Reconstruction of Random Fields Concentrated on an Unknown Curve using Irregularly Sampled Data

Guillaume Perrin, Christian Soize

Reconstruction of random fields concentrated on an unknown curve using irregularly sampled data

Guillaume Perrin[1*] and Christian Soize[2]

[1*]Université Gustave Eiffel, COSYS, Marne-la-Vallée, 77454, France.
[2]Université Gustave Eiffel, MSME UMR 8208 CNRS, 5 bd Descartes, Marne-la-Vallée, 77454, France.

*Corresponding author(s). E-mail(s): guillaume.perrin@univ-eiffel.fr;
Contributing authors: christian.soize@univ-eiffel.fr;

Abstract

In the world of connected automated objects, increasingly rich and structured data are collected daily (positions, environmental variables, etc.). In this work, we are interested in the characterization of the variability of the trajectories of one of these objects (robot, drone, or delivery droid for example) along a particular path from irregularly sampled data in time and space. To do so, we model the position of the considered object by a random field indexed in time, whose distribution we try to estimate (for risk analysis for example). This distribution being by construction concentrated on an unknown curve, two phases are proposed for its reconstruction: a phase of identification of this curve, by clustering and polynomial smoothing techniques, then a phase of statistical inference of the random field orthogonal to this curve, by spectral methods and kernel reconstructions. The efficiency of the proposed approach, both in terms of computation time and reconstruction quality, is illustrated on several numerical applications.

Keywords: Statistical inference, manifold learning, nonparametric representation, data-driven sampling

1 Introduction

An increasing number of physical systems are equipped with sensors measuring their positions (trains, cars, drones, etc.). Often, this information is presented in the form of a sequence of vectors in dimension 2 (in the plane) or 3 (in space) indexed by time. The different recordings of these spatial positions between two particular points can then be considered as particular realizations of a random field $\boldsymbol{X}$. Characterizing this unknown probability distribution plays a central role in uncertainty quantification, operational safety, and data analysis, and is the main objective of this paper.

Of particular interest is the case when this probability distribution is spatially concentrated around a mean curve (i.e. concentrated around the standard path of the system), and when the available information consists of a finite set of realizations of $\boldsymbol{X}$, which are assumed to be non-uniformly discretized in time into a variable number of points. For this work, it is important to note that this mean curve is *a priori* unknown, and that the fluctuations of the observations around this curve will not be due to measurement noise, but rather to the fact that the realizations of $\boldsymbol{X}$ will not follow the same path exactly. And in addition to the mean curve, it is also these fluctuations that we would like to characterize by estimating the distribution of $\boldsymbol{X}$.

Given these irregularly sampled data, one possible method to address this identification problem is to suppose that the searched probability distribution belongs to an algebraic class of probability distributions, which can be mapped from a relatively small number of parameters (for instance, the set of Gaussian random fields whose mean and covariance functions belong to specific parametric classes) [1–3]. Generating new sample paths of $\boldsymbol{X}$ amounts then at identifying the parameters that best suit the available data and, in a second step, at sampling independent realizations associated with the identified parametric probability distribution. However, proposing parameterisations that are sufficiently sophisticated to include the true distribution without being too complicated to allow its estimation, both for the mean curve, and for the fluctuations around the mean curve, can be very difficult in the general case. In this case, one generally prefers to turn to non-parametric approaches [4, 5]. Such approaches may indeed be able to identify the hidden curve behind the data, as was done in [6] or in [7]. However, to the best of the authors' knowledge, these methods most often consider the available information on $\boldsymbol{X}$ as a point cloud rather than as a set of discretized realizations, and in this sense, have difficulty to really exploit the statistical dependencies within the elements of the data set that are associated with the same realization of $\boldsymbol{X}$. This is mainly due to the fact that the speeds at which the system evolves have no reason to be the same from one path to another one, nor to be constant as a function of time along the same path. As a consequence, the realizations of $\boldsymbol{X}$, which are assumed to be discretized in time, will be generally difficult to compare if we plot them as a function of time, even if they share strong similarities once represented in space. And the methods previously introduced were not designed

to move from this parameterisation in time to another parameterisation in space, which we can hope to be more suitable.

To circumvent this difficulty, a two-step approach is proposed in this work. The first step is to identify the mean curve on which the distribution of $\boldsymbol{X}$ is concentrated, using clustering [8] and spline approximation techniques [9]. It is important to note that this first step is not motivated by a desire to reduce the size of the data, but by a desire to break down the problem of estimating the probability distribution of $\boldsymbol{X}$. In this sense, even if they may share some tools, the method proposed for this identification differs from standard nonlinear dimension reduction methods, such as Locally Linear Embedding (LLE) or t-distributed Stochastic Neighbor Embedding techniques [10, 11], in that we are here more interested in the geometric characterization of the curve than in the mapping from the high-dimensional space to the low-dimensional embedding. In a second step, after projection of the data on the identified curve and on its orthogonal space, the statistical properties of the random field to be identified are estimated by spectral decomposition [12–14] and kernel density estimation [6]. As indicated previously, an important difficulty here comes from the choice of the index on which these fluctuations depend. Indeed, to exploit the fact that the probability distribution of $\boldsymbol{X}$ is concentrated on a curve, it seems natural to index $\boldsymbol{X}$ by the curvilinear abscissa along this curve. But by introducing this non-intrinsic parametrisation, it is very likely that we end up with a non-trivial probability distribution for $\boldsymbol{X}$. For example, even if the probability distribution of $\boldsymbol{X}$ is at the beginning relatively simple (stationary in time for example), its approximation from an approximated curve, and thus from a new index, has every chance of no longer verifying these simplifying hypotheses (by becoming non-stationary in the chosen curvilinear abscissa for example).

The outline of this work is as follows. Section 2 introduces the general framework for carrying out this identification in inverse. The coupling between clustering techniques and spline approximation is then described in Section 3, and Section 4 deals with the statistical inference. Since the scope of this work is mainly methodological, it is limited in terms of applications in Section 5 to two analytical configurations that are intended to be the most representative of what we could have to deal with in configurations for reconstructing the trajectories (and their variability) of physical systems from measurements discretized in time and space. Concluding remarks and prospects for this work are finally given in Section 6.

2 Theoretical framework

2.1 Notations

Let d be in $\{2, 3\}$, $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $\boldsymbol{X}$ be the random field whose probability distribution is to be identified. This random field can, for example, characterize a set of realistic and representative trajectories of a particular physical system between two relatively well identified points. This random field is assumed to be a second-order random field defined on $(\Omega, \mathcal{A}, \mathbb{P})$,

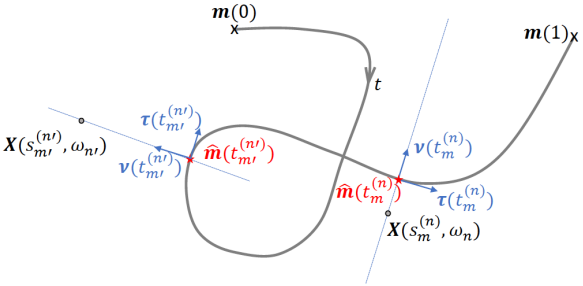


Figure 1 Graphical illustration of the notations, the black solid line being one potential $\mathcal{M} \in \mathcal{M}_1(\mathbb{R}^d)$.

whose trajectories are in the set $\mathcal{C}^2([0,1], \mathbb{R}^d)$ of all the twice continuously differentiable functions from $[0,1]$ to $\mathbb{R}^d$ almost-surely:

$$\boldsymbol{X} := \left\{ \boldsymbol{X}(s,\omega) \in \mathbb{R}^d,\ s \in [0,1],\ \omega \in \Omega \right\}.$$

The probability distribution of $\boldsymbol{X}$ is moreover assumed to be concentrated on an unknown curve $\mathcal{M}^\star$, which is assumed to belong to the set $\mathcal{M}_1(\mathbb{R}^d)$ of simply connected and twice-differentiable curves. It is important to note the strong link between $\mathcal{C}^2([0,1], \mathbb{R}^d)$ and $\mathcal{M}_1(\mathbb{R}^d)$. Indeed, each function in $\mathcal{C}^2([0,1], \mathbb{R}^d)$ defines a unique curve of $\mathcal{M}_1(\mathbb{R}^d)$. And we obtain an element of $\mathcal{C}^2([0,1], \mathbb{R}^d)$ by indexing each point of a curve of $\mathcal{M}_1(\mathbb{R}^d)$ by the curvilinear abscissa defined along this manifold, which has been normalized so that it is equal to 0 at one end of the curve and 1 at the other.

To identify the probability distribution of $\boldsymbol{X}$, which is *a priori* unknown, we assume that we have access to $N > 1$ discrete projections of N independent trajectories of $\boldsymbol{X}$, which are denoted by

$$\left\{ \boldsymbol{X}(s_1^{(n)}, \omega_n), \ldots, \boldsymbol{X}(s_{M_n}^{(n)}, \omega_n) \right\}_{n=1}^{N}, \tag{1}$$

where M_n corresponds to the number of observation points of the n^{th} trajectory. It is important to note that the values of $s_m^{(n)}$, which can be seen as deterministic quantities, are all unknown here. It is nevertheless reasonable to assume that they are sorted in increasing order (in the case of path discretization, this is equivalent to assuming that the components of the same vector are ordered chronologically, which is generally the case).

2.2 Two-step statistical inference

As explained in Introduction, a two-step procedure is proposed for the identification of the probability distribution of $\boldsymbol{X}$. First, we look for the best approximation $\widehat{\mathcal{M}} \in \mathcal{M}_1(\mathbb{R}^d)$ of $\mathcal{M}^\star$, in the sense that

$$\max_{\boldsymbol{x} \in \mathcal{M}^\star} \min_{\boldsymbol{y} \in \widehat{\mathcal{M}}} \|\boldsymbol{x} - \boldsymbol{y}\| \tag{2}$$

is minimum, where we denote by $\langle \cdot, \cdot \rangle$, $\|\cdot\|$, and $\cdot \times \cdot$ the Euclidean scalar product, the Euclidean norm, and the cross product in $\mathbb{R}^d$ respectively. The way in which this problem will be addressed will be the focus of Section 3. The very important contribution of this estimation is to allow the indexation of $\boldsymbol{X}$ by a new curvilinear abscissa that we can now manipulate. Indeed, once $\widehat{\mathcal{M}}$ is defined, we can orient it and associate with it a normalized curvilinear abscissa noted $t \in [0,1]$, as well as a function $\widehat{\boldsymbol{m}}$ in $\mathcal{C}^2([0,1], \mathbb{R}^d)$ such that $\widehat{\boldsymbol{m}}(t)$ is the point of $\widehat{\mathcal{M}}$ that we find at abscissa t. Given these notations, for all $t \in [0,1]$, we can gather in the matrix $\boldsymbol{B}(t) := [\boldsymbol{\tau}(t)\ \boldsymbol{\nu}(t)]$ (respectively $\boldsymbol{B}(t) := [\boldsymbol{\tau}(t)\ \boldsymbol{\nu}(t)\ \boldsymbol{\kappa}(t)]$ when $d = 3$) a local basis for $\widehat{\mathcal{M}}$, where $\boldsymbol{\tau} := \frac{d\widehat{\boldsymbol{m}}}{dt} / \left\| \frac{d\widehat{\boldsymbol{m}}}{dt} \right\|$ is the unit tangent vector, $\boldsymbol{\nu} := \frac{d\boldsymbol{\tau}}{dt} / \left\| \frac{d\boldsymbol{\tau}}{dt} \right\|$ is the normal unit vector (and $\boldsymbol{\kappa} := \boldsymbol{\tau} \times \boldsymbol{\nu}$ is the binormal unit vector). See Figure 1 for a graphical illustration of these notations. From this local representation, we can define $\mathbb{S}(\widehat{\mathcal{M}})$ as the set of random fields $\widehat{\boldsymbol{X}}$ that can be written under the form

$$\widehat{\boldsymbol{X}}(t, \omega) := \widehat{\boldsymbol{m}}(t) + \boldsymbol{B}(t)\widehat{\boldsymbol{Y}}(t, \omega),\ \omega \in \Omega, \tag{3}$$

where $\widehat{\boldsymbol{Y}}$ is a centered random field indexed by $t \in [0,1]$ with trajectories in $\mathcal{C}^2([0,1], \mathbb{R}^{d-1})$ almost-surely. In the same manner, each observation point $\boldsymbol{X}(s_m^{(n)}, \omega_n)$ of Eq. (1) can be decomposed as the sum of its projection on $\widehat{\mathcal{M}}$, noted $\widehat{\boldsymbol{m}}(t_m^{(n)})$, and a remaining term $\boldsymbol{X}_m^{\perp}(\omega_n) := \boldsymbol{X}(s_m^{(n)}, \omega_n) - \widehat{\boldsymbol{m}}(t_m^{(n)})$, where:

$$\widehat{\boldsymbol{m}}(t_m^{(n)}) := \underset{\boldsymbol{x} \in \widehat{\mathcal{M}}}{\arg\min} \left\| \boldsymbol{x} - \boldsymbol{X}(s_m^{(n)}, \omega_n) \right\|, \tag{4}$$

$$t_m^{(n)} := \int_{\widehat{\boldsymbol{m}}(0)}^{\widehat{\boldsymbol{m}}(t_m^{(n)})} dt \Bigg/ \int_{\widehat{\boldsymbol{m}}(0)}^{\widehat{\boldsymbol{m}}(1)} dt. \tag{5}$$

By construction, for each $1 \leq n \leq N$ and $1 \leq m \leq M_n$ so that $t_m^{(n)} \notin \{0,1\}$, $\boldsymbol{X}_m^{\perp}(\omega_n)$ is in the orthogonal space at $\widehat{\boldsymbol{m}}(t_m^{(n)})$, which means that it can be written $\boldsymbol{B}(t_m^{(n)})\boldsymbol{Y}(t_m^{(n)}, \omega_n)$, with

$$\boldsymbol{Y}(t_m^{(n)}, \omega_n) := \boldsymbol{B}(t_m^{(n)})^T \boldsymbol{X}_m^{\perp}(\omega_n). \tag{6}$$

Finally, the objective of the second step of reconstruction of the probability distribution of $\boldsymbol{X}$ is to search in $\mathbb{S}(\widehat{\mathcal{M}})$ for the random field that is the closest to $\boldsymbol{X}$ in distribution, using the information gathered in the observations coefficients $\{\boldsymbol{Y}(t_m^{(n)}, \omega_n)\}_{1 \leq m \leq M_n, 1 \leq n \leq N}$. This will be the subject of Section 4.

3 Manifold learning

The objective of this section is to propose a method to solve, using limited and disperse data, the optimization problem defined by Eq. (2). As it stands, since $\mathcal{M}^\star$ is unknown, the problem we are trying to solve is ill-defined. To rely only on the available points, rather than directly minimizing the maximum

distance between $\mathcal{M}^\star$ and a candidate curve, we propose to construct a kind of spatial average of the observation points, which are listed in Eq. (1). To this end, let us gather (without considering the statistical dependencies between points) all these points in $\mathcal{X}^L := \{\boldsymbol{x}_\ell,\ 1 \leq \ell \leq L\}$, with $L := \sum_{n=1}^{N} M_n$, and:

$$\boldsymbol{x}_1 := \boldsymbol{X}(s_1^{(1)}, \omega_1), \ldots, \boldsymbol{x}_L := \boldsymbol{X}(s_{M_N}^{(N)}, \omega_N). \tag{7}$$

The square distance between each $\mathcal{M} \in \mathcal{M}_1(\mathbb{R}^d)$ and $\mathcal{X}^L$ can then be defined as:

$$d(\mathcal{M}, \mathcal{X}^L)^2 := \sum_{\boldsymbol{x}_\ell \in \mathcal{X}^L} \left\| \boldsymbol{x}_\ell - \arg\min_{\boldsymbol{y} \in \mathcal{M}} \|\boldsymbol{x}_\ell - \boldsymbol{y}\| \right\|^2. \tag{8}$$

However, seeking to minimize this distance is not sufficient to approach $\mathcal{M}^\star$. Indeed, any curve passing through each points of $\mathcal{X}^L$ exactly makes this pseudo-distance be zero. In order to avoid such overlearning, it is necessary to penalize the complexity of the search set of candidate functions. In this prospect, a progressive approach is proposed in the following, based on four main steps (see Figure 2 for a graphical illustration of these steps).

Step 1 : Quantization

The first step is to define a small number of $K > 1$ representative points in the dataset, noted $\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_K$. The identification of these points is carried out by clustering, and more precisely by K-means clustering [15], due to the fact that we are interested in an average reconstruction with respect to the Euclidean distance. For the three first steps, K is assumed to be predetermined. For choosing these K representative points, we therefore look for the optimal partition $\widehat{S}_1, \ldots, \widehat{S}_K$ of $\mathcal{X}^L$, which minimizes the Within-Cluster Sum of Squares (WCSS), and we define:

$$\widehat{\boldsymbol{x}}_k := \frac{1}{|\widehat{S}_k|} \sum_{i \in \widehat{S}_k} \boldsymbol{x}_i, \quad 1 \leq k \leq K,$$

where $|\widehat{S}_k|$ is the number of elements of $\widehat{S}_k$. Alternative techniques could be used to choose these representative points, such as the ones presented in [16] or [17], without much impact on the following.

Step 2 : Piecewise segment approximation

Once the representative points have been identified, it is possible to construct a first approximation of $\mathcal{M}^\star$ by connecting the points $\widehat{\boldsymbol{x}}_k$ by segments. Since we are interested in curves, we expect each point to be connected to at most two other points. However, if one imagines that the curve to be identified loops one or more times (that is the case for the proposed application), it is possible that some points serve as crossing points, and must be connected to more than two points. In order to identify a single 1D path connecting all representative points, it is then proposed to proceed as follows.

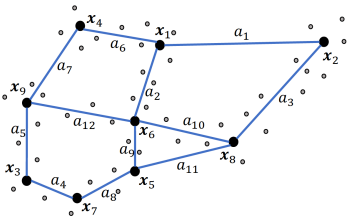
First, each point $\widehat{\boldsymbol{x}}_k$ is connected to its $N_{\text{nn}} = 2$ nearest neighbors (in the sense of the Euclidean distance). We note $a_1, \ldots, a_{N_s}$ the segments obtained, with N_s the total number of segments. If the union of the obtained segments does not form a connected graph, we increase the value of N_{nn} (see Figure 2-a). The suitability of the curve made of these N_s segments, which is denoted by $\mathcal{A}$, to represent the available data, can then be measured by $d(\mathcal{A}, \mathcal{X}^L)^2$.
Curve $\mathcal{A}$ is nevertheless likely to gather too many segments. In order to propose a sufficiently regular curve, we then make the assumption that the length of a segment is small compared to the inverse of the curvature of $\mathcal{M}^\star$ in the neighborhood of this segment. In principle, this prohibits the presence of completely connected groups of more than two points (all points in this group are connected by segments to all other points). To remove this excess of segments, we propose to sequentially remove the segment $a_{j^\star}$ such that:

$$j^\star \in \underset{1 \leq j \leq q}{\arg\min} \frac{d(\mathcal{A} \backslash \{a_j\}, \mathcal{X}^L)^2}{\text{Len}(a_j)^2}, \tag{9}$$

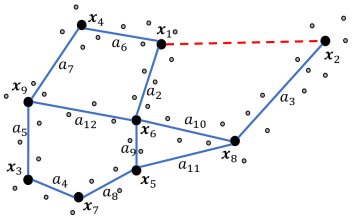
where $a_1, \ldots, a_q$ correspond to q segments of a completely connected group, and $\text{Len}(a_j)$ is the length of segment a_j. Thus, the idea is to remove the segment with the smallest length-weighted effect on the projection error. At the end of this sequential approach, it is then assumed that any point $\widehat{\boldsymbol{x}}_k$ is connected to only two neighboring points, except in looping cases where the crossing points will have more connected neighbors (see Figures 2-a,b).
The last phase of this step consists of an orientation of the resulting graph. To do this, we start from a simply connected point (if there is one, otherwise, we pick at random a point in the graph), then we go up the graph by following the connectivities. In the case of a point connected to more than two points, we propose to choose for the next point the neighbor whose segment will be the most aligned with the segment followed to reach this point (we thus favor the most regular trajectories). We also propose to duplicate the point that turns out to be connected to more than two neighbors, so that in the final connection graph, any point is only connected to one or two points (see Figure 2-b). We finally note $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{\widetilde{K}}$ (with $\widetilde{K}$ potentially larger than K in the case of duplication) the oriented sequence of elements of $\{\widehat{\boldsymbol{x}}_1, ..., \widehat{\boldsymbol{x}}_K\}$ such that $\widetilde{\boldsymbol{x}}_k$ and $\widetilde{\boldsymbol{x}}_{k+1}$ are connected by a segment for each $1 \leq k \leq \widetilde{K} - 1$.

Step 3 : Points projection and spline approximation

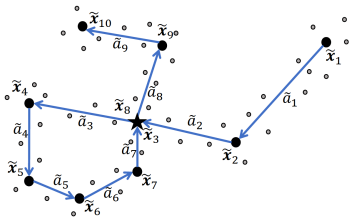
In order to reconstruct a sufficiently regular curve from the points $\widetilde{\boldsymbol{x}}_1$ to $\widetilde{\boldsymbol{x}}_{\widetilde{K}}$, a piecewise polynomial reconstruction is now proposed.
To begin, for each $1 \leq k \leq \widetilde{K} - 1$, we denote by $\widetilde{a}_k$ the segment connecting $\widetilde{\boldsymbol{x}}_k$ and $\widetilde{\boldsymbol{x}}_{k+1}$ (these segments are the same than the ones identified in Step 2 up to a modification of the numbering). Curve $\widetilde{M}^{\widetilde{K}} := \bigcup_{k=1}^{\widetilde{K}-1} \widetilde{a}_k$ can therefore be seen as a piecewise linear approximation of the searched curve $\mathcal{M}^\star$. We can then project each point $\boldsymbol{x}_\ell$ of $\mathcal{X}^L$ on $\widetilde{M}^{\widetilde{K}}$, and associate with each of them a segment index i_ℓ and a curvilinear abscissa t_ℓ, such that for all $1 \leq \ell \leq L$,



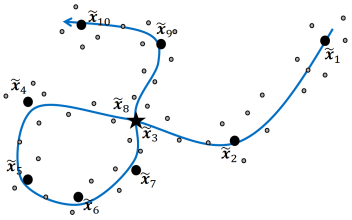
(a) Quantization and segment approximation



(b) Graph reduction



(c) Point duplication and graph orientation



(d) Polynomial approximation

Figure 2 Graphical representation of steps 1, 2 and 3 for curve learning. The grey points correspond to the observation points, the big black points are the representative points, the black star indicates a point that is duplicated, the segments connecting the representative points are in blue, and the the blue curve is the final curve.

$$t_\ell := \left\| \widetilde{\boldsymbol{x}}_{i_\ell} - \Pi(\boldsymbol{x}_\ell;\ \widetilde{M}^{\widetilde{K}}) \right\| + \sum_{k=1}^{i_\ell - 1} \mathrm{Len}(\widetilde{a}_k), \tag{10}$$

where $\Pi(\boldsymbol{x}_\ell;\ \widetilde{M}^{\widetilde{K}})$ is the projection of $\boldsymbol{x}_\ell$ on $\widetilde{a}_{i_\ell}$.
To recover the expected regularity (we restrict ourselves to twice-differentiable functions), we finally replace each segment $\widetilde{a}_k$ by a parabola, $\mathrm{pa}(\boldsymbol{b}^{(k)})$, parameterized by a $3d$-dimensional vector $\boldsymbol{b}^{(k)}$ (the parabola being characterized in each dimension by a second-order polynomial based on 3 independent constants) so that $\mathcal{M}^\star$ is now searched as the concatenation of $\widetilde{K} - 1$ pieces of parabolas (see Figure 2-c):

$$\widetilde{M}_{\boldsymbol{B}}^{(\widetilde{K})} := \bigcup_{k=1}^{\widetilde{K}-1} \mathrm{pa}(\boldsymbol{b}^{(k)}) = \{\boldsymbol{x}(t;\ \boldsymbol{B}),\ 0 \le t \le 1\}, \tag{11}$$

where the value of $\boldsymbol{B} = \left[\boldsymbol{b}^{(1)}\ \cdots\ \boldsymbol{b}^{(\widetilde{K}-1)}\right]$ is *a priori* unknown, but needs to be estimated from data. In that prospect, we propose to choose $\boldsymbol{B}$ that minimizes

$$e^2(\boldsymbol{B}) := \sum_{\ell=1}^{L} \|\boldsymbol{x}_\ell - \boldsymbol{x}(t_\ell;\ \boldsymbol{B})\|^2 \tag{12}$$

under the constraint that the reconstructed curve is twice continuously differentiable. By construction, this only requires that $\boldsymbol{B}$ guarantees the continuity of the curve at the junction indices, as well as the continuity of its derivatives with respect to t. Without going into too much details, we underline that error $e^2(\boldsymbol{B})$ can be written as the sum of a constant independent of $\boldsymbol{B}$ and $\widetilde{K}-1$ terms $e_k^2(\boldsymbol{b}^{(k)})$ depending quadratically in $\boldsymbol{b}^{(k)}$. As a consequence, if we choose to impose the continuity constraints using standard ℓ^2 penalty techniques [18], we obtain a constrained least squares problem that admits an explicit solution (to the value of the penalty constant λ). Denoting by $\widetilde{\boldsymbol{B}}(\lambda)$ this solution, the curve $\widetilde{M}^{(\widetilde{K})}_{\widetilde{\boldsymbol{B}}(\lambda)}$ can finally be chosen for the approximation of $\mathcal{M}^\star$ in $\mathcal{M}_1(\mathbb{R}^d)$.

Step 4 : Choice of K and λ

The former construction depends on two constants: the penalty constant λ, and the number K (or $\widetilde{K}$ if we consider the duplicated points) of representative points. Whereas λ is chosen as large as possible, the value of K results from a convergence analysis. For this purpose, we plot the mean-square distance, $d(\widetilde{M}^{(\widetilde{K})}_{\widetilde{\boldsymbol{B}}(\lambda)}, \mathcal{X}^L)$, with respect to K, and we consider the "elbow" method (see [19] fore more details about this selection criterion) to determine the optimal value of K. The intuition behind this heuristic is that by increasing the number of representative points, we naturally reduce the projection error, since we are projecting the points on a more complex (and *a priori* longer) curve, but at some point, there is over-fitting: we focus on data rather than on the curve we are interested in.

Remarks

- At step 3, one may or may not want to make the estimated curve pass through the representative points. From the few numerical examples we have dealt with, we have nevertheless observed a better reconstruction when this passage through the representative points is not imposed. We will therefore not consider this constraint in the numerical examples that will be presented in the following.
- On its own, this curve identification approach can be seen as a particular nonlinear dimension reduction technique (going from a dimension d to a dimension one), or seen as a denoising method under the assumption that the fluctuations around $\mathcal{M}^\star$ correspond to noise.

4 Statistical inference

4.1 Approximation class and learning set

As explained in Section 2, the estimation of the probability distribution of $\boldsymbol{X}$ relies on the estimation of two quantities: the approximation $\widehat{\mathcal{M}}$ of $\mathcal{M}^\star$, on which the probability distribution of $\boldsymbol{X}$ is assumed to be concentrated, and the probability distribution of the projection coefficients of $\boldsymbol{X}$ on the

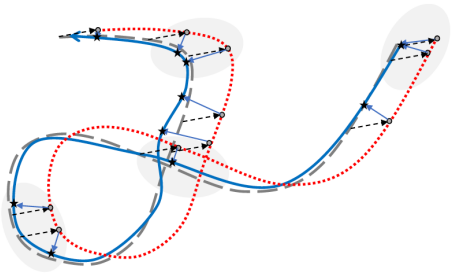


Figure 3 Graphical illustration of several difficulties related to the reconstruction of the probability distribution of $\widehat{\boldsymbol{Y}}$: potential presence of local oscillations, difficulties at the crossings and extremities, distortions related to the projection operation. For this example, the true curve $\mathcal{M}^\star$ is shown in grey dashed line and its oriented approximation $\widehat{\mathcal{M}}$ in blue solid line. A potential realization of $\boldsymbol{X}$ associated with a translation of $\mathcal{M}^\star$ is shown in red dotted line, and the black stars correspond to the projections of the grey points on $\widehat{\mathcal{M}}$.

orthogonal space of $\widehat{\mathcal{M}}$. Indeed, once $\widehat{\mathcal{M}}$ has been chosen (unless otherwise stated, $\widehat{\mathcal{M}} = \widetilde{M}^{(\tilde{K})}_{\widetilde{\boldsymbol{B}}(\lambda)}$ in the following), we can orient it, associate with it a function $\widehat{\boldsymbol{m}}$ in $\mathcal{C}^2([0,1],\mathbb{R}^d)$, and gather in $\boldsymbol{B}(t)$ a local basis of it. In line with Eq. (3), this allows us to search $\boldsymbol{X}$ under the form

$$\widehat{\boldsymbol{m}}(t) + \boldsymbol{B}(t)\widehat{\boldsymbol{Y}}(t,\omega),\ 0 \leq t \leq 1,\ \omega \in \Omega,$$

where we recall that the knowledge of $\widehat{\mathcal{M}}$ implies the knowledge of $\widehat{\boldsymbol{m}}$ and $\boldsymbol{B}$. Characterizing the probability distribution of $\boldsymbol{X}$ amounts therefore at estimating the probability distribution of $\widehat{\boldsymbol{Y}}$.

This estimation remains particularly difficult for several reasons (see Figure 3 for a graphical illustration of some of these difficulties). First, the identified curve is likely to introduce parasitic oscillations with respect to the true (but unknown) curve $\mathcal{M}^\star$. In a second step, the identification of the probability distribution of $\widehat{\boldsymbol{Y}}$ relies on the post-processing of the projections of the observation points. However, such a projection is itself a source of complications. In particular, there is the problem of the projections at the extremities of $\widehat{\mathcal{M}}$, which are likely to concentrate entire sections of realizations of $\boldsymbol{X}$ at a single point. The presence of crossings in the curve to be identified poses another problem, in the sense that a direct projection based on the search of the nearest point of $\widehat{\mathcal{M}}$ may be flawed. Indeed, at the crossings, two points that are close from a Euclidean norm point of view can be associated with very different curvilinear abscissas, and thus be very far from each other when travelling along the curve. Thirdly, the probability distribution of $\widehat{\boldsymbol{Y}}$ may well be much more complicated than that of $\boldsymbol{X}$. For example, one can imagine that the realizations of $\boldsymbol{X}$ are particular translations of $\mathcal{M}^\star$. In this case, the probability distribution of $\boldsymbol{X}$ is then of very low statistical dimension, when the one of

$\widehat{\boldsymbol{Y}}$, by the game of projections on the approximate curve, is likely to be non-stationary (two points projected on the curve can in particular move closer or further away depending on the local curvature).
Keeping these difficulties in mind, we first assume that $\widehat{\boldsymbol{Y}}$ is centered, and we choose to neglect the problems at the extremities of the curve, in the sense that all the (*a priori* few) observation points whose projection will be at one of the extremities of the estimated curve will be discarded (the integration of these points is thus left as a working perspective). Hence, the learning set for the estimation of the probability distribution of $\widehat{\boldsymbol{Y}}$ gathers the elements of the set $\{\boldsymbol{Y}(t_m^{(n)}, \omega_n),\ 1 \leq m \leq M_n\}_{n=1}^N$ (defined by Eq. (6)) that verify $0 < t_m^{(n)} < 1$. These points are then considered as (potentially statistically dependent) realizations of $\widehat{\boldsymbol{Y}}$.

4.2 Spectral decomposition

In this part, we try to estimate the probability distribution of the random field $\widehat{\boldsymbol{Y}}$ with, as maximum knowledge, the observation points listed in previous section. To this end, we propose to consider a spectral approach, which consists, first, in projecting $\widehat{\boldsymbol{Y}}$ on a finite set of deterministic functions $\boldsymbol{f}^{(1)}, \ldots, \boldsymbol{f}^{(Q)}$, and secondly, in inferring the joint probability distribution of the projection coefficients.
Assuming that $\widehat{\boldsymbol{Y}}$ is a $\mathbb{R}^{d-1}$-valued second-order random field, and that its covariance function, which is noted $\boldsymbol{C}_Y$, is continuous on $[0,1] \times [0,1]$, it is well known (projection theorem in Hilbert space) that the best choice for the projection functions is the set of eigenfunctions of $\boldsymbol{C}_Y$ associated with the highest eigenvalues, in the sense that is allows minimizing the signal energy of the difference between $\widehat{\boldsymbol{Y}}$ and its projection at any Q.
To estimate these projection functions, we therefore need to estimate $\boldsymbol{C}_Y$. Such an approximation using projections of $\widehat{\boldsymbol{Y}}$ at distinct values of t is often based on the assumption that $\boldsymbol{C}_Y$ belongs to a chosen parametric class. Unfortunately, there is no reason why $\widehat{\boldsymbol{Y}}$ should be parameterized by a reduced number of parameters. In particular, the dependence structure between the components of $\widehat{\boldsymbol{Y}}$ is non-trivial due to the projection. For all these reasons, we focus instead on the following empirical approximation $\widehat{\boldsymbol{C}}_Y(t,t')$ of $\boldsymbol{C}_Y(t,t')$:

$$\widehat{\boldsymbol{C}}_Y(t,t') := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{Y}_{\mathrm{int}}(t,\omega_n)\boldsymbol{Y}_{\mathrm{int}}(t',\omega_n)^T, \tag{13}$$

where for each $1 \leq n \leq N$, $t \mapsto \boldsymbol{Y}_{\mathrm{int}}(t,\omega_n)$ is an interpolation of $t \mapsto \widehat{\boldsymbol{Y}}(t,\omega_n)$ over $[0,1]$ such that $\boldsymbol{Y}_{\mathrm{int}}(t_m^{(n)},\omega_n) = \boldsymbol{Y}(t_m^{(n)},\omega_n)$ for each $1 \leq m \leq M_n$ as long as $t_m^{(n)} \notin \{0,1\}$. For this work, we limit ourselves to interpolations based on a Gaussian process regression, because of their flexibility and their very good properties for the approximation of functions defined on low-dimensional compacts [20]. But this interpolation phase can be carried out in other ways [21, 22]. Let $\{\widehat{\boldsymbol{f}}^{(q)},\ 1 \leq q \leq Q\}$ be the set gathering the Q eigenfunctions

associated with the Q largest eigenvalues of $\widehat{\boldsymbol{C}}_Y$, and $\widehat{\boldsymbol{Y}}^{(Q)}$ be the projection of $\boldsymbol{Y}_{\text{int}}$ on it. As this family is orthonormal, we can write:

$$\boldsymbol{Y}_{\text{int}} \approx \widehat{\boldsymbol{Y}}^{(Q)} := \sum_{q=1}^{Q} \xi_q \widehat{\boldsymbol{f}}^{(q)}, \tag{14}$$

where $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_Q)$ is a centered random vector (remind that $\widehat{\boldsymbol{Y}}$ is assumed centered), whose components are uncorrelated. In addition, using this formalism, N independent realizations of $\boldsymbol{\xi}$ can be computed by projecting each function $\boldsymbol{Y}_{\text{int}}(\cdot, \omega_n)$ on $\{\widehat{\boldsymbol{f}}^{(q)},\ 1 \leq q \leq Q\}$.
Note that if $\boldsymbol{Y}_{\text{int}}$ was Gaussian, the components of $\boldsymbol{\xi}$ would be Gaussian and therefore would be statistically independent. But in the general case, although uncorrelated, the components of $\boldsymbol{\xi}$ are likely to be statistically dependent. Once again, this dependence structure may be complex, which leads us to focus on nonparametric approaches for the estimation of the probability distribution of $\boldsymbol{\xi}$. Among these methods, the multidimensional Gaussian kernel-density estimation (G-KDE) method proposes to write the probability density function (PDF) of $\boldsymbol{\xi}$, if it exists, as a sum of N multidimensional Gaussian PDFs with the same covariance matrix $\boldsymbol{H}$ (generally called "the bandwidth matrix"), which are centered at the available realizations of $\boldsymbol{\xi}$. Correctly adapting the value of $\boldsymbol{H}$ is particularly important for this approximation, as this matrix controls the influence of each realization of $\boldsymbol{\xi}$ on the final PDF approximation. Many contributions can be found in the literature on this subject (see for instance [6, 23, 24]).
In summary, once projection family $\{\widehat{\boldsymbol{f}}^{(1)}, \ldots, \widehat{\boldsymbol{f}}^{(Q)}\}$ and matrix $\boldsymbol{H}$ have been estimated, we have a statistical model for $\boldsymbol{\xi}$ (a very large number of independent realizations of $\boldsymbol{\xi}$ can be quickly generated), and therefore one for $\widehat{\boldsymbol{Y}}^{(Q)}$ thanks to Eq. (14), and therefore one for the approximation $\widehat{\boldsymbol{X}} := \widehat{\boldsymbol{m}} + \boldsymbol{B}\widehat{\boldsymbol{Y}}^{(Q)}$ of $\boldsymbol{X}$, which was the initial goal of this paper.

Relying on eigenvalue decay and restored variance is a classic choice for selecting the number of eigenvectors. Other considerations could also be taken into account for this choice. Indeed, the larger Q is, the closer the covariance matrix of the reconstructed vector $\widehat{\boldsymbol{Y}}^{(Q)}$ approaches $\widehat{\boldsymbol{C}}_Y$. But since $\widehat{\boldsymbol{C}}_Y$ is itself an approximation of $\boldsymbol{C}_Y$, we could also choose Q so that the loss of information due to truncation is of the same order of magnitude as the loss of information due to the approximation of $\boldsymbol{C}_Y$ by $\widehat{\boldsymbol{C}}_Y$ (assuming that we are able to estimate this difference). Furthermore, it's important to note that the larger Q is, the larger the projection vector will be, and therefore the more difficult the PDF estimation phase will be, which could again be incorporated into this choice of Q. These considerations will be taken into account for the examples discussed in Section 5: we will be looking for a value for Q that is large enough to limit the loss of information on the covariance of the projection vector, without being too large to allow a satisfactory estimation of the joint distribution, in the

sense that further increasing Q would have only a very small impact on the statistical properties of the reconstructed random field.

4.3 Validation attempt

Validating the method of reconstructing the probability distribution of $\boldsymbol{X}$ is a delicate question. The random field $\boldsymbol{X}$ being *a priori* non-Gaussian, its probability distribution is indeed a complex mathematical object. In contrast, the maximal information on $\boldsymbol{X}$ is very partial, consisting of a supposed reduced number N of realizations (N being nevertheless *a priori* much larger than Q for identifiability reasons), which are themselves discretized into an also reduced number of curvilinear abscissae, the values of these curvilinear abscissae being unknown.

What seems more reasonable to do is to find out if the distance between two sets of trajectories of the estimated process is of the same order of magnitude as the distance between the training base and one of these sets of generated trajectories. In this prospect, let $\Omega_1, \ldots, \Omega_P$ be P independent sets of N independent realizations of $\widehat{\boldsymbol{X}}$, which was defined at the end of Section 4.2, such that for all $1 \leq p \leq P$,

$$\Omega_p := \{\widehat{\boldsymbol{X}}(\cdot, \omega_{n,p}),\ 1 \leq n \leq N,\ \omega_{n,p} \in \Omega\}, \tag{15}$$

and Ω^{ref} be the set gathering the continuous reconstructions of the N realizations of $\boldsymbol{X}$, which are noted $\boldsymbol{X}_{\mathrm{rec}}^{(n)}$ and which are constructed from the interpolated process $\boldsymbol{Y}_{\mathrm{int}}$ defined in Section 4.2:

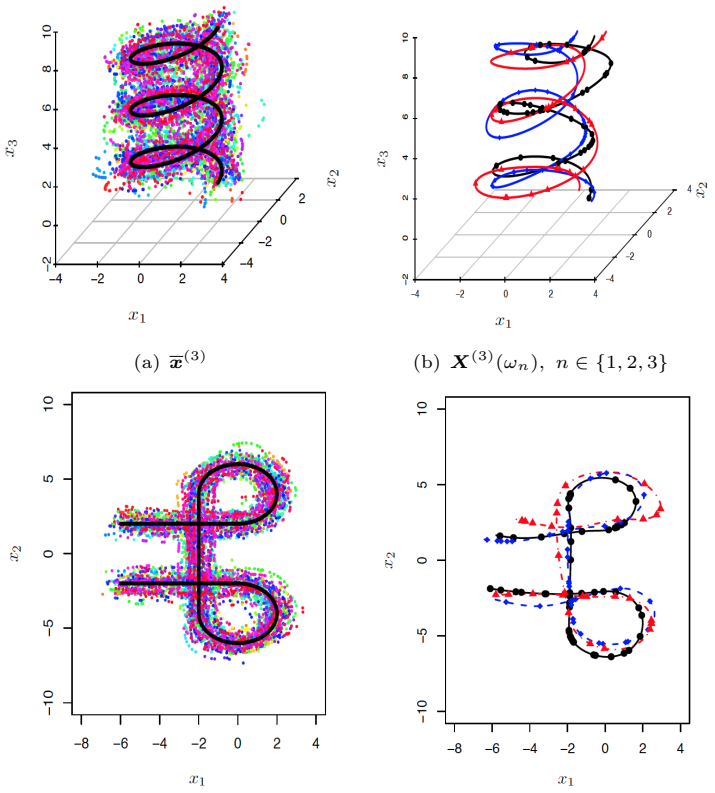
$$\boldsymbol{X}_{\mathrm{rec}}^{(n)}(t) = \widehat{\boldsymbol{m}}(t) + \boldsymbol{B}(t)\boldsymbol{Y}_{\mathrm{int}}(t, \omega_n),\ 0 \leq t \leq 1. \tag{16}$$

Under these notations, Ω^{ref} should be seen as the reference set, and Ω_p as a candidate set, whose statistical content is hopefully close to that of Ω^{ref}. Two types of criteria are considered to compare the statistical content of these two sets. First, we will compare the mean and variance functions as a function of t. In a second step, somewhat like the non-asymptotic Kolmogorov test of law adequacy [25], we propose to consider the following statistic for comparing two different sets $\Omega_1 := \{\boldsymbol{Z}_n^{(1)},\ 1 \leq n \leq N\}$ and $\Omega_2 := \{\boldsymbol{Z}_n^{(2)},\ 1 \leq n \leq N\}$:

$$\zeta_N(\Omega_1, \Omega_2) := \max_{\boldsymbol{h} \in \mathcal{B}(1), z \in \mathbb{R}} \left| F^1(\boldsymbol{h}, z) - F^2(\boldsymbol{h}, z) \right|, \tag{17}$$

$$F^i(\boldsymbol{h}, z) := \frac{1}{N} \sum_{n=1}^{N} 1_{(\boldsymbol{h}, \boldsymbol{Z}_n^{(i)}) \leq z},\ 1 \leq i \leq 2, \tag{18}$$

where $(\cdot, \cdot)$ is the scalar product in $L^2([0,1], \mathbb{R}^d)$, and $\mathcal{B}(1)$ defines the set of realizations of the restriction to $[0,1]$ of the Gaussian white noise indexed by $\mathbb{R}$ with values in $\mathbb{R}^d$, whose components are independent, and whose integral of the square of its values over $[0,1]$ is equal to 1. To evaluate the relevance of the proposed approach to construct a good approximation of the probability distribution of $\boldsymbol{X}$, we can therefore put the values of $\left(\zeta_N(\Omega^{\mathrm{ref}}, \Omega_p)\right)_{1 \leq p \leq P}$ in perspective with those of $(\zeta_N(\Omega_p, \Omega_{p'}))_{1 \leq p \neq p' \leq P}$.



(a) $\overline{\boldsymbol{x}}^{(3)}$ (b) $\boldsymbol{X}^{(3)}(\omega_n),\ n \in \{1,2,3\}$

(c) $\overline{\boldsymbol{x}}^{(2)}$ (d) $\boldsymbol{X}^{(2)}(\omega_n),\ n \in \{1,2,3\}$

Figure 4 Graphical representation of the test cases studied. Figures (a) and (c) show the statistical dispersion of all the available points (two points of different trajectories are represented with two different colors) around their mean functions $\overline{\boldsymbol{x}}^{(2)}$ and $\overline{\boldsymbol{x}}^{(3)}$ in black solid line. Figures (b) and (d) represent the graphs of three independent realizations of $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$, as well as their discretizations (black points, red triangles and blue squares) to be used for the statistical inference.

5 Application

5.1 Presentation of the test cases

In this section, two numerical applications are introduced to show to what extent the proposed method proves to be efficient. As indicated in the introduction, these examples, although analytical, are intended to be at the same time challenging (presence of loops, significant dispersion of the trajectories around their mean, etc.) and representative of real situations in dimensions 2 and 3, where one would seek to quantify the dispersion of trajectories of physical systems designed to connect two points according to an *a priori* route.

In the first case, we place ourselves in dimension $d = 2$, and we are interested in the reconstruction of the probability distribution of $\boldsymbol{X}^{(2)}$, while in the second

case, we place ourselves in dimension $d = 3$, and we are interested in the probability distribution of $\boldsymbol{X}^{(3)}$. For $d = 2, 3$, $\boldsymbol{X}^{(d)}$ is decomposed as:

$$\boldsymbol{X}^{(d)} = \overline{\boldsymbol{x}}^{(d)} + \boldsymbol{Z}^{(d)},$$

where $\overline{\boldsymbol{x}}^{(d)}$ is a deterministic function in $\mathcal{C}^2([0,1], \mathbb{R}^d)$, and $\boldsymbol{Z}^{(d)}$ is the restriction to $[0,1]$ of a centered and stationary Gaussian process indexed by $s \in \mathbb{R}$ with values in $\mathbb{R}^d$. For $d \in \{2,3\}$, the graphs of $\overline{\boldsymbol{x}}^{(d)}$ are shown in Figure 4, and the covariance functions of $\boldsymbol{Z}^{(d)}$, noted $\boldsymbol{C}^{(d)}$, are given, for all s, s' in $[0,1]$ and all $1 \leq i, j \leq d$, by:

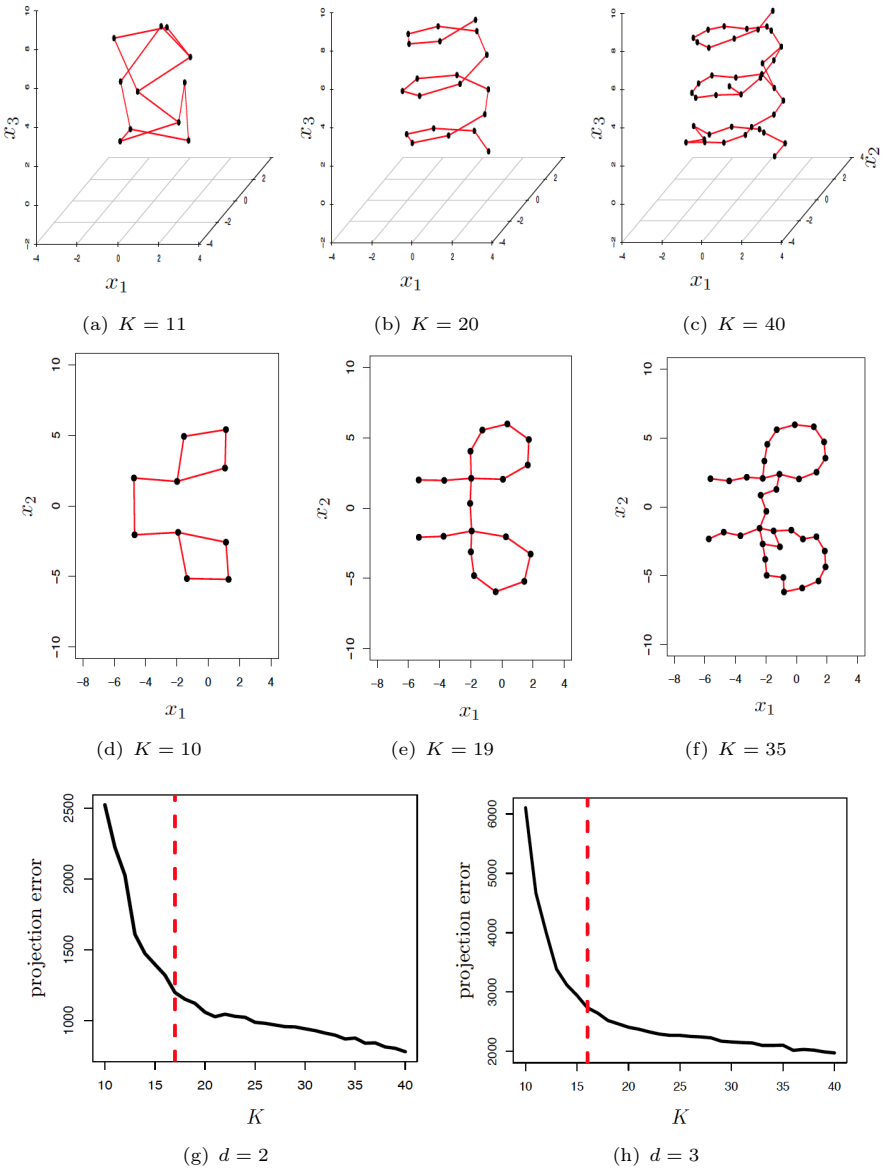
$$(\boldsymbol{C}^{(d)}(s,s'))_{ij} = \delta_{ij}\sigma_i^2 \exp\{-(s-s')^2/\ell_i^2\}, \tag{19}$$

with $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.2, 0.1, 0.03)$ and $(\ell_1, \ell_2, \ell_3) = (0.2, 0.1, 0.15)$. In each case, we randomly generate $N = 100$ independent trajectories of $\boldsymbol{X}^{(d)}$, which we discretize into M_n values of s randomly and uniformly chosen between 0 and 1. The values of M_n are also randomly and uniformly chosen between 50 and 100, which leads, for the two cases considered, to a total number of available points equal to $L = 7737$ for $d = 2$, and to $L = 7529$ for $d = 3$.

5.2 Estimation of the mean function

The first step to reconstruct the probability distribution of $\boldsymbol{X}^{(d)}$ is to construct an approximation of its mean function. This approximation is based on the identification of K representative points, which we want to be sufficiently numerous to precisely describe the function to be approximated, but not too numerous to avoid overlearning. As explained in Section 3, we proceed sequentially, progressively increasing the value of K, until the decrease of the projection error, which was noted $d(\widetilde{M}^{(\widetilde{K})}_{\widetilde{\boldsymbol{B}}(\lambda)}, \mathcal{X}^L)$, slows down significantly ("elbow" method). For $d \in \{2,3\}$, the influence of the value of K on the piecewise linear approximation of $\overline{\boldsymbol{x}}^{(d)}$ and the evolution of the projection error are represented in Figure 5, and the comparison between the curve identified for $K = K^\star$ and the mean functions of $\boldsymbol{X}^{(d)}$ is shown in Figure 7. From these two sets of figures, we can visualize the two anticipated phases: a first phase ($K \leq K^\star$) where the increase of K results in a finer description of $\overline{\boldsymbol{x}}^{(d)}$, and a second phase ($K > K^\star$) where the increase of K adds irregularities that are not necessarily informative, and which can be associated with over-learning. Focusing on Figure 7, we note the very close proximity between the functions $\overline{\boldsymbol{x}}^{(d)}$ and the identified curves, which reflects the relevance of the proposed approach for these two test cases. Note that the approximated curves do not necessarily pass through the representative points, as indicated at the end of Section 3.

In order to quantify the numerical efficiency of the proposed method, Figure 6 indicates the time needed to estimate the mean curve for the two studied cases and for different values of L and K, which are respectively the total number

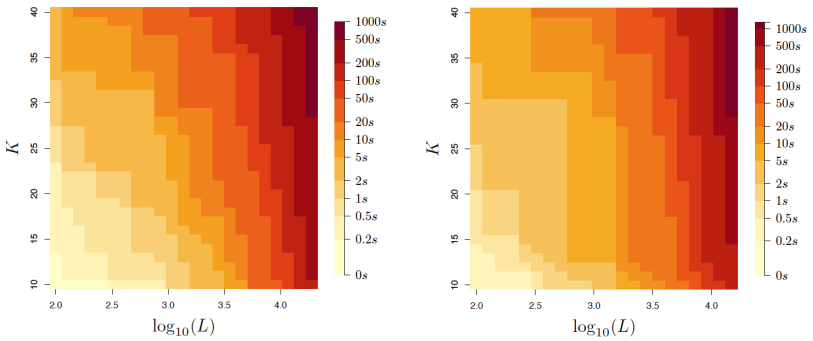


(a) $K = 11$ (b) $K = 20$ (c) $K = 40$

(d) $K = 10$ (e) $K = 19$ (f) $K = 35$

(g) $d = 2$ (h) $d = 3$

Figure 5 Illustration of the influence of the value of K on the piecewise linear approximation of the mean functions of $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$ (the representative points correspond to the big black points). Figures (g) and (h) represent the projection-error decrease with respect to K for the two considered test cases. The vertical line corresponds to $K = K^\star$, which is the value of K we recommend.



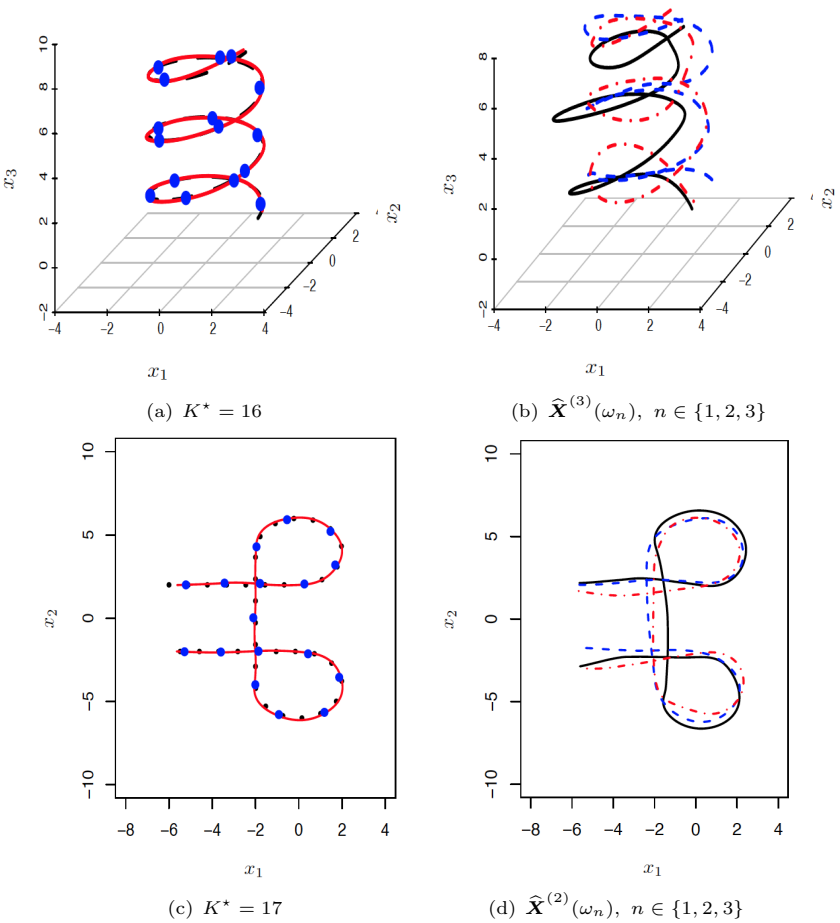
(a) $d = 2$ (b) $d = 3$

Figure 6 Evolution of the time needed (in seconds s) to estimate the mean curve with respect to the total number of available points L and the number of representative points K.

of available points and the number of considered representative points. For illustration purposes, these values can be chosen higher or smaller than the ones chosen for the two cases studied in the rest of Section 5. These results were obtained on a standard laptop (2.7 GHz Intel Core i5 with 8MB memory). Unsurprisingly, the larger L and K are, the longer it takes to estimate this mean curve. These estimation times are moreover relatively independent of d, and quite reasonable, in spite of the polynomial evolution in L associated with the local spline approximation.

5.3 Statistical inference

Once one has identified a satisfactory approximation of the mean of $\boldsymbol{X}^{(d)}$, its available realizations can be projected onto it, and one can focus on approximating the probability distribution of the statistical fluctuations of $\boldsymbol{X}^{(d)}$ around it. This identification is done in three steps. First, we empirically estimate the covariance function of the part of $\boldsymbol{X}^{(d)}$ that is locally orthogonal to the identified curve using interpolated approximations of its trajectories. We then project $\boldsymbol{X}^{(d)}$ on the Q eigenfunctions associated with the highest eigenvalues of this approximated covariance function. For the test cases considered here, $Q = 22$ for $d = 2$, and $Q = 36$ for $d = 3$, which correspond to a restitution of 99.9% of the total sum of the eigenvalues. The probability distribution of the projection coefficients is then reconstructed by using the Gaussian-kernel density estimation, as detailed in Section 4.2. Once these steps are done, we have identified an approximation of $\boldsymbol{X}^{(d)}$, which is noted $\widehat{\boldsymbol{X}}^{(d)}$. As an illustration, Figures 7-b,d show three independent realizations of $\widehat{\boldsymbol{X}}^{(d)}$. A visual validation of the relevance of the inference step can thus be made by comparing these figures to Figures 4-b,d. To compare these results in a more quantitative way, we gather in Ω^{ref} the continuous reconstructions of the $N = 100$ realizations of $\boldsymbol{X}^{(d)}$, and we generate $P = 100$ sets $\Omega_1, \ldots, \Omega_P$ each containing $N = 100$



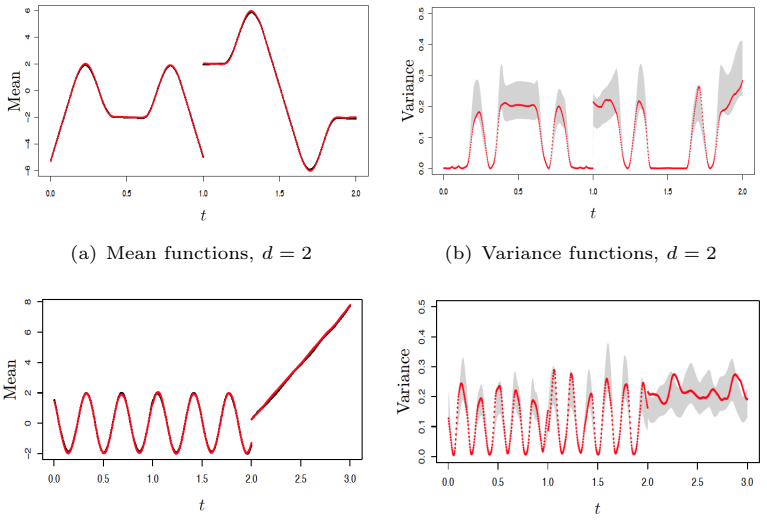
(a) $K^\star = 16$ (b) $\widehat{\boldsymbol{X}}^{(3)}(\omega_n),\ n \in \{1,2,3\}$

(c) $K^\star = 17$ (d) $\widehat{\boldsymbol{X}}^{(2)}(\omega_n),\ n \in \{1,2,3\}$

Figure 7 Figures (a) and (c) compare the graphs of $\overline{\boldsymbol{x}}^{(d)}$ (in black dotted lines) and $\widetilde{M}^{(K^\star)}_{\widehat{\boldsymbol{B}}(\lambda)}$ (in red solid lines), which is the curve approximation we propose for $K = K^\star$. On these two figures, the big blue points correspond to the positions of the $K^\star$ identified representative points. Figures (b) and (d) show three particular trajectories of the approximated field $\widehat{\boldsymbol{X}}^{(d)}$, for $d \in \{2,3\}$.

independent continuous realizations of $\widehat{\boldsymbol{X}}^{(d)}$. We refer to Figures 8 and 9 for a comparison of the statistical content of these sets. On the one hand, Figure 8 shows a very good accuracy between the component by component mean and variance functions, whether they are calculated empirically from Ω^{ref} or from the Ω_p.

On the other hand, Figure 9 compares the dispersion of the values of $\left(\zeta_N(\Omega^{\text{ref}}, \Omega_p)\right)_{1\le p\le P}$ with those of $\left(\zeta_N(\Omega_p, \Omega_{p'})\right)_{1\le p\ne p'\le P}$, where statistic ζ_N is defined by Eq. (17). This dispersion is represented in the form of a probability density function (PDF) estimated by a kernel method, and we recall that a large value for this statistic indicates a significant difference in the statistical



(a) Mean functions, $d = 2$ (b) Variance functions, $d = 2$

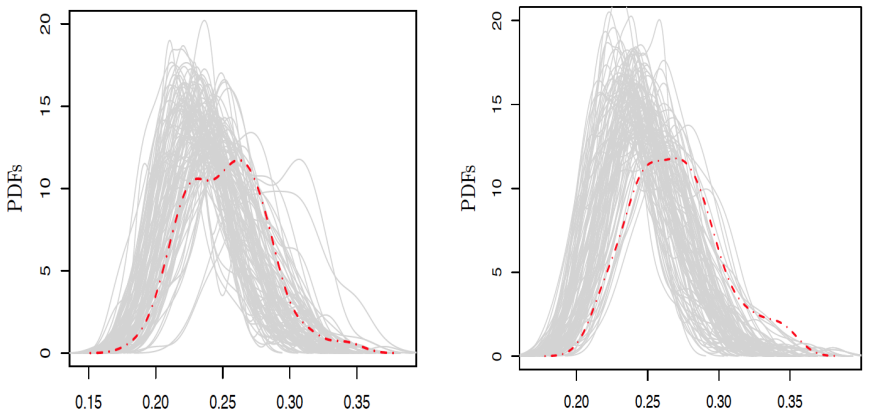
(c) Mean functions, $d = 3$ (d) Variance functions, $d = 3$

Figure 8 Comparison of the mean and variance functions computed from Ω^{ref} (in black dotted lines for the mean functions, and in shaded areas for the 95% confidence intervals for the variance functions) and from $\Omega_1, \ldots, \Omega_P$ (in red solid line). For the sake of brevity, the functions associated with each component of the random fields are juxtaposed with each other in the different figures.

content of the two sets being compared. The confrontation of reference and generated values is represented as a red solid line, when the confrontations of each generated set to the other ones are represented in light grey. The generation model being by construction imperfect, we observe without surprise for the cases $d = 2$ and $d = 3$ higher values of the statistics on average when the generated trajectories are compared to the reference trajectories. But this increase remains quite reasonable, which makes these results very encouraging, in the sense that out of the $P = 100$ generated sets, several sets lead to even higher values of the statistic on average.

6 Conclusions and prospects

Identifying reasonable approximations of the probability distribution (and thus the statistical dependencies) of random fields from their observations in a finite and often reduced number of points in space is a difficult challenge. In this work, we place ourselves in the particular case where this probability distribution is concentrated on an unknown curve. This is in particular the configuration in which we find ourselves when we try to characterize the set of probable trajectories that a certain system could follow to connect two points. The approach we propose for this identification decomposes the problem into two steps. First, we are interested in the estimation of the mean of the random field, which is likely to provide an interesting approximation of the curve on



(a) Case $d = 2$ (b) Case $d = 3$

Figure 9 Evaluation of the representativeness of the generated data with respect to the reference data. The red dashed line characterizes the dispersion of the statistic ζ_N when confronting the reference set and $P = 100$ generated sets of trajectories. The light grey lines are associated with the dispersion of the same statistic but when confronting one generated set to the $P - 1$ other sets of generated trajectories. The higher the statistic, the more statistical differences in the compared sets.

which the random field is concentrated. The developments presented for this first step, based on clustering, graph manipulations, and polynomial smoothing, are quite generalizable to dimension reduction or automatic denoising issues. In a second step, we build a model for the statistical fluctuations of the random process around its mean, using interpolation techniques, spectral methods, and kernel reconstruction. Two important assets of the proposed approach are its robustness and its reasonable numerical cost for the analysis of configurations close to the numerical test cases presented. Indeed, the different bricks of the proposed method are essentially based on matrix computation which is fast to execute.

The perspectives for this work are numerous, whether it is to tackle configurations that could exploit a much larger number of observations, using automatic domain decomposition for instance, or to be able to consider larger dimensions, both at the level of the space in which the points evolve and at the level of the dimension of the manifold on which the points are concentrated. And at the time of the Internet of Things, there is no doubt that the industrial applications related to these issues will multiply. From a more theoretical point of view, proving the consistency of the approach presented when the number of observations tends towards infinity would also be a very interesting continuation for this work, whether at the level of the reconstruction of the mean curve, than the estimation of the statistical properties of the process orthogonal to this mean curve.

Declarations

- Funding : No funds, grants, or other support was received.
- Conflict of interest/Competing interests : The authors declare they have no financial interests. They have no conflicts of interest to declare that are relevant to the content of this article.
- Ethics approval : The authors approve the Springer ethics policy.
- Consent to participate : Not applicable
- Consent for publication : Not applicable
- Availability of data and materials : Not applicable
- Code availability : Not applicable
- Authors' contributions : Guillaume Perrin and Christian Soize wrote the whole manuscript.

References

[1] Ghanem, R., Spanos, P.D.: Polynomial Chaos in Stochastic Finite Elements. Journal of Applied Mechanics **57**(1), 197–202 (1990)

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

[3] Perrin, G., Soize, C., Duhamel, D., Funfschilling, C.: Identification of polynomial chaos representations in high dimension from a set of realizations. SIAM J. Sci. Comput. **34(6)**, 2917–2945 (2012)

[4] Wand, M.P., Jones, M.C.: Kernel smoothing. Encyclopedia of Statistics in Behavioral Science **60**(60), 212 (1995)

[5] Scott, D.W., Sain, S.R.: Multidimensional density estimation. In: Rao, C.R., Wegman, E.J., Solka, J.L. (eds.) Data Mining and Data Visualization. Handbook of Statistics, vol. 24, pp. 229–261. Elsevier, Amsterdam (2005)

[6] Perrin, G., Soize, C., Ouhbi, N.: Data-driven kernel representations for sampling with an unknown block dependence structure under correlation constraints. Journal of Computational Statistics and Data Analysis **119**, 139–154 (2018)

[7] Soize, C., Ghanem, R.: Probabilistic learning on manifolds (PLoM) with partition. International Journal for Numerical Methods in Engineering **123**(1), 268–290 (2022)

[8] Rokach, L., Maimon, O.: In: Maimon, O., Rokach, L. (eds.) Clustering Methods, pp. 321–352. Springer, Boston, MA (2005)

[9] Nürnberger, G.: Approximation by spline functions. Springer, Heidelberg (1989)

[10] Donoho, D.L., Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences **100**(10), 5591–5596 (2003) https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.1031596100. https://doi.org/10.1073/pnas.1031596100

[11] van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[12] Ghanem, R., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach, Rev. Ed. Dover Publications, New York (2003)

[13] Le Maître, O., Knio, O.M.: Spectral Methods for Uncertainty Quantification. Springer, Dordrecht (2010)

[14] Soize, C.: Uncertainty Quantification. Interdisciplinary Applied Mathematics, vol. 47, pp. 1–327. Springer, Cham (2017)

[15] Wu, J.: Cluster Analysis and K-means Clustering: An Introduction, pp. 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[16] Mak, S., Joseph, V.R.: Support points. Annals of Statistics **46**, 2562–2592 (2018)

[17] Teymur, O., Gorham, J., Riabiz, M., Oates, C.J.: Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) **130** (2021)

[18] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY, USA (2001)

[19] Aldenderfer, M.S., Blashfield, R.K.: Cluster Analysis. SAGE Publications, Inc (1984)

[20] Santner, T.J., Williams, B.J., Notz, W.I.: The Design and Analysis of Computer Experiments. Springer, New York (2003)

[21] Maday, Y., Nguyen, N.C., Patera, A.T., Pau, S.H.: A general multipurpose interpolation procedure: the magic points. Communications on Pure & Applied Analysis **8**(1), 383–404 (2009)

[22] Haberstich, C., Nouy, A., Perrin, G.: Boosted optimal weighted least-squares. Mathematics of Computation (2022)

[23] Silverman, B.W.: Density estimation for statistics and data analysis. In: Monographs on Statistics and Applied Probability vol. 37, p. 120 (1986)

[24] Duong, T., Cowling, A., Koch, I., Wand, M.P.: Feature significance for multivariate kernel density estimation. Computational Statistics and Data Analysis **52**(9), 4225–4242 (2008)

[25] Shorack, G.R., Wellner, J.A.: Empirical processes with applications to statistics. Philadelphia: Society for Industrial and Applied Mathematics, Boston, MA (2009)