Polynomial chaos representation of databases on manifolds

Christian Soize, Roger Ghanem

Polynomial chaos representation of databases on manifolds

C. Soize[a,*], R. Ghanem[b]

[a] *Université Paris-Est, Laboratoire Modélisation et Simulation Multi-Echelle, MSME UMR 8208 CNRS, 5 bd Descartes, 77454 Marne-La-Vallée, Cedex 2, France*
[b] *University of Southern California, 210 KAP Hall, Los Angeles, CA 90089, United States*

Abstract

Characterizing the polynomial chaos expansion (PCE) of a vector-valued random variable with probability distribution concentrated on a manifold is a relevant problem in data-driven settings. The probability distribution of such random vectors is multimodal in general, leading to potentially very slow convergence of the PCE. In this paper, we build on a recent development for estimating and sampling from probabilities concentrated on a diffusion manifold. The proposed methodology constructs a PCE of the random vector together with an associated generator that samples from the target probability distribution which is estimated from data concentrated in the neighborhood of the manifold. The method is robust and remains efficient for high dimension and large datasets. The resulting polynomial chaos construction on manifolds permits the adaptation of many uncertainty quantification and statistical tools to emerging questions motivated by data-driven queries.

Keywords: Polynomial chaos expansion, Arbitrary probability measure, Concentration of probability, Measure concentration, Generator, Probability distribution on manifolds, Random sampling generator, MCMC generator, Diffusion maps, Statistics on manifolds

Notations

A lower case letter such as x, η, or u, is a real deterministic variable.
A boldface lower case letter such as $\mathbf{x}$, $\boldsymbol{\eta}$, or $\mathbf{u}$ is a real deterministic vector.

*Corresponding author: C. Soize, christian.soize@univ-paris-est.fr
Email addresses: christian.soize@univ-paris-est.fr (C. Soize), ghanem@usc.edu (R. Ghanem)

An upper case letter such as X, H, or U, is a real random variable.
A boldface upper case letter, **X**, **H**, or **U**, is a real random vector.
A lower case letter between brackets such as $[x]$, $[\eta]$, or $[u]$), is a real deterministic matrix.
A boldface upper case letter between brackets such as [**X**], [**H**], or [**U**], is a real random matrix.

$\mathbb{N} = \{0, 1, 2, \ldots\}$: set of all the null and positive integers.
$\mathbb{R}$: set of all the real numbers.
$\mathbb{R}^n$: Euclidean vector space on $\mathbb{R}$ of dimension n.
$\|\mathbf{x}\|_{\mathbb{R}^n}$: usual Euclidean norm in $\mathbb{R}^n$.
$<\mathbf{x}\,,\mathbf{y}>_{\mathbb{R}^n}$: usual Euclidean inner product in $\mathbb{R}^n$.
$\|\mathbf{x}\|$: represents $\|\mathbf{x}\|_{\mathbb{R}^n}$ if no confusion is possible.
$<\mathbf{x}\,,\mathbf{y}>$: represents $<\mathbf{x}\,,\mathbf{y}>_{\mathbb{R}^n}$ if no confusion is possible.
$\mathbb{M}_{n,N}$: set of all the $(n \times N)$ real matrices.
$\mathbb{M}_{\nu}$: set of all the square $(\nu \times \nu)$ real matrices.
$[x]_{kj}$: entry of matrix $[x]$.
$[x]^T$: transpose of matrix $[x]$.
tr$\{[x]\}$: trace of a square matrix $[x]$.
$\|[x]\|_F$: Frobenius norm of matrix $[x]$ such that $\|x\|_F^2 = \text{tr}\{[x]^T\,[x]\}$.
$[I_\nu]$: identity matrix in $\mathbb{M}_\nu$.
$\delta_{kk'}$: Kronecker's symbol such that $\delta_{kk'} = 0$ if $k \neq k'$ and $= 1$ if $k = k'$.
E: Mathematical expectation.
$L^2(\Theta, F)$: Hilbert space of all the F-valued second-order random variables defined on $(\Theta, \mathcal{T}, \mathcal{P})$.

1. Introduction

This work is a continuation of a recent paper [1] in which a methodology has been proposed for identifying, from a database made up of N samples of a $\mathbb{R}^\nu$-valued random variable (possibly for a high value of ν), its non-Gaussian probability distribution that is assumed to be unknown and that is concentrated on an unknown subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$. The method proposes an identification of subset $\mathcal{S}_\nu$ and the construction of generator of samples, based on an Itô stochastic differential equation (ISDE), for the unknown probability distribution that allows for preserving the concentration on $\mathcal{S}_\nu$ and consequently, avoiding the scattering of the generated samples. The main contribution of the present paper is to propose a mathematical formulation and an algorithm for constructing an intrinsic analytical

representation based on a polynomial chaos representation (PCE) of the database that is concentrated on subset $\mathcal{S}_\nu$. The proposed PCE and its associated generator of samples preserves the concentration over subset $\mathcal{S}_\nu$. In order to obtain an efficient algorithm for constructing the PCE in high dimension, a non classical construction is proposed for which the random germ of the chaos is not constituted of independent normalized Gaussian random variables but depends on the correlation structure of the germ used by the ISDE-based generator of samples. Consequently, the orthonormal multivariate polynomials are constructed with an efficient algorithm for a multivariate probability density function that is not separable with respect to the coordinates.

Constructing the high-dimensional polynomial chaos expansion (PCE) of a random vector $\mathbf{H}$ with values in the Euclidean space $\mathbb{R}^\nu$ remains a challenging problem when the probability distribution of the random vector is concentrated on a subset (a manifold) $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$. In such cases, the probability distribution on $\mathbb{R}^\nu$ is in general multimodal resulting in a potentially very slow convergence of the PCE.

In a recent paper [1], a new methodology was proposed for constructing a generator of samples from a given normalized dataset related to a second-order $\mathbb{R}^\nu$-valued random variable $\mathbf{H} = (H_1, \ldots, H_\nu)$ for which the probability density function (pdf), $\boldsymbol{\eta} \mapsto p_{\mathbf{H}}(\boldsymbol{\eta})$, with respect to the Lebesgue measure $\mathrm{d}\boldsymbol{\eta}$ on $\mathbb{R}^\nu$ is unknown and is concentrated on an unknown subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$. The method consists of first delineating a diffusion manifold for the available data [2], and then developing a reduced-order Itô stochastic differential equation (ISDE) by the projection on this manifold of an ISDE that admits $p_{\mathbf{H}}(\boldsymbol{\eta})\, d\boldsymbol{\eta}$ as a unique invariant measure. Specifically, we first introduce a random matrix $[\mathbf{H}] = [\mathbf{H}^1 \ldots \mathbf{H}^N]$ with values in $\mathbb{M}_{\nu,N}$, whose columns are independent copies $\mathbf{H}^1, \ldots, \mathbf{H}^N$ of random vector $\mathbf{H}$ and for which the number N of columns is equal to the number of independent data points in an initial dataset. This dataset, suitably normalized, is represented by a matrix $[\eta_d]$ given in $\mathbb{M}_{\nu,N}$, which is taken as a sample of random matrix $[\mathbf{H}]$. The approach presented in [1] is then summarized using the following steps, with an expanded summary provided in section 2 below:

- Multidimensional kernel-density estimation methods [3, 4] are used to construct the pdf $[\eta] \mapsto p_{[\mathbf{H}]}([\eta])$ with respect to the volume element $d[\eta]$ on $\mathbb{M}_{\nu,N}$ of random matrix $[\mathbf{H}]$, which is assumed to be a second-order random variable. This estimate of the pdf is related to the data and can be used for generating samples of $[\mathbf{H}]$. It should be noted that the method proposed in

[1] differs from the MCMC methods on Riemann manifolds that have recently been presented in the very good paper [8], for which the manifold is the locus of density functions and not of the data itself.

- A generator for random matrix $[\mathbf{H}]$ is constructed using the approach proposed in [5, 6] belonging to the class of Hamiltonian Monte Carlo methods [7, 8], which is an MCMC algorithm [9, 10, 11]. These samples are obtained by solving an ISDE corresponding to a stochastic nonlinear dissipative Hamiltonian dynamical system, for which $p_{\mathbf{H}}(\boldsymbol{\eta})\, d\boldsymbol{\eta}$ is the unique invariant measure.

- A diffusion-map approach [2, 12, 13] is then used to discover and characterize the local geometry structure of the normalized dataset concentrated in the neighborhood of $\mathcal{S}_\nu$. The diffusion-map vectors $[g] = [\mathbf{g}^1 \ldots \mathbf{g}^m] \in \mathbb{M}_{N,m}$ are thus constructed. They are associated with the first m eigenvalues of the transition matrix relative to the local geometric structure of the given normalized dataset.

- A reduced-order representation $[\mathbf{H}] = [\mathbf{Z}]\,[g]^T$ is constructed on the manifold in which $[\mathbf{Z}]$ is a random matrix with values in $\mathbb{M}_{\nu,m}$ (with $m \ll N$).

- The reduced-order ISDE is obtained by projecting the ISDE from the second step above onto the diffusion manifold by using the reduced-order basis represented by matrix $[g]^T$. The constructed reduced-order ISDE is then used for generating additional samples $[z^1_{\text{ar}}], \ldots, [z^{n_{\text{MC}}}_{\text{ar}}]$ of random matrix $[\mathbf{Z}]$, and therefore, for deducing the additional samples $[\eta^1_{\text{ar}}], \ldots, [\eta^{n_{\text{MC}}}_{\text{ar}}]$ of random matrix $[\mathbf{H}]$.

A numerical validation of this procedure was obtained through a number of applications ranging from standard templates to real experimental datasets. These results show the efficiency and the robustness of the method proposed, which allows for concentrating the additional generated samples in the neighborhood of subset $\mathcal{S}_\nu$.

In the present paper, we propose the construction of a polynomial chaos expansion (PCE) of the second-order random matrix $[\mathbf{Z}]$, with the associated generator of samples, which preserves the concentration of the probability measure of the column $\mathbf{H}^1, \ldots, \mathbf{H}^N$ of $[\mathbf{H}] = [\mathbf{Z}]\,[g]^T$ around $\mathcal{S}_\nu$. Such a construction will give an intrinsic analytical representation of random matrix $[\mathbf{H}]$ through the polynomial chaos expansion of random matrix $[\mathbf{Z}]$, which is very useful, for instance,

in uncertainty quantification and statistical data analysis, in stochastic modeling and associated statistical inverse problems for boundary value problems. In particular, such a representation could be used for response surface methodology, for computing extreme value statistics, for constructing parameterized reduced-order models, for accelerating robust updating, robust optimization, and robust design.

A systematic construction of PCE of second-order random fields and their use for analyzing the uncertainty propagation in boundary value problems was initiated in [14, 15]. Since then, PCE and their use in spectral approaches for solving linear and nonlinear stochastic boundary value problems, and some associated statistical inverse problems has rapidly grown [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]). Several extensions have been proposed concerning the generalized chaos expansions, the PCE for an arbitrary probability measure, and the PCE with random coefficients [34, 35, 36, 37, 38, 39], and recently, the construction of a basis adaptation in homogeneous chaos spaces [40].

Significant work has also been devoted to the acceleration of stochastic convergence of the PCE [41, 42, 43, 40, 6]. Nevertheless, the construction of a PCE of a second-order random vector for which the probability measure is concentrated on a subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$, remains a difficult and challenging problem.

In [1], it has been shown that a direct sampling of $\mathbf{H}$ using, for instance, a nonparametric estimation of the pdf of $\mathbf{H}$ constructed with the given normalized dataset and an MCMC method for generating samples, yields samples that are not concentrated around subset $\mathcal{S}_\nu$, but that are scattered. This was indeed the motivation for the method proposed in [1] using the diffusion maps and summarized above. In order to preserve the concentration of the probability measure for the analytical representation of the second-order random matrix $[\mathbf{H}]$, constraining associated samples of its columns $\mathbf{H}^1, \ldots, \mathbf{H}^N$ to be concentrated in the neighborhood of $\mathcal{S}_\nu$, the following methodology is proposed.

- The PCE of random $\mathbb{M}_{\nu,m}$-valued random variable $[\mathbf{Z}]$ is performed with respect to a $\mathbb{M}_{N_g,\mu}$-valued random germ $[\boldsymbol{\Xi}]$, with $1 \leq N_g \leq \nu$ and $1 \leq \mu \leq m$. The random germ $[\boldsymbol{\Xi}]$ is constructed as a linear mapping of the matrix-valued Wiener process that is used in the reduced-order ISDE that allows the samples of random matrix $[\mathbf{Z}]$ to be generated on the manifold.

- The PCE of random matrix $[\mathbf{Z}]$ is constructed relative to a basis of multivariate polynomials $\Psi_{\boldsymbol{\beta}}([\boldsymbol{\Xi}])$ orthonormal with respect to the Gaussian centered measure of random matrix $[\boldsymbol{\Xi}]$ for which the given covariance tensor

depends on the diffusion-map vectors related to the manifold. The quantity $\Psi_{\boldsymbol{\beta}}([\boldsymbol{\Xi}])$ is a shortand notation for the non-separable orthonormal multivariate polynomials $\Psi_{\boldsymbol{\beta}^1,\ldots\boldsymbol{\beta}^\mu}(\boldsymbol{\Xi}^1,\ldots,\boldsymbol{\Xi}^\mu)$, in which $\boldsymbol{\Xi}^1,\ldots,\boldsymbol{\Xi}^\mu$ are the μ columns of random matrix $[\boldsymbol{\Xi}]$ and are mutually dependent Gaussian $\mathbb{R}^{N_g}$-valued random variables, $\boldsymbol{\beta}^\alpha = \{\beta_1^\alpha,\ldots,\beta_{N_g}^\alpha\}$. Denoting the maximum degree of the polynomials as N_d (i.e. $|\boldsymbol{\beta}^1| + \ldots + |\boldsymbol{\beta}^\mu| \leq N_d$ with $|\boldsymbol{\beta}^\alpha| = \beta_1^\alpha + \ldots + \beta_{N_g}^\alpha$), we denote the PCE of random matrix $[\mathbf{Z}]$ by $[\mathbf{Z}(N_d, N_g, \mu)]$.

- Since the germ of the reduced-order ISDE (which allows for generating the samples for $[\mathbf{Z}]$) is statistically dependent on the germ $[\boldsymbol{\Xi}]$ that is used for constructing $[\mathbf{Z}(N_d, N_g, \mu)]$, the coefficients of $[\mathbf{Z}(N_d, N_g, \mu)]$ can easily be computed.

- The analytical representation of random matrix $[\mathbf{H}]$ with values in $\mathbb{M}_{\nu,N}$ is then obtained by replacing $[\mathbf{Z}]$ by $[\mathbf{Z}(N_d, N_g, \mu)]$ in the reduced-order representation $[\mathbf{H}] = [\mathbf{Z}]\,[g]^T$, for which the germ is the random matrix $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$.

- The PCE $[\mathbf{Z}(N_d, N_g, \mu)]$ of $[\mathbf{Z}]$ allows for generating the samples $[z_{\text{ar}}^1],\ldots,[z_{\text{ar}}^{n_{\text{MC}}}]$ of random matrix $[\mathbf{Z}]$ and therefore, for deducing the samples $\{[\boldsymbol{\eta}_{\text{ar}}^\ell] = [z_{\text{ar}}^\ell]\,[g]^T, \ell = 1,\ldots,n_{\text{MC}}\}$ of $[\mathbf{H}]$, whose columns will be concentrated on $\mathcal{S}_\nu$.

- The $\mathbb{R}^\nu$-valued random variable $\mathbf{H}$ is an application-specific normalization of a $\mathbb{R}^n$-valued random variable $\mathbf{X}$ (with $\nu \leq n$) for which N samples constitute the given dataset. According to our formulation, these N samples of $\mathbf{X}$ constitute one sample of random matrix $[\mathbf{X}]$ with values in $\mathbb{M}_{n,N}$.

Remark on the methodology proposed. Another approach could be envisaged for generating additional samples without recourse to the reduced-order ISDE. Such an approach would consist in performing a direct generation of samples of random matrix $[\mathbf{H}]$ from the multidimensional Gaussian kernel estimation method for which the concentration would be lost due to unavoidable scattering of the generated samples. These scattered samples would then be projected on the subspace spanned by the $m \ll N$ diffusion-maps vectors represented by matrix $[g]$ in order to generate the corresponding samples of $[\mathbf{Z}]$. With such an approach, the stochastic germ, which would be used by this Gaussian kernel generator, would not live on the "manifold" identified by the diffusion maps. The corresponding generator would produce samples that would belong to the big set $\mathbb{M}_{\nu,N}$, before projecting

them on the identified "manifold". Such an approach has not been investigated in this work for the following theoretical reason. The proposed method has been developed with two constraints in mind. The first one concerns the stochastic germ of the reduced-order ISDE that is used for generating the samples of $[\mathbf{Z}]$, which must live on the "manifold" identified by the diffusion maps, that is to say, which must live on a subset of the set $\mathbb{M}_{\nu,m}$, which has a small dimension because $m \ll N$. The second criterion requires the samples of $[\mathbf{Z}]$ to be directly generated by the reduced-order ISDE on the "manifold" which is a subset of $\mathbb{M}_{\nu,m}$ with a small dimension. The proposed generator based on a reduced-order ISDE, which allows these two constraints to be easily implemented. These two constraints are significant since they allows for constructing the PCE on the "manifold" without ambiguity and with a very efficient algorithm, as proposed in the manuscript.

Organization of the paper. The paper is organized as follows. Section 2 is devoted to a brief review of the methodology and algorithm proposed in [1] for generating samples that follow the pdf of the dataset, in the neighborhood of $\mathcal{S}_\nu$. Section 3 deals with the PCE of random matrix $[\mathbf{Z}]$, which capitalizes on the results summarized in Section 2. In Section 4, using the PCE of random matrix $[\mathbf{Z}]$, we present the generation of additional samples of random matrix $[\mathbf{X}]$ for which the dataset constitutes one sample. Finally, Section 5 is devoted to numerical examples.

2. Short summary of the methodology and algorithm proposed in [1] for a concentrated probability measure

In this section, we give a brief summary of the methodology presented in [1], which underlies the proposed PCE construction. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector, defined on a probability space $(\Theta, \mathcal{T}, \mathcal{P})$, with values in $\mathbb{R}^n$, for which the probability density function (pdf) $p_{\mathbf{X}}$ on $\mathbb{R}^n$ is unknown but is concentrated on an unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$, and for which N data points $\mathbf{x}^{d,1}, \ldots, \mathbf{x}^{d,N}$ in $\mathbb{R}^n$ correspond to N statistically independent samples of $\mathbf{X}$, which are given and which are represented by the matrix $[x_d] = [\mathbf{x}^{d,1} \ldots \mathbf{x}^{d,N}]$ in $\mathbb{M}_{n,N}$. Only using $[x_d]$ as available information, this method allows for generating samples of random vector $\mathbf{X}$ that are concentrated in the neighborhood of $\mathcal{S}_n$. The steps detailed in [1] are summarized hereinafter.

2.1. Scaling and normalizing the given dataset $[x_d]$

From a given unscaled dataset represented by a matrix $[x_d^{uns}]$ in $\mathbb{M}_{n,N}$, the (scaled) dataset $[x_d]$ in $\mathbb{M}_{n,N}$ is easily deduced from $[x_d^{uns}]$. The normalization of $\mathbf{X}$ is obtained by a principal component analysis. Introducing the random matrix $[\mathbf{X}] = [\mathbf{X}^1, \ldots, \mathbf{X}^N]$ with values in $\mathbb{M}_{n,N}$, whose columns are N independent copies of random vector $\mathbf{X}$, we introduce the corresponding normalized random matrix $[\mathbf{H}] = [\mathbf{H}^1, \ldots, \mathbf{H}^N]$ with values in $\mathbb{M}_{\nu,N}$, whose columns are N independent copies of random vector $\mathbf{H}$, with $\nu \leq n$, such that

$$[\mathbf{X}] = [\underline{x}] + [\varphi]\,[\lambda]^{1/2}\,[\mathbf{H}]\,, \tag{1}$$

in which $[\lambda]$ is the $(\nu \times \nu)$ diagonal matrix of the ν positive eigenvalues of the empirical estimation of the covariance matrix of $\mathbf{X}$, where $[\varphi]$ is the $(n \times \nu)$ matrix of the associated eigenvectors such $[\varphi]^T\,[\varphi] = [I_\nu]$, and where $[\underline{x}]$ is the matrix in $\mathbb{M}_{n,N}$ for which each column is the empirical estimation of the mean value of random vector $\mathbf{X}$. The sample $[\eta_d] = [\boldsymbol{\eta}^{d,1} \ldots \boldsymbol{\eta}^{d,N}] \in \mathbb{M}_{\nu,N}$ of $[\mathbf{H}]$ (associated with the sample $[x_d]$ of $[\mathbf{X}]$) is thus computed by

$$[\eta_d] = [\lambda]^{-1/2}[\varphi]^T\,([x_d] - [\underline{x}])\,. \tag{2}$$

Consequently, the empirical estimates of the mean value and of the covariance matrix of random vector $\mathbf{H}$ are $\mathbf{0}_\nu$ and $[I_\nu]$, respectively.

2.2. Constructing the diffusion-maps vectors for $\boldsymbol{H}$

Let $k_\varepsilon(\boldsymbol{\eta}, \boldsymbol{\eta}') = \exp(-\frac{1}{4\varepsilon}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|^2)$ be the kernel defined on $\mathbb{R}^\nu \times \mathbb{R}^\nu$, depending on a real smoothing parameter $\varepsilon > 0$. It should be noted that this kernel could be replaced by another one satisfying the symmetry, the positivity preserving, and the positive semi-definiteness properties. For $m \leq N$, let $[g] = [\mathbf{g}^1 \ldots \mathbf{g}^m] \in \mathbb{M}_{N,m}$ be the "diffusion-maps basis" associated with kernel k_ε, which is defined and constructed in Appendix A (for $m = N$, $[g]$ is an algebraic basis of $\mathbb{R}^N$). For $\alpha = 1, \ldots, m$, the diffusion-maps vector $\mathbf{g}^\alpha \in \mathbb{R}^N$ is defined by Eq. (A.3). The subspace of $\mathbb{R}^N$ spanned by the vector basis $\{\mathbf{g}^\alpha\}_\alpha$ allows for characterizing the local geometry structure of the dataset concentrated in the neighborhood of $\mathcal{S}_\nu$.

2.3. Introducing the reduced-order representation of random matrix $[\boldsymbol{H}]$

The reduced-order representation is obtained in projecting each column of the $\mathbb{M}_{N,\nu}$-valued random matrix $[\mathbf{H}]^T$ on the subspace of $\mathbb{R}^N$, spanned by $\{\mathbf{g}^1 \ldots \mathbf{g}^m\}$.

Introducing the random matrix $[\mathbf{Z}]$ with values in $\mathbb{M}_{\nu,m}$, the following reduced-order representation of $[\mathbf{H}]$ is introduced,

$$[\mathbf{H}] = [\mathbf{Z}]\,[g]^T\,. \tag{3}$$

Since the matrix $[g]^T\,[g] \in \mathbb{M}_m$ is invertible, Eq. (3) yields the least squares approximation to $\mathbf{Z}$ in the form,

$$[\mathbf{Z}] = [\mathbf{H}]\,[a] \quad , \quad [a] = [g]\,([g]^T\,[g])^{-1} \in \mathbb{M}_{N,m}\,. \tag{4}$$

In particular, matrix $[\eta_d] \in \mathbb{M}_{\nu,N}$ can be written as $[\eta_d] = [z_d]\,[g]^T$ in which the matrix $[z_d] \in \mathbb{M}_{\nu,m}$ is given by

$$[z_d] = [\eta_d]\,[a] \in \mathbb{M}_{\nu,m}\,. \tag{5}$$

Consequently, from Eqs. (1) and (4), it can be deduced the following representation of random matrix $[\mathbf{X}]$ as a function of random matrix $[\mathbf{Z}]$,

$$[\mathbf{X}] = [\underline{x}] + [\varphi]\,[\lambda]^{1/2}\,[\mathbf{Z}]\,[g]^T\,. \tag{6}$$

2.4. Estimating dimension m of the reduced-order representation of random matrix $[\mathbf{Z}]$

For a given value of integer ζ related to the analysis scale of the local geometric structure of the dataset (see Eq. (A.3) in Appendix A) and for a given value of smoothing parameter $\varepsilon > 0$, the decreasing of the graph $\alpha \mapsto \Lambda_\alpha$ of the positive eigenvalues of transition matrix $[\mathbb{P}]$ (see Appendix A) yields a criterion for choosing the value of m that allows the local geometric structure of the dataset represented by $[\eta_d]$ to be discovered. Nevertheless, this criterion may not be sufficient, and the L^2-convergence may need to be enforced by increasing, as required, the value of m. However, if the value of m is chosen too large, the localization of the geometric structure of the dataset is lost. Consequently, a compromise must be reached between the very small value of m given by the decreasing criteria of the eigenvalues of matrix $[\mathbb{P}] \in \mathbb{M}_N$ and a larger value of m which is necessary for obtaining a reasonable mean-square convergence. A criterion for estimating an optimal value of m is given in Appendix B.

2.5. Reduced-order ISDE for generating additional samples $[z^1_{\text{ar}}], \ldots, [z^{n_{\text{MC}}}_{\text{ar}}]$ of random matrix $[\mathbf{Z}]$

For m, ε, and ζ fixed, we introduce the Markov stochastic process $\{([\boldsymbol{\mathcal{Z}}(r)], [\boldsymbol{\mathcal{Y}}(r)]), r \in \mathbb{R}^+\}$, defined on $(\Theta, \mathcal{T}, \mathcal{P})$, indexed by $\mathbb{R}^+ = [0\,, +\infty[$, with values in

$\mathbb{M}_{\nu,m} \times \mathbb{M}_{\nu,m}$, which is the unique second-order stationary (for the shift semi-group on $\mathbb{R}^+$) and ergodic diffusion stochastic process, of the following reduced-order ISDE, for $r > 0$,

$$d[\boldsymbol{\mathcal{Z}}(r)] = [\boldsymbol{\mathcal{Y}}(r)]\, dr\,, \tag{7}$$

$$d[\boldsymbol{\mathcal{Y}}(r)] = [\mathcal{L}([\boldsymbol{\mathcal{Z}}(r)])]\, dr - \frac{1}{2} f_0\, [\boldsymbol{\mathcal{Y}}(r)]\, dr + \sqrt{f_0}\, [d\boldsymbol{\mathcal{W}}(r)]\,, \tag{8}$$

with the initial condition

$$[\boldsymbol{\mathcal{Z}}(0)] = [\mathbf{H}_d]\, [a] \quad , \quad [\boldsymbol{\mathcal{Y}}(0)] = [\boldsymbol{\mathcal{N}}]\, [a] \quad a.s\,, \tag{9}$$

in which the random matrices $[\mathcal{L}([\boldsymbol{\mathcal{Z}}(r)])]$ and $[d\boldsymbol{\mathcal{W}}(r)]$ with values in $\mathbb{M}_{\nu,m}$ are such that

$$[\mathcal{L}([\boldsymbol{\mathcal{Z}}(r)])] = [L([\boldsymbol{\mathcal{Z}}(r)]\, [g]^T)]\, [a] \quad , \quad [d\boldsymbol{\mathcal{W}}(r)] = [d\mathbf{W}(r)]\, [a]\,. \tag{10}$$

(i) For all $[u] = [\mathbf{u}^1 \dots \mathbf{u}^N]$ in $\mathbb{M}_{\nu,N}$ with $\mathbf{u}^\ell = (u_1^\ell, \dots, u_\nu^\ell)$ in $\mathbb{R}^\nu$, the matrix $[L([u])]$ in $\mathbb{M}_{\nu,N}$ is defined, for all $k = 1, \dots, \nu$ and for all $\ell = 1, \dots, N$, by

$$[L([u])]_{k\ell} = \frac{1}{q(\mathbf{u}^\ell)}\, \{\boldsymbol{\nabla}_{\mathbf{u}^\ell}\, q(\mathbf{u}^\ell)\}_k\,, \tag{11}$$

$$q(\mathbf{u}^\ell) = \frac{1}{N} \sum_{j=1}^{N} \exp\{-\frac{1}{2\widehat{s}_\nu^2} \| \frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell \|^2\}\,, \tag{12}$$

$$\boldsymbol{\nabla}_{\mathbf{u}^\ell}\, q(\mathbf{u}^\ell) = \frac{1}{\widehat{s}_\nu^2} \frac{1}{N} \sum_{j=1}^{N} (\frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell)\, \exp\{-\frac{1}{2\widehat{s}_\nu^2} \| \frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell \|^2\}\,, \tag{13}$$

$$s_\nu = \left\{ \frac{4}{N(2+\nu)} \right\}^{1/(\nu+4)} \quad , \quad \widehat{s}_\nu = \frac{s_\nu}{\sqrt{s_\nu^2 + \frac{N-1}{N}}}\,. \tag{14}$$

(ii) The stochastic process $\{[d\mathbf{W}(r)], r \geq 0\}$ with values in $\mathbb{M}_{\nu,N}$ is such that $[d\mathbf{W}(r)] = [d\mathbf{W}^1(r) \dots d\mathbf{W}^N(r)]$ in which the columns $\mathbf{W}^1, \dots, \mathbf{W}^N$ are N independent copies of the normalized Wiener process $\mathbf{W} = (W_1, \dots, W_\nu)$ defined on $(\Theta, \mathcal{T}, \mathcal{P})$, indexed by $\mathbb{R}^+$ with values in $\mathbb{R}^\nu$. The matrix-valued autocorrelation function $[R_{\mathbf{W}}(r, r')] = E\{\mathbf{W}(r)\, \mathbf{W}(r')^T\}$ of $\mathbf{W}$ is then written as $[R_{\mathbf{W}}(r, r')] = \min(r, r')\, [I_\nu]$.

(iii) The probability distribution of the random matrix $[\mathbf{H}_d]$ with values in $\mathbb{M}_{\nu,N}$ is $p_{[\mathbf{H}]}([\eta])\, d[\eta]$. A known sample of $[\mathbf{H}_d]$ is matrix $[\eta_d]$ defined by Eq. (2). The random matrix $[\boldsymbol{\mathcal{N}}]$ with values in $\mathbb{M}_{\nu,N}$ is written as $[\boldsymbol{\mathcal{N}}] = [\boldsymbol{\mathcal{N}}^1 \ldots \boldsymbol{\mathcal{N}}^N]$ in which the columns $\boldsymbol{\mathcal{N}}^1, \ldots, \boldsymbol{\mathcal{N}}^N$ are N independent copies of the normalized Gaussian vector $\boldsymbol{\mathcal{N}}$ with values in $\mathbb{R}^\nu$ (this means that $E\{\boldsymbol{\mathcal{N}}\} = \mathbf{0}$ and $E\{\boldsymbol{\mathcal{N}}\boldsymbol{\mathcal{N}}^T\} = [I_\nu]$). The random matrices $[\mathbf{H}_d]$ and $[\boldsymbol{\mathcal{N}}]$, and the normalized Wiener process $\{\mathbf{W}(r), r \geq 0\}$ are assumed to be independent.

(iv) The free parameter $f_0 > 0$ allows the dissipation term of the nonlinear second-order dynamical system (dissipative Hamiltonian system) to be controlled.

(v) The algorithm for solving Eqs. (7) to (9) is detailed in [1] and is summarized in Appendix C.

2.6. Generating additional samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ of random matrix $[\boldsymbol{X}]$

For θ fixed in Θ, the deterministic quantities $\{[\boldsymbol{\mathcal{W}}(r;\theta)], r \geq 0\}$, $[\boldsymbol{\mathcal{Z}}(0;\theta)] = [\eta_d]\,[a]$, and $[\boldsymbol{\mathcal{Y}}(0;\theta)] = [\boldsymbol{\mathcal{N}}(\theta)]\,[a]$ are independent samples of the stochastic process $\{[\boldsymbol{\mathcal{W}}(r)], r \geq 0\}$, the random matrix $[\boldsymbol{\mathcal{Z}}(0)]$, and the random matrix $[\boldsymbol{\mathcal{Y}}(0)]$. Let $\{([\boldsymbol{\mathcal{Z}}(r;\theta)], [\boldsymbol{\mathcal{Y}}(r;\theta)]), r \in \mathbb{R}^+\}$ be the corresponding sample of the unique stationary diffusion process $\{([\boldsymbol{\mathcal{Z}}(r)], [\boldsymbol{\mathcal{Y}}(r)]), r \in \mathbb{R}^+\}$ of the problem defined by Eqs. (7) to (9)). For $\rho = M_0\,\Delta r$, in which Δr is the sampling step of the continuous index parameter r used in the integration scheme (see Appendix C), and where M_0 is a positive integer greater or equal to 1 such that $M = M_0 \times n_{\text{MC}}$, the additional samples $[\tilde{z}_{\text{ar}}^1], \ldots, [\tilde{z}_{\text{ar}}^{n_{\text{MC}}}]$ of random matrix $[\mathbf{Z}]$ and the corresponding samples $[\tilde{\eta}_{\text{ar}}^1], \ldots, [\tilde{\eta}_{\text{ar}}^{n_{\text{MC}}}]$ of random matrix $[\mathbf{H}]$ are given by

$$[\tilde{z}_{\text{ar}}^\ell] = [\boldsymbol{\mathcal{Z}}(\ell \times \rho\,; \theta)] \quad , \quad [\tilde{\eta}_{\text{ar}}^\ell] = [\tilde{z}_{\text{ar}}^\ell]\,[g]^T \quad , \quad \ell = 1, \ldots, n_{\text{MC}}\,. \tag{15}$$

- If $M_0 = 1$, then $\rho = \Delta r$ and the n_{MC} additional samples are dependent, but the ergodic property of $\{[\boldsymbol{\mathcal{Z}}(r)], r \in \mathbb{R}^+\}$ can be used for obtaining the convergence of statistics constructed using $[\tilde{z}_{\text{ar}}^1], \ldots, [\tilde{z}_{\text{ar}}^{n_{\text{MC}}}]$ for random matrix $[\mathbf{Z}]$.

- If integer M_0 is chosen sufficiently large (such that ρ is much larger than the relaxation time of the dissipative Hamiltonian dynamical system), then $[\tilde{z}_{\text{ar}}^1], \ldots, [\tilde{z}_{\text{ar}}^{n_{\text{MC}}}]$ can approximatively be considered as independent samples of random matrix $[\mathbf{Z}]$.

- The representation defined by Eq. (1) has been constructed in order that the empirical estimates of the mean value and the covariance matrix of random vector $\mathbf{H}$ are $\mathbf{0}_\nu$ and $[I_\nu]$, respectively. As the generator of random matrix $[\mathbf{H}]$ defined by Eqs. (7) to (14) can introduce a bias induced by the integration scheme defined by Eqs. (C.4) to (C.7) in Appendix C, the samples $\{[\tilde{\eta}_{\text{ar}}^\ell], \ell = 1, \ldots n_{\text{MC}}\}$ defined by Eq. (15) are re-normalized using the following algorithm:

  (i) Computation of the empirical estimate $\widetilde{\mathbf{m}} = (\widetilde{m}_1, \ldots, \widetilde{m}_\nu) \in \mathbb{R}^\nu$ of the mean vector of $\mathbf{H}$ and of the empirical estimate $[\widetilde{c}] \in \mathbb{M}_\nu$ of the covariance matrix of $\mathbf{H}$, such that, for all k and k' in $\{1, \ldots, \nu\}$,

$$\widetilde{m}_k = \frac{1}{N \times n_{\text{MC}}} \sum_{j=1}^{N} \sum_{\ell=1}^{n_{\text{MC}}} [\tilde{\eta}_{\text{ar}}^\ell]_{kj} \,, \tag{16}$$

$$[\widetilde{c}]_{kk'} = \frac{1}{N \times n_{\text{MC}} - 1} \sum_{j=1}^{N} \sum_{\ell=1}^{n_{\text{MC}}} ([\tilde{\eta}_{\text{ar}}^\ell]_{kj} - \widetilde{m}_k)([\tilde{\eta}_{\text{ar}}^\ell]_{k'j} - \widetilde{m}_{k'}) \,. \tag{17}$$

  (ii) Assuming that matrix $[\widetilde{c}]$ is positive definite, computing the Cholesky factorization $[\widetilde{c}] = [\widetilde{L}]^T [\widetilde{L}]$ and computing the re-normalized samples $\{[\eta_{\text{ar}}^\ell], \ell = 1, \ldots n_{\text{MC}}\}$ such that, for all $\ell = 1, ..., n_{\text{MC}}$, $j = 1, \ldots, N$, and $k = 1, \ldots, \nu$,

$$[\eta_{\text{ar}}^\ell]_{kj} = \sum_{k'=1}^{\nu} \{[\widetilde{L}]^{-T}\}_{kk'} ([\tilde{\eta}_{\text{ar}}^\ell]_{k'j} - \widetilde{m}_{k'}) \,. \tag{18}$$

  Consequently, the empirical estimates of the mean value and covariance matrix performed with the re-normalized samples $\{[\eta_{\text{ar}}^\ell], \ell = 1, \ldots n_{\text{MC}}\}$ yield the zero vector and the identity matrix for random vector $\mathbf{H}$, respectively.

- Using Eqs. (4), it is deduced that the samples $[z_{\text{ar}}^1], \ldots, [z_{\text{ar}}^{n_{\text{MC}}}]$ associated with the re-normalized samples $[\eta_{\text{ar}}^1], \ldots, [\eta_{\text{ar}}^{n_{\text{MC}}}]$ defined by Eq. (18), can be computed by

$$[z_{\text{ar}}^\ell] = [\eta_{\text{ar}}^\ell] \, [a] \quad , \quad \ell = 1, \ldots n_{\text{MC}} \,. \tag{19}$$

Using Eq. (6), the additional samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ of random matrix $[\mathbf{X}]$ can be generated by

$$[x_{\text{ar}}^\ell] = [\underline{x}] + [\varphi] \, [\lambda]^{1/2} \, [z_{\text{ar}}^\ell] \, [g]^T \quad , \quad \ell = 1, \ldots n_{\text{MC}} \,. \tag{20}$$

3. Polynomial chaos expansion of random matrix [**Z**]

3.1. Construction of an adapted Gaussian germ for the PCE on the manifold

In general, $N_g \leq \nu$ and $\mu \leq m$. However, for explaining the construction of the Gaussian germ $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$, we construct the germ $[\boldsymbol{\Xi}(\nu,m)]$ with values in $\mathbb{M}_{\nu,m}$ and then, for $N_g < \nu$ and $\mu < m$, the germ $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$ will be obtained by extracting the block $(N_g \times \mu)$ from $[\boldsymbol{\Xi}(\nu,m)]$.

The idea is to construct a Gaussian random matrix $[\boldsymbol{\Xi}(\nu,m)]$, defined on $(\Theta,\mathcal{T},\mathcal{P})$ with values in $\mathbb{M}_{\nu,m}$ that is statistically dependent of the $\mathbb{M}_{\nu,m}$-valued stochastic process $\{[\boldsymbol{\mathcal{W}}(r)], r \geq 0\}$, in order to perform the computation of the coefficients of the PCE of random matrix $[\mathbf{Z}]$ by a direct use of the projection formula. Nevertheless, in order to avoid potential numerical difficulties for computing the samples of the PCE with respect to $[\boldsymbol{\Xi}]$, we need to introduce a germ whose statistical fluctuations have unit variance. Consequently, we will perform the construction by normalizing $\{[\boldsymbol{\mathcal{W}}(r)], r \geq 0\}$ for which the statistical fluctuations can have a variance greater than 1 due to the presence of matrix $[a]$ related to the manifold.

Let $\{[\Delta\boldsymbol{\mathcal{W}}_{\ell'}], \ell' = 1,\ldots,M\}$ be the sequence of the independent Gaussian centered random matrices with values in $\mathbb{M}_{\nu,m}$, which are defined by Eq. (C.2) of Appendix C,

$$[\Delta\boldsymbol{\mathcal{W}}_{\ell'}] = [\Delta\mathbf{W}_{\ell'}]\,[a]\,. \tag{21}$$

Taking into account Eq. (C.3) of Appendix C, it can easily be seen that, for all ℓ' fixed in $\{1,\ldots,M\}$, the fourth-order covariance tensor of random matrix $[\Delta\boldsymbol{\mathcal{W}}_{\ell'}]$, is independent of ℓ', and is such that, for all α and α' in $\{1,\ldots,m\}$, and for all k and k' in $\{1,\ldots,\nu\}$,

$$E\{\,[\Delta\boldsymbol{\mathcal{W}}_{\ell'}]_{k\alpha}\,[\Delta\boldsymbol{\mathcal{W}}_{\ell'}]_{k'\alpha'}\} = \delta_{kk'}\,[\mathcal{A}]_{\alpha\alpha'}\,, \tag{22}$$

in which the positive-definite matrix $[\mathcal{A}]$ is written as

$$[\mathcal{A}] = \Delta r\,[a]^T\,[a] \in \mathbb{M}^+_m\,. \tag{23}$$

We then defined the random matrix $[\boldsymbol{\Xi}(\nu,m)]$ as the Gaussian, centered, random matrix with values in $\mathbb{M}_{\nu,m}$, for which the fourth-order covariance tensor is such that, for all k and k' in $\{1,\ldots,\nu\}$ and for all α and α' in $\{1,\ldots,m\}$,

$$E\{\,[\boldsymbol{\Xi}(\nu,m)]_{k\alpha}\,[\boldsymbol{\Xi}(\nu,m)]_{k'\alpha'}\} = \delta_{kk'}\,\frac{[\mathcal{A}]_{\alpha\alpha'}}{\sqrt{[\mathcal{A}]_{\alpha\alpha}[\mathcal{A}]_{\alpha'\alpha'}}}\,, \tag{24}$$

in which the matrix $[\mathcal{A}]$ is defined by Eq. (23). By comparing Eq. (22) with Eq. (24), and in taking ρ and the sample θ in Θ used in Eq. (15), it can be concluded that n_{MC} samples, $\{[\xi^\ell(\nu,m)], \ell = 1,\ldots,n_{\text{MC}}\}$ of random matrix $[\boldsymbol{\Xi}(\nu,m)]$ can be constructed, for $\ell = 1,\ldots,n_{\text{MC}}$, as

$$[\xi^\ell(\nu,m)] = [\Delta \boldsymbol{\mathcal{W}}_{\ell\times\rho}(\theta)]\,[d\,]\,, \tag{25}$$

in which $[d\,]$ is the $(m\times m)$ real diagonal positive-definite matrix which is written as,

$$[d\,]_{\alpha\alpha'} = \frac{\delta_{\alpha\alpha'}}{\sqrt{[\mathcal{A}]_{\alpha\alpha}}}\,. \tag{26}$$

For fixed $N_g \leq \nu$ and $\mu \leq m$, the n_{MC} samples, $\{[\xi^\ell], \ell = 1,\ldots,n_{\text{MC}}\}$ of random matrix $[\boldsymbol{\Xi}]$ are then given by extracting, for $\ell = 1,\ldots,n_{\text{MC}}$, the block $(N_g\times\mu)$ from $[\xi^\ell(\nu,m)]$,

$$[\xi^\ell]_{k\alpha} = [\xi^\ell(\nu,m)]_{k\alpha} \quad , \quad k = 1,\ldots,N_g \quad , \quad \alpha = 1,\ldots,\mu\,. \tag{27}$$

3.2. Orthonormal multivariate polynomials

Definition of the multi-indices for the PCE of $[\mathbf{Z}]$ *with respect to* $[\boldsymbol{\Xi}]$. Let N_g and μ be fixed such that $1 \leq N_g \leq \nu$ and $1 \leq \mu \leq m$. For $\alpha \in \{1,\ldots,\mu\}$, let $\boldsymbol{\beta}^\alpha = (\beta_1^\alpha,\ldots,\beta_{N_g}^\alpha)$ be the multi-index such that $\boldsymbol{\beta}^\alpha \in \mathbb{N}^{N_g}$ and let be $|\boldsymbol{\beta}^\alpha| = \beta_1^\alpha + \ldots + \beta_{N_g}^\alpha$. Relatively to the columns $\boldsymbol{\Xi}^1,\ldots,\boldsymbol{\Xi}^\mu$ of matrix $[\boldsymbol{\Xi}]$, we introduce the multi-index $\boldsymbol{\beta} = (\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^\mu)$ that belongs to $\mathbb{N}^{\mu N_g} = \mathbb{N}^{N_g}\times\ldots\times\mathbb{N}^{N_g}$ such that $|\boldsymbol{\beta}| = |\boldsymbol{\beta}^1| + \ldots + |\boldsymbol{\beta}^\mu|$. For a fixed value of N_d that is the maximum degree of the orthonormal multivariate polynomials, such that $N_d \geq 1$, the following set $\mathcal{B}_{N_d,N_g,\mu}$ of multi-indices is introduced,

$$\mathcal{B}_{N_d,N_g,\mu} = \{\boldsymbol{\beta} \in \mathbb{N}^{\mu N_g} \quad | \quad 0 \leq |\boldsymbol{\beta}| \leq N_d\}\,. \tag{28}$$

The K elements of set $\mathcal{B}_{N_d,N_g,\mu}$, which depend on N_d, N_d, and μ (and also written as $K(N_d,N_g,\mu)$), are denoted by $\boldsymbol{\beta}^{(1)},\ldots\,\boldsymbol{\beta}^{(K)}$ in which $\boldsymbol{\beta}^{(1)}$ is the multi-index $(\mathbf{0},\ldots,\mathbf{0})$, and where the integer $K(N_d,N_g,\mu)$ is such that

$$K(N_d,N_g,\mu) = \frac{(\mu N_g + N_d)!}{(\mu N_g)!\,N_d!}\,. \tag{29}$$

Orthonormal multivariate polynomials. Let $p_{[\boldsymbol{\Xi}]}([\xi])\,d[\xi]$ be the Gaussian, centered, probability measure of the random matrix $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$, for

which the fourth-order covariance tensor, which is derived from Eq. (24), is written, for all k and k' in $\{1, \ldots, N_g\}$, and for all α and α' in $\{1, \ldots, \mu\}$, as

$$E\{\, [\boldsymbol{\Xi}]_{k\alpha}\, [\boldsymbol{\Xi}]_{k'\alpha'}\} = \delta_{kk'}\, \frac{[\mathcal{A}]_{\alpha\alpha'}}{\sqrt{[\mathcal{A}]_{\alpha\alpha}[\mathcal{A}]_{\alpha'\alpha'}}}\,. \tag{30}$$

Note that the probability measure $p_{[\boldsymbol{\Xi}]}([\xi])\, d[\xi]$ is such that

$$p_{[\boldsymbol{\Xi}]}([\xi])\, d[\xi] = p_{\boldsymbol{\Xi}^1,\ldots,\boldsymbol{\Xi}^\mu}(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^\mu)\, d\boldsymbol{\xi}^1 \ldots d\boldsymbol{\xi}^\mu\,. \tag{31}$$

in which $\boldsymbol{\xi}^\alpha \in \mathbb{R}^{N_g}$ and where $d\boldsymbol{\xi}^\alpha$ is the Lebesgue measure on $\mathbb{R}^{N_g}$. For all multi-indices $\boldsymbol{\beta} \in \mathbb{N}^{\mu N_g}$, we introduce the real-valued multivariate polynomials $\Psi_{\boldsymbol{\beta}}([\xi])$ on $\mathbb{M}_{N_g,\mu}$, which is defined by

$$\Psi_{\boldsymbol{\beta}}([\xi]) = \Psi_{\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^\mu}(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^\mu)\,, \tag{32}$$

in which $[\xi] = [\boldsymbol{\xi}^1 \ldots \boldsymbol{\xi}^\mu]$. Let $\{\Psi_{\boldsymbol{\beta}}([\xi]), \boldsymbol{\beta} \in \mathbb{N}^{\mu N_g}\}$ be the family of real-valued multivariate polynomials orthonormal with respect to $p_{[\boldsymbol{\Xi}]}([\xi])\, d[\xi]$. For all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ in $\mathbb{N}^{\mu N_g}$,

$$\int_{\mathbb{M}_{N_g,\mu}} \Psi_{\boldsymbol{\beta}}([\xi])\, \Psi_{\boldsymbol{\beta}'}([\xi])\, p_{[\boldsymbol{\Xi}]}([\xi])\, d[\xi] = E\{\Psi_{\boldsymbol{\beta}}([\boldsymbol{\Xi}])\, \Psi_{\boldsymbol{\beta}'}([\boldsymbol{\Xi}])\} = \delta_{\boldsymbol{\beta}\boldsymbol{\beta}'}\,, \tag{33}$$

where $\delta_{\boldsymbol{\beta}\boldsymbol{\beta}'}$ is the Kronecker symbol. By convention, for $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)} = (\mathbf{0}, \ldots, \mathbf{0})$, $\Psi_{\boldsymbol{\beta}^{(1)}}([\xi]) = 1$ is the constant normalized multivariate polynomial. Since the random vectors $\boldsymbol{\Xi}^1, \ldots, \boldsymbol{\Xi}^\mu$ are mutually dependent, the probability density function $p_{\boldsymbol{\Xi}^1,\ldots,\boldsymbol{\Xi}^\mu}(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^\mu)$ is not separable in $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^\mu$ and consequently, the orthonormal multivariate polynomials $\Psi_{\boldsymbol{\beta}}([\xi])$ cannot be written as $\Psi_{\boldsymbol{\beta}^1}(\boldsymbol{\xi}^1) \times \ldots \times \Psi_{\boldsymbol{\beta}^\mu}(\boldsymbol{\xi}^\mu)$. From a theoretical point of view, the orthonormal multivariate polynomials can be viewed as the result of a Gram-Schmidt orthonormalization algorithm of the multivariate monomials (see Section 3.5) defined, for $\boldsymbol{\beta} \in \mathbb{N}^{\mu N_g}$, by

$$\mathcal{M}_{\boldsymbol{\beta}}([\xi]) = \mathcal{M}_{\boldsymbol{\beta}^1}(\boldsymbol{\xi}^1) \times \ldots \times \mathcal{M}_{\boldsymbol{\beta}^\mu}(\boldsymbol{\xi}^\mu)\,, \tag{34}$$

in which for all α in $\{1, \ldots, \mu\}$,

$$\mathcal{M}_{\boldsymbol{\beta}^\alpha}(\boldsymbol{\xi}^\alpha) = (\xi_1^\alpha)^{\beta_1^\alpha} \times \ldots \times (\xi_{N_g}^\alpha)^{\beta_{N_g}^\alpha}\,. \tag{35}$$

3.3. Polynomial chaos expansion of random matrix $[\mathbf{Z}]$

In this section, we construct the PCE of random matrix $[\mathbf{Z}]$. Since $[\mathbf{Z}]$ is a second-order random variable with values in $\mathbb{M}_{\nu,m}$, random matrix $[\mathbf{Z}]$ can be expanded in polynomial chaos by using the classical theory. What is different in our case, it is the fact that we propose to construct the truncated PCE of random matrix $[\mathbf{Z}]$ with respect to the germ $[\boldsymbol{\Xi}]$ defined in Section 3.2, which is related to the correlation structure of the germ on the manifold. We then have to analyze the problem related to the L^2 convergence of the truncated PCE of $[\mathbf{Z}]$.

(i) Truncated PCE and convergence. For $N_g \leq \nu$, $\mu \leq m$, and for $N_d > 1$ fixed, the truncated PCE denoted by $[\mathbf{Z}(N_d, N_g, \mu)]$ of random matrix $[\mathbf{Z}]$ with values in $\mathbb{M}_{\nu,m}$, with respect to random matrix $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$, is written as

$$[\mathbf{Z}(N_d, N_g, \mu)] = \sum_{\kappa=1}^{K(N_d,N_g,\mu)} [y^{(\kappa)}]\, \Psi_{\boldsymbol{\beta}^{(\kappa)}}([\boldsymbol{\Xi}]) \quad , \quad [y^{(\kappa)}] \in \mathbb{M}_{\nu,m}\,, \tag{36}$$

in which the integer $K(N_d, N_g, \mu)$ and the multi-indices $\boldsymbol{\beta}^{(\kappa)}$ are defined in Section 3.2 and where $\Psi_{\boldsymbol{\beta}}([\xi])$ is defined by Eq. (32). The family $\{[y^{(\kappa)}]\}_\kappa$ of the PCE coefficients are matrices in $\mathbb{M}_{\nu,m}$, which are computed by using the formula,

$$[y^{(\kappa)}] = E\{[\mathbf{Z}]\, \Psi_{\boldsymbol{\beta}^{(\kappa)}}([\boldsymbol{\Xi}])\}\,. \tag{37}$$

The L^2-convergence,

$$[\mathbf{Z}] = \lim_{\substack{\mu\to m \\ N_g\to\nu \\ N_d\to+\infty}} [\mathbf{Z}(N_d, N_g, \mu)]\,, \tag{38}$$

is analyzed in paragraph (ii) below.

It should be noted that the statistical dependence between random matrix $[\boldsymbol{\Xi}]$ and random matrix $[\mathbf{Z}]$ allows the development of very efficient algorithms for computing the coefficients $[y^{(\kappa)}]$ of the PCE of $[\mathbf{Z}]$ using Eq. (37). Indeed, if $[\boldsymbol{\Xi}]$ had been chosen as an independent random matrix of $[\mathbf{Z}]$, then Eq. (37) would give $[y^{(\kappa)}] = [0]$ for all κ, because $E\{[\mathbf{Z}]\} = [0]$, and consequently, could not be used. The maximum likelihood method should then be used for identifying the $K(N_d, N_g, \mu)$ matrix-valued coefficients. This would entail an optimization problem with $\nu \times \mu \times K(N_d, N_g, \mu)$ variables, which would quickly become computationally prohibitive.

(ii) Comments about the L^2- convergence of the truncated PCE of $[\mathbf{Z}]$ when the germ is chosen as $[\boldsymbol{\Xi}]$.

(ii-1) As ν and m are finite, for $N_g = \nu$ and $\mu = m$, the L^2-convergence of the sequence $\{[\mathbf{Z}(N_d, \nu, m)]\}_{N_d}$ of random matrices can be proved as follows. Let $[\boldsymbol{\Xi}(\nu, m)]$ be the Gaussian random matrix with values in $\mathbb{M}_{\nu,m}$ defined in Section 3.2. From the construction of the $\mathbb{M}_{\nu,m}$-valued random matrix $[\mathbf{Z}]$ presented in Section 2.5, $[\mathbf{Z}]$ can be written as $[\mathbf{Z}] = f^{(\nu,m)}([\boldsymbol{\Xi}(\nu, m)])$ in which $[\xi] \mapsto f^{(\nu,m)}([\xi])$ is a measurable mapping from $\mathbb{M}_{\nu,m}$ into $\mathbb{M}_{\nu,m}$. As $[\mathbf{Z}]$ is a second-order random variable, we have

$$\int_{\mathbb{M}_{\nu,m}} \|f^{(\nu,m)}([\xi])\|_F^2 \, p_{[\boldsymbol{\Xi}]}([\xi]) \, d[\xi] \; < \; +\infty \, ,$$

in which $\| \cdot \|_F$ is the Frobenius norm and where $p_{[\boldsymbol{\Xi}]}$ is defined by Eq. (31). On the other hand, it can easily be seen that, for any multi-index $\boldsymbol{\beta}^{(\kappa)}$ with κ in $\mathbb{N}$, the multivariate monomial $\mathcal{M}_{\boldsymbol{\beta}^{(\kappa)}}([\xi])$ defined by Eq. (34) is such that

$$\int_{\mathbb{M}_{\nu,m}} |\mathcal{M}_{\boldsymbol{\beta}^{(\kappa)}}([\xi])| \, p_{[\boldsymbol{\Xi}]}([\xi]) \, d[\xi] \; < \; +\infty \, .$$

Consequently, the mathematical results proved in [35] allow us to conclude that the sequence $\{[\mathbf{Z}(N_d, \nu, m)]\}_{N_d}$ is convergent in L^2 to $[\mathbf{Z}] = f^{(\nu,m)}([\boldsymbol{\Xi}(\nu, m)])$ when $N_d \to +\infty$, that is to say,

$$[\mathbf{Z}] = \lim_{N_d \to +\infty} [\mathbf{Z}(N_d, \nu, m)] \, . \tag{39}$$

(ii-2) The correlation structure that we have retained (see Eq. (30) deduced from Eq. (24)) for $[\boldsymbol{\Xi}]$ shows that the components $(\Xi_1^\alpha, \ldots, \Xi_{N_g}^\alpha)$ of each column $\boldsymbol{\Xi}^\alpha$ with values in $\mathbb{R}^{N_g}$ of random matrix $[\boldsymbol{\Xi}]$ are mutually independent (Gaussian, centered, and not correlated), but the columns $\boldsymbol{\Xi}^1 \ldots \boldsymbol{\Xi}^\mu$ are mutually dependent Gaussian vectors. Consequently, for $\mu = m$, the truncated PCE $[\mathbf{Z}(N_d, N_g, m)]$ of $[\mathbf{Z}]$ with respect to $[\boldsymbol{\Xi}]$ yields a corresponding truncated PCE $\mathbf{Z}^\alpha(N_d, N_g, m)$ of each column $\mathbf{Z}^\alpha$ of $[\mathbf{Z}]$ that depends on $[\boldsymbol{\Xi}] = [\boldsymbol{\Xi}^1 \ldots \boldsymbol{\Xi}^m]$. In particular, the truncated PCE $\mathbf{Z}^\alpha(N_d, N_g, m)$ of column $\mathbf{Z}^\alpha$ with values in $\mathbb{R}^\nu$ depends on $\boldsymbol{\Xi}^\alpha$ with values in $\mathbb{R}^{N_g}$, for which all the components $(\Xi_1^\alpha, \ldots, \Xi_{N_g}^\alpha)$ are independent (because $\mathbf{Z}^\alpha$ depends on $\boldsymbol{\Xi}^1, \ldots, \boldsymbol{\Xi}^\mu$ and consequently, depends on $\boldsymbol{\Xi}^\alpha$). This property and Eq. (39) guarantee that such a truncated PCE $\mathbf{Z}^\alpha(N_d, N_g, m)$ of $\mathbf{Z}^\alpha$ with values in $\mathbb{R}^\nu$ is convergent in L^2 when $N_g \to \nu$ and $N_d \to +\infty$. Consequently, if $[\mathbf{Z}(N_d, N_g, \mu)]$ denotes the truncated PCE of random matrix $[\mathbf{Z}]$, it can

be deduced that, for $\mu = m$, we have the following L^2-convergence,

$$[\mathbf{Z}] = \lim_{\substack{N_g \to \nu \\ N_d \to +\infty}} [\mathbf{Z}(N_d, N_g, m)] \,. \tag{40}$$

(ii-3) However, for μ fixed such that $\mu < m$, due to the dependence of the Gaussian random vectors $\boldsymbol{\Xi}^1, \ldots, \boldsymbol{\Xi}^\mu$, Eqs. (39) and (40) do not prove the convergence in L^2 of the truncated PCE $[\mathbf{Z}(N_d, N_g, \mu)]$ of $[\mathbf{Z}]$ when $N_g \to \nu$ and $N_d \to +\infty$. As the explicit expression of the probability distribution of random matrix $[\mathbf{Z}]$ is unknown, it seems difficult to prove such a convergence. In this case we settle for a numerical exploration using an adapted criterion such that the one defined by Eq. (50)) below. Nevertheless, due to Eq. (40), for $\mu = m$, the given theoretical proof ensures convergence.

3.4. Estimation of the matrix-valued coefficients

In this section, integers n_{MC}, N_g, μ, and N_d are fixed. The construction of the truncated PCE $[\mathbf{Z}(N_d, N_g, \mu)]$ of random matrix $[\mathbf{Z}]$ requires two steps. The first one is the numerical calculation of the samples of the polynomial chaos. The second one is the computation of the PCE coefficients of random matrix $[\mathbf{Z}]$. These steps are detailed next and for simplifying the notation, $K(N_d, N_g, \mu)$ is sometimes simply written as K.

(i) Samples of the polynomial chaos for the computation of the PCE coefficients. Let $\boldsymbol{\Psi}([\boldsymbol{\Xi}]) = (\Psi_{\boldsymbol{\beta}^{(1)}}([\boldsymbol{\Xi}]), \ldots, \Psi_{\boldsymbol{\beta}^{(K)}}([\boldsymbol{\Xi}]))$ be the random vector with values in $\mathbb{R}^K$. The n_{MC} samples of the K multivariate polynomials are represented by the matrix $[\Psi] \in \mathbb{M}_{K,n_{\text{MC}}}$ such that

$$[\Psi]_{\kappa\ell} = \Psi_{\boldsymbol{\beta}^{(\kappa)}}([\xi^\ell]) \quad , \quad \kappa = 1, \ldots, K \quad , \quad \ell = 1, \ldots, n_{\text{MC}} \,, \tag{41}$$

where $[\xi^\ell]$ is defined by Eq. (27) (because we are computing the coefficients $\{[y^{(\kappa)}]\}_\kappa$). Due to Eq. (33), matrix $[\Psi]$ must be such that

$$\lim_{n_{\text{MC}} \to +\infty} \frac{1}{(n_{\text{MC}} - 1)} [\Psi]\, [\Psi]^T = [I_K] \,. \tag{42}$$

(ii) Computation of the coefficients. From Eqs. (37), (41), and (42), it can be deduced that, for n_{MC} sufficiently large (greater than K), the entries $[y^{(\kappa)}]_{k\alpha}$ of an estimation of matrix-valued coefficient $[y^{[\kappa)}]$, can be written, for $\kappa = 1, \ldots, K$, for $k = 1, \ldots, \nu$, and for $\alpha = 1, \ldots, m$, as

$$[y^{(\kappa)}]_{k\alpha} = [\widehat{y}^\alpha]_{k\kappa} \,, \tag{43}$$

$$[\widehat{y}^{\alpha}] \simeq \frac{1}{(n_{\text{MC}}-1)} [\widehat{z}^{\alpha}_{\text{ar}}] \, [\Psi]^T \, . \tag{44}$$

In Eq. (44), the matrix $[\widehat{z}^{\alpha}_{\text{ar}}] \in \mathbb{M}_{\nu, n_{\text{MC}}}$ is defined by

$$[\widehat{z}^{\alpha}_{\text{ar}}]_{k\ell} = [z^{\ell}_{\text{ar}}]_{k\alpha} \quad , \quad \ell = 1, \ldots, n_{\text{MC}} \, , \tag{45}$$

in which $[z^1_{\text{ar}}], \ldots, [z^{n_{\text{MC}}}_{\text{ar}}]$ are the samples of random matrix $[\mathbf{Z}]$, computed with Eq. (19).

3.5. Computation of matrix $[\Psi]$

For the reasons detailed in [44], matrix $[\Psi]$ is computed with the algorithm based on the Cholesky factorization proposed in [39] (we refer the reader to [44] for a method involving the singular value decomposition and to [45] for a method involving the QR factorization). This algorithm preserves the orthogonality property defined by Eq. (42) for the high dimensions and requires that $n_{\text{MC}} > K$.

Let $[\mathbb{M}]$ be the $(K \times n_{\text{MC}})$ real matrix of samples of the monomials defined by Eq. (34) such that

$$[\mathbb{M}]_{\kappa\ell} = \mathcal{M}_{\boldsymbol{\beta}^{(\kappa)}}([\xi^{\ell}]) \quad , \quad \kappa = 1 \ldots, K \quad , \quad \ell = 1, \ldots, n_{\text{MC}} \, , \tag{46}$$

in which $[\xi^{\ell}]$ is defined by Eq. (27) and where $\mathcal{M}_{\boldsymbol{\beta}^{(\kappa)}}([\xi^{\ell}])$ is computed by using Eqs. (34) and (35). The algorithm is then as follows:

- Compute matrix $[\mathbb{M}]$ defined by Eq. (46), and then compute the $(K \times K)$ real matrix $[\mathbb{F}] = \frac{1}{n_{\text{MC}}-1} [\mathbb{M}] \, [\mathbb{M}]^T$. Since $n_{\text{MC}} > K$, then matrix $[\mathbb{F}]$ to be positive definite.

- Compute the lower triangular $(K \times K)$ real matrix $[\mathbb{L}]$ from the Cholesky decomposition $[\mathbb{L}] \, [\mathbb{L}]^T$ of positive-definite symmetric matrix $[\mathbb{F}]$.

- Compute the $(K \times n_{\text{MC}})$ real matrix $[\Psi]$ as the solution of the linear matrix equation $[\mathbb{L}] \, [\Psi] = [\mathbb{M}]$.

It should be noted that, when the rank of matrix $[\mathbb{M}]$ is greater or equal to $K(N_d, N_g, \mu)$, which corresponds to $n_{\text{MC}} > K$, the proposed algorithm shows that the constructed polynomials are exactly orthonormal.

3.6. Computation of n_{MC} samples of $[\mathbf{Z}]$ using its PCE

Let $N_d^{\max} \geq 1$ be the greatest value of N_d, which is considered for performing the L^2-convergence analysis of the PCE of $[\mathbf{Z}]$. We thus consider N_d, N_g, and μ fixed such that $1 \leq N_d \leq N_d^{\max}$, $1 \leq N_g \leq \nu$, and $1 \leq \mu \leq m$ (see Section 3.3-(i)). We are thus interested in computing n_{MC} samples $[z_{\text{PCE}}^1], \ldots, [z_{\text{PCE}}^{n_{\text{MC}}}]$ of the random matrix $[\mathbf{Z}(N_d, N_g, \mu)]$ such that, for $\ell = 1, \ldots, n_{\text{MC}}$,

$$[z_{\text{PCE}}^{\ell}]_{k\alpha} = [\widehat{z}_{\text{PCE}}^{\alpha}]_{k\ell} \quad , \quad k = 1, \ldots, \nu \quad , \quad \alpha = 1, \ldots, m \, , \tag{47}$$

in which, for $\alpha = 1, \ldots, m$, the matrix $[\widehat{z}_{\text{PCE}}^{\alpha}] \in \mathbb{M}_{\nu, n_{\text{MC}}}$ is computed by

$$[\widehat{z}_{\text{PCE}}^{\alpha}] = [\widehat{y}^{\alpha}] \, [\Psi] \, . \tag{48}$$

The matrix $[\widehat{y}^{\alpha}] \in \mathbb{M}_{m,K}$ is given by

$$[\widehat{y}^{\alpha}]_{k\kappa} = [y^{(\kappa)}]_{k\alpha} \quad , \quad k = 1, \ldots, \nu \quad , \quad \kappa = 1, \ldots, K \, , \tag{49}$$

in which matrix $[y^{(\kappa)}] \in \mathbb{M}_{\nu,m}$ has been computed by using Eqs. (43) to (45) and where The matrix $[\Psi] \in \mathbb{M}_{K,n_{\text{MC}}}$ has been computed by using Eq. (41).

3.7. Quantification of the L^2-convergence

The quantification of the L^2-convergence of the sequence of random matrices $\{[\mathbf{Z}(N_d, N_g, \mu)]\}_{N_d, N_g, \mu}$ towards $[\mathbf{Z}]$ (see Eq. (38)) is performed by constructing the error function $(N_d, N_g, \mu) \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ defined on $[1, N_d^{\max}] \times [1, \nu] \times [1, m]$ by

$$\text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu) = \sqrt{\frac{\sum_{\ell=1}^{n_{\text{MC}}} \| [z_{\text{ar}}^{\ell}] - [z_{\text{PCE}}^{\ell}] \|_F^2}{\sum_{\ell=1}^{n_{\text{MC}}} \| [z_{\text{ar}}^{\ell}] \|_F^2}} \, . \tag{50}$$

4. Generating additional samples of random vector X using its analytical representation and L^2-error

For N_d, N_g, and μ fixed to the values identified by the analysis of the error function $(N_d, N_g, \mu) \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ defined by Eq. (50), the number $K = K(N_d, N_g, \mu)$ of coefficients is known. For a given number n_{sim} of additional samples, the methodology consists of

- computing n_{sim} samples of the analytical representation $\mathbf{X}(N_d, N_g, \mu)$ of random vector $\mathbf{X}$ by using Eq. (6) and the PCE $[\mathbf{Z}(N_d, N_g, \mu)]$ of random matrix $[\mathbf{Z}]$;
- estimating the L^2-error for $\mathbf{X}(N_d, N_g, \mu)$ with respect to random vector $\mathbf{X}$.

*4.1. Algorithm for computing samples of random vector **X** using the PCE of random matrix [**Z**]*

For $\alpha = 1, \ldots, m$, the matrices $[\widehat{y}^{\alpha}] \in \mathbb{M}_{m,K}$ are known (computed with Eq. (44)). The algorithm for generating n_{sim} independent samples of the analytical representation $\mathbf{X}(N_d, N_g, \mu)$ of random vector $\mathbf{X}$ is deduced from the developments presented in Section 3 and is detailed hereinafter.

- By using the results presented in Section 3.1, n_{sim} independent samples $\{[\xi^{\ell}_{\text{sim}}], \ell = 1, \ldots, n_{\text{sim}}\}$ of random matrix $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$ are calculated as follows:
  1. Computation of the $(N_g \times N)$ real matrix $[u^{\ell}] = \sqrt{\Delta r}\,[\gamma^{\ell}]$ in which $[\gamma^{\ell}]$ is an independent sample of a normalized Gaussian random matrix $[\Gamma]$ with values in $\mathbb{M}_{N_g,N}$. This means that all the entries $[\Gamma]_{kj}$ are independent normalized Gaussian real-valued random variables. For instance by using Matlab, the generation is written as $[\gamma^{\ell}] = \text{randn}(N_g, N)$.
  2. Computation of the $(N_g \times m)$ real matrix $[w^{\ell}] = [u^{\ell}]\,[a]$ in which $[a]$ is the $(N \times m)$ real matrix defined by Eq. (4).
  3. Computation of the sample $[\xi^{\ell}_{\text{sim}}]$ of random matrix $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$ such that, for $k = 1, \ldots, N_g$ and $\alpha = 1, \ldots, \mu$, $[\xi^{\ell}_{\text{sim}}]_{k\alpha} = \{[w^{\ell}]\,[d\,]\}_{k\alpha}$ in which matrix $[d\,]$ is defined by Eq. (26).

- Computation of $[\Psi_{\text{sim}}] \in \mathbb{M}_{K,n_{\text{sim}}}$ by using Section 3.5 in replacing n_{MC} by n_{sim} and $[\xi^{\ell}]$ by $[\xi^{\ell}_{\text{sim}}]$).

- Computation of the samples $[z^1_{\text{PCE,sim}}], \ldots, [z^{n_{\text{sim}}}_{\text{PCE,sim}}]$ in $\mathbb{M}_{\nu,m}$ of the random matrix $[\mathbf{Z}(N_d, N_g, \mu)]$ with values in $\mathbb{M}_{\nu,m}$, such that, for $\ell = 1, \ldots, n_{\text{sim}}$, for $k = 1, \ldots, \nu$, and for $\alpha = 1, \ldots, m$, we have $[z^{\ell}_{\text{PCE,sim}}]_{k\alpha} = [\widehat{z}^{\alpha}_{\text{PCE,sim}}]_{k\ell}$ with $[\widehat{z}^{\alpha}_{\text{PCE,sim}}] = [\widehat{y}^{\alpha}]\,[\Psi_{\text{sim}}]$ (see Eqs. (47) and (48)), in which $[\widehat{y}^{\alpha}] \in \mathbb{M}_{m,K}$ has previously been computed with Eq. (44).

- Computation of the additional samples $[x^1_{\text{PCE,sim}}], \ldots, [x^{n_{\text{sim}}}_{\text{PCE,sim}}]$ in $\mathbb{M}_{n,N}$ of random matrix $[\mathbf{X}(N_d, N_g, \mu)]$ with values in $\mathbb{M}_{n,N}$ such that, for $\ell = 1, \ldots n_{\text{sim}}$, we have $[x^{\ell}_{\text{PCE,sim}}] = [\underline{x}] + [\varphi]\,[\lambda]^{1/2}\,[z^{\ell}_{\text{PCE,sim}}]\,[g]^T$ (see Eq. (6)).

- Deducing the $n_{\text{sim}} \times N$ additional samples $\mathbf{x}^1_{\text{PCE,sim}}, \ldots, \mathbf{x}^{n_{\text{sim}} \times N}_{\text{PCE,sim}}$ in $\mathbb{R}^n$ of random vector $\mathbf{X}(N_d, N_g, \mu)$ with values in $\mathbb{R}^n$ such that, for $\ell = 1, \ldots, n_{\text{sim}} \times N$ and for $k = 1, \ldots, n$, the component $x^{\ell}_{\text{PCE,sim},k}$ of $\mathbf{x}^{\ell}_{\text{PCE,sim}}$ is such that $x^{\ell}_{\text{PCE,sim},k} = [\mathbb{x}^{\ell}_{\text{PCE,sim}}]_k$ in which $[\mathbb{x}^{\ell}_{\text{PCE,sim}}]$ is the matrix in $\mathbb{M}_{n,n_{\text{sim}} \times N}$ defined by $[\mathbb{x}^{\ell}_{\text{PCE,sim}}] = [\,[x^1_{\text{PCE,sim}}] \ \ldots \ [x^{n_{\text{sim}}}_{\text{PCE,sim}}]\,]$.

4.2. Estimation of the L^2-error for $\boldsymbol{X}(N_d, N_g, \mu)$ with respect to random vector $\boldsymbol{X}$

For given N_d, N_g, μ, and n_{sim}, the L^2-error of the random vector $\mathbf{X}(N_d, N_g, \mu)$ with respect to random vector $\mathbf{X}$ is estimated by

$$\text{error}_{\mathbf{X}}(N_d, N_g, \mu) = \frac{\| [\text{cov}_{\mathbf{X}(N_d,N_g,\mu)}] - [\text{cov}_{\mathbf{X}}] \|_F}{\| [\text{cov}_{\mathbf{X}}] \|_F}, \tag{51}$$

in which the covariance matrix $[\text{cov}_{\mathbf{X}(N_d,N_g,\mu)}] \in \mathbb{M}_n$ of the random vector $\mathbf{X}(N_d, N_g, \mu)$ is estimated with the $n_{\text{sim}} \times N$ samples $\mathbf{x}^1_{\text{PCE,sim}}, \ldots, \mathbf{x}^{n_{\text{sim}} \times N}_{\text{PCE,sim}}$, and where the covariance matrix $[\text{cov}_{\mathbf{X}}] \in \mathbb{M}_n$ of the random vector $\mathbf{X}$ is estimated with the $n_{\text{MC}} \times N$ samples $\mathbf{x}^1_{\text{ar}}, \ldots, \mathbf{x}^{n_{\text{MC}} \times N}_{\text{ar}}$ that are the columns of the matrices $[x^1_{\text{ar}}], \ldots [x^{n_{\text{MC}}}_{\text{ar}}]$, that are computed using Section 2 (see Eq. (20)). It should be noted that the number of samples for random vectors $\mathbf{X}$ and $\mathbf{X}(N_d, N_g, \mu)$ are not the same and, in addition, $\mathbf{X}$ and $\mathbf{X}(N_d, N_g, \mu)$ are statistically independent. Consequently, the classical L^2-norm cannot be introduced for quantifying the convergence, and since $\mathbf{X}$ and $\mathbf{X}(N_d, N_g, \mu)$ have the same mean vectors, the distance between the covariance matrices is used.

5. Applications

Three applications presented in [1] are partially reused for illustrating the methodology proposed. The three examples consist of obtaining samples of random vector $\mathbf{X}$ with values in $\mathbb{R}^n$ such that:

1. in example 1, the dimension is $n = 2$ and there are $N = 230$ given data points in subset $\mathcal{S}_n$, for which the mean value is made up of two circles in the plane).
2. in example 2, the dimension is $n = 3$ and there are $N = 400$ given data points in subset $\mathcal{S}_n$, for which the mean value is made up of a helix in three-dimensional space).
3. the third example corresponds to a petro-physics database that is made up of experimental measurements (downloaded from [46]) and detailed in [47], for which the dimension is $n = 35$ and for which $N = 13,056$ given data points are concentrated in an unknown "complex" subset $\mathcal{S}_n$ of $\mathbb{R}^n$, which cannot be easily described once discovered.

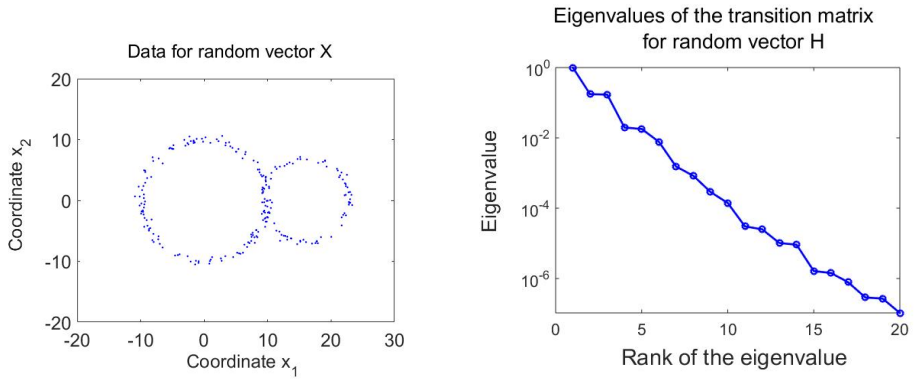

Figure 1: 230 given data points (left). Eigenvalues of the transition matrix for random vector **H** (right).

5.1. Application 1: Dimension $n = 2$ with $N = 230$ given data points

For this first application, the statistical fluctuations are around two circles. The number of given data points is $N = 230$, no scaling of data is performed, but the normalization defined in Section 2.1 is done. Fig. 1 (left) displays the 230 given data points for random vector $\mathbf{X} = (X_1, X_2)$ of the dataset represented by matrix $[x^d]$ in $\mathbb{M}_{2,230}$, and shows that the given data points are concentrated in the neighborhood of two circles. The kernel is defined in Section 2.2, the value used for the smoothing parameter is $\varepsilon = 100 \times 2\pi/N = 2.73$, parameter ζ defined in Appendix A is chosen to 1, and the graph of the eigenvalues of the transition matrix for random vector **H** is displayed in Fig. 1 (right). This graph shows that dimension m can be chosen to be 3, for which the value of $e_{\text{red}}(m)$ (defined by Eq. (B.4)) is 6.96×10^{-4} (it can thus be considered that a reasonable mean-square convergence has been reached). Fig. 2 (left) displays the pdf for random variables X_1 and X_2 computed with a nonparametric estimation from the data points. For generating $9,200$ additional samples, the numerical values of the parameters are $f_0 = 1.5$, $\Delta r = 0.1179$, $M_0 = 110$, and $n_{\text{MC}} = 40$, yielding $4,400$ for the parameter M defined in Appendix C. The additional samples $[x^1_{\text{ar}}], \ldots, [x^{n_{\text{MC}}}_{\text{ar}}]$ (with $n_{\text{MC}} = 40$) are computed by using Section 2.6 and are displayed in Fig. 2 (right), which shows the 230 given data points and the $9,200$ additional samples generated using the reduced-order ISDE. It can be seen that the additional samples are effectively concentrated in subset $\mathcal{S}_n$ (as it has be shown in [1], if a direct simulation was used instead of the method proposed, then the $9,200$ additional samples would not be concentrated in subset $\mathcal{S}_n$, but would be scattered).

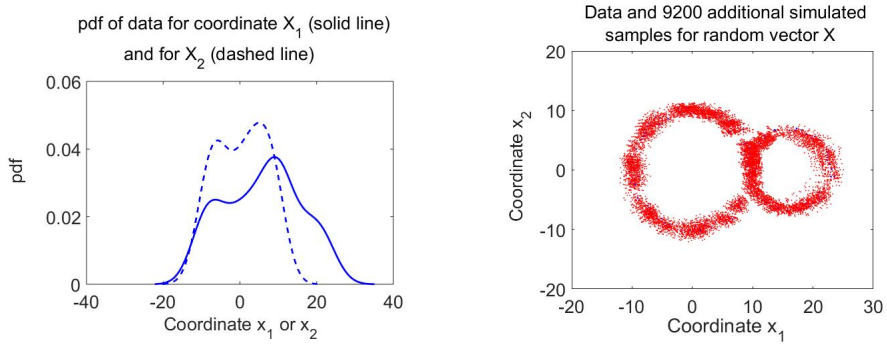

Figure 2: pdf for random variables X_1 (solid line) and X_2 (dashed line) obtained by a nonparametric estimation from data points (left). 230 given data points (blue symbols) and $9,200$ additional samples (red symbols) generated using the reduced-order ISDE with $m = 3$ (right).

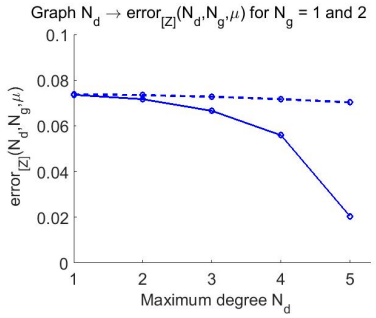

Figure 3: Graph of the error function $N_d \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ for $N_g = 1$ (dashed line) and for $N_g = 2$ (solid line), and for $\mu = m = 3$.

The analytical representation of $[\mathbf{X}]$ is constructed by using the methodology presented in Sections 3 and 4. All the computations are performed with $f_0 = 1.5$, $\Delta r = 0.1179$, $M_0 = 110$, and $n_{\text{MC}} = 500$. The value of μ is fixed such that $\mu = m = 3$. Fig. 3 displays the graph of the error function $(N_d, N_g) \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ defined by Eq. (50) for $N_d = 1, \ldots, 5$, for $N_g = 1$ and 2, and for $\mu = 3$. For $N_d = 5$, for $N_g = 2$, and for $\mu = 3$, the number of vector-valued coefficients is $K(N_d, N_g, \mu) = 462$, and the value of the error function (defined by Eq. (51)) is $\text{error}_{\mathbf{X}}(N_d, N_g, \mu) = 7.58 \times 10^{-4}$. The convergence in probability distribution of the components X_1 and X_2 of random vector $\mathbf{X}$ is shown in Fig. 4, which displays the pdf of X_1 (left figure) and the pdf of X_2 (right figure), estimated by using $[x^1_{\text{ar}}], \ldots, [x^{n_{\text{MC}}}_{\text{ar}}]$ (reference pdf, computed with $n_{\text{MC}} = 40$) and estimated using $[x^1_{\text{PCE,sim}}], \ldots, [x^{n_{\text{MC}}}_{\text{PCE,sim}}]$ (pdf computed with the an-

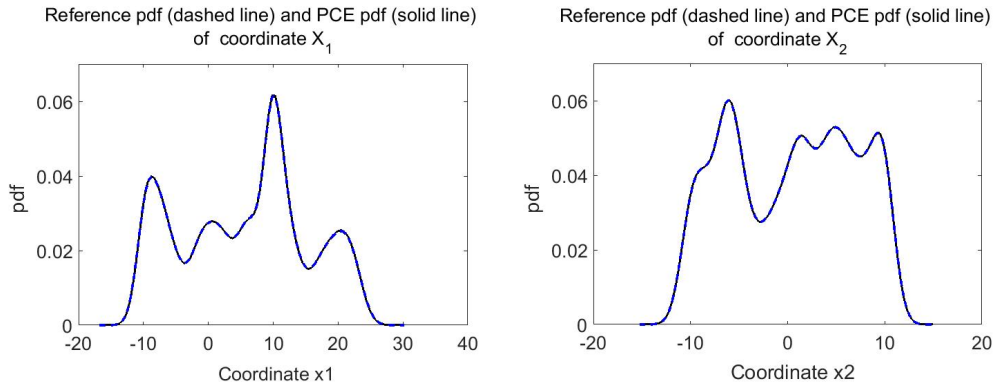

Figure 4: For $N_d = 5$, $N_g = 2$, and $\mu = m = 3$, graphs of the reference pdf (dashed line) and of the pdf computed with the analytical representation (solid line) for random variables X_1 (left figure) and X_2 (right figure). In each figure, the dashed line and the solid line are superimposed.

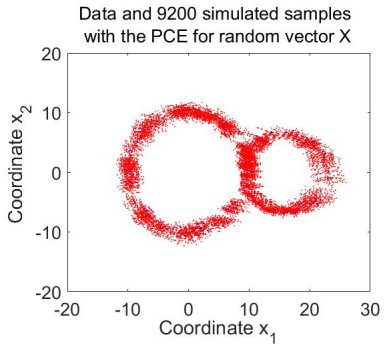

Figure 5: 230 given data points (blue symbols) and $9,200$ additional samples (red symbols) generated using the analytical representation.

alytical representation for $n_{\text{MC}} = 500$). It can be seen that the convergence is excellent. The 40-first additional samples $[x^1_{\text{PCE,sim}}], \ldots, [x^{40}_{\text{PCE,sim}}]$ are displayed in Fig. 5, which shows the 230 initial data points and the $9,200$ additional samples generated using the analytical representation. It can be seen that the samples are effectively concentrated in subset $\mathcal{S}_n$.

5.2. Application 2: Dimension $n = 3$ with $N = 400$ given data points

The number of given data points is $N = 400$, no scaling of data is performed, but the normalization defined in Section 2.1 is done. Fig. 6 (left) displays the 400 given data points for random vector $\mathbf{X} = (X_1, X_2, X_3)$ of the dataset represented

by matrix $[x^d]$ in $\mathbb{M}_{3,400}$, and shows that the given data points are concentrated in the neighborhood of a helical. The kernel is defined in Section 2.2, the value of the smoothing parameter that is retained is $\varepsilon = 100 \times 2\pi/N = 1.57$, parameter ζ defined in Appendix A is set to 1, and the graph of the eigenvalues of the transition matrix for random vector $\mathbf{H}$ is displayed in Fig. 6 (right). This graph shows that dimension m can be chosen to be 4 for which the value of $e_{\text{red}}(m)$ (defined by Eq. (B.4)) is 5.48×10^{-4} (it can thus be considered that a reasonable mean-square convergence has been reached). Fig. 7 (left) displays the pdf for random variables

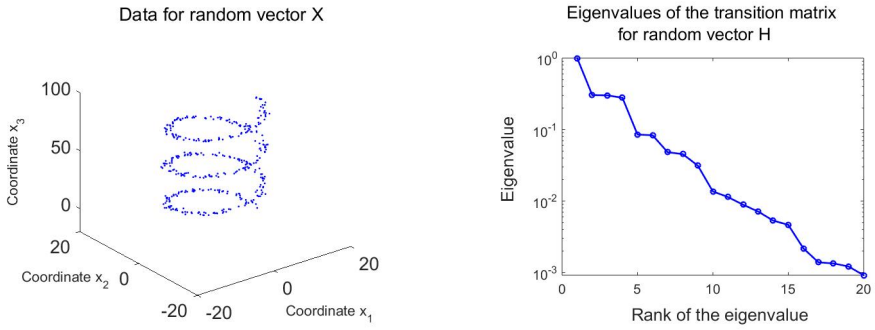

Figure 6: 400 given data points (left). Eigenvalues of the transition matrix for random vector $\mathbf{H}$(right).

X_1, X_2, and X_3 computed with a nonparametric estimation from the data points. For generating $8,000$ additional samples, the numerical values of the parameters are $f_0 = 1.5$, $\Delta r = 0.1179$, $M_0 = 110$, and $n_{\text{MC}} = 20$, yielding $2,200$ for the parameter M defined in Appendix C. The additional samples $[x^1_{\text{ar}}], \ldots, [x^{n_{\text{MC}}}_{\text{ar}}]$ (with $n_{\text{MC}} = 20$) are computed by using Section 2.6 and are displayed in Fig. 7 (right), which shows the 400 given data points and the $8,000$ additional samples generated using the reduced-order ISDE. It can be seen that the additional samples are effectively concentrated in subset $\mathcal{S}_n$ (as it has be shown in [1], if a direct simulation was used instead of the method proposed, then the $8,000$ additional samples would not be concentrated in subset $\mathcal{S}_n$, but would be scattered).

The analytical representation of $[\mathbf{X}]$ is constructed using the methodology presented in Sections 3 and 4. All the computations are performed with $f_0 = 1.5$, $\Delta r = 0.1197$, $M_0 = 110$, and $n_{\text{MC}} = 2,000$. The value of μ is fixed such that $\mu = m = 4$. Fig. 8 displays the graph of the error function $(N_d, N_g) \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ defined by Eq. (50) for $N_d = 1, \ldots, 4$, for $N_g = 1, 2, 3$, and

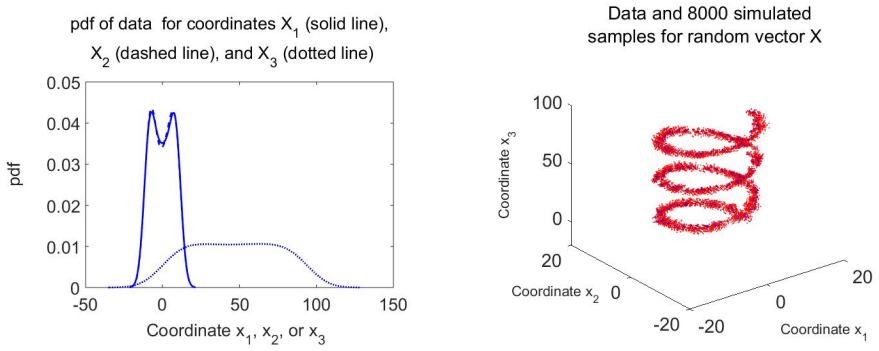

Figure 7: pdf for random variables X_1 (solid line), X_2 (dashed line), and X_3 (dotted line) obtained by a nonparametric estimation from data points (left). 400 given data points (blue symbols) and $8,000$ additional samples (red symbols) generated using the reduced-order ISDE with $m = 4$ (right).

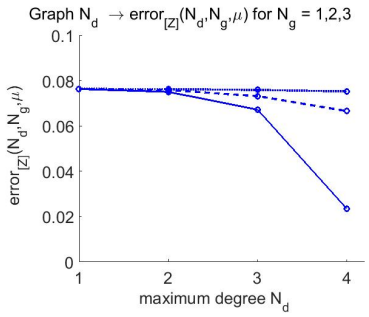

Figure 8: Graph of the error function $N_d \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ for $N_g = 1$ (dotted line), $N_g = 2$ (dashed line), $N_g = 3$ (solid line), and for $\mu = m = 4$. In each figure, the dashed line and the solid line are superimposed.

for $\mu = 4$. For $N_d = 4$, for $N_g = 3$, and for $\mu = 4$, the number of vector-valued coefficients is $K(N_d, N_g, \mu) = 1,820$, the value of the error function (defined by Eq. (51)) is $\text{error}_{\mathbf{X}}(N_d, N_g, \mu) = 2.54 \times 10^{-4}$, and the convergence in probability distribution of the components X_1, X_2, and X_3 of random vector $\mathbf{X}$ is shown in Fig. 9, which displays the pdf of X_1 (left figure), the pdf of X_2 (central figure), and the pdf of X_3 (right figure), estimated using $[x^1_{\text{ar}}], \ldots, [x^{n_{\text{MC}}}_{\text{ar}}]$ (reference pdf, computed with $n_{\text{MC}} = 20$) and estimated using $[x^1_{\text{PCE,sim}}], \ldots, [x^{n_{\text{MC}}}_{\text{PCE,sim}}]$ (pdf computed with the analytical representation for $n_{\text{MC}} = 2,000$). It can be seen that the convergence is good. The 20-first additional samples $[x^1_{\text{PCE,sim}}], \ldots, [x^{20}_{\text{PCE,sim}}]$ are displayed in Fig. 10, which shows the 400 given data points and the $8,000$ additional sam-

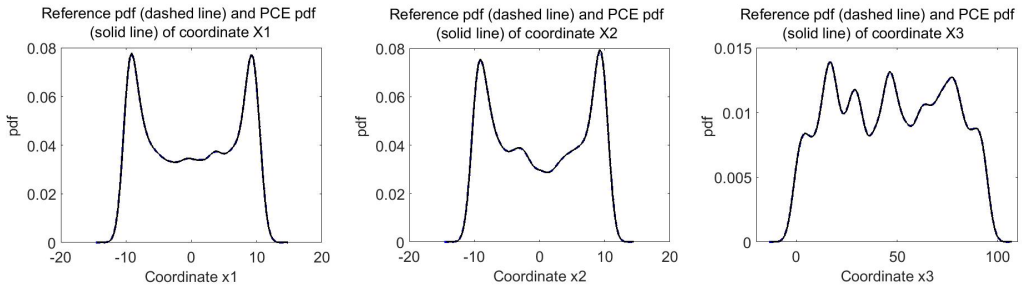

Figure 9: For $N_d = 4$, $N_g = 3$, and $\mu = m = 4$, graphs of the reference pdf (dashed line) and of the pdf computed with the analytical representation (solid line) for random variables X_1 (left figure), X_2 (central figure), and X_3 (right figure).

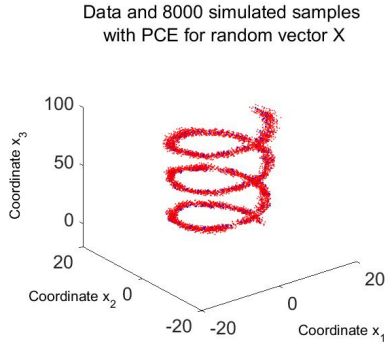

Figure 10: 400 given data points (blue symbols) and $8,000$ additional samples (red symbols) generated using the analytical representation.

ples generated using the analytical representation. It can be seen that the samples are effectively concentrated in subset $\mathcal{S}_n$.

5.3. Application 3: Dimension $n = 35$ with $N = 13,056$ given data points

The database used corresponds to petro-physics field observations. The dimension of random vector $\mathbf{X}$ is $n = 35$ and the number of given data points is $N = 13,056$. The scaling defined in Section 2.1 is necessary and has been done. The scaled data points are then represented by the matrix $[x_d]$ in $\mathbb{M}_{35,13056}$. The normalization is performed using Eq. (1) for which $\nu = 32$. Figure 11 displays the $13,056$ given data points viewed from coordinates x_{16} and x_{28}, from coordinates x_{27} and x_{28}, and from coordinates x_{30}, x_{32}, and x_{33}. Although only a partial representation of the data points are shown in these three figures, it can be seen that $\mathcal{S}_n$ is certainly a complex subset of $\mathbb{R}^n$. The kernel is defined in Section 2.2, the value of the

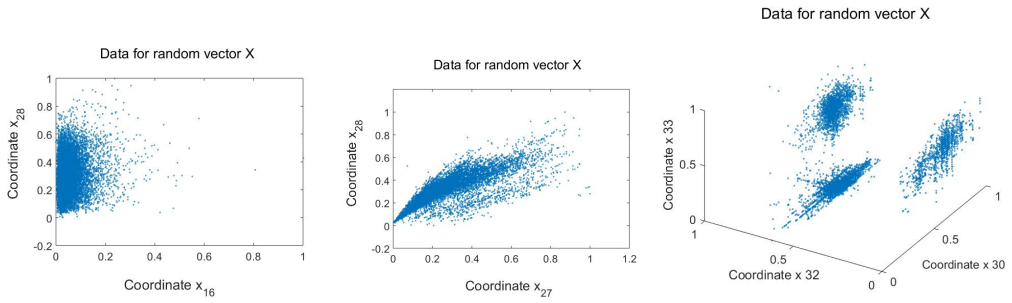

Figure 11: $13,056$ given data points viewed from coordinates x_{16} and x_{28} (left), viewed from coordinates x_{27} and x_{28} (center), and viewed from coordinates x_{30}, x_{32}, and x_{33} (right).

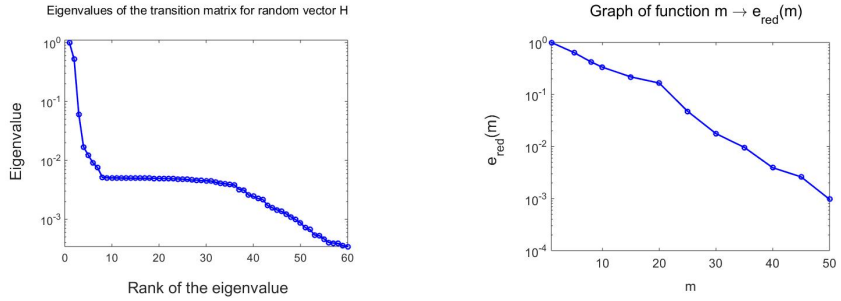

Figure 12: Eigenvalues of the transition matrix for random vector $\mathbf{H}$ (left). Graph $m \mapsto e_{\text{red}}(m)$ in $\log_{10}$ scale (right).

smoothing parameter that is retained is $\varepsilon = 150$, parameter ζ defined in Appendix A is chosen to 1, and the graph of the eigenvalues of the transition matrix for random vector $\mathbf{H}$ is displayed in Fig. 12 (left). This figure shows that the value $m = 8$ could potentially be used, because it corresponds to a good value for the identification of the geometry of the support. However, for $m = 8$, the value of $e_{\text{red}}(m)$ (defined by Eq. (B.4)) is 0.422 indicating that mean-square convergence is not reached, which means that the convergence is not yet obtained with respect to the probability distribution of $[\mathbf{H}]$. Fig. 12 (right) displays the graph of the function $m \mapsto e_{\text{red}}(m)$. The mean-square convergence is reasonably reached for $m = 50$, for which the value of $e_{\text{red}}(m)$ is 9.69×10^{-4}. Consequently, the value $m = 50$ has been selected and random matrix $[\mathbf{Z}]$ has values in $\mathbb{M}_{\nu,m}$. The construction of germ $[\boldsymbol{\Xi}]$ with values in $\mathbb{M}_{N_g,\mu}$ is performed with $\mu = 8 \ll m = 50$, a consistent choice which limits the numerical cost of performing a convergence analysis with respect to N_g and N_d. As noted and explained below, good convergence with respect to

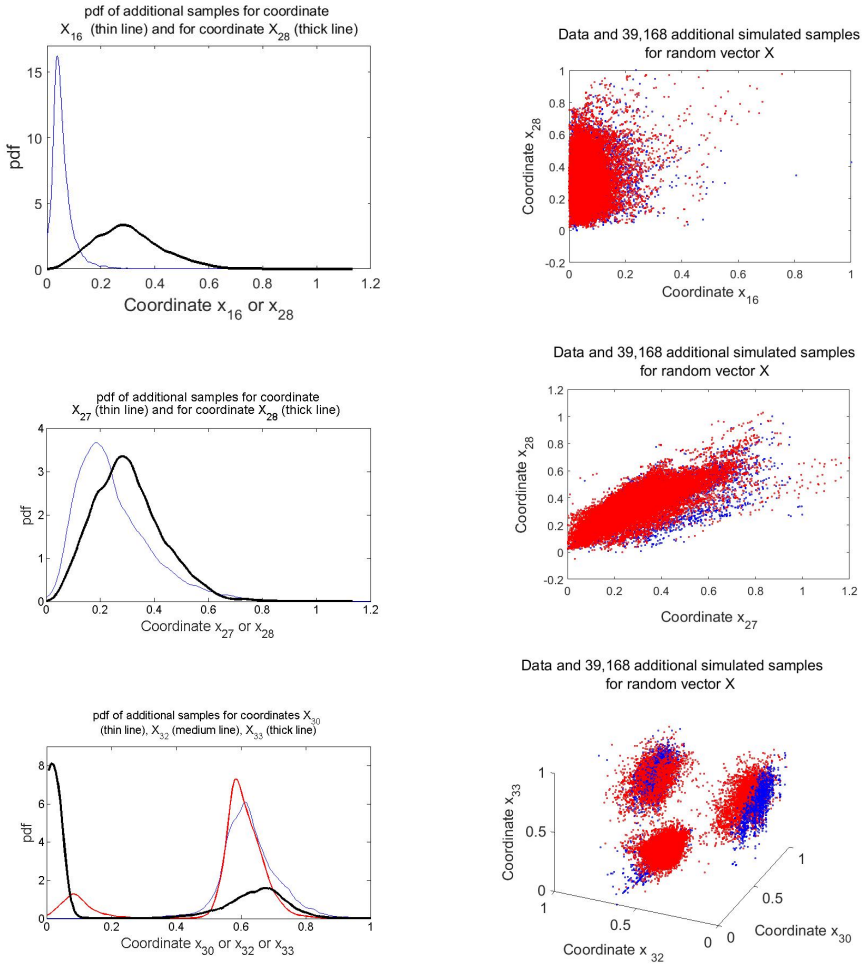

Figure 13: Left figures: pdf for some components of random vector $\mathbf{X}$ estimated from the data points $[x_d]$ (dashed lines) and estimated with the simulated samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ (solid lines). Right figures: $13,056$ given data points (blue symbols) and $39,168$ simulated samples (red symbols).

N_g and N_d is achieved for $\mu = 8$. No additional computation are thus performed for $8 < \mu \leq 50$.

In order to limit the volume of data presented in the figures, only $n_{\text{MC}} = 3$ samples of random matrix $[\mathbf{X}]$ with values in $\mathbb{M}_{35,13056}$ are displayed, yielding $39,168$ samples of random vector $\mathbf{X}$. The simulated samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ of $[\mathbf{X}]$ are

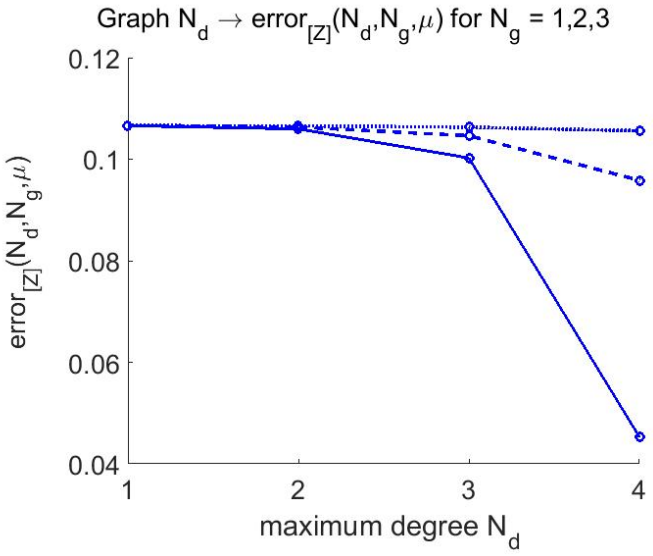

Figure 14: Graph of the error function $N_d \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ for $N_g = 1$ (dotted line), $N_g = 2$ (dashed line), $N_g = 3$ (thin solid line), and for $\mu = 8$.

computed by using Eq. (20) for which the numerical values of the parameters (defined in Appendix C for solving the reduced-order ISDE) are $\Delta r = 0.06142$, $M_0 = 330$, and $n_{\text{MC}} = 3$, yielding $M = 990$. For the same coordinates as those introduced in Fig. 11, the left figures in Fig. 13 display the pdf of the considered components of random vector $\mathbf{X}$ estimated using the data points $[x_d]$ and estimated with the simulated samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ (with $n_{\text{MC}} = 3$). The right figures in Fig. 13 display the $13,056$ given data points $[x_d]$ and the $39,168$ simulated samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ (with $n_{\text{MC}} = 3$). It can be seen that these simulated samples are effectively concentrated in subset $\mathcal{S}_n$.

The analytical representation is constructed using the methodology presented in Section 3. All the computations are performed with $f_0 = 1.5$, $\Delta r = 0.06142$, $M_0 = 1$, and $n_{\text{MC}} = 25,000$ (the ergodic property is used for estimating the coefficients of the PCE). Fig. 14 displays the graph of the error function $(N_d, N_g) \mapsto \text{error}_{[\mathbf{Z}]}(N_d, N_g, \mu)$ defined by Eq. (50) for $N_d = 1, \ldots, 4$, for $N_g = 1, \ldots, 3$, and for $\mu = 8$ (for the reason given before). For $N_d = 4$, $N_g = 3$, and $\mu = 8$, the number of vector-valued coefficients is $K(N_d, N_g, \mu) = 20,475$ and the value of the error function (defined by Eq. (51)) is $\text{error}_{\mathbf{X}}(N_d, N_g, \mu) = 6.6 \times 10^{-3}$. The same components X_{16}, X_{27}, X_{28}, X_{30}, X_{32}, and X_{33} that those used in Fig. (13)

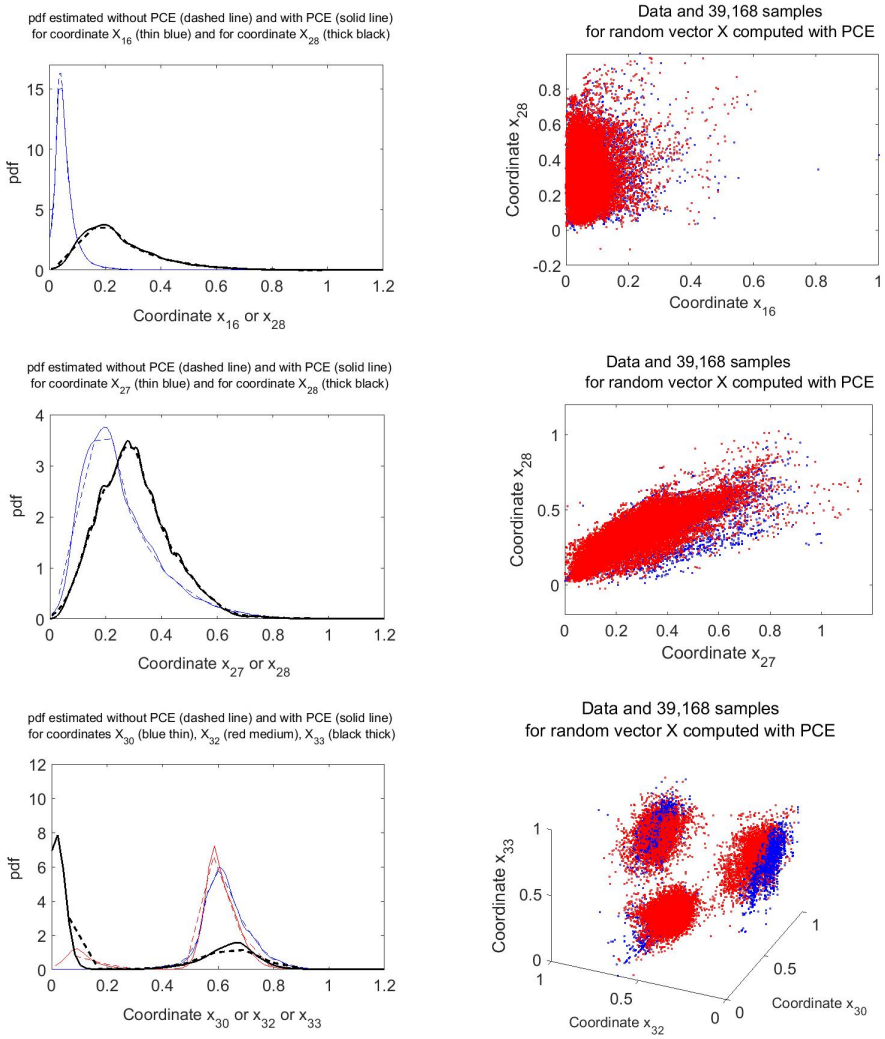

Figure 15: Left figures: pdf for some components of random vector $\mathbf{X}$ estimated without the analytical representation (dashed lines) and estimated with the simulated samples $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ (solid lines). Right figures: $13,056$ given data points (blue symbols) and $39,168$ simulated samples (red symbols).

are observed for presenting the results obtained by using the analytical representation. Figure (15) displays the pdf of X_{16} and X_{28} (left top figure), the pdf of X_{27} and X_{28} (left central figure), and the pdf of X_{30}, X_{32}, and X_{33} (left bottom figure), estimated using $[x_{\text{ar}}^1], \ldots, [x_{\text{ar}}^{n_{\text{MC}}}]$ (reference pdf, computed with $n_{\text{MC}} = 3$) and estimated using $[x_{\text{PCE,sim}}^1], \ldots, [x_{\text{PCE,sim}}^{n_{\text{MC}}}]$ (pdf computed with the analytical repre-

sentation for $n_{\text{MC}} = 25,000$). It can be seen that the convergence (in measure) of the analytical representation of $\mathbf{X}$ is good (the dashed lines are superimposed to the solid lines). In Fig. 15, each one of the three right figures displays the $13,056$ given data points $[x_d]$ and the $39,168$ samples corresponding to the 3-first samples $[x^1_{\text{PCE,sim}}], \ldots, [x^3_{\text{PCE,sim}}]$ generated with the analytical representation. It can be seen that the samples computed with the analytical representation are effectively concentrated in subset $\mathcal{S}_n$.

6. Conclusions

Starting with a dataset concentrated around a manifold, a new methodology has been presented and validated for constructing a probabilistic characterization of the dataset in the form of a polynomial chaos expansion. Mathematically, this takes the form of a known affine transformation of a matrix-valued random variable for which its polynomial chaos expansion is constructed and is concentrated on the manifold that is identified from the dataset. The proposed methodology is robust and can be used for high dimension and for large given datasets. In the first article, we have proposed a method for constructing a generator of samples on the manifold on which the dataset lives. In this paper, we have proposed an analytical characterization of the dataset on the manifold, which also allows for generating samples on the manifold, which is easy to implement. In addition, a germ is characterized on the manifold, which allows other processes to be constructed on the manifold and which also allows for implementing the spectral methods for any nonlinear transformations defined on the manifold.

7. Acknowledgments

This research was supported by the ScramJet-UQ project funded under DARPA's EQUIPS Program.

Appendix A. Construction of the diffusion-maps basis

In this appendix, we summarize the construction of the diffusion map-basis based on [2, 12] and detailed in [1]. Let $[b]$ be the positive-definite diagonal real matrix in $\mathbb{M}_N$ such that $[b]_{ij} = \delta_{ij} \sum_{j'=1}^{N} [K]_{ij'}$ in which $[K]_{ij'} = k_\varepsilon(\boldsymbol{\eta}^{d,i}, \boldsymbol{\eta}^{d,j'})$. Let $[\mathbb{P}]$ be the transition matrix in $\mathbb{M}_N$ such that $[\mathbb{P}] = [b]^{-1}[K]$ and let $[\mathbb{P}_S]$ be the symmetric matrix in $\mathbb{M}_N$ such that $[\mathbb{P}_S] = [b]^{1/2}[\mathbb{P}][b]^{-1/2} = [b]^{-1/2}[K][b]^{-1/2}$. Let m be an integer such that $1 < m \leq N$. The eigenvalues of $[\mathbb{P}_S]\,\boldsymbol{\phi}^\alpha = \Lambda_\alpha\,\boldsymbol{\phi}^\alpha$

are positive and such that $1 = \Lambda_1 > \Lambda_2 \geq \ldots \geq \Lambda_m$. Let $[\phi]$ be the matrix in $\mathbb{M}_{N,m}$ such that $[\phi]^T [\phi] = [I_m]$, whose columns are the m orthonormal eigenvectors $\boldsymbol{\phi}^1, \ldots, \boldsymbol{\phi}^m$ associated with $\Lambda_1, \ldots, \Lambda_m$. The right eigenvectors $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^m$ associated with $\Lambda_1, \ldots, \Lambda_m$, which are such that $[\mathbb{P}]\, \boldsymbol{\psi}^\alpha = \Lambda_\alpha\, \boldsymbol{\psi}^\alpha$, are written as

$$\boldsymbol{\psi}^\alpha = [b]^{-1/2}\, \boldsymbol{\phi}^\alpha \in \mathbb{R}^N \quad , \quad \alpha = 1, \ldots, m\,, \tag{A.1}$$

and consequently, the matrix $[\psi] = [\boldsymbol{\psi}^1 \ldots \boldsymbol{\psi}^m] = [b]^{-1/2}\, [\phi] \in \mathbb{M}_{N,m}$ is such that

$$[\psi]^T\, [b]\, [\psi] = [I_m]\,, \tag{A.2}$$

which defined the normalization of the right eigenvectors of $[\mathbb{P}]$. A "diffusion-maps basis" is defined by $[g] = [\mathbf{g}^1 \ldots \mathbf{g}^m] \in \mathbb{M}_{N,m}$ (which is an algebraic basis of $\mathbb{R}^N$ for $m = N$) such that

$$\mathbf{g}^\alpha = \Lambda_\alpha^\zeta\, \boldsymbol{\psi}^\alpha \in \mathbb{R}^N \quad , \quad \alpha = 1, \ldots, m\,, \tag{A.3}$$

in which ζ is an integer that is chosen for fixing the analysis scale of the local geometric structure of the dataset. It should be noted that the family $\{\Psi_\zeta\}_\zeta$ of diffusion maps are defined [2, 12] by the vector $\Psi_\zeta = (\Lambda_1^\zeta\, \boldsymbol{\psi}^1, \ldots, \Lambda_m^\zeta\, \boldsymbol{\psi}^m)$ in order to construct a diffusion distance, and integer ζ is thus such that the probability of transition is in ζ steps. However, as it has been explained in [1], we do not use such a diffusion distance.

Appendix B. Criterion for estimating an optimal value of m

In this Appendix, we recall the criterion introduced in [1] for estimating a value of dimension m. Let $[x_d] \in \mathbb{M}_{n,N}$ be the matrix of the dataset introduced in Section 2.1 and let $[\eta_d] \in \mathbb{M}_{\nu,N}$ be the matrix computed with Eq. (2). We then introduced the matrix $[x_{\text{red}}(m)] \in \mathbb{M}_{n,N}$ such that (see Eqs. (4) and (6)),

$$[x_{\text{red}}(m)] = [\underline{x}] + [\varphi]\, [\lambda]^{1/2}\, [z_d]\, [g]^T \quad , \quad [z_d] = [\eta_d]\, [a]\,. \tag{B.1}$$

Let $\mathbf{x}^1_{\text{red}}(m), \ldots, \mathbf{x}^N_{\text{red}}(m)$ be the N vectors in $\mathbb{R}^n$, which constitute the columns of matrix $[x_{\text{red}}(m)] \in \mathbb{M}_{n,N}$. We then introduced the empirical estimates $\mathbf{m}_{\text{red}}(m) \in \mathbb{R}^n$ and $[c_{\text{red}}(m)] \in \mathbb{M}_n$ of the mean value and the covariance matrix calculated with the sample $[x_{\text{red}}(m)] \in \mathbb{M}_{n,N}$ such that

$$\mathbf{m}_{\text{red}}(m) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}^j_{\text{red}}(m)\,, \tag{B.2}$$

$$[c_{\text{red}}(m)] = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{x}^j_{\text{red}}(m) - \mathbf{m}_{\text{red}}(m))\,(\mathbf{x}^j_{\text{red}}(m) - \mathbf{m}_{\text{red}}(m))^T\,. \tag{B.3}$$

A criterion for the mean-square convergence can then chosen as

$$e_{\text{red}}(m) = \frac{\|[c_{\text{red}}(m)] - [c]\|_F}{\|[c]\|_F}\,. \tag{B.4}$$

in which $[c]$ is defined by

$$\mathbf{m} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}^{d,j} \quad , \quad [c] = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{x}^{d,j} - \mathbf{m})\,(\mathbf{x}^{d,j} - \mathbf{m})^T\,. \tag{B.5}$$

Since $[x_{\text{red}}(N)] = [x_d]$, it can be deduced that $e_{\text{red}}(m) \to 0$ when m goes to N. For a fixed reasonable value $\epsilon_0 > 0$ of the relative tolerance $e_{\text{red}}(m)$, an estimate of m will consist in looking for the smallest value of m such that $e_{\text{red}}(m) \leq \varepsilon_0$.

Appendix C. Algorithm for solving the reduced-order ISDE

The algorithm for solving the reduced-order ISDE defined by Eqs. (7) to (9) is detailed in [1] and is summarized hereinafter. The Störmer-Verlet scheme is used. Let $M = n_{\text{MC}} \times M_0$ be the positive integer in which n_{MC} and M_0 have been introduced in Section 2.6. The reduced-order ISDE is solved on the finite interval $\mathcal{R} = [0\,, M\,\Delta r]$, in which Δr is the sampling step of the continuous index parameter r. The integration scheme is based on the use of the $M+1$ sampling points $r_{\ell'}$ such that $r_{\ell'} = \ell'\,\Delta r$ for $\ell' = 0, \ldots, M$. The following notations are introduced: $[\boldsymbol{\mathcal{Z}}_{\ell'}] = [\boldsymbol{\mathcal{Z}}(r_{\ell'})]$, $[\boldsymbol{\mathcal{Y}}_{\ell'}] = [\boldsymbol{\mathcal{Y}}(r_{\ell'})]$, and $[\boldsymbol{\mathcal{W}}_{\ell'}] = [\boldsymbol{\mathcal{W}}(r_{\ell'})]$, for $\ell' = 0, \ldots, M$, with

$$[\boldsymbol{\mathcal{Z}}_0] = [\mathbf{H}_d]\,[a] \quad , \quad [\boldsymbol{\mathcal{Y}}_0] = [\boldsymbol{\mathcal{N}}]\,[a] \quad , \quad [\boldsymbol{\mathcal{W}}_0] = [0_{\nu,m}] \quad a.s\,. \tag{C.1}$$

For $\ell' = 0, \ldots, M-1$, let

$$[\Delta\boldsymbol{\mathcal{W}}_{\ell'+1}] = [\Delta\mathbf{W}_{\ell'+1}]\,[a]\,, \tag{C.2}$$

be the sequence of random matrices with values in $\mathbb{M}_{\nu,m}$, in which $[\Delta\mathbf{W}_{\ell'+1}] = [\mathbf{W}_{\ell'+1}] - [\mathbf{W}'_{\ell}]$. The increments $[\Delta\mathbf{W}_1], \ldots, [\Delta\mathbf{W}_M]$ are M independent random matrices with values in $\mathbb{M}_{\nu,N}$. For all $k = 1, \ldots, \nu$ and for all $j = 1, \ldots, N$, the

real-valued random variables $\{[\Delta \mathbf{W}_{\ell'+1}]_{kj}\}_{kj}$ are independent, Gaussian, second-order, and centered random variables such that

$$E\{[\Delta \mathbf{W}_{\ell'+1}]_{kj}[\Delta \mathbf{W}_{\ell'+1}]_{k'j'}\} = \Delta r\, \delta_{kk'}\, \delta_{jj'}\,. \tag{C.3}$$

For $\ell' = 0, \ldots, M-1$, the Störmer-Verlet scheme applied to Eqs. (7) and (8) yields

$$[\boldsymbol{\mathcal{Z}}_{\ell'+\frac{1}{2}}] = [\boldsymbol{\mathcal{Z}}_{\ell'}] + \frac{\Delta r}{2}\,[\boldsymbol{\mathcal{Y}}_{\ell'}]\,, \tag{C.4}$$

$$[\boldsymbol{\mathcal{Y}}_{\ell'+1}] = \frac{1-b}{1+b}\,[\boldsymbol{\mathcal{Y}}_{\ell'}] + \frac{\Delta r}{1+b}\,[\boldsymbol{\mathcal{L}}_{\ell'+\frac{1}{2}}] + \frac{\sqrt{f_0}}{1+b}\,[\Delta \boldsymbol{\mathcal{W}}_{\ell'+1}]\,, \tag{C.5}$$

$$[\boldsymbol{\mathcal{Z}}_{\ell'+1}] = [\boldsymbol{\mathcal{Z}}_{\ell'+\frac{1}{2}}] + \frac{\Delta r}{2}\,[\boldsymbol{\mathcal{Y}}_{\ell'+1}]\,, \tag{C.6}$$

with the initial condition defined by (9), where $b = f_0\,\Delta r\,/4$, and where $[\boldsymbol{\mathcal{L}}_{\ell'+\frac{1}{2}}]$ is the $\mathbb{M}_{\nu,m}$-valued random variable such that

$$[\boldsymbol{\mathcal{L}}_{\ell'+\frac{1}{2}}] = [\mathcal{L}([\boldsymbol{\mathcal{Z}}_{\ell'+\frac{1}{2}}])] = [L([\boldsymbol{\mathcal{Z}}_{\ell'+\frac{1}{2}}]\,[g]^T)]\,[a]\,, \tag{C.7}$$

in which, for all $[u] = [\mathbf{u}^1 \ldots \mathbf{u}^N]$ in $\mathbb{M}_{\nu,N}$ with $\mathbf{u}^{\ell'} = (u_1^{\ell'}, \ldots, u_\nu^{\ell'})$ in $\mathbb{R}^\nu$, the entries of matrix $[L([u])]$ in $\mathbb{M}_{\nu,N}$ are defined by Eqs. (11) to (14).

References

[1] C. Soize, R. Ghanem, Data-driven probability concentration and sampling on manifold, Journal of Computational Physics 321 (2016) 242-258.

[2] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, PNAS 102(21) (2005) 7426-7431.

[3] A.W. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis, Oxford University Press, Oxford, UK, 1997.

[4] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition, John Wiley and Sons, 2015.

[5] 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) 1583-1611.

[6] C. Soize, Polynomial chaos expansion of a multimodal random vector, SIAM/ASA Journal on Uncertainty Quantification 3(1) (2015) 3460.

[7] R.M. Neal, MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo (eds S. Brooks, A. gelman, G. Jones, and X.-L. Meng), Chapman and Hall-CRC Press, Boca Raton, 2010

[8] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, Journal of the Royal Statistics Society 73(Part 2) (2011) 123-214.

[9] J. Kaipio, E. Somersalo, Statistical and Computational Inverse Problems, Springer-Verlag, New York, 2005.

[10] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer-Verlag, New York, 2005.

[11] J.C. Spall, Introduction to Stochastic Search and Optimization, John Wiley and Sons, Hoboken, New Jersey, 2003.

[12] R.R. Coifman, S. Lafon, Diffusion maps, Applied and Computational Harmonic Analysis 21(1) (2006) 5-30.

[13] R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Applied and Computational Harmonic Analysis 39(1) (2015) 138-160.

[14] R. Ghanem, P.D. Spanos, Polynomial chaos in stochastic finite elements, Journal of Applied Mechanics - Transactions of the ASME 57(1) (1990) 197-202.

[15] R. Ghanem, P.D. Spanos, Stochastic Finite Elements: A spectral Approach, Springer-verlag, New-York, 1991 (revised edition, Dover Publications, New York, 2003).

[16] R. Ghanem, R.M. Kruger, Numerical solution of spectral stochastic finite element systems, Computer Methods in Applied Mechanics and Engineering 129(3) (1996), 289-303.

[17] D.B. Xiu, G.E. Karniadakis, Wiener-Askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24(2) (2002) 619-644.

[18] P. Frauenfelder, C. Schwab, R.A. Todor, Finite elements for elliptic problems with stochastic coefficients, Computer Methods in Applied Mechanics and Engineering, 194(2-5) (2005) 205-228.

[19] B.J. Debusschere, H.N. Najm, P.P. Pebay, O.M. Knio, R. Ghanem, O.P. Le Maitre, Numerical challenges in the use of polynomial chaos representations for stochastic processes, SIAM Journal on Scientific Computing 26(2) (2004) 698-719.

[20] C. Desceliers, R. Ghanem, C. Soize, Maximum likelihood estimation of stochastic chaos representations from experimental data, International Journal for Numerical Methods in Engineering 66(6) (2006) 978-1001.

[21] I. Babuska, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Journal on Numerical Analysis, 45(3) (2007) 1005-1034.

[22] A. Doostan, R. Ghanem, J. Red-Horse, Stochastic model reduction for chaos representations, Computer Methods in Applied Mechanics and Engineering 196(37-40) (2007) 3951-3966.

[23] B. Ganapathysubramanian, N. Zabaras, Sparse grid collocation schemes for stochastic natural convection problems, Journal of Computational Physics 225(1) (2007) 652-685.

[24] A. Nouy, A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering 196 (45-48) (2007) 4521-4537.

[25] S. Das, R. Ghanem, J. Spall, Asymptotic sampling distribution for polynomial chaos representation of data: A maximum-entropy and fisher information approach, SIAM Journal on Scientific Computing 30(5) (2008) 2207-2234.

[26] R. Ghanem, R. Doostan, J. Red-Horse, A probability construction of model validation, Computer Methods in Applied Mechanics and Engineering 197(29-32) (2008) 2585-2595.

[27] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering & System Safety 93(7) (2008) 964-979.

[28] H.N. Najm, Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics, Journal Review of Fluid Mechanics 41 (2009) 35-52.

[29] M. Arnst, R. Ghanem, C. Soize, Identification of Bayesian posteriors for coefficients of chaos expansion, Journal of Computational Physics 229(9) (2010) 3134-3154.

[30] O.P. Le Maitre, O.M. Knio, Spectral Methods for Uncertainty Quantification with Applications to Computational Fluid Dynamics, Springer, Heidelberg, 2010.

[31] A. Doostan, H. Owhadi, A non-adapted sparse approximation of PDEs with stochastic inputs, Journal of Computational Physics 230(8) (2011) 3015-3034.

[32] C. Soize, Identification of high-dimension polynomial chaos expansions with random coefficients for non-Gaussian tensor-valued random fields using partial and limited experimental data, Computer Methods in Applied Mechanics and Engineering 199(33-36) (2010) 2150-2164.

[33] A. Nouy, C. Soize, Random fields representations for stochastic elliptic boundary value problems and statistical inverse problems, European Journal of Applied Mathematics 25(3) (2014) 339-373.

[34] D. Lucor, C.H. Su, G.E. Karniadakis, Generalized polynomial chaos and random oscillators, International Journal for Numerical Methods in Engineering 60(3) (2004) 571-596.

[35] C. Soize, R. Ghanem, Physical systems with random uncertainties: Chaos representation with arbitrary probability measure, SIAM Journal on Scientific Computing 26(2) (2004) 395-410.

[36] X.L. Wan, G.E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM Journal on Scientific Computing 28(3) (2006) 901-928.

[37] C. Soize, R. Ghanem, Reduced chaos decomposition with random coefficients of vector-valued random variables and random fields, Computer Methods in Applied Mechanics and Engineering 198(21-26) (2009) 1926-1934.

[38] O.G. Ernst, A. Mugler, H.J. Starkloff, E. Ullmann, On the convergence of generalized polynomial chaos expansions, ESAIM: Mathematical Modelling and Numerical Analysis 46(2) (2012) 317-339.

[39] G. Perrin, C. Soize, D. Duhamel, C. Funfschilling, Identification of polynomial chaos representations in high dimension from a set of realizations, SIAM Journal on Scientific Computing 34(6) (2012) A2917-A2945.

[40] R. Tipireddy, R. Ghanem, Basis adaptation in homogeneous chaos spaces, Journal of Computational Physics 259 (2014) 304-317.

[41] D. Ghosh, R. Ghanem, Stochastic convergence acceleration through basis enrichment of polynomial chaos expansions, International Journal for Numerical Methods in Engineering 73(2) (2008) 162-184.

[42] V. Keshavarzzadeh, R. Ghanem, S.F. Masri, O.J. Aldraihem, Convergence acceleration of polynomial chaos solutions via sequence transformation, Computer Methods in Applied Mechanics and Engineering 271 (2014) 167-184.

[43] Y.M. Marzouk, H.N. Najm, Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, Journal of Computational Physics 228(6) (2009) 1862-1902.

[44] C. Soize, C. Desceliers, Computational aspects for constructing realizations of polynomial chaos in high dimension, SIAM Journal On Scientific Computing 32(5) (2010) 2820-2831.

[45] M. Arnst, C. Soize, R. Ghanem, Hybrid sampling/spectral method for solving stochastic coupled problems, SIAM/ASA Journal on Uncertainty Quantification, 1(1) (2013) 218243.

[46] Data Center BOEM, Bureau of Ocean Energy Management, http://www.data.boem.gov/.

[47] C. Thimmisetty, A. Khodabakhshnejad, N. Jabbari, F. Aminzadeh, R. Ghanem, K. Rose, J. Bauer, C. Disenhof, Multiscale stochastic representation in high-dimensional data using Gaussian processes with implicit diffusion metrics, Lecture Notes in Computer Science, Vol. 8964, 2015 (Proceedings of the Dynamic Data-driven Environmental Systems Science Conference, MIT, Cambridge, MA, Nov 5-7, 2014.)