Prior Algebraic Stochastic Modeling**.

From the $\nu^{\text{exp}}$ independent realizations of $\boldsymbol{\eta}$, the $N_\eta$ marginal cumulative distributions $F_{\eta_k}$ of $\eta_k$, with $1 \leq k \leq N_\eta$, are estimated using a non parametric statistical method. In addition, a Gaussian copula $C^{\text{gauss}}_{\text{rank}}$ (see [6] for more details about the copula) based on the rank correlation is chosen (this type of copula has been chosen as it is the most commonly used in the PASM approaches):

$$C^{\text{gauss}}_{\text{rank}}(x_1, \cdots, x_{N_\eta}) = \phi^{N_\eta}_{\text{rank}}\left(\phi^{-1}(x_1), \cdots, \phi^{-1}(x_{N_\eta})\right), \tag{4.17}$$

$$\phi^{N_\eta}_{\text{rank}}(\boldsymbol{u}) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_{N_\eta}} \frac{1}{(2\pi)^{N_\eta/2}\sqrt{\det\left([R^{\text{rank}}]\right)}} \exp\left(-\frac{1}{2}\boldsymbol{u}^T[R^{\text{rank}}]\boldsymbol{u}\right) du_1 \cdots du_{N_\eta}, \tag{4.18}$$

$$\phi(v) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} \exp\left(-\frac{v^2}{2}\right) dv, \tag{4.19}$$

$$[R^{\text{rank}}]_{ij} = 2\sin\left(\frac{\pi}{6}\rho^S_{ij}\right), \tag{4.20}$$

where $\rho^S_{ij}$ is the Spearman correlation coefficient between $\eta_i$ and $\eta_j$. Let $\boldsymbol{\eta}^{\text{cop}}$ be the random vector characterized by the copula $C^{\text{gauss}}_{\text{rank}}$ and the marginal cumulative distributions $\{F_{\eta_k},\ 1 \leq k \leq N_\eta\}$. $Q$ independent realizations of $\boldsymbol{\eta}^{\text{cop}}$ are thus gathered in the matrix $[\eta^{\text{cop}}]$.

- **Polynomial chaos expansion**.

Finally, using the matrices $[y^{\text{class}}]$ and $[y^{\text{new}}]$ of Section 4.1.3, and a new $(6 \times Q)$ real matrix $[\Psi^Q]$ of realizations, $Q$ independent realizations of $\boldsymbol{\eta}^{\text{class}}(6)$ and $\boldsymbol{\eta}^{\text{new}}(6)$ are gathered in the matrix $[\eta^{\text{class}}] = [y^{\text{class}}][\Psi^Q]$ and $[\eta^{\text{new}}] = [y^{\text{new}}][\Psi^Q]$.

**Relevance of the PCE modeling when identifying multidimensional PDFs from a limited amount of independent realizations.** Using the results of Parametrical Statistics, this section assesses the relevance of the four methods to construct multidimensional PDFs. Three kinds of analysis are achieved: adequacy tests, 2D graphical representations, and multidimensional likelihood computations.

- **Adequacy tests.**

From the matrices of independent realizations $[\eta^{\text{ker}}]$, $[\eta^{\text{cop}}]$, $[\eta^{\text{class}}]$ and $[\eta^{\text{new}}]$, the estimations $\left\{\widehat{F}^{\text{ker}}_k,\ 1 \leq k \leq N_\eta\right\}$, $\left\{\widehat{F}^{\text{cop}}_k,\ 1 \leq k \leq N_\eta\right\}$, $\left\{\widehat{F}^{\text{class}}_k,\ 1 \leq k \leq N_\eta\right\}$ and $\left\{\widehat{F}^{\text{new}}_k,\ 1 \leq k \leq N_\eta\right\}$ of the cumulative distribution functions (CDF) of each components of $\boldsymbol{\eta}^{\text{ker}}$, $\boldsymbol{\eta}^{\text{cop}}$, $\boldsymbol{\eta}^{\text{class}}(6)$ and $\boldsymbol{\eta}^{\text{new}}(6)$ are respectively assessed. Let $\widetilde{\boldsymbol{\eta}}^{(1)}, \cdots, \widetilde{\boldsymbol{\eta}}^{(N_\eta)}$ be the $1 \times \nu^{\text{exp}}$-dimension linear forms corresponding to the lines of $[\eta^{\text{exp}}]$. For

$1 \leq k \leq N_\eta$, $\widetilde{\boldsymbol{\eta}}^{(k)}$ gathers therefore the $\nu^{\text{exp}}$ independent realizations of the component $\eta_k$ of $\boldsymbol{\eta}$, which have been used to compute the statistical modelings. For $1 \leq k \leq N_\eta$, the Kolmogorov-Smirnov adequacy tests are then performed. For each component $\eta_k$ of $\boldsymbol{\eta}$, the null distribution of the Kolmogorov-Smirnov statistics is computed under the null hypothesis that the $\nu^{\text{exp}}$ independent realizations of $\widetilde{\boldsymbol{\eta}}^{(k)}$ are drawn from the distribution of the chosen stochastic model. Table 4.1 gives the $\beta$-value for each stochastic model, which is defined as the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. Without surprise, this table allows us to verify that the modeling based on the Gaussian copula and the empirical PDFs of each components of $\boldsymbol{\eta}$ gives the best results. Moreover, with an error level of 5%, only the tests for the copula model and the PCE identification based on the new formulation are positive. The classical PCE and the Kernel mixture modelings are indeed less relevant to characterize the marginal PDFs of $\boldsymbol{\eta}$.

| CDF | $\widehat{F}_1^{\text{class}}$ | $\widehat{F}_1^{\text{new}}$ | $\widehat{F}_1^{\text{ker}}$ | $\widehat{F}_1^{\text{cop}}$ |
|---|---|---|---|---|
| $\beta$-value | 0.3779 | 0.6331 | 0.2142 | 0.9996 |
| CDF | $\widehat{F}_2^{\text{class}}$ | $\widehat{F}_2^{\text{new}}$ | $\widehat{F}_2^{\text{ker}}$ | $\widehat{F}_2^{\text{cop}}$ |
| $\beta$-value | 0.0000 | 0.0967 | 0.0000 | 0.4573 |
| CDF | $\widehat{F}_3^{\text{class}}$ | $\widehat{F}_3^{\text{new}}$ | $\widehat{F}_3^{\text{ker}}$ | $\widehat{F}_3^{\text{cop}}$ |
| $\beta$-value | 0.0000 | 0.8692 | 0.0411 | 0.9849 |

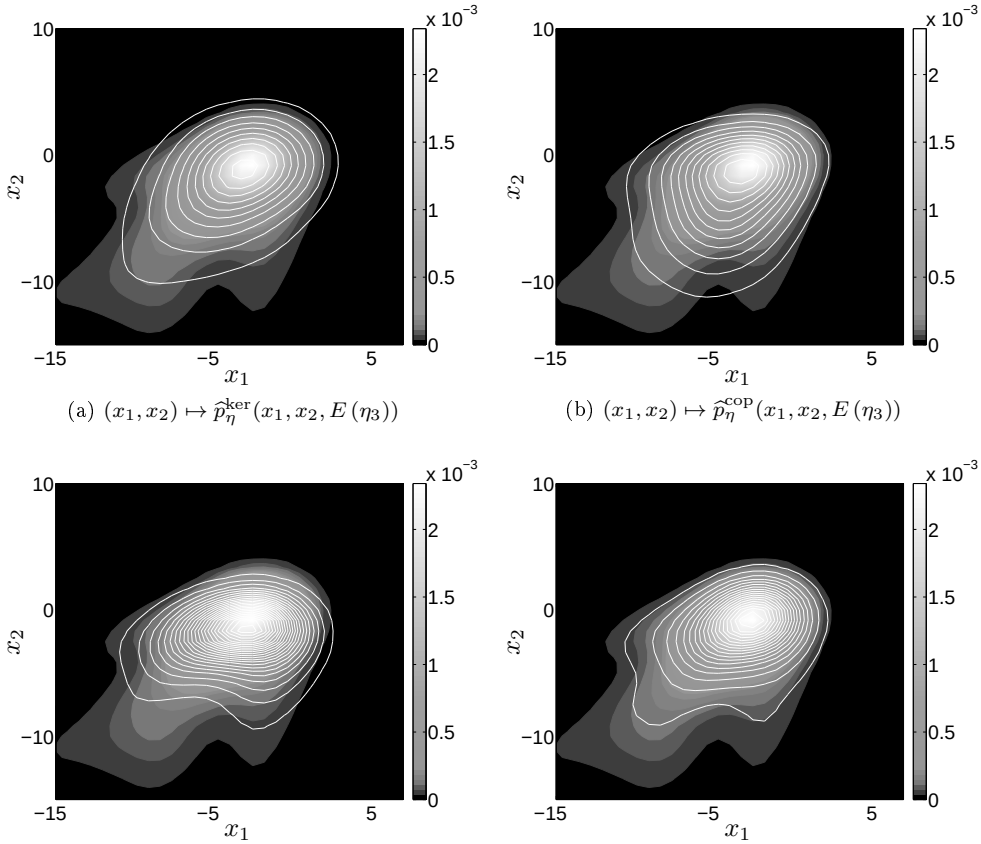
Table 4.1: Computation of the $\beta$-values corresponding to the different stochastic models.

- **Two-dimensions graphical analysis.**

From $[\eta^{\text{ref}}]$, $[\eta^{\text{ker}}]$, $[\eta^{\text{cop}}]$, $[\eta^{\text{class}}]$ and $[\eta^{\text{new}}]$, the estimations $\boldsymbol{x} \mapsto \widehat{p}_\eta^{\text{ref}}(\boldsymbol{x})$, $\boldsymbol{x} \mapsto \widehat{p}_\eta^{\text{ker}}(\boldsymbol{x})$, $\boldsymbol{x} \mapsto \widehat{p}_\eta^{\text{cop}}(\boldsymbol{x})$, $\boldsymbol{x} \mapsto \widehat{p}_\eta^{\text{class}}(\boldsymbol{x})$ and $\boldsymbol{x} \mapsto \widehat{p}_\eta^{\text{new}}(\boldsymbol{x})$ of the multidimensional PDF of $\boldsymbol{\eta}$, $\boldsymbol{\eta}^{\text{ker}}$, $\boldsymbol{\eta}^{\text{cop}}$, $\boldsymbol{\eta}^{\text{class}}(6)$, $\boldsymbol{\eta}^{\text{new}}(6)$ are respectively computed using the non parametric statistical estimation defined by Eq. (2.18). Projections of these functions are presented in Figures 4.5, 4.6 and 4.7. In each figure, the surface plot characterizes the reference PDF (based on the $\nu^{\text{ref}} = 2 \times 10^6$ independent realizations), and the contour plot refers to isovalues of the projected PDF of interest. It can therefore be seen that the new formulation of the PCE gives very good results in identifying multidimensional PDFs. In addition, in this example, the Kernel mixture model is more adapted than the copula based model to characterize the multidimensional PDFs.

- **Likelihood estimations.**

From Eq. (2.10), the multidimensional log-likelihood functions $\mathcal{L}_{\boldsymbol{\eta}^{\text{ker}}}([\eta^{\text{exp}}])$, $\mathcal{L}_{\boldsymbol{\eta}^{\text{cop}}}([\eta^{\text{exp}}])$, $\mathcal{L}_{\boldsymbol{\eta}^{\textbf{class}}}([\eta^{\text{exp}}])$ and $\mathcal{L}_{\boldsymbol{\eta}^{\textbf{new}}}([\eta^{\text{exp}}])$ are estimated from the realizations matrices $[\eta^{\text{exp}}]$, $[\eta^{\text{ker}}]$, $[\eta^{\text{cop}}]$, $[\eta^{\text{class}}]$ and $[\eta^{\text{new}}]$, in order to evaluate the multidimensional relevance of the different stochastic models. In the same manner, $[\eta^{\text{ref}}]_{1000}$ is defined as the 1000 first columns of $[\eta^{\text{ref}}]$, and the log-likelihood functions $\mathcal{L}_{\boldsymbol{\eta}^{\text{ker}}}\left([\eta^{\text{ref}}]_{1000}\right)$, $\mathcal{L}_{\boldsymbol{\eta}^{\text{cop}}}\left([\eta^{\text{ref}}]_{1000}\right)$, $\mathcal{L}_{\boldsymbol{\eta}^{\textbf{class}}}\left([\eta^{\text{ref}}]_{1000}\right)$ and $\mathcal{L}_{\boldsymbol{\eta}^{\textbf{new}}}\left([\eta^{\text{ref}}]_{1000}\right)$ are computed. These values are gathered in Table 4.2. It can thus be verified that the new formulation of the PCE identification gives the best results when considering the maximization of the log-likelihood.




(a) $(x_1, x_2) \mapsto \widehat{p}_{\eta}^{\mathrm{ker}}(x_1, x_2, E(\eta_3))$

(b) $(x_1, x_2) \mapsto \widehat{p}_{\eta}^{\mathrm{cop}}(x_1, x_2, E(\eta_3))$

(c) $(x_1, x_2) \mapsto \widehat{p}_{\eta}^{\mathrm{class}}(x_1, x_2, E(\eta_3))$

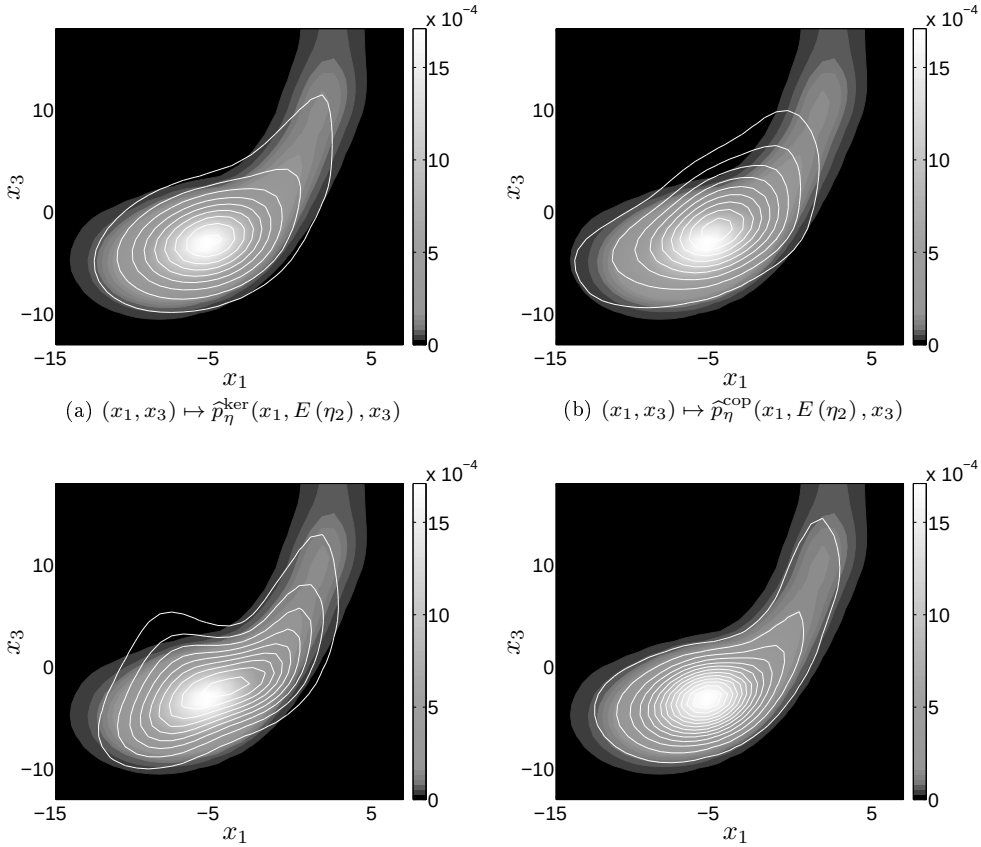
(d) $(x_1, x_2) \mapsto \widehat{p}_{\eta}^{\mathrm{new}}(x_1, x_2, E(\eta_3))$

Figure 4.5: Comparison of 2D contours plots in the plane $[x_3 = E(\eta_3)]$.

| $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{ker}}}([\eta^{\mathrm{exp}}])$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{cop}}}([\eta^{\mathrm{exp}}])$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{class}}}([\eta^{\mathrm{exp}}])$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{new}}}([\eta^{\mathrm{exp}}])$ |
|---|---|---|---|
| $-8.0712.10^3$ | $-8.7530.10^3$ | $-8.1844.10^3$ | $-7.8624.10^3$ |
| $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{ker}}}\left([\eta^{\mathrm{ref}}]_{1000}\right)$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{cop}}}\left([\eta^{\mathrm{ref}}]_{1000}\right)$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{class}}}\left([\eta^{\mathrm{ref}}]_{1000}\right)$ | $\mathcal{L}_{\boldsymbol{\eta}^{\mathrm{new}}}\left([\eta^{\mathrm{ref}}]_{1000}\right)$ |
| $-8.1933.10^3$ | $-8.5535.10^3$ | $-8.1797.10^3$ | $-7.8457.10^3$ |

Table 4.2: Computation of the multidimensional log-likelihood corresponding to the different stochastic models.

As a conclusion for this example, in low dimension, it can be seen that the new formulation of the PCE identification is very relevant when trying to identify multidimensional distributions from a limited number of measurements. Indeed, it allows us to build multidimensional distributions that are still relevant for experimental data that have not been used in the identification process.




(a) $(x_1, x_3) \mapsto \widehat{p}_\eta^{\text{ker}}(x_1, E(\eta_2), x_3)$

(b) $(x_1, x_3) \mapsto \widehat{p}_\eta^{\text{cop}}(x_1, E(\eta_2), x_3)$

(c) $(x_1, x_3) \mapsto \widehat{p}_\eta^{\text{class}}(x_1, E(\eta_2), x_3)$

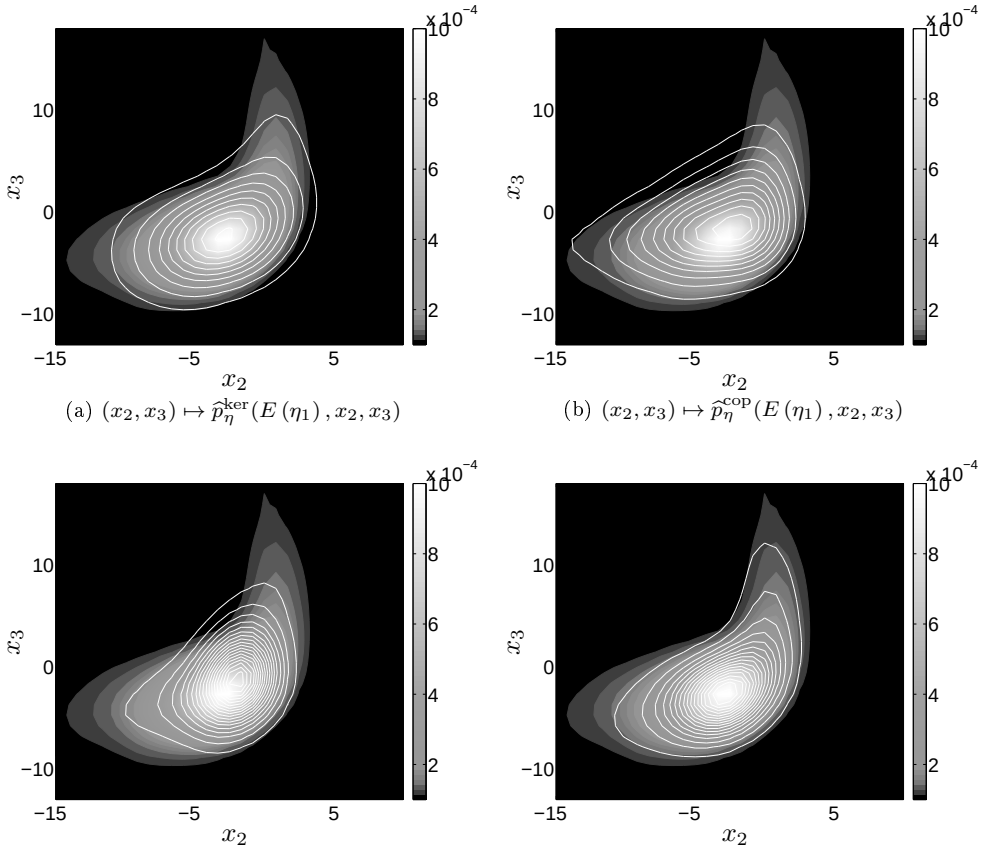
(d) $(x_1, x_3) \mapsto \widehat{p}_\eta^{\text{new}}(x_1, E(\eta_2), x_3)$

Figure 4.6: Comparison of 2D contours plots in the plane $[x_2 = E(\eta_2)]$.

**4.2. Application in high dimension.** The idea of this second application is to underline the ability of the new PCE formulation to carry out convergence analysis in high dimension. Indeed, as it has been shown in Section 3, for a given value of $\nu^{\text{chaos}}$, when the size $N$ of the polynomial basis increases, and more specially when the maximum degree $p$ of the polynomial basis becomes high, the difference $\frac{1}{\nu^{\text{chaos}}}[\Psi][\Psi]^T - [I_N]$ introduces a significant numerical bias which perturbs the classical PCE identification. In opposite, the new PCE formulation, which avoids computational autocorrelation errors, allows the numerical algorithms to be much more stable and to give more relevant results.

**4.2.1. Generation of a high dimension random vector.** Using the same notations than in Sections 2.1 and 4.1.1, let $[X^{\text{HD}}]$ be a $(N_\eta \times N)$ real matrix whose entries are randomly generated, such that random vector $\boldsymbol{\eta}$ is given by:

$$\boldsymbol{\eta} = [X^{\text{HD}}]\boldsymbol{\Psi}\left(\boldsymbol{\xi}^{\text{exp}}, p^{\text{exp}}\right), \tag{4.21}$$




(a) $(x_2, x_3) \mapsto \widehat{p}_\eta^{\,\mathrm{ker}}(E(\eta_1), x_2, x_3)$

(b) $(x_2, x_3) \mapsto \widehat{p}_\eta^{\,\mathrm{cop}}(E(\eta_1), x_2, x_3)$

(c) $(x_2, x_3) \mapsto \widehat{p}_\eta^{\,\mathrm{class}}(E(\eta_1), x_2, x_3)$

(d) $(x_2, x_3) \mapsto \widehat{p}_\eta^{\,\mathrm{new}}(E(\eta_1), x_2, x_3)$

Figure 4.7: Comparison of 2D contours plots in the plane $[x_1 = E(\eta_1)]$.

$$\boldsymbol{\xi}^{\mathrm{exp}} = \left(\xi_1^{\mathrm{exp}}, \xi_2^{\mathrm{exp}}, \cdots, \xi_{N_g}^{\mathrm{exp}}\right), \tag{4.22}$$

where $\{\xi_k^{\mathrm{exp}},\ 1 \leq k \leq N_\eta\}$ is a set of $N_\eta$ independent normalized Gaussian random variables. As in Section 4.1, we define a $(N_\eta \times \nu^{\mathrm{exp}})$ real matrix $[\eta^{\mathrm{exp}}]$, which gathers $\nu^{\mathrm{exp}}$ independent realizations of $\boldsymbol{\eta}$:

$$[\eta^{\mathrm{exp}}] = [X^{\mathrm{HD}}][\Psi^{\mathrm{exp}}], \tag{4.23}$$

$$[\Psi^{\mathrm{exp}}] = [\boldsymbol{\Psi}\left(\boldsymbol{\xi}^{\mathrm{exp}}(\theta_1), p^{\mathrm{exp}}\right) \quad \cdots \quad \boldsymbol{\Psi}\left(\boldsymbol{\xi}^{\mathrm{exp}}(\theta_{\nu^{\mathrm{exp}}}), p^{\mathrm{exp}}\right)]\,. \tag{4.24}$$

The components of the random vector $\boldsymbol{\eta}$ are again strongly dependent. As a numerical illustration, it is supposed that $\nu^{\mathrm{exp}} = 1000$, $p^{\mathrm{exp}} = 9$, $N_g^{\mathrm{exp}} = 3$, $N = (9+3)!/(9!\,3!) = 220$, $N_\eta = 50$. A high value of $p^{\mathrm{exp}}$ is deliberately chosen, in order to emphasize the

difficulties of the classical PCE formulation to carry out convergence analysis in high dimension. Nevertheless, this high value of $p$ implies an ill-conditionning of $[\Psi^{\text{exp}}]$, such that $\boldsymbol{\eta}$ can have very high values.

**4.2.2. Identification of the PCE truncation parameters.** According to Eq. (2.3), the truncated PCE, $\boldsymbol{\eta}^{\text{chaos}}(N)$, of $\boldsymbol{\eta}$ is given by:

$$\boldsymbol{\eta}^{\text{chaos}}(N) = [y]\boldsymbol{\Psi}(\boldsymbol{\xi}, p). \tag{4.25}$$

Eq. (2.29) implies that the number $N_y$ of elements of $[y]$ has to be higher than $N_\eta\,(N_\eta + 1)$. When $N_\eta$ is large, this leads us to the identification of thousands of coefficients. However, as it has been said in Section 2.4, the higher $N_y$ is, the less precise is the PCE identification, for a given computational cost $M$. In addition, Section 3 has emphasized the ill-conditionning of matrix $[\Psi]$ for high values of $p$. This motivates the definition of a new set $\widetilde{\mathcal{Q}}(p^{\max}, N^{\max})$, such that the optimal values $p^{\text{opt}}$ and $N_g^{\text{opt}}$ are given by:

$$\widetilde{\mathcal{Q}}(p^{\max}, N^{\max}) = \{(p, N_g)\,,\ N_g \leq N_\eta,\ p \leq p^{\max},\ (N_g + p)!/\,(N_g!\ p!) \leq N^{\max}\}\,, \tag{4.26}$$

$$(p^{\text{opt}}, N_g^{\text{opt}}) = \arg \min_{(p,N_g)\in\widetilde{\mathcal{Q}}(p^{\max},N^{\max})} err(N_g, p), \tag{4.27}$$

where $err(N_g, p)$ is computed from $M$ independent matrices in $\mathcal{O}_\eta$. For a fixed value $\nu^{\text{chaos}} = 1000$, the detrimental influence of the autocorrelation errors $err^2$ and $err^3$ of Eqs. (3.5) and (3.6) can then be noticed in Figure 4.8, when high values of $N$ (and more specially high values of $p$) are considered. The error functions $err^{\text{class}}(N_g, p)$ and $err^{\text{new}}(N_g, p)$ correspond, respectively, to the classical formulation and the new formulation of the PCE identification. It can be seen that for $p \geq 8$, the ratio $err^{\text{class}}(N_g, p)/err^{\text{new}}(N_g, p)$ becomes greater than five, whereas the two methodologies are globally similar for low values of $p$. Hence, the accuracy of the classical method seems to be limited to low values of $p$ and is therefore less relevant for convergence analysis which handle high polynomial orders. At last, the five lowest values of the numerical assessments of $err^{\text{new}}(N_g, p)$ are gathered in Table 4.3. It can be seen that the new formulation allows finding back the couple $(p^{\text{exp}}, N_g^{\text{exp}})$ as the minimum of the error function. Nevertheless, keeping in mind that the lowest N is, the easiest the identification is, this result also shows that using the couple $(p, N_g) = (11, 2)$ could be interesting.

| couples $(p, N_g)$ | (11,2) | (9,3) | (7,4) | (6,5) | (2,27) |
|---|---|---|---|---|---|
| values of N | 78 | 220 | 330 | 462 | 406 |
| $err^{\text{new}}(N_g, p)$ | 0.06104 | 0.06005 | 0.06228 | 0.06301 | 0.06521 |

Table 4.3: Lowest values of $err^{\text{new}}(N_g, p)$ with respect to $(p, N_g)$.

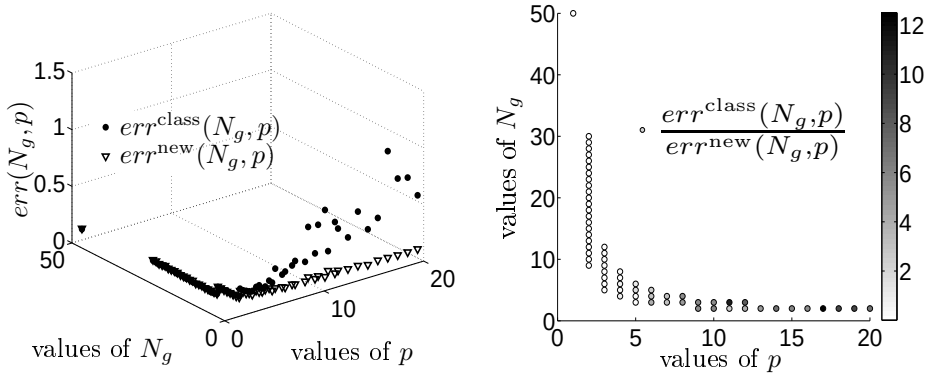



Figure 4.8: Comparison of the results for the convergence analysis of the two PCE identification formulations.

**4.3. PCE Identification.** From the $\nu^{\text{exp}}$ independent realizations of $\boldsymbol{\eta}$, a PCE identification using the new formulation can be computed for the truncation parameters $p = 9$ and $N_g = 3$, which correspond to $N = 220$. The results of the numerical identification with a computational cost of $M = 1000$ are given in Figure 4.9. The value of $M$ has been chosen for the PCE error function $err(N_g, p)$ to be independent of $M$. In this figure, the marginal PDFs $\widehat{p}^{\text{chaos}}_{\eta_{41}}$ and $\widehat{p}^{\text{chaos}}_{\eta_{39}}$ of $\eta^{\text{new}}_{41}(220)$ and $\eta^{\text{new}}_{39}(220)$ are compared to the experimental estimations $\widehat{p}^{\text{exp}}_{\eta_{41}}$ and $\widehat{p}^{\text{exp}}_{\eta_{39}}$ of the components $\eta_{41}$ and $\eta_{39}$, respectively. The values $\eta^{\text{new}}_{41}(220)$ and $\eta^{\text{new}}_{39}(220)$ correspond to the minimum and to the maximum values of the unidimensional error function $err_k(3, 11)$, for $1 \leq k \leq 50$, which is defined by Eq. (2.25). In order to evaluate the distance between these estimations and the true marginal PDFs of $\boldsymbol{\eta}$, the marginal PDFs estimated by the non parametric statistical Kernel method, with $\nu^{\text{ref}} = 2 \times 10^5$ independent realizations of $\eta_{41}$ and $\eta_{39}$, are added to the figures. These PDFs are considered as the reference. These figures therefore emphasize that the new PCE identification method allows building a stochastic model of the distribution of $\boldsymbol{\eta}$ that suits the experimental marginal PDFs.

**5. Conclusion.** In the last decade, the increasing computational power has encouraged the development of computational models with increasing degrees of freedom. Hence, developing computational methods which can be applied to very high dimension cases is currently of great interest.

In this concern, this paper emphasized the efficiency of the PCE when building multidimensional distributions. After having quantified the detrimental influence of a numerical bias in the usual PCE identification methods in high dimension, this paper proposed a new formulation to allow performing relevant convergence analysis and PCE identification with respect to an arbitrary measure for the high dimension case. Finally, the method proposed allows making the PCE range reachable for many engineering applications with many degrees of freedom.

**Acknowledgments.** This work was supported by SNCF (Innovation and Research Department) and by the French Research Agency (Grant No: ANR-2010-BLAN-0904).

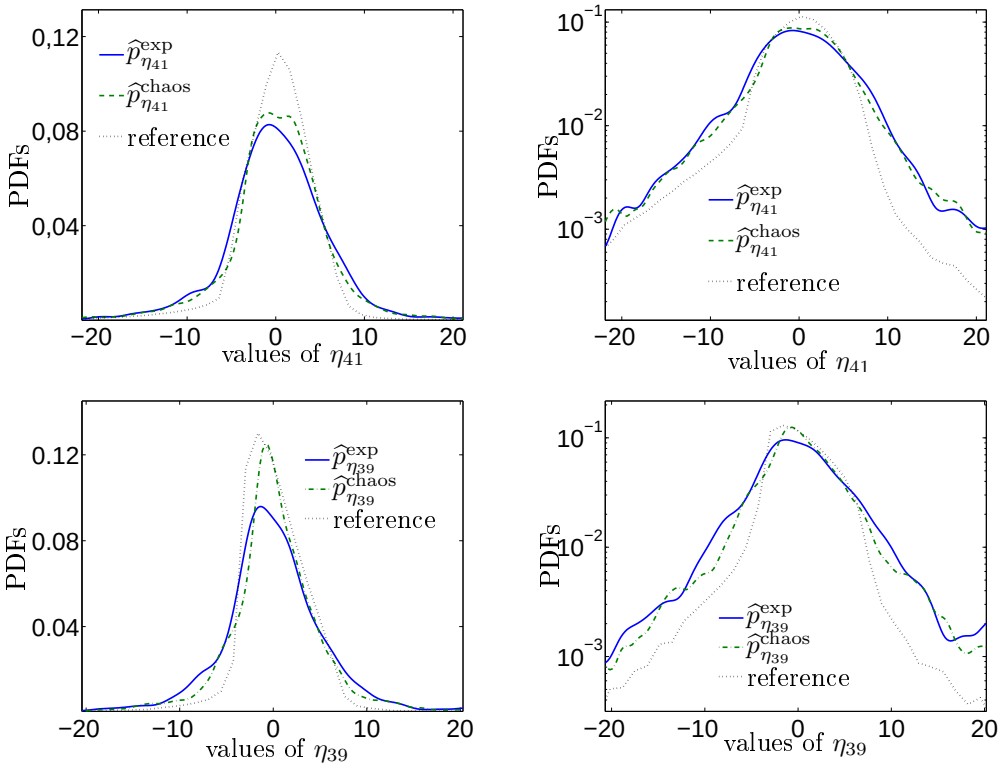



Figure 4.9: Graphs of the estimated marginal PDFs for two particular components of $\boldsymbol{\eta}$.

## REFERENCES


[1] M. Arnst, R. Ghanem, and C. Soize, *Identification of bayesian posteriors for coefficients of chaos expansions*, Journal of Computational Physics, 229 (9) (2010), pp. 3134–3154.

[2] S. Das, R. Ghanem, and S. Finette, *Polynomial chaos representation of spatio-temporal random field from experimental measurements*, J. Comput. Phys., 228 (2009), pp. 8726–8751.

[3] B. Debusschere, H. Najm, P. Pébay, O. M. Knio, R. G. Ghanem, and O. P. L. Maître, *Numerical challenges in the use of polynomial chaos representations for stochastic processes*, SIAM J. Sci. Comput., 26 (2004), pp. 698–719.

[4] C. Desceliers, R. Ghanem, and C. Soize, *Maximum likelihood estimation of stochastic chaos representations from experimental data*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 978–1001.

[5] C. Desceliers, C. Soize, and R. Ghanem, *Identification of chaos representations of elastic properties of random media using experimental vibration tests*, Comput. Mech., 39 (2007), pp. 831–838.

[6] A. Dutfoy, *Reference guide*, tutorial, Open TURNS version 0.11.3, 2008.

[7] R. Ghanem and R. Doostan, *Characterization of stochastic system parameters from experimental data: A bayesian inference approach*, J. Comput. Phys., 217 (2006), pp. 63–81.

[8] R. Ghanem, S. Masri, M. Pellissetti, and R. Wolfe, *Identification and prediction of stochastic dynamical systems in a polynomial chaos basis*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1641–1654.

[9] R. Ghanem and J. Red-Horse, *Propagation of uncertainty in complex physical systems using a stochastic finite element approach*, Phys. D, 133 (1999), pp. 137–144.

[10] R. Ghanem and P. Spanos, *Polynomial chaos in stochastic finite elements*, Journal of Applied Mechanics, Transactions of teh ASME 57 (1990), pp. 197–202.

[11] R. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach, rev. ed.*, Dover Publications, New York, 2003.

[12] D. Ghosh and R. Ghanem, *Stochastic convergence acceleration through basis enrichment of polynomial chaos expansions*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 162–184.

[13] E. T. Jaynes, *Information theory and statistical mechanics*, The Physical Review, 106 (4) (1963), pp. 620–630.

[14] O. M. Knio and O. P. L. Maître, *Uncertainty propagation in cfd using polynomial chaos decomposition*, Fluid Dynam. Res., 38 (2006), pp. 616–640.

[15] O. Le Maître and O. Knio, *Spectral Methods for Uncertainty Quantification*, Springer, 2010.

[16] D. Lucor, C. H. Su, and G. E. Karniadakis, *Generalized polynomial chaos and random oscillators*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 571–596.

[17] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems*, J. Comput. Phys., 228 (2009), pp. 1862–1902.

[18] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *spectral methods for efficient bayesian solution of inverse problems*, J. Comput. Phys., 224 (2007), pp. 560–586.

[19] H. Matthies, *Stochastic finite elements: Computational approaches to stochastic partial differential equations*, Zamm-Zeitschrift für Angewandte Mathematik und Mechanik, 88 (11) (2008), pp. 849–873.

[20] A. Nataf, *Détermination des distributions de probabilité dont les marges sont données*, Comptes Rendus de l'Académie des Sciences, 225 (1986), pp. 42–43.

[21] A. Nouy, A. Clement, F. Schoefs, and N. Moes, *An extended stochastic finite element method for solving stochastic partial differential equations on random domains*, Methods Appl. Mech. Engrg., 197 (2008), pp. 4663–4682.

[22] J. R. Red-Horse and A. S. Benjamin, *A probabilistic approach to uncertainty quantification with limited information*, Reliability Engrg. System Safety, 85 (2004), pp. 183–190.

[23] M. Rosenblatt, *Remarks on a multivariate transformation*, Annals of Mathematical Statistics, 23 (1952), pp. 470–472.

[24] S. Sakamoto and R. Ghanem, *Polynomial chaos decomposition for the simulation of non-gaussian nonstationary stochastic processes*, J. Engrg. Mechanics, 128 (2002), pp. 190–201.

[25] S. Sarkar, S. Gupta, and I. Rychlik, *Wiener chaos expansions for estimating rain-flow fatigue damage in randomly vibrating structures with uncertain parameters.*, Probabilistic Engineering Mechanics, 26 (2011), pp. 387–398.

[26] G. I. Schueller, *On the treatment of uncertainties in structural mechanics and analysis*, Computational and Structures, 85 (2007), pp. 235–243.

[27] C. Soize, *Construction of probability distributions in high dimension using the maximum entropy principle. applications to stochastic processes, random fields and random matrices.*, International Journal for Numerical Methods in Engineering, 76(10) (2008), pp. 1583–1611.

[28] C. Soize, *Generalized probabilistic approach of uncertainties in computational dynamics using random matrices and polynomial chaos decompositions.*, Internat. J. Numer. Methods Engrg., 81 (2010), pp. 939–970.

[29] C. Soize, *Identification of high-dimension polynomial chaos expansions with random coefficients for non-gausian tensor-valued random fields using partial and limited experimental data.*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 2150–2164.

[30] C. Soize and C. Desceliers, *Computational aspects for constructing realizations of polynomial chaos in high dimension.*, SIAM Journal on Scientific Computing, 32(5) (2010), pp. 2820–2831.

[31] C. Soize and R. Ghanem, *Physical systems with random uncertainties: Chaos representations with arbitrary probability measure*, SIAM J. Sci. Comput., 26 (2004).

[32] G. Stefanou, A. Nouy, and A. Clement, *Identification of random shapes from images through polynomial chaos expansion of random level set functions.*, Internat. J. Numer. Methods Engrg., 79 (2009), pp. 127–155.

[33] N. Wiener, *The homogeneous chaos*, American Journal of Mathematics, 60 (1938), pp. 897–936.

[34] D. Xiu and G. E. Karniadakis, *The wiener-askey polynomial chaos for stochastic differential equations*, SIAM Journal on Scientific Computing, 24, No. 2 (2002), pp. 619–644.