Data-driven probability concentration and sampling on manifold

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

A new methodology is proposed for generating realizations of a random vector with values in a finite-dimensional Euclidean space that are statistically consistent with a dataset of observations of this vector. The probability distribution of this random vector, while a-priori not known, is presumed to be concentrated on an unknown subset of the Euclidean space. A random matrix is introduced whose columns are independent copies of the random vector and for which the number of columns is the number of data points in the dataset. The approach is based on the use of (i) the multidimensional kernel-density estimation method for estimating the probability distribution of the random matrix, (ii) a MCMC method for generating realizations for the random matrix, (iii) the diffusion-maps approach for discovering and characterizing the geometry and the structure of the dataset, and (iv) a reduced-order representation of the random matrix, which is constructed using the diffusion-maps vectors associated with the first eigenvalues of the transition matrix relative to the given dataset. The convergence aspects of the proposed methodology are analyzed and a numerical validation is explored through three applications of increasing complexity. The proposed method is found to be robust to noise levels and data complexity as well as to the intrinsic dimension of data and the size of experimental datasets. Both the methodology and the underlying mathematical framework presented in this paper contribute new capabilities and perspectives at the interface of uncertainty quantification, statistical data analysis, stochastic modeling and associated statistical inverse problems.

*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)

Preprint submitted to Journal of Computational Physics, accepted 28 May 2016 May 31, 2016

Keywords: Concentration of probability, Measure concentration, Probability distribution on manifolds, Random sampling generator, MCMC generator, Diffusion maps, Statistics on manifolds, Design of experiments for random parameters

1. Introduction

The construction of a generator of realizations from a given dataset related to a $\mathbb{R}^n$-valued random vector, for which the probability distribution is unknown and is concentrated on an unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$, is a central and difficult problem in uncertainty quantification and statistical data analysis, in stochastic modeling and associated statistical inverse problems for boundary value problems, in the design of experiments for random parameters, and certainly, in signal processing and machine learning. A common situation, addressed in the last example in the paper pertains to the availability of a limited number of high-dimensional samples (i.e. each sample has many attributes). In such cases it is often desirable to carry out a statistical analysis of the data for the purpose of inference. Acknowledging the local structure of the data, when such structure exists, provides additional knowledge that should be valuable for an efficient characterization and sampling schemes. While the last example in the paper presents a problem in petrophysics, similar problems abound in all branches of science and engineering including biology, astronomy, and nuclear physics.

Two fundamental tools serve as building blocks for addressing this problem. First, nonparametric statistical methods [1, 2] can be effectively used to construct probability distribution on $\mathbb{R}^n$ of a random vector given an initial dataset of its samples. Multidimensional Gaussian kernel-density estimation is one efficient subclass of these methods. Markov chain Monte Carlo (MCMC) procedures can then be used to sample additional realizations from the resulting probability model, and which are thus statistically consistent with the initial dataset [3, 4, 5]. The second building block consists of manifold embedding algorithms, where low-dimensional structure is characterized within a larger vector space. Diffusion maps [6, 7, 8] is a powerful tool for characterizing and delineating $\mathcal{S}_n$ using the initial dataset and concepts of geometric diffusion.

The first tool described above, consisting of using nonparametric density estimation with MCMC, does not allow, in general, the restriction of new samples to the subset $\mathcal{S}_n$ on which the probability distribution is concentrated. The scatter of generated samples outside of $\mathcal{S}_n$ is more pronounced the more complex and

disconnected this set is.
The second tool consisting of diffusion maps, while effectively allowing for the discovery and characterization of subset $\mathcal{S}_n$ on which the probability distribution is concentrated, does not give a direct approach for generating additional realizations in this subset that are drawn from a target distribution consistent with the initial dataset.
These two fundamental tools have been used independently and quite successfully to address problems of sampling from complex probability models and detecting low-dimensional manifolds in high-dimensional settings. An analysis of MCMC methods on Riemann manifolds has been presented recently [9] where the manifold is the locus of density functions and not of the data itself. This paper addresses the still open challenge of efficient statistical sampling on manifolds defined by limited data.
It should be noted that the PCA [10] yields a statistical reduction method for second-order random vectors in finite dimension, similarly to the Karhunen-Loève expansion (KLE) [11, 12], which yields a statistical reduction method for second-order stochastic processes and random fields, and which has been used for obtaining an efficient construction [13, 14] of the polynomial chaos expansion (PCE) of stochastic processes and random fields [15], and for which some ingredients have more recently been introduced for analyzing complex problems encountered in uncertainty quantification [16, 17]. *A priori* and in general, the PCA or the KLE, which use a nonlocal basis with respect to the dataset (global basis related to the covariance operator estimated with the dataset) does not allow for discovering and characterizing the subset on which the probability law is concentrated. The present work can be viewed as an extension and generalization of previous work by the authors where the low-dimensional manifold was unduly restricted [18, 19, 20].
After formalizing the problem in Section 2, the proposed methodology is presented in Section 3 and developed in Section 4. Section 5 deals with three applications: the first two applications correspond to analytical examples in dimension 2 with 230 given data points and in dimension 3 with 400 data points. The third application is related to a petro-physics database made up of experimental measurements for which the dimension is 35 with $13,056$ given data points.

Comments concerning the motivation, the objectives, and the methodology

(i) As it has been previously explained, the fundamental hypothesis of this paper is that the solely available information are described by a given dataset of N independent realizations for the random vector $\mathbf{H}$ with values in $\mathbb{R}^{\nu}$ (which is as-

sumed to be second-order and not Gaussian). Consequently, the given dataset is represented by a given $(\nu \times N)$ real matrix $[\eta_d]$. The objective of this paper is to construct a generator of new additional realizations in using the diffusion maps that allows for discovering the geometry of the subset $\mathcal{S}_\nu \subset \mathbb{R}^\nu$ in which the unknown probability distribution is concentrated and consequently, permitting the enrichment of the knowledge that we have from the data. For constructing such a generator, a probability distribution (that is non-Gaussian and that must be coherent with the dataset) has to be constructed using what may be referred to as a indirect approach or a direct approach. An indirect approach consists in introducing a parameterized stochastic model that has the capability of generating the required realizations. For instance, a polynomial chaos expansion (PCE) can be introduced for which the coefficients must be identified by solving a statistical inverse problem. A direct approach consists in constructing an estimation of the probability distribution directly from the dataset, either by using parametric statistics (and then, by solving a statistical inverse problem for identifying the parameters) or by using nonparametric statistics. Concerning the parametric statistics, as it is assumed that no information is available in addition to the data set, information theory is not very useful for constructing a parameterized prior informative probability measure in the framework of parametric statistics. In any case, the method that would be selected must be able to take into account the information concerning the geometry of the subset $\mathcal{S}_\nu$ on which the probability distribution is concentrated (constructed with the diffusion maps), and must be computationally efficient for problems in high dimension. In this framework, the PCE is surely an attractive representation, but which cannot be easily implemented, because the statistical inverse problem for identifying the coefficients must be coupled with the formalism of the diffusion maps methodology, a non-trivial task. This motivates the approach followed in the present paper where nonparametric statistics are used to construct the probability density function of $\mathbf{H}$ for which a generator that belongs to the class of the MCMC methods is then developed. Concerning the choice of the MCMC method, we propose to use the one that is based on an Itô stochastic differential equation. This choice allows us to capitalize on the geometry of $\mathcal{S}_\nu$ and to construct via projections, a restriction of the MCMC to the manifold. In addition, a very efficient and scalable discretization scheme can be used which remains efficient in high dimensions. Clearly, the proposed methodology hinges on the dataset in question being supported by a diffusion manifold.

(ii) Concerning the choice of the nonparametric statistical estimator of the probability distribution, we propose to use a modification of the classical multi-dimen-

sional Gaussian kernel-density estimation method. It should be noted that the methodology presented in this paper, for taking into account the geometry of $\mathcal{S}_\nu$ constructed with diffusion maps, is independent of the choice of the kernel and therefore, is independent of the choice of the bandwidth. The proposed methodology is based on a projection in a subspace related to the geometry of $\mathcal{S}_\nu$ discovered by the diffusion maps. This implies that the concentration on $\mathcal{S}_\nu$ is independent of the choice of the bandwidth. This also means that even if a strong smoothing is required (high dimension with a relatively small number of realizations), the generator that is constructed will concentrate the realizations in $\mathcal{S}_\nu$ the description of which is independent of the chosen estimator for the probability distribution. The only implicit assumption in the foregoing is that the observed data points are independent identically distributed. Kernel density estimation methods are then used to estimate their common probability density function. Certainly, for a given dataset, the quality of the results depends of the choice of the kernel and its associated bandwidth, and this quality highly depends on the number N of realizations. This problem has extensively been studied in the literature and is outside the scope of the paper. However, any progress obtained in this field can be reused without modifying the methodology proposed. In particular, the geometry of $\mathcal{S}_\nu$ could be used for improving the bandwidth of the kernel, such as the implementation of an appropriate adaptive bandwidth method, but such an extension is out the scope of the present paper. Finally, it is very important to note that the generator uses two important sources of information related to the given dataset, but cannot include information that are not available. If the number N of realizations is not sufficiently large, it is well known that the non-Gaussian probability distribution can only be roughly estimated, but in the absence of more information, the methodology proposed uses all the possible available information. If a reference is known (as for the applications presented in Section 5 for validating the methodology) and if the number N of realizations is not sufficiently large, then it is usual to obtain a difference between the estimates of the probability distribution and the referenced one. Nevertheless, the two following arguments must be kept in mind. The generator presented can be viewed as the one of a constructed non-Gaussian random vector that admits the realizations given by the dataset, but is not the one of the hypothetical non-Gaussian random vector for which its exact probability distribution is unknown and, in general, will never be known. Clearly, the classical mathematical argument related to the convergence of the sequence of estimators with respect to N can always be used, independently from the methodology proposed, but in practice, N is fixed, has a specified fixed value for a given application, and cannot be increase for assessing convergence.

(iii) In order to properly define the mathematical projection whose vector basis is constructed by using the diffusion maps, a compatibility of the dimensions must be respected. For that, it is necessary to introduce a random $(\nu \times N)$ real matrix $[\mathbf{H}]$ for which the N columns are N independent copies of random vector $\mathbf{H}$. The ν independent realizations of the given dataset for $\mathbf{H}$ are thus represented by the given realization $[\eta_d]$ of $[\mathbf{H}]$. The MCMC generator for $[\mathbf{H}]$ is therefore an Itô stochastic differential equation in a stochastic process $\{[\mathbf{U}(r)], r \geq 0\}$ with values in the set of all the $(\nu \times N)$ real matrices. The natural choice of the initial condition is thus $[\mathbf{U}(0)] = [\eta_d]$ almost surely. This is the reason why the new additional realizations generated by the method start from each point of the given dataset (as the reader will be able to see it in Section 5) and stay confined in $\mathcal{S}_\nu$ thanks to the projection of $[\mathbf{H}]$ on the vector basis calculated with the diffusion maps.

(iv) The given dataset is made up of N given realizations of random vector $\mathbf{H}$ with values in $\mathbb{R}^\nu$ which can result from a pre-processing of N independent realizations of a random vector $\mathbf{X}$ with values in $\mathbb{R}^n$ with $\nu \leq n$. Such a pre-processing is performed using a principal component analysis (PCA). The reason for this pre-processing is not to reduce the statistical dimension but rather to normalize the dataset (which means that $\mathbf{H}$ is constructed as a random vector with a zero mean and with an identity covariance matrix). Such a pre-processing is necessary for guarantying a good numerical behavior when using multi-dimensional Gaussian kernel-density estimation methods. If an accurate mean-square convergence is obtained for $\nu < n$ a statistical reduction can be introduced; if not, $\nu = n$. Clearly, the PCA can modify the local structure of the dataset, but this is not a problem. In the theory proposed, this is the local structure of random vector $\mathbf{H}$ (and not the local structure of random vector $\mathbf{X}$) that is analyzed with the diffusion maps. The theory proposed introduces an integer m that is the dimension of the projection vector basis. If $m \ll N$ then there is a subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$ on which the probability distribution of $\mathbf{H}$ is concentrated.

(v) The computational cost is less or equal to the computational cost induced by the classical methodology for generating new additional realizations to a given dataset using the nonparametric statistics and a MCMC algorithm. If $m \ll N$, then the numerical cost is reduced. It should be noted, however, that, unlike the method proposed in this paper, these classical methodologies will not generate realizations that are concentrated on $\mathcal{S}_\nu$.

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.
An upper case letter such as X, H, or U, is a real random variable.
A boldface upper case letter, $\mathbf{X}$, $\mathbf{H}$, or $\mathbf{U}$, 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 $[\mathbf{X}]$, $[\mathbf{H}]$, or $[\mathbf{U}]$, is a real random matrix.

E: Mathematical expectation.
$\mathbb{M}_{n,N}$: set of all the $(n \times N)$ real matrices.
$\mathbb{M}_{\nu}$: $\mathbb{M}_{\nu,\nu}$.
$\|\mathbf{x}\|$: Euclidean norm of vector $\mathbf{x}$.
$[x]_{kj}$: 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 $k = k'$.

2. Problem set-up

The following four ingredients serve to set the stage for the mathematical analysis required for constructing the target probability distribution and sampling from it.
(i) Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a generic point in $\mathbb{R}^n$ and let $d\mathbf{x} = dx_1 \ldots dx_n$ be the Lebesgue measure. A family of N vectors in $\mathbb{R}^n$ will be written as $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$.

(ii) 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 distribution is defined by a probability density function (pdf) on $\mathbb{R}^n$ (*a priori* and in general, the probability distribution is not Gaussian). This pdf is unknown but is assumed to be concentrated on an unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$. A specific realization of random vector $\mathbf{X}$ will be denoted by $\mathbf{X}(\theta)$ where $\theta \in \Theta$.

(iii) The available information consists of a given set of N data points specified by N vectors $\mathbf{x}^{d,1}, \ldots, \mathbf{x}^{d,N}$ in $\mathbb{R}^n$. These will be assumed to constitute N statistically independent realizations (or samples) $\mathbf{X}(\theta_1), \ldots, \mathbf{X}(\theta_N)$ of random vector $\mathbf{X}$. For $j = 1, \ldots, N$, the vector $\mathbf{x}^{d,j}$ in $\mathbb{R}^n$ is written as $\mathbf{x}^{d,j} = (x_1^{d,j}, \ldots, x_n^{d,j})$. The N data points can then be represented by the matrix $[x_d]$ in $\mathbb{M}_{n,N}$ such that $[x_d]_{kj} = x_k^{d,j}$.

(iv) The local structure of the given dataset is captured via random matrix $[\mathbf{X}]$, defined on $(\Theta, \mathcal{T}, \mathcal{P})$, with values in $\mathbb{M}_{n,N}$. Specifically, $[\mathbf{X}] = [\mathbf{X}^1 \ldots \mathbf{X}^N]$ in which the columns $\mathbf{X}^1, \ldots, \mathbf{X}^N$ are N independent copies of random vector $\mathbf{X}$. Consequently, matrix $[x_d]$ can be viewed as one realization of random matrix $[\mathbf{X}]$.

The objective of this paper then is to construct a generator of realizations of random matrix $[\mathbf{X}]$ in $\mathbb{M}_{n,N}$, for which the unknown probability distribution is directly deduced from the unknown probability distribution of random vector $\mathbf{X}$, which is concentrated on the unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$, and for which only one realization $[x_d]$ is given.

The unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$ can be viewed as a manifold, which corresponds to the structure of data $[x_d]$, and on which the unknown probability measure is concentrated. Consequently, the objective of the paper is to perform "data-driven probability concentration and sampling on a manifold".

3. Summary of the methodology proposed

To enhance the utility of the present paper and to clarify the inter-relation between a number of intricate mathematical steps, the proposed methodology is summarized in the following seven steps.

1. In general, the given data are heterogeneous and badly conditioned. Consequently, the first step consists in performing a scaling of the given data, which yields the matrix $[x_d]$ in $\mathbb{M}_{n,N}$ of the scaled given data (the matrix introduced in Section 2), and simply called the given dataset (removing the word "scaled"). The given dataset are then normalized by using a principal component analysis (but without trying to introduce a statistical reduced-order representation). Therefore, the random matrix $[\mathbf{X}]$ (corresponding to scaled data $[x_d]$) is written as an affine transformation of a random matrix $[\mathbf{H}]$ with values in $\mathbb{M}_{\nu,N}$ with $1 < \nu \leq n$ (in general, $\nu = n$, but sometimes

some eigenvalues (of the empirical estimate of the covariance matrix of $\mathbf{X}$) exhibits zeros eigenvalues that are removed, yielding $\nu < n$). Random matrix $[\mathbf{H}]$ can then be written as $[\mathbf{H}] = [\mathbf{H}^1 \dots \mathbf{H}^N]$ in which the columns $\mathbf{H}^1, \dots, \mathbf{H}^N$ are N independent copies of a random vector $\mathbf{H}$ with values in $\mathbb{R}^\nu$, whose probability density function on $\mathbb{R}^\nu$ is unknown and is concentrated on an unknown subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$. The given data $[x_d]$ in $\mathbb{M}_{n,N}$ (related to $[\mathbf{X}]$) are then transformed into given data, $[\eta_d]$ in $\mathbb{M}_{\nu,N}$, related to random matrix $[\mathbf{H}]$. The data represented by $[\eta_d]$ are thus normalized. Let $p_{\mathbf{H}}$ be the nonparametric estimate of the probability density function of random vector $\mathbf{H}$, which is performed by using $[\eta_d]$ (note that $p_{\mathbf{H}}$ is not the pdf of $\mathbf{H}$ but is the nonparametric estimate of the pdf of $\mathbf{H}$). Consequently, the nonparametric estimate of the probability distribution on $\mathbb{M}_{\nu,N}$ of random matrix $[\mathbf{H}]$ is written as $p_{[\mathbf{H}]}([\eta])\, d[\eta] = p_{\mathbf{H}}(\boldsymbol{\eta}^1) \times \dots \times p_{\mathbf{H}}(\boldsymbol{\eta}^N)\, d\boldsymbol{\eta}^1 \dots d\boldsymbol{\eta}^N$ in which $[\eta]$ is any matrix in $\mathbb{M}_{\nu,N}$ such that $[\eta] = [\boldsymbol{\eta}^1 \dots \boldsymbol{\eta}^N]$ with $\boldsymbol{\eta}^j \in \mathbb{R}^\nu$.

2. The second step consists in constructing the nonparametric statistical estimate $p_{\mathbf{H}}$ of the probability density function of $\mathbf{H}$ using data $[\eta_d] \in \mathbb{M}_{\nu,N}$. This is an usual problem that will be performed by using the classical multidimensional Gaussian kernel-density estimation method. Nevertheless, we will use the modification proposed in [21] (instead of the classical method) in order that the nonparametric estimate $p_{\mathbf{H}}$ yields, for the estimation of the covariance matrix of $\mathbf{H}$ (using $[\eta_d]$), the identity matrix $[I_\nu]$ in $\mathbb{M}_\nu$. This construction is directly used in the following third step.
3. The third step consists in introducing an adapted generator of realizations for random matrix $[\mathbf{H}]$, which belongs to the class of the MCMC methods such as the Metropolis-Hastings algorithm [22, 23] (that requires the definition of a good proposal distribution), the Gibbs sampling [24] (that requires the knowledge of the conditional distribution) or the slice sampling [25] (that can exhibit difficulties related to the general shape of the probability distribution, in particular for multimodal distributions). This adapted generator will be the one derived from [21], which is based on a nonlinear Itô stochastic differential equation (ISDE) formulated for a dissipative Hamiltonian dynamical system [26], which admits $p_{[\mathbf{H}]}([\eta])\, d[\eta]$ as an invariant measure, and for which the initial condition depends on matrix $[\eta_d]$.
4. The fourth step of the methodology consists in characterizing the subset $\mathcal{S}_\nu$ from scaled and normalized data $[\eta_d]$. This will be done using the formulation of the diffusion maps, which is a very powerful mathematical tool for doing that. It should be noted that the diffusion-maps method explores a given dataset using a local kernel while the PCA explores the same dataset

using global averages that, in general, cannot see the local geometric structure of the given dataset. However, the diffusion distance, which has been introduced in [6] for discovering and characterizing $\mathcal{S}_\nu$, and which is constructed using the diffusion maps, does not allow for constructing a generator of realizations of random matrix $[\mathbf{H}]$ for which data $[\eta_d]$ are given but for which its probability measure and the subset $\mathcal{S}_\nu$ of concentration are unknown. This step is introduced for constructing an algebraic vector basis $\{\mathbf{g}^1, \ldots, \mathbf{g}^N\}$ of $\mathbb{R}^N$, depending on two parameters that are a smoothing parameter $\varepsilon > 0$ and an integer κ related to the analysis scale of the local geometric structure of the dataset. For $\alpha = 1, \ldots, N$, the vectors $\mathbf{g}^\alpha = (g_1^\alpha, \ldots, g_N^\alpha) \in \mathbb{R}^N$ are directly related to the diffusion maps. A subset of this basis will be able to characterize the subset $\mathcal{S}_\nu$ of $\mathbb{R}^\nu$ on which the probability measure of $\mathbf{H}$ is concentrated. We will then introduce the matrix $[g]$ in $\mathbb{M}_{N,m}$ made up of the first m vectors $\{\mathbf{g}^1, \ldots, \mathbf{g}^m\}$ of the diffusion-maps basis, with $1 < m \ll N$.

5. The fifth step consists in estimating an adapted value of m in order to capture the local geometric structure of $\mathcal{S}_\nu$ and to obtain a reasonable mean-square convergence.
6. Using the first m vectors (represented by matrix $[g]$) of the diffusion-maps basis, the sixth step consists in constructing a reduced-order ISDE, which allows for generating some additional realizations of the reduced-order representation of random matrix $[\mathbf{H}]$, by introducing the random matrix $[\mathbf{Z}]$ with values in $\mathbb{M}_{\nu,m}$ such that $[\mathbf{H}] = [\mathbf{Z}]\,[g]^T$.
7. The last step consists in numerically solving the reduced-order ISDE for computing the additional realizations $[z_s^1], \ldots, [z_s^{n_{\text{MC}}}]$ of random matrix $[\mathbf{Z}]$ and then to deduce the additional realizations $[x_s^1], \ldots, [x_s^{n_{\text{MC}}}]$ of random matrix $[\mathbf{X}]$ for which only one realization $[x_d]$ was given.

4. Formulation

In this section, a detailed presentation of the methodology is given that parallels the steps described in Section 3.

4.1. Scaling and normalizing the given dataset

Let $[x_d^{uns}]$ be the matrix in $\mathbb{M}_{n,N}$ of the unscaled given dataset. The matrix $[x_d]$ in $\mathbb{M}_{n,N}$ of the scaled given dataset (simply called the given dataset) is constructed (if the data effectively require such a scaling, which will be the case for the third

application presented in Section 5.3) such that, for all $k = 1, \ldots, n$ and $j = 1, \ldots, N$,

$$[x_d]_{kj} = \frac{[x_d^{uns}]_{kj} - \min_{j'}[x_d^{uns}]_{kj'}}{\max_{j'}[x_d^{uns}]_{kj'} - \min_{j'}[x_d^{uns}]_{kj'}} + \epsilon_s \,. \tag{1}$$

The quantity ϵ_s is added to the scaled data in order to avoid the scalar 0 in the nonparametric statistical estimation of the pdf. Let $\mathbf{m}$ and $[c]$ be the empirical estimates of the mean vector $E\{\mathbf{X}\}$ and the covariance matrix $E\{(\mathbf{X}-E\{\mathbf{X}\})\,(\mathbf{X}-E\{\mathbf{X}\})^T\}$, such that

$$\mathbf{m} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}^{d,j} \quad , \quad [c] = \frac{1}{N-1}\sum_{j=1}^{N} (\mathbf{x}^{d,j} - \mathbf{m})\,(\mathbf{x}^{d,j} - \mathbf{m})^T \,. \tag{2}$$

We consider the eigenvalue problem $[c]\,\boldsymbol{\varphi}^k = \mu_k\,\boldsymbol{\varphi}^k$. Noting that matrix $[c]$ is often of rank $\nu \leq n$, denote its ν positive eigenvalues by $\{\mu_i\}_{i=1}^{\nu}$ with $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_\nu$ and let $[\varphi]$ be the $(n \times \nu)$ matrix such $[\varphi]^T\,[\varphi] = [I_\nu]$, whose columns are the associated orthonormal eigenvectors $\boldsymbol{\varphi}^1, \ldots, \boldsymbol{\varphi}^\nu$. Consequently, random matrix $[\mathbf{X}]$ can be rewritten as

$$[\mathbf{X}] = [\underline{x}] + [\varphi]\,[\mu]^{1/2}\,[\mathbf{H}]\,, \tag{3}$$

in which $[\underline{x}]$ is the matrix in $\mathbb{M}_{n,N}$ for which each column is vector $\mathbf{m}$ and where $[\mu]$ is the positive diagonal $(\nu \times \nu)$ real matrix such that $[\mu]_{kk'} = \delta_{kk'}\mu_k$. The realization $[\eta_d] \in \mathbb{M}_{\nu,N}$ of $[\mathbf{H}]$ associated with the realization $[x_d]$ of $[\mathbf{X}]$ is thus computed by

$$[\eta_d] = [\mu]^{-1/2}[\varphi]^T\,([x_d] - [\underline{x}])\,. \tag{4}$$

Let $\boldsymbol{\eta}^{d,1}, \ldots, \boldsymbol{\eta}^{d,N}$ be the N vectors in $\mathbb{R}^\nu$ such that $[\boldsymbol{\eta}^{d,1} \ldots \boldsymbol{\eta}^{d,N}] = [\eta_d]$ (the columns of $[\eta_d]$). It can easily be seen that the empirical estimates $\mathbf{m}'$ of the mean vector $E\{\mathbf{H}\}$ and $[c']$ of the covariance matrix $E\{(\mathbf{H} - E\{\mathbf{H}\})\,(\mathbf{H} - E\{\mathbf{H}\})^T\}$ of random vector $\mathbf{H}$ are such that

$$\mathbf{m}' = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{\eta}^{d,j} = \mathbf{0} \quad , \quad [c'] = \frac{1}{N-1}\sum_{j=1}^{N} \boldsymbol{\eta}^{d,j}\,(\boldsymbol{\eta}^{d,j})^T = [\,I_\nu]\,. \tag{5}$$

4.2. Construction of a nonparametric estimate $p_{\boldsymbol{H}}$ of the pdf of $\boldsymbol{H}$

The estimation $p_{\mathbf{H}}$ on $\mathbb{R}^\nu$ of the pdf of random vector $\mathbf{H}$ is carried out by using the Gaussian kernel-density estimation method and the N independent realizations $\boldsymbol{\eta}^{d,1}, \ldots, \boldsymbol{\eta}^{d,N}$ represented by matrix $[\eta_d]$ computed with Eq. (4). As proposed in [21], a modification of the classical Gaussian kernel-density estimation

method is used in order that the mean vector and the covariance matrix (computed with the nonparametric estimate $p_{\mathbf{H}}$) are equal to $\mathbf{0}$ and $[\, I_\nu]$ respectively (see Eq. (5)). The positive-valued function $p_{\mathbf{H}}$ on $\mathbb{R}^\nu$ is then defined, for all $\boldsymbol{\eta}$ in $\mathbb{R}^\nu$, by

$$p_{\mathbf{H}}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{j=1}^{N} \pi_{\nu,\widehat{s}_\nu} \left(\frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \boldsymbol{\eta} \right), \tag{6}$$

in which $\pi_{\nu,\widehat{s}_\nu}$ is the positive function from $\mathbb{R}^\nu$ into $]0\,, +\infty[$ defined, for all $\boldsymbol{\eta}$ in $\mathbb{R}^\nu$, by

$$\pi_{\nu,\widehat{s}_\nu}(\boldsymbol{\eta}) = \frac{1}{(\sqrt{2\pi}\,\widehat{s}_\nu\,)^\nu} \, \exp\{-\frac{1}{2\widehat{s}_\nu^2} \|\boldsymbol{\eta}\|^2\}\,, \tag{7}$$

with $\|\boldsymbol{\eta}\|^2 = \eta_1^2 + \ldots + \eta_\nu^2$ and where the positive parameters s_ν and $\widehat{s}_\nu$ are defined by

$$s_\nu = \left\{ \frac{4}{N(2+\nu)} \right\}^{1/(\nu+4)} \quad , \quad \widehat{s}_\nu = \frac{s_\nu}{\sqrt{s_\nu^2 + \frac{N-1}{N}}}\,. \tag{8}$$

Parameter s_ν is the usual multidimensional optimal Silverman bandwidth (in taking into account that the empirical estimate of the standard deviation of each component is unity), and parameter $\widehat{s}_\nu$ has been introduced in order that the second equation in Eq. (5) holds. Using Eqs. (6) to (8), it can easily be verified that

$$\int_{\mathbb{R}^\nu} \boldsymbol{\eta}\, p_{\mathbf{H}}(\boldsymbol{\eta})\, d\boldsymbol{\eta} = \frac{\widehat{s}_\nu}{s_\nu}\, \mathbf{m}' = \mathbf{0}\,, \tag{9}$$

$$\int_{\mathbb{R}^\nu} \boldsymbol{\eta}\, \boldsymbol{\eta}^T\, p_{\mathbf{H}}(\boldsymbol{\eta})\, d\boldsymbol{\eta} = \widehat{s}_\nu^2\, [\, I_\nu] + (\frac{\widehat{s}_\nu}{s_\nu})^2\, \frac{(N-1)}{N} [c'] = [\, I_\nu]\,. \tag{10}$$

A nonparametric estimate $p_{[\mathbf{H}]}$ on $\mathbb{M}_{\nu,N}$ of the probability density function of random matrix $[\mathbf{H}]$ is then written as

$$p_{[\mathbf{H}]}([\eta]) = p_{\mathbf{H}}(\boldsymbol{\eta}^1) \times \ldots \times p_{\mathbf{H}}(\boldsymbol{\eta}^N)\,, \tag{11}$$

in which $p_{\mathbf{H}}$ is defined by Eqs. (6) to (8).

4.3. Construction of an ISDE for generating realizations of random matrix $[\boldsymbol{H}]$

The probability density function defined by Eqs. (6) to (8) is directly used for constructing the Itô stochastic differential equation. Let $\{([\mathbf{U}(r)], [\mathbf{V}(r)]), r \in \mathbb{R}^+\}$ be the Markov stochastic process defined on the probability space $(\Theta, \mathcal{T}, \mathcal{P})$,

indexed by $\mathbb{R}^+ = [0\,, +\infty[$, with values in $\mathbb{M}_{\nu,N} \times \mathbb{M}_{\nu,N}$, satisfying, for all $r > 0$, the following ISDE

$$d[\mathbf{U}(r)] = [\mathbf{V}(r)]\, dr\,, \tag{12}$$

$$d[\mathbf{V}(r)] = [L([\mathbf{U}(r)])]\, dr - \frac{1}{2} f_0\, [\mathbf{V}(r)]\, dr + \sqrt{f_0}\, [d\mathbf{W}(r)]\,, \tag{13}$$

with the initial condition

$$[\mathbf{U}(0)] = [\mathbf{H}_d] \quad , \quad [\mathbf{V}(0)] = [\boldsymbol{\mathcal{N}}] \quad a.s\,. \tag{14}$$

In Eqs. (13) and (14), the different quantities are defined as follows.

(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

$$[L([u])]_{k\ell} = -\frac{\partial}{\partial u_k^\ell} \mathcal{V}(\mathbf{u}^\ell)\,, \tag{15}$$

in which the potential $\mathcal{V}(\mathbf{u}^\ell)$ defined on $\mathbb{R}^\nu$ with values in $\mathbb{R}^+$, is defined by

$$\mathcal{V}(\mathbf{u}^\ell) = -\log\{q(\mathbf{u}^\ell)\}\,, \tag{16}$$

where $\mathbf{u}^\ell \mapsto q(\mathbf{u}^\ell)$ is the continuously differentiable function from $\mathbb{R}^\nu$ into $]0\,, +\infty[$ such that

$$q(\mathbf{u}^\ell) = \frac{1}{N} \sum_{j=1}^{N} \exp\{-\frac{1}{2\widehat{s}_\nu^2} \|\frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell\|^2\}\,. \tag{17}$$

From Eqs. (16) and (17), it can be deduced that,

$$[L([u])]_{k\ell} = \frac{1}{q(\mathbf{u}^\ell)}\, \{\boldsymbol{\nabla}_{\mathbf{u}^\ell}\, q(\mathbf{u}^\ell)\}_k\,, \tag{18}$$

$$\boldsymbol{\nabla}_{\mathbf{u}^\ell}\, q(\mathbf{u}^\ell) = \frac{1}{\widehat{s}_\nu^2} \frac{1}{N} \sum_{j=1}^{N} (\frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell)\, \exp\{-\frac{1}{2\widehat{s}_\nu^2} \|\frac{\widehat{s}_\nu}{s_\nu} \boldsymbol{\eta}^{d,j} - \mathbf{u}^\ell\|^2\}\,. \tag{19}$$

(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 realization of $[\mathbf{H}_d]$ is matrix $[\eta_d]$. 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.

Since the columns $\mathbf{H}^1, \ldots, \mathbf{H}^N$ of random matrix $[\mathbf{H}]$ are independent copies of random vector $\mathbf{H}$, and since the pdf of random matrix $[\mathbf{H}_d]$ is $p_{[\mathbf{H}]}$, using Theorems 4 to 7 in pages 211 to 216 of Ref. [27], in which the Hamiltonian is taken as $\mathcal{H}(\mathbf{u},\mathbf{v}) = \|\mathbf{v}\|^2/2 + \mathcal{V}(\mathbf{u})$, and using [28, 29] for proving the ergodic property, it can be proved that the problem defined by Eqs. (12) to (14) admits a unique invariant measure and a unique solution $\{([\mathbf{U}(r)], [\mathbf{V}(r)]), r \in \mathbb{R}^+\}$ that is a second-order diffusion stochastic process, which is stationary (for the shift semi-group on $\mathbb{R}^+$ defined by the positive shifts $r \mapsto r + \tau$, $\tau \geq 0$) and ergodic, and such that, for all r fixed in $\mathbb{R}^+$, the probability distribution of random matrix $[\mathbf{U}(r)]$ is $p_{[\mathbf{H}]}([\eta])\,d[\eta]$ in which $p_{[\mathbf{H}]}$ is defined by Eq. (11).

Remarks.

1. It should be noted that the invariant measure is independent of f_0.

2. If the initial condition $[\mathbf{U}(0)]$ was not $[\mathbf{H}_d]$ but was any other random matrix whose pdf is not $p_{[\mathbf{H}]}$, then the unique diffusion process $\{([\mathbf{U}(r)], [\mathbf{V}(r)]), r \in \mathbb{R}^+$ would not be stationary, but would be asymptotic (for $r \to +\infty$) to a stationary diffusion process $\{([\mathbf{U}_{\text{st}}(r_{\text{st}})], [\mathbf{V}_{\text{st}}(r_{\text{st}})]), r_{\text{st}} \geq 0\}$ such that, for all $r_{\text{st}} > 0$, $[\mathbf{H}] = [\mathbf{U}_{\text{st}}(r_{\text{st}})] = \lim_{r\to+\infty}[\mathbf{U}(r)]$ in probability distribution (this implies that, for all $r_{\text{st}} > 0$, the pdf of random matrix $[\mathbf{U}_{\text{st}}(r_{\text{st}})]$ is $p_{[\mathbf{H}]}$). In such a case, the free parameter $f_0 > 0$ allows the transient response generated by the initial condition to be rapidly killed in order to get more rapidly the asymptotic behavior corresponding to the stationary and ergodic solution associated with the invariant measure.

3. As the nonparametric estimate $p_{[\mathbf{H}]}$ of the pdf of $[\mathbf{H}]$ does not explicitly take into account the local structure of dataset $[\eta_d]$, if the pdf of $\mathbf{H}$ is concentrated on $\mathcal{S}_\nu$, then the generator of realizations constructed by the MCMC method defined by Eqs. (12) to (14) (or by any other MCMC method), will not give some realizations localized in the subset $\mathcal{S}_\nu$ (see the applications in Section 5).

4. As explained in [21], a variant of Eq. (13) could be introduced in replacing it by $d[\mathbf{V}(r)] = [L([\mathbf{U}(r)])]\, dr - \frac{1}{2} f_0\, [D_0]\, [\mathbf{V}(r)]\, dr + \sqrt{f_0}\, [S_0]\, [d\mathbf{W}(r)]$ in which $[S_0]$ would belong to $\mathbb{M}_\nu$ and where $[D_0]$ would be a positive symmetric matrix such that $[D_0] = [S_0]\, [S_0]^T$ with $1 \leq \text{rank}[D_0] \leq \nu$. In the present case, such an extension would not allow for improving the methodology proposed because the initial condition for $[\mathbf{U}(0)]$ is the given matrix $[\eta_d]$ that follows $p_{[\mathbf{H}]}$.

5. For θ fixed in Θ, let $\{[\mathbf{W}(r;\theta)], r \geq 0\}$, $[\mathbf{H}_d(\theta)] = [\eta_d]$, and $[\boldsymbol{\mathcal{N}}(\theta)]$ be independent realizations of the stochastic process $\{[\mathbf{W}(r)], r \geq 0\}$, the random matrix $[\mathbf{H}_d]$, and the random matrix $[\boldsymbol{\mathcal{N}}]$. Let $\{([\mathbf{U}(r;\theta)], [\mathbf{V}(r;\theta)]), r \in \mathbb{R}^+\}$ be the corresponding realization of the unique stationary diffusion process $\{([\mathbf{U}(r)], [\mathbf{V}(r)]), r \in \mathbb{R}^+\}$ of the ISDE problem defined by Eqs. (12) to (14)). Then additional realizations $[\eta_s^1], \ldots, [\eta_s^{n_{\text{MC}}}]$ of random matrix $[\mathbf{H}]$ can be generated by

$$[\eta_s^\ell] = [\mathbf{U}(\ell\rho;\theta)] \quad , \quad \rho = M_0\, \Delta r \quad , \quad \ell = 1, \ldots, n_{\text{MC}}\,, \tag{20}$$

in which Δr is the sampling step of the continuous index parameter r used in the integration scheme (see Section 4.7.1) and where M_0 is a positive integer:

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

- 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 $[\eta_s^1], \ldots, [\eta_s^{n_{\text{MC}}}]$ can approximatively be considered as independent realizations of random matrix $[\mathbf{H}]$. We underscore here that each sample of random matrix $[\eta_s]$ consists of N simultaneous samples of random vector $\mathbf{H}$ inherit additional statistical properties from the matrix structure of $[\mathbf{H}]$ to ensure their coalescence around the low-dimensional structure $\mathcal{S}_n$.

4.4. Construction of a diffusion-maps basis $[g]$

Let $k_\varepsilon(\boldsymbol{\eta}, \boldsymbol{\eta}')$ be the kernel defined on $\mathbb{R}^\nu \times \mathbb{R}^\nu$, depending on a real smoothing parameter $\varepsilon > 0$, which verifies the following properties:

- $k_\varepsilon(\boldsymbol{\eta}, \boldsymbol{\eta}') = k_\varepsilon(\boldsymbol{\eta}', \boldsymbol{\eta})$ (symmetry).
- $k_\varepsilon(\boldsymbol{\eta}, \boldsymbol{\eta}') \geq 0$ (positivity preserving).
- k_ε is positive semi-definite.

A classical choice (that we will use in Section 5) for the kernel that satisfies the above three properties is the Gaussian kernel specified as,

$$k_\varepsilon(\boldsymbol{\eta}, \boldsymbol{\eta}') = \exp(-\frac{1}{4\varepsilon}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|^2)\,. \tag{21}$$

Let $[K]$ be the symmetric matrix in $\mathbb{M}_N$ with positive entries such that

$$[K]_{ij} = k_\varepsilon(\boldsymbol{\eta}^{d,i}, \boldsymbol{\eta}^{d,j}) \quad , \quad i \text{ and } j \in \{1, \ldots, N\}\,. \tag{22}$$

Let $[b]$ be the positive-definite diagonal real matrix in $\mathbb{M}_N$ such that

$$[b]_{ij} = \delta_{ij} \sum_{j'=1}^{N} [K]_{ij'}\,, \tag{23}$$

and let $[\mathbb{P}]$ be the matrix in $\mathbb{M}_N$ such that

$$[\mathbb{P}] = [b]^{-1}\,[K]\,. \tag{24}$$

Consequently, matrix $[\mathbb{P}]$ has positive entries and satisfies $\sum_{j=1}^{N} [\mathbb{P}]_{ij} = 1$ for all $i = 1, \ldots, N$. It can thus be viewed as the transition matrix of a Markov chain that yields the probability of transition in one step. Let $[\mathbb{P}_S]$ be the symmetric matrix in $\mathbb{M}_N$ such that

$$[\mathbb{P}_S] = [b]^{1/2}\,[\mathbb{P}]\,[b]^{-1/2} = [b]^{-1/2}\,[K]\,[b]^{-1/2}\,. \tag{25}$$

We consider the eigenvalue problem $[\mathbb{P}_S]\,\boldsymbol{\phi}^\alpha = \lambda_\alpha\,\boldsymbol{\phi}^\alpha$. Let m be an integer such that $1 < m \leq N$. It can easily be proved that the associated eigenvalues are real, positive, and such that

$$1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_m\,. \tag{26}$$

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 eigenvalues of matrix $[\mathbb{P}]$ are the same as the eigenvalues of matrix $[\mathbb{P}_S]$. 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

$$\boldsymbol{\psi}^\alpha = [b]^{-1/2} \boldsymbol{\phi}^\alpha \in \mathbb{R}^N \quad , \quad \alpha = 1, \ldots, m \,, \tag{27}$$

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

$$[\psi]^T [b] [\psi] = [I_m] \,, \tag{28}$$

which defines the normalization of the right eigenvectors of $[\mathbb{P}]$.

We then define a "diffusion-maps basis" 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

$$\mathbf{g}^\alpha = \lambda_\alpha^\kappa \boldsymbol{\psi}^\alpha \in \mathbb{R}^N \quad , \quad \alpha = 1, \ldots, m \,, \tag{29}$$

in which κ 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 $\{\Psi_\kappa\}_\kappa$ of diffusion maps are defined [6, 7] by the vector $\Psi_\kappa = (\lambda_1^\kappa \boldsymbol{\psi}^1, \ldots, \lambda_m^\kappa \boldsymbol{\psi}^m)$ in order to construct a diffusion distance, and integer κ is thus such that the probability of transition is in κ steps. However, as we have previously explained, we do not use such a diffusion distance, but we use the "diffusion-maps basis" $\{\mathbf{g}^1 \ldots \mathbf{g}^N\}$ that we have introduced for performing a projection of 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 $\mathbb{M}_{\nu,m}$, we can then construct the following reduced-order representation of $[\mathbf{H}]$,

$$[\mathbf{H}] = [\mathbf{Z}] [g]^T \,. \tag{30}$$

Since the matrix $[g]^T [g] \in \mathbb{M}_m$ is invertible, Eq. (30) yields

$$[\mathbf{Z}] = [\mathbf{H}] [a] \quad , \quad [a] = [g] ([g]^T [g])^{-1} \in \mathbb{M}_{N,m} \,. \tag{31}$$

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 written as

$$[z_d] = [\eta_d] [a] \in \mathbb{M}_{\nu,m} \,. \tag{32}$$

4.5. Estimating dimension m of the reduced-order representation of random matrix $[\boldsymbol{H}]$

Because an estimation of the value of the order-reduction dimension m must be known before beginning the generation of additional realizations of random matrix $[\mathbf{Z}]$ using the reduced-order representation of random matrix $[\mathbf{H}]$, we propose a methodology which is only based on the use of the known dataset represented by matrix $[\eta_d]$ that is a realization of random matrix $[\mathbf{H}]$.

For a given value of integer κ and for a given value of smoothing parameter $\varepsilon > 0$, the decay of the graph $\alpha \mapsto \lambda_\alpha$ of the eigenvalues of transition matrix $[\mathbb{P}]$, 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 can be misleading as it does not capture statistical fluctuations around the embedded manifold. An additional mean-square convergence must be verified, and if necessary, the value of m must be increased. 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 applied 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.

Using Eqs. (30) to (32) allows for calculating the reduced-order representation $[\eta_{\text{red}}(m)] \in \mathbb{M}_{\nu,N}$ of $[\eta_d]$ such that $[\eta_{\text{red}}(m)] = [\eta_d]\,[a]\,[g]^T$ in which $[a]$ and $[g]$ depend on m. It should be noted that if $m = N$, then $[a]\,[g]^T = [I_N]$ and therefore, $[\eta_{\text{red}}(m)] = [\eta_d]$. In such a case, the "reduced-order" representation would correspond to a simple change of vector basis in $\mathbb{R}^N$ and the localization of the geometric structure of the dataset would be lost. This implies that m must be much more less than N for preserving the capability of the approach to localize the geometric structure of the dataset, and must be chosen as the smallest possible value that yields a reasonable mean-square convergence. Let $[x_{\text{red}}(m)] \in \mathbb{M}_{n,N}$ be the matrix $[x_d]$ of the dataset, calculated using Eq. (3) with $[\eta_{\text{red}}(m)]$. We then have

$$[x_{\text{red}}(m)] = [\underline{x}] + [\varphi]\,[\mu]^{1/2}\,[\eta_d]\,[a]\,[g]^T\,. \tag{33}$$

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 $\mathbf{m}_{\text{red}} \in \mathbb{R}^N$ and $[c_{\text{red}}] \in \mathbb{M}_N$ of the mean value and the covariance matrix calculated with the

realization $[x_{\text{red}}(m)] \in \mathbb{M}_{n,N}$ such that

$$\mathbf{m}_{\text{red}}(m) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_{\text{red}}^{j}(m)\,, \tag{34}$$

$$[c_{\text{red}}(m)] = \frac{1}{N-1} \sum_{j=1}^{N} (\mathbf{x}_{\text{red}}^{j}(m) - \mathbf{m}_{\text{red}})\,(\mathbf{x}_{\text{red}}^{j}(m) - \mathbf{m}_{\text{red}})^T\,. \tag{35}$$

The mean-square convergence criterion is then defined by

$$e_{\text{red}}(m) = \frac{\|[c_{\text{red}}(m)] - [c]\|_F}{\|[c]\|_F}\,. \tag{36}$$

in which $[c]$ is defined by Eq. (2). 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 $e_{\text{red}}(m) \leq \varepsilon_0$. An illustration of the use of this criterion will be given in the third application presented in Section 5.3.

*4.6. Reduced-order ISDE for generation of additional realizations of random matrix [**X**]*

For m, ε, and κ fixed, the reduced-order representation $[\mathbf{H}] = [\mathbf{Z}]\,[g]^T$ of random matrix $[\mathbf{H}]$, defined by Eq. (30), is used for constructing the reduced-order ISDE associated with Eqs. (12) to (14). Introducing the change of stochastic processes $[\mathbf{U}(r)] = [\boldsymbol{\mathcal{Z}}(r)]\,[g]^T$ and $[\mathbf{V}(r)] = [\boldsymbol{\mathcal{Y}}(r)]\,[g]^T$ into these equations, then right multiplying the obtained equations by matrix $[a]$, and taking into account Eq. (31), it can be seen that $\{([\boldsymbol{\mathcal{Z}}(r)], [\boldsymbol{\mathcal{Y}}(r)]), r \in \mathbb{R}^+\}$ is a Markov stochastic process defined on the probability space $(\Theta, \mathcal{T}, \mathcal{P})$, indexed by $\mathbb{R}^+ = [0\,, +\infty[$, with values in $\mathbb{M}_{\nu,m} \times \mathbb{M}_{\nu,m}$, satisfying, for all $r > 0$, the following reduced-order ISDE,

$$d[\boldsymbol{\mathcal{Z}}(r)] = [\boldsymbol{\mathcal{Y}}(r)]\,dr\,, \tag{37}$$

$$d[\boldsymbol{\mathcal{Y}}(r)] = [\mathcal{L}([\boldsymbol{\mathcal{Z}}(r)])]\,dr - \frac{1}{2} f_0\,[\boldsymbol{\mathcal{Y}}(r)]\,dr + \sqrt{f_0}\,[d\boldsymbol{\mathcal{W}}(r)]\,, \tag{38}$$

with the initial condition

$$[\boldsymbol{\mathcal{Z}}(0)] = [\mathbf{H}_d]\,[a] \quad , \quad [\boldsymbol{\mathcal{Y}}(0)] = [\boldsymbol{\mathcal{N}}]\,[a] \quad a.s\,, \tag{39}$$

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

$$[\mathcal{L}([\boldsymbol{\mathcal{Z}}(r)])] = [L([\boldsymbol{\mathcal{Z}}(r)]\,[g]^T)]\,[a]\,, \tag{40}$$

$$[d\boldsymbol{\mathcal{W}}(r)] = [d\mathbf{W}(r)]\,[a]\,. \tag{41}$$

From Section 4.3, it can be deduced that the problem defined by Eqs. (37) to (41) admits a unique invariant measure and a unique solution $\{([\boldsymbol{\mathcal{Z}}(r)], [\boldsymbol{\mathcal{Y}}(r)]), r \in \mathbb{R}^+\}$ that is a second-order diffusion stochastic process, which is stationary (for the shift semi-group on $\mathbb{R}^+$) and ergodic.

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 realizations 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 realization of the unique stationary diffusion process $\{([\boldsymbol{\mathcal{Z}}(r)], [\boldsymbol{\mathcal{Y}}(r)]), r \in \mathbb{R}^+\}$ of the reduced-order ISDE problem defined by Eqs. (37) to (39)). Then, using Eq. (30), some additional realizations $[\eta_s^1], \ldots, [\eta_s^{n_{\text{MC}}}]$ of random matrix $[\mathbf{H}]$ can be generated by

$$[\eta_s^\ell] = [\boldsymbol{\mathcal{Z}}(\ell\rho;\theta)]\,[g]^T \quad , \quad \rho = M_0\,\Delta r \quad , \quad \ell = 1, \ldots n_{\text{MC}}\,, \tag{42}$$

and using Eq. (3), some additional realizations $[x_s^1], \ldots, [x_s^{n_{\text{MC}}}]$ of random matrix $[\mathbf{X}]$ can be generated (using the reduced-order representation defined by Eq. (3)) by

$$[x_s^\ell] = [\underline{x}] + [\varphi]\,[\mu]^{1/2}\,[\eta_s^\ell] \quad , \quad \ell = 1, \ldots n_{\text{MC}}\,. \tag{43}$$

4.7. Solving the reduced-order ISDE and computing the additional realizations for random matrix $[\boldsymbol{X}]$

For numerically solving the reduced-order ISDE defined by Eqs. (37) to (39), a discretization scheme must be used. For general surveys on discretization schemes for Itô stochastic differential equations, we refer the reader to [30, 31, 32]. Concerning the particular cases related to Hamiltonian dynamical systems (which have also been analyzed in [33] using an implicit Euler scheme), we propose to use the Störmer-Verlet scheme, which is a very efficient scheme that preserves energy for nondissipative Hamiltonian dynamical systems (see [34] for reviews about this scheme in the deterministic case, and see [35] and the references therein for the stochastic case).

4.7.1. Discretization scheme of the reduced-order ISDE

We then propose to reuse hereinafter the Störmer-Verlet scheme, introduced and validated in [36, 37, 21] for weakly dissipative stochastic Hamiltonian dynamical system.

Let $M = n_{\text{MC}} \times M_0$ be the positive integer in which n_{MC} and M_0 have been introduced in Remark 5 of Section 4.3. The reduced-order Itô stochastic differential equation defined by Eqs. (37) and (38) with the initial condition defined by Eq. (39), 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_ℓ 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

$$[\boldsymbol{\mathcal{Z}}_0] = [\mathbf{H}_d]\,[a] \quad , \quad [\boldsymbol{\mathcal{Y}}_0] = [\boldsymbol{\mathcal{N}}]\,[a] \quad , \quad [\boldsymbol{\mathcal{W}}_0] = [0_{\nu,m}] \quad a.s\,. \tag{44}$$

For $\ell = 0, \ldots, M-1$, let $[\Delta\boldsymbol{\mathcal{W}}_{\ell+1}] = [\Delta\mathbf{W}_{\ell+1}]\,[a]$ 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. 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 $E\{[\Delta\mathbf{W}_{\ell+1}]_{kj}[\Delta\mathbf{W}_{\ell+1}]_{k'j'}\} = \Delta r\,\delta_{kk'}\,\delta_{jj'}$. For $\ell = 0, \ldots, M-1$, the Störmer-Verlet scheme applied to Eqs. (37) and (38) yields

$$[\boldsymbol{\mathcal{Z}}_{\ell+\frac{1}{2}}] = [\boldsymbol{\mathcal{Z}}_\ell] + \frac{\Delta r}{2}\,[\boldsymbol{\mathcal{Y}}_\ell]\,, \tag{45}$$

$$[\boldsymbol{\mathcal{Y}}_{\ell+1}] = \frac{1-b}{1+b}\,[\boldsymbol{\mathcal{Y}}_\ell] + \frac{\Delta r}{1+b}\,[\boldsymbol{\mathcal{L}}_{\ell+\frac{1}{2}}] + \frac{\sqrt{f_0}}{1+b}\,[\Delta\boldsymbol{\mathcal{W}}_{\ell+1}]\,, \tag{46}$$

$$[\boldsymbol{\mathcal{Z}}_{\ell+1}] = [\boldsymbol{\mathcal{Z}}_{\ell+\frac{1}{2}}] + \frac{\Delta r}{2}\,[\boldsymbol{\mathcal{Y}}_{\ell+1}]\,, \tag{47}$$

with the initial condition defined by (44), 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

$$[\boldsymbol{\mathcal{L}}_{\ell+\frac{1}{2}}] = [\mathcal{L}([\boldsymbol{\mathcal{Z}}_{\ell+\frac{1}{2}}])] = [L([\boldsymbol{\mathcal{Z}}_{\ell+\frac{1}{2}}]\,[g]^T)]\,[a]\,, \tag{48}$$

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. (18) and (19).

4.7.2. Remarks about the estimation of the numerical integration parameters of the reduced-order ISDE

Some estimations of the values of the parameters $f_0, \Delta r$, and M_0, which are used in the discretization scheme of the ISDE (with and without reduced-order representation of random matrix $[\mathbf{H}]$, and introduced in Sections 4.3 and 4.7.1) are described below.

(i) Parameter Δr is written as $\Delta r = 2\pi\,\widehat{s}_\nu/\text{Fac}$ in which $\text{Fac} > 1$ is an oversampling that has to be estimated for getting a sufficient accuracy of the Störmer-Verlet scheme (for instance, $\text{Fac} = 20$). This means that a convergence analysis of the solution must be carried out with respect to Fac.

(ii) As the accuracy of the Störmer-Verlet scheme is finite, a small numerical integration error is unavoidably introduced. Although that the initial conditions are chosen in order to directly construct the stationary solution (associated with the unique invariant measure), a small transient response can occur and be superimposed to the stationary stochastic solution. Therefore, f_0 is chosen in order that the damping in the dissipative Hamiltonian system is sufficiently large to rapidly kill such a small transient response (a typical value that is retained in the applications presented in Section 5 is $f_0 = 1.5$).

(iii) Using an estimation of the relaxation time of the underlying linear second-order dynamical system, and choosing an attenuation of $1/100$ for the transient response, parameter M_0 must be chosen larger than $2\log(100)\text{Fac}/(\pi f_0\,\widehat{s}_\nu)$. A typical value that is retained in the applications presented in Section 5 is $M_0 = 110$ or 330).

5. Applications

Three applications are presented for random vector $\mathbf{X}$ with values in $\mathbb{R}^n$ for which:

- 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).
- 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).

- the third example corresponds to a petro-physics database that is made up of experimental measurements (downloaded from [38]) and detailed in [20], 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 it is discovered.

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

For this first application, two cases are considered: small (case 1.1) and medium (case 1.2) statistical fluctuations around the two circles. For every case, the number of given data points is $N = 230$, no scaling of data is performed, but the normalization defined in Section 4.1 is done and yields $\nu = 2$. In Figs. 1 to 5, the left figures are relative to case 1.1 and the right ones to case 1.2. Fig. 1 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, with small (case 1.1) and medium (case 1.2) statistical fluctuations. The kernel is defined by Eq. (21), the value

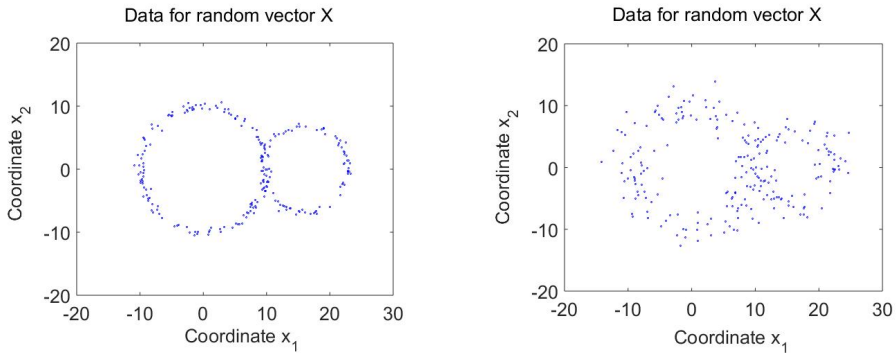


Figure 1: 230 given data points: case 1.1 (left), case 1.2 (right).

of the smoothing parameter that is retained is $\varepsilon = 2.7318$, κ is chosen to 1, and the graph of the eigenvalues of the transition matrix for random vector $\mathbf{H}$ is displayed in Fig. 2. These two graphs show that dimension m can be chosen to 3, and for $m = 3$, the value of $e_{\text{red}}(m)$ (defined by Eq. (36)) is 6.34×10^{-4} for case 1.1 and 9.28×10^{-4} for case 1.2. It can thus be considered that a reasonable mean-square convergence is reached for these two cases. Fig. 3 displays the pdf for random variables X_1 and X_2 computed with a nonparametric estimation from the data points. For all the computation, the numerical values of the parameters for generating $9,200$ additional realizations are $\Delta r = 0.1179$, $M_0 = 110$, and

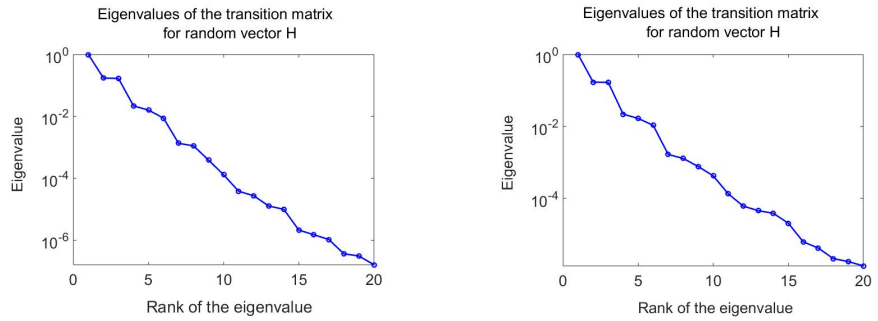


Figure 2: Eigenvalues in $\log_{10}$-scale of the transition matrix for random vector $\mathbf{H}$: case 1.1 (left), case 1.2 (right).

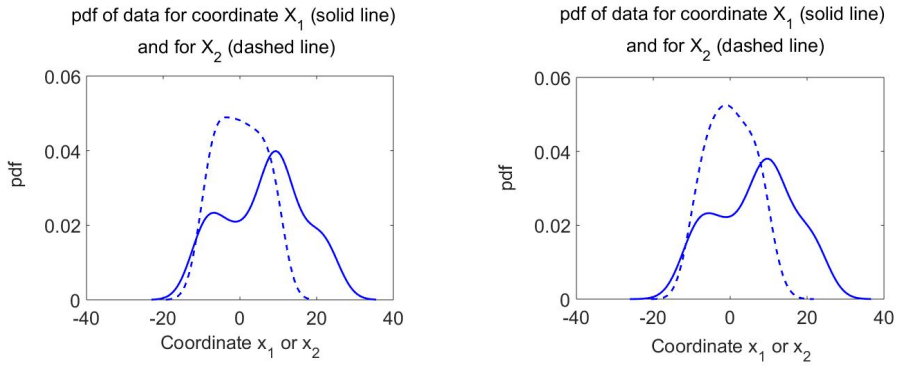


Figure 3: pdf for random variables X_1 (solid line) and X_2 (dashed line) obtained by a nonparametric estimation from data points: case 1.1 (left), case 1.2 (right).

$n_{\text{MC}} = 40$, yielding $M = 4,400$. The results obtained with the reduced-order ISDE (for which the first $m = 3$ vectors of the diffusion-maps basis are used) are displayed in Fig. 4, which shows the 230 given data points and the $9,200$ additional realizations generated using the reduced-order ISDE. It can be seen that the additional realizations are effectively concentrated in subset $\mathcal{S}_n$. Fig. 5 displays the 230 given data points and the $9,200$ additional realizations generated using a direct simulation of the ISDE presented in Section 4.3. It can be seen that the realizations are not concentrated in subset $\mathcal{S}_n$, but are scattered.

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

As previously, two cases are considered: small (case 2.1) and medium (case 2.2) statistical fluctuations around the helical. For every case, the number of given

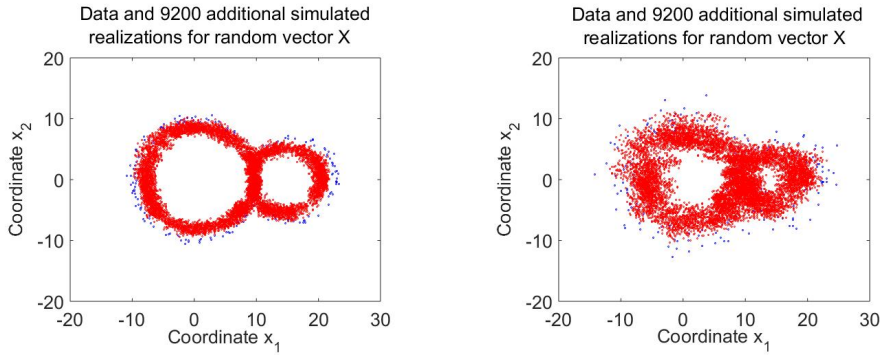


Figure 4: 230 given data points (blue symbols) and $9,200$ additional realizations (red symbols) generated using the reduced-order ISDE with $m = 3$: case 1.1 (left), case 1.2 (right).

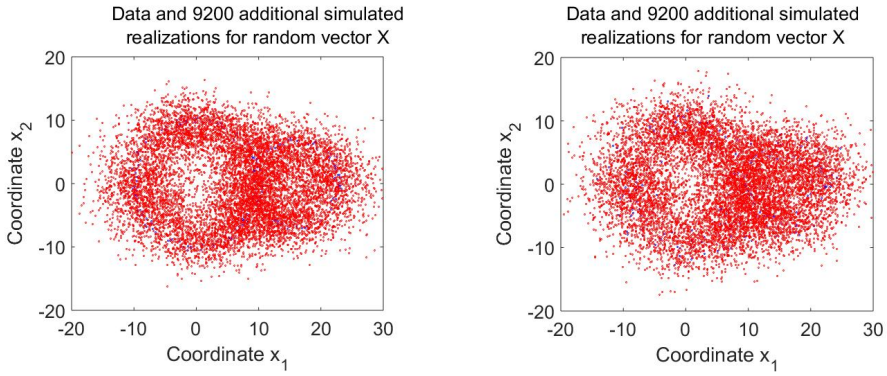


Figure 5: 230 given data points (blue symbols) and $9,200$ additional realizations (red symbols) generated using the ISDE: case 1.1 (left), case 1.2 (right).

data points is $N = 400$, no scaling of data is performed, but the normalization defined in Section 4.1 is done and yields $\nu = 3$. In Figs. 6 to 10, the left figures are relative to case 2.1 and the right ones to case 2.2. Fig. 6 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 $\mathbb{M}_{3,400}$. Fig. 6 shows that the given data points are concentrated in the neighborhood of the helical, with small (case 2.1) and medium (case 2.2) statistical fluctuations. The kernel is defined by Eq. (21), the value of the smoothing parameter that is retained is $\varepsilon = 1.57$, κ is chosen to 1, and the graph of the eigenvalues of the transition matrix for random vector $\mathbf{H}$ is displayed in Fig. 7. These two graphs show that dimension m can be chosen to 4, and for $m = 4$, the value of $e_{\text{red}}(m)$ (defined by Eq. (36)) is 5.53×10^{-4} for case 2.1

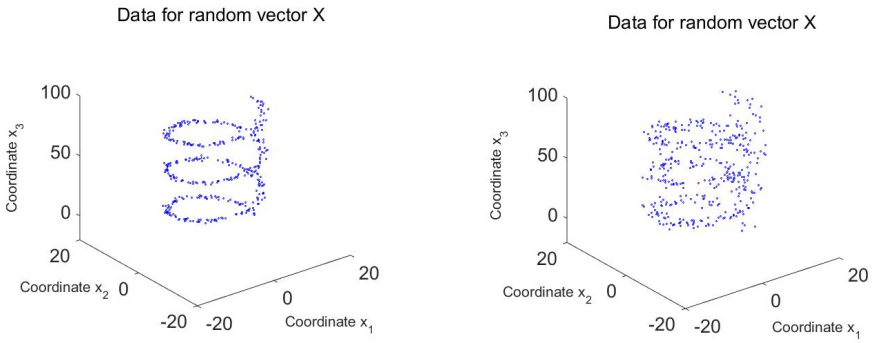


Figure 6: 400 given data points: case 2.1 (left), case 2.2 (right).

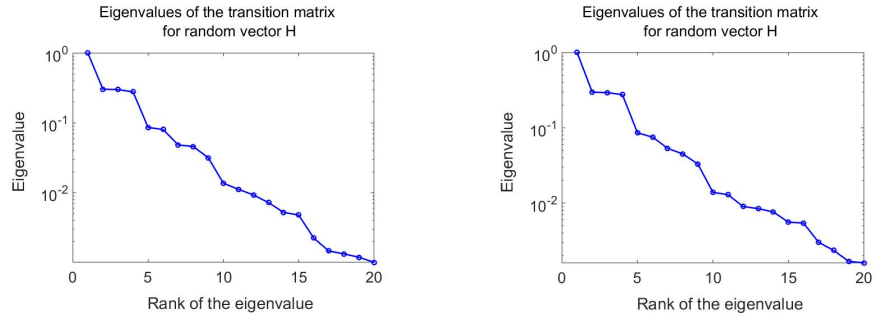


Figure 7: Eigenvalues in $\log_{10}$-scale of the transition matrix for random vector $\mathbf{H}$: case 2.1 (left), case 2.2 (right).

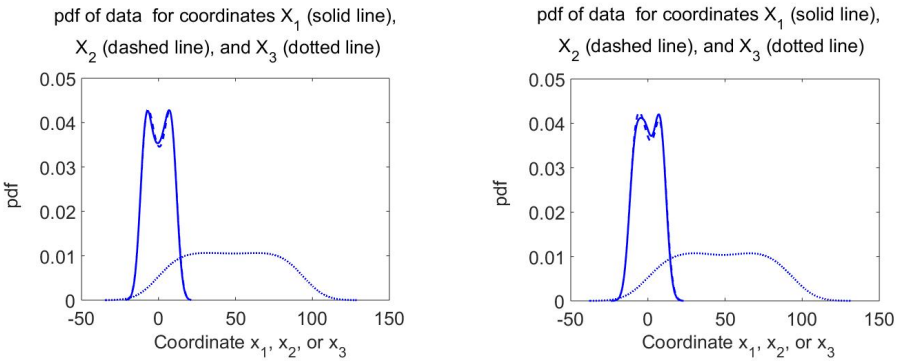


Figure 8: pdf for random variables X_1 (solid line), X_2 (dashed line), and X_3 (dotted line) obtained by a nonparametric estimation from data points: case 2.1 (left), case 2.2 (right).

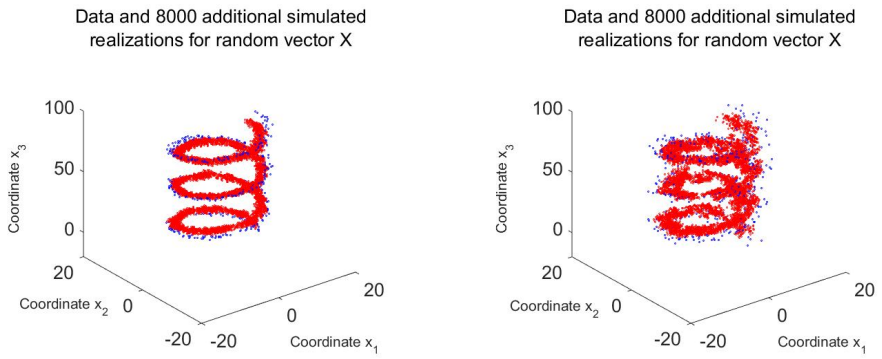


Figure 9: 400 given data points (blue symbols) and $8,000$ additional realizations (red symbols) generated using the reduced-order ISDE with $m = 4$: case 2.1 (left), case 2.2 (right).

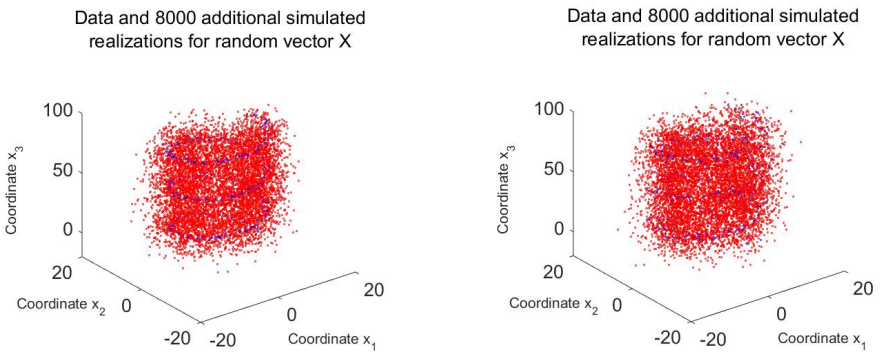


Figure 10: 400 given data points (blue symbols) and $8,000$ additional realizations (red symbols) generated using the ISDE: case 2.1 (left), case 2.2 (right).

and 4.28×10^{-4} for case 2.2. It can thus be considered that a reasonable mean-square convergence is reached for these two cases. Fig. 8 displays the pdf for random variables X_1, X_2, and X_3 computed with a nonparametric estimation from the data points. For all the computation, the numerical values of the parameters for generating $9,200$ additional realizations are $\Delta r = 0.1196$, $M_0 = 110$, and $n_{\text{MC}} = 20$, yielding $M = 2,200$. The results obtained with the reduced-order ISDE (for which the first $m = 4$ vectors of the diffusion-maps basis are used) are displayed in Fig. 9, which shows the 400 given data points and the $8,000$ additional realizations generated using the reduced-order ISDE. It can be seen that the additional realizations are effectively concentrated in subset $\mathcal{S}_n$. Fig. 10 displays the 400 given data points and the $8,000$ additional realizations generated

using a direct simulation with the ISDE presented in Section 4.3. It can be seen that the realizations are not concentrated in subset $\mathcal{S}_n$, but are scattered.

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

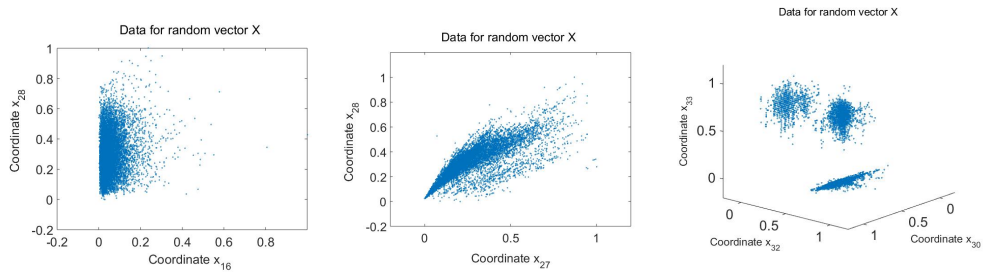


Figure 11: $13,056$ given data points viewed from coordinates x_{16} and x_{28} (up left), viewed from coordinates x_{27} and x_{28} (up right), and viewed from coordinates x_{30}, x_{32}, and x_{33} (down).

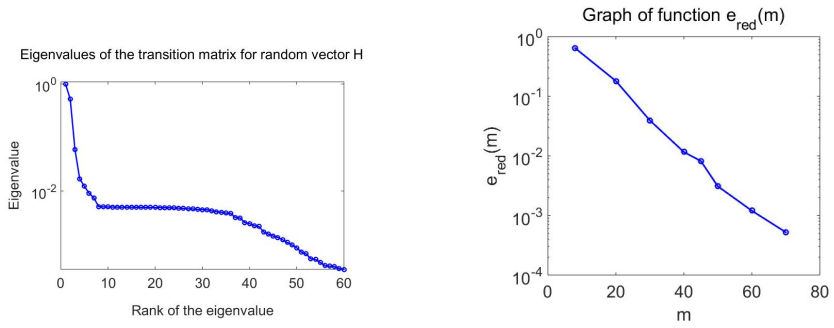


Figure 12: Eigenvalues in $\log_{10}$-scale of the transition matrix for random vector $\mathbf{H}$ (left). Graph $m \mapsto e_{\text{red}}(m)$ in $\log_{10}$ scale (right).

The data base used corresponds to a petro-physics data base of experimental experiments. The dimension of random vector $\mathbf{X}$ is $n = 35$ and the number of given data points is $N = 13,056$. The scaling and the normalization defined in Section 4.1 are necessary, have been done, and yield $\nu = 32$. Fig. 11 displays $13,056$ 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 $13,056$ data points for the $\mathbb{R}^n$-valued random vector $\mathbf{X}$ is given, this figure shows that $\mathcal{S}_n$ is certainly a complex subset of $\mathbb{R}^n$. The kernel is defined by Eq. (21), the value of the smoothing parameter that has been used is $\varepsilon = 100$,

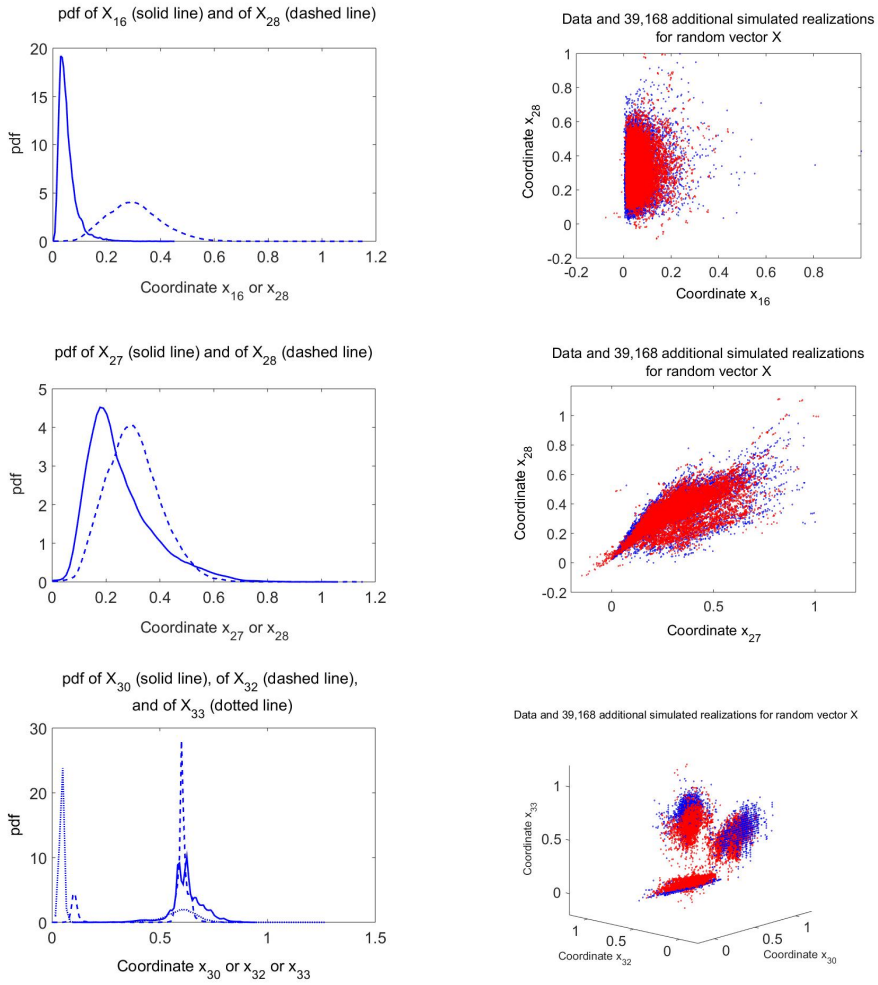


Figure 13: Left figures: Illustration of the pdf for some components of random vector $\mathbf{X}$ obtained by a nonparametric estimation from the data points and the simulated data points. Right figures: $13,056$ given data points (blue symbols) and $39,168$ additional realizations (red symbols) generated using the reduced-order representation of $[\mathbf{H}]$ with $m = 50$, viewed from different components of random vector $\mathbf{X}$.

and κ has also been chosen to 1. The graph of the eigenvalues (of the transition matrix relative to random vector $\mathbf{H}$) displayed in Fig. 12 (left) shows that the value $m = 8$ could potentially be a good choice for the value of m. However, for $m = 8$, the value of $e_{\text{red}}(m)$ is 0.99 that shows that the mean-square convergence is not reached. Consequently, an analysis has been performed in constructing the

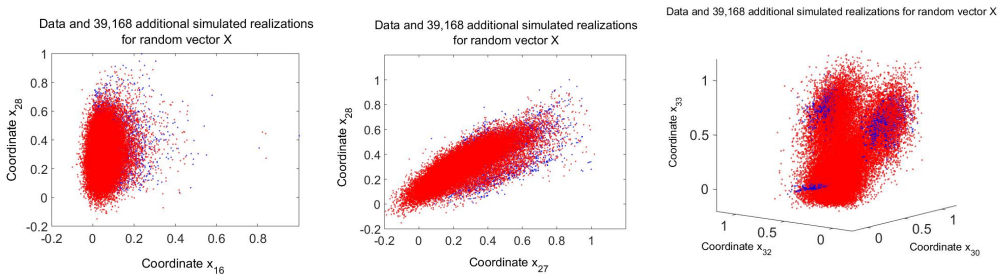


Figure 14: $13,056$ given data points (blue symbols) and $39,168$ additional realizations (red symbols) generated without using the reduced-order representation, viewed from different components of random vector $\mathbf{X}$.

graph of function $m \mapsto e_{\text{red}}(m)$ in order to identify the smallest value of m for which the mean-square convergence is reasonably reached. The graph displays in Fig. 12 (right) clearly shows that a good choice is $m = 50$ for which the value of $e_{\text{red}}(m)$ is 3.08×10^{-3} that can thus be considered as a reasonable mean-square convergence. For all the computation, the numerical values of the parameters for generating $39,168$ additional realizations are $\Delta r = 0.06142$, $M_0 = 330$, and $n_{\text{MC}} = 3$, yielding $M = 990$.

For the same coordinates that those introduced in Fig. 11, the left figures in Fig. 13 display the pdf of the considered components of random vector $\mathbf{X}$ obtained by a nonparametric estimation from the data points and the simulated data points obtained with the reduced-order ISDE, and the right figures display the $13,056$ given data points and the $39,168$ additional realizations generated using the reduced-order ISDE using the first $m = 50$ vectors of the diffusion-maps basis. It can be seen that the additional realizations are effectively concentrated in subset $\mathcal{S}_n$. Fig. 14 displays the $13,056$ given data points and the $39,168$ additional realizations generated using a direct simulation with the ISDE presented in Section 4.3. It can be seen that the realizations are not concentrated in subset $\mathcal{S}_n$, but are scattered. In particular, the positivity of random variable X_{16} is not satisfied.

6. Conclusions

A new methodology has been presented and validated for generating realizations of an $\mathbb{R}^n$-valued random vector, for which the probability distribution is unknown and is concentrated on an unknown subset $\mathcal{S}_n$ of $\mathbb{R}^n$. Both the probability distribution and the subset $\mathcal{S}_n$ are constructed to be statistically consistent with a specified dataset construed as providing initial realizations of the random vector.

The proposed method is robust and can be used for high dimension and for large initial datasets. It is expected that the proposed method will contribute to open new possibilities of developments in many areas of uncertainty quantification and statistical data analysis, in particular in the design of experiments for random parameters.

7. Acknowledgment

Part of this research was supported by the U.S. Department of Energy Office of Advanced Scientific Computing Research.

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

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

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

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

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

[6] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, . 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.

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

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

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

[10] I.T. Jolliffe, Principal Component Analysis (Second Edition), Springer-Verlag, New York, 2002.

[11] K. Karhunen, Über lineare methoden in der wahrscheinlichkeits-rechnung, Annals of Academic Science Fennicade Series A1, Mathematical Physics, 37 (1946) 3-79.

[12] M. Loève, Probability Theory, D. Van Nostrand, Princeton, New Jersey, 1955.

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

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

[15] R.H. Cameron, W.T. Martin, The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals, The Annals of Mathematics, Second Series 48(2) (1947) 385-392.

[16] 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.

[17] G. Perrin, C. Soize, D. Duhamel, C. Funfschilling, Karhunen-Loève expansion revisited for vector-valued random fields: scaling, errors and optimal basis, Journal of Computational Physics 242(1) (2013) 607-622.

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

[19] R. Ghanem, C. Soize, Remarks on stochastic properties of materials through finite deformations, International Journal for Multiscale Computational Engineering, doi: 10.1615/IntJMultCompEng.2015013959, in press (2015).

[20] 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 (Pro-

ceedings of the Dynamic Data-driven Environmental Systems Science Conference, MIT, Cambridge, MA, Nov 5-7, 2014.)

[21] C. Soize, Polynomial chaos expansion of a multimodal random vector, SIAM/ASA Journal on Uncertainty Quantification 3(1) (2015) 34-60.

[22] N. Metropolis, S. Ulam, The Monte Carlo method, Journal of the American Statistical Association 44 (1949) 335-341.

[23] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 109 (1970) 57-97.

[24] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian distribution of images, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAM I-6 (1984) 721-741.

[25] R.M. Neal, Slice sampling, Annals of Statistics 31 (2003) 705-767.

[26] 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.

[27] C. Soize, The Fokker-Planck Equation for Stochastic Dynamical Systems and its Explicit Steady State Solutions, World Scientific, Singapore, 1994.

[28] J.L. Doob, Stochastic Processes, John Wiley and Sons, New York, 1990.

[29] R. Khasminskii, Stochastic Stability of Differential Equations, Series: Stochastic Modelling and Applied Probability, Vol. 66, 2nd edition, Springer, Heidelberg, 2012. Originally published in Russian, by Nauka, Moskow, 1969. First English edition published in 1980 under R.Z. Has'minski in the series Mechanics: Analysis by Sijthoff & Noordhoff.

[30] P.E. Kloeden, E. Platen, Numerical Solution of Stochastic Differentials Equations, Springer-Verlag, Heidelberg, 1992.

[31] D. Talay, L. Tubaro, Expansion of the global error for numerical schemes solving stochastic differential equation, Stochastic Analysis and Applications 8(4) (1990) 94-120.

[32] D. Talay, Simulation and numerical analysis of stochastic differential systems, pp. 54-96, in Probabilistic Methods in Applied Physics, Lecture Notes in Physics, 451, P. Kree and W. Wedig, eds., Springer-Verlag, Heidelberg, 1995.

[33] D. Talay, Stochastic Hamiltonian system: exponential convergence to the invariant measure and discretization by the implicit Euler scheme, Markov Processes and Related Fields 8 (2002) 163-198.

[34] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, Springer-Verlag, Heidelberg, 2002.

[35] K. Burrage, I. Lenane, G. Lythe, Numerical methods for second-order stochastic differential equations, SIAM Journal on Scientific Computing 29 (2007) 245-264.

[36] C. Soize, I.E. Poloskov, Time-domain formulation in computational dynamics for linear viscoelastic media with model uncertainties and stochastic excitation, *Computers and Mathematics with Applications*, doi:10.1016/j.camwa. 2012.09.010, **64**(11), 3594-3612 (2012).

[37] J. Guilleminot, C. Soize, Stochastic model and generator for random fields with symmetry properties: application to the mesoscopic modeling of elastic random media, Multiscale Modeling and Simulation (A SIAM Interdisciplinary Journal) 11(3) (2013) 840-870.

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