Journal of Computational Physics 321 (2016) 242–258

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

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

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 14 May 2016
Accepted 22 May 2016
Available online 27 May 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

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.

1. Introduction

The construction of a generator of realizations from a given dataset related to 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 $\mathbb{R}^n$, 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

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

http://dx.doi.org/10.1016/j.jcp.2016.05.044
0021-9991/© 2016 Elsevier Inc. All rights reserved.

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 [3–5]. The second building block consists of manifold embedding algorithms, where low-dimensional structure is characterized within a larger vector space. Diffusion maps [6–8] 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 [18–20].

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 assumed 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 an 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