Accepted Manuscript

An unsupervised data completion method for physically-based data-driven models

Jacobo Ayensa-Jiménez, Mohamed H. Doweidar, Jose A. Sanz-Herrera, Manuel Doblaré

PII: S0045-7825(18)30488-2
DOI: https://doi.org/10.1016/j.cma.2018.09.035
Reference: CMA 12095

To appear in: *Comput. Methods Appl. Mech. Engrg.*

Received date : 22 December 2017
Revised date : 21 September 2018
Accepted date : 24 September 2018

Please cite this article as: J. Ayensa-Jiménez, et al., An unsupervised data completion method for physically-based data-driven models, *Comput. Methods Appl. Mech. Engrg.* (2018), https://doi.org/10.1016/j.cma.2018.09.035

An unsupervised data completion method for physically-based data-driven models

Jacobo Ayensa-Jiménez[a], Mohamed H. Doweidar[a], Jose A. Sanz-Herrera[b], Manuel Doblaré[a,*]

[a]*Mechanical Engineering Department, University of Zaragoza, Spain; Aragón Institute of Engineering Research (I3A), University of Zaragoza, Spain; Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain. Campus Río Ebro, Edificio I+D, Mariano Esquillor s.n., 50018 Zaragoza (Spain)*
[b]*School of Engineering (ETSI), University of Seville, Spain*

Abstract

Data-driven methods are an innovative model-free approach for engineering and sciences, still in process of maturation. The idea behind is the combination of data analytics techniques, to handle the huge amount of data derived from continuous monitoring or experimental measurements, and of the constraints imposed by universal physical laws, particular to the field in hands. A well-known problem in the former corresponds to the quality and completeness of the available data that, sometimes, are so poor that make the predictions useless. In data-driven simulation-based engineering and sciences (DDSBES), the intrinsic physical constraints may help in completing the missing data in a more precise manner, by forcing them to remain in the manifold defined by the physical laws. In this work, a suitable imputation method to complete incomplete data that preserves the data context-dependent structure is presented. This is accomplished by enforcing the set of physical constraints, specific to the problem. For this purpose, a generalization of the weighted mean concept is proposed, where the distance to the admissible points (in a physical sense) is used as a weighting function to get the optimal candidate. The method is evaluated in a classical regression problem, where it is compared with other standard methods, showing

*Corresponding author
Email address: mdoblare@unizar.es (Manuel Doblaré)

better results. Then, its application is illustrated in two data-driven problems, where no filling data procedure has been yet proposed, showing good predictive capability, provided that the data are close enough to the actual system state.

Keywords: Data-driven methods, Data completion, Statistical imputation, Weighted mean, Computational Mechanics

1. Introduction

Data are everywhere around us. An incredibly huge amount of sensors and transducers get measurements from the physical world. Businesses of every kind search and collect data across the globe related to consumer preferences and trends. Governments regularly collect all sorts of data from census information to incident reports in police departments. According to the 2016 IDC directives presented in its yearly event in San Jose (US), this deluge of data is set to rise steeply from the estimated world total amount of 4,4 zettabytes of data in 2013 to 180 zettabytes by 2025 (one zettabyte is equivalent to one trillion of gigabytes). The advent of the Internet of Things will likely make to surpass these figures by far [1].

More and more complicated strategies are used to extract patterns and/or relevant knowledge from this massive amount of available structured and unstructured data. In fact, the framework, in which it is easier to get predictions directly derived from available data than from tedious, complicated and sometimes inaccurate mathematical models, is progressively changing the paradigm of predictive Physics [2, 3].

Since the main ideas and concepts were introduced at the beginning of the century, an extensive literature may be found on this broad area of Data Analytics and Artificial Intelligence [4, 5, 6]. The main trend today is the constant improvement of the accuracy of predictions by continuous "learning" from a non-stop input of new data, thus progressively refining the predictions by comparing the predicted and actual responses.

Since the seminal idea of the *perceptron* [7], artificial neural networks has

been another pushful field where new concepts as Deep Learning and Dynamic Networks are in continuous development. Today, these methods allow extracting abstract features and solving very complex problems, many times not fully formalized [8, 9]. These techniques try to mimic the process of human knowledge acquisition and structuring and have become amenable after remarkable advances in sensoring; data acquisition, transfer, storage and management; enormous improvements in the performance of computers; and continuous contributions in their theoretical and algorithmic foundations.

In an engineering context, a straightforward application of all these techniques is the so-called dynamic data-driven assimilation systems (DDDAS) [10], in which the idea is providing both predictive and learning capabilities to a control system from data acquired from a set of sensors. This paradigm was settled down by Kalman [11] in the sixties with his groundbreaking filter. Nowadays, it is still a hot topic of research [12].

In the last years, a new approach to simulation-based engineering and sciences that uses the power of data-science methods has been proposed. This approach, of increasing importance, and known as data-driven simulation-based engineering and sciences (DDSBES), combines physical constraints and raw data. In the absence of physical constraints, the standard Data Science and Machine Learning framework is recovered, while the use of an *a priori* parametric model, that fits the experimental data, recovers the classical SBES. Actually, all linear and nonlinear phenomenological constitutive models can be formulated in terms of parametric mathematical equations, where the variables of interest are forced to remain within a given pre-established manifold. This manifold is derived from observation and experience (empiricism) eventually by means of a trial-error fitting procedure.

One idea in this direction was started by Chinesta and coworkers [13] who defined a strategy for data-driven Computational Mechanics, combining manifold learning techniques and a (possibly optimized) directional search strategy inspired in the LaTin method [14]. Ortiz and his group [15] presented a model-free method based on the minimization of the distance between the searched

solution and a set of experimental data, using a proper energy norm. The solution was also forced to remain in the equilibrium manifold, by means of a well-posed penalty approach. This work was extended by several groups to take into account the uncertainty of the data in what are now called as reliability based data-driven solvers [16, 17].

However, the referred works are often restricted to the frame of perfect information, that is, the data-set is complete. In other words, all state variables are assumed to be known for a given measure. This is not always the case in practical situations, being this one of the main concerns in Data Science [18, 19, 20, 21, 22, 23].

There are many methods that have been developed to address this problem, both model-free or model-based. Among the model-free, the most fundamental are Listwise Deletion (LD) and Pairwise Deletion (PD) [22] that consist of discarding incomplete data. These methods, however, decrease statistical power [24] and introduce bias [25] if the missing process is Missing Not At Random (MNAR) [18, 26, 27], what is obviously the case when dealing with data obtained by experiments or measurements. Another common approach is Single Imputation (SI), based on a filling strategy for the missing data that uses values obtained from complete data, for example, Mean or Mode Imputation [19, 20]. This technique, however, reduces the variability and weakens the covariance between variables. Another approach is to create dummy variables accounting for the missing data variables (Dummy Variable Adjustment) [19, 20]. This results in biased estimators and is not theoretically based. Finally, it is also possible to replace missing values with a predicted score from a regression equation [20, 19]. This weakens the variance and overestimates model fit and correlation estimates. Moreover, these methods do not take into account the local structure and geometry of the data, which is critical when the data have some underlying physics. In order to solve this problem, interpolation (linear interpolation, nearest interpolation or spline interpolation, [28]) is a common technique. In this approach, only the physics inherent to the data is learned in the imputation process.

Other model-based methods have been developed to deal with the missing data problem, such as Multiple Imputation [29] using the regression method [30], the Predictive Mean Matching Method [31] or the Markov Chain Monte-carlo Method (MCMC) [25, 32], Full Information Maximum-Likelihood (FIML) estimation [33, 34, 35] and Expectation-maximization [36]. The problem of all these model-based methods is that they assume, one way or the other, a statistical model for the data (e.g. normality). This is usually the case in social and economical sciences [37] but is not the general case for physical problems, where variables follow some fundamental laws incompatible with normality or other distributional assumptions.

The amount of missing data is not, however, the sole criterion to assess the quality of the available data, especially if they correspond to a problem that relies on some physical laws [38]. Our aim in this work is then to establish a framework in which both the local structure of the data and the supplementary physics, not explicitly included in the data structure, are used to improve the imputation procedure. In this context, the imputation method can be compatible with any DDSBES method. The presented technique is based on the mean concept and, therefore, could be interpreted as a generalization of the Mean Imputation Method. On one hand, the local structure of the empirical data-set is preserved since the data are forced to belong to specific manifolds, which depend on the problem nature. On the other hand, the underlying physics of the problem is imposed via supplementary constraints on the data. The imputation procedure is then performed by using an unsupervised learning algorithm that finds the point that minimizes an, in general context-dependent, weighted quadratic error, while preserving the local and global physics of the problem.

2. Mathematical formulation

2.1. Data-driven simulation-based engineering and sciences (DDSBES): A general framework

Our aim is to present a methodology that fits within the context of data-driven problems. In particular, problems in which the governing equations of the system are fundamental laws of Physics and the modeling strategy is replaced by data, that, in general, may be incomplete. This means that computations will be carried out directly from data without an *a priori* parametrization step of the state equations (e.g. constitutive model in Continuum Mechanics).

In fact, any physical system can be defined as a manifold $\mathcal{M}$, the state space, that corresponds to the admissible states that fulfil a set of equations defining the particular physical problem in hands [39]. Usually, the state space is treated as an embedded manifold $\mathcal{M} \subset \mathbb{R}^n$ in a higher Euclidean dimensional space and is defined in terms of a set of state governing equations $F(\mathbf{x}) = 0, \mathbf{x} \in \mathbb{R}^n$. Observables of the system are magnitudes that are related to the state variables by means of geometric or physical relations. For example, in Continuum Mechanics, forces are observable variables related to the stress components, while displacements are observable variables associated with strains. The relationship between forces and stresses is defined by means of the equilibrium equations, while the relation between displacements and strains is derived from kinematic conditions. A set of measurements of the system is a set of observables obtained in particular conditions.

For instance, in Continuum Mechanics, the state variables are the stress and the strain tensors with $6 + 6 = 12$ components in 3D problems for each spatial point (assuming balance of angular momentum yields). But these components are not independent. Indeed, they are related to the observables by kinematic and equilibrium (linear momentum conservation) constraints (note that the standard constitutive relations are implicit in the measurements so they are not considered explicitly).

The fundamental problem in Physics is stated as: from a set of measure-

ments in the physical system given by the values of several observables, derive the value of another observable of interest or/and the rest of state variables. Unfortunately, this inference strongly depends on the quality of measurements as well as on the particular complexity of the system.

As we have seen, measurements (observables) are related to the state-space variables. However, the information given by these measurements is frequently less informative than the state variables themselves. In other words, measurements are known variables living in a lower dimensional space. For example, the displacements along a given direction do not give us the information about the complete strain tensor while the forces over a surface do not characterize the whole stress tensor. In a data-driven framework, the state of the system as well as any desired observable has to be derived from a set of measurements. However, in practical situations, we do not have complete information about the state variables, but only particular measurements in particular states. That is, we have an incomplete set of data. As these measurements are related to the state variables, they may be formulated in terms of manifolds. For the Continuum Mechanics problem, for instance, the knowledge of the stress associated with a given plane orientation tells us the relationship between components of the stress tensor, so we have a set of measurements with a reduced dimension of the state space. It is essential, therefore, to use these values as points of measure sets being then the goal to properly complete these incomplete measurements to perform data-driven computation. The filling data strategy should take into account therefore the following assumptions:

- It should take into account all the measurements, that are formulated in terms of manifolds embedded in the state space (incomplete measurements) better than in terms of points (complete measurements). We call this condition the *generalization* assumption.
- It should respect the physics of the system: the new derived data obtained from the incomplete data should be consistent with the geometric structure defined by the data manifolds. We call this condition the *consistency*

assumption.

- It should guarantee accuracy for the states associated with the input observables used as starting points for predictions. Among all observations in our data-set, those closer to our physical constraints or to our actual observables should be overweighted. This is done by means of an *appropriate weighting strategy*.

The methodology for incomplete data processing presented herein should be defined to be in accordance with this framework for DDSBES.

First, it is fundamental to define an averaging technique that takes into account complete and incomplete measurements according to the generalization assumption and that respects the data structure, according to the consistency assumption. Next, an appropriate weighting strategy for this kind of problems will be defined, i.e, how to compute a set of weighting values in the averaging process in order to make good predictions on the state of the system. Consequently, the data completion step should take into account how far are the measurements from the known observables and/or the *physical manifolds*.

2.2. Averaging procedure: Generalization and consistency

In this section, the filling-data strategy is presented, using a formal mathematical framework. Let us consider a set of complete measurements in the measurement space, $p \in \mathbb{R}^n$, being n the dimension of the space, and a set of incomplete measurements that, in general, will be embedded manifolds $\mathcal{M} \subset \mathbb{R}^n$, with $\mathcal{M} = \{x \in \mathbb{R}^n | \Phi(x) = 0\}$, $\Phi : \mathbb{R}^n \to \mathbb{R}^k, k < n$, the map defining the manifold, and $\dim(\mathcal{M}) = n - k = m$. A very particular example of these maps is the one of orthogonal projections on a linear manifold. In this particular case, if $\Phi = \pi_{\mathcal{V}}$ is a linear projection over a given linear manifold $\mathcal{V}$, then $\mathcal{M} = \mathcal{V}^{\perp}$. In the field of continuum mechanics, normal stresses associated with a given plane, strains associated with a given direction or the mean pressure at a given point are examples of these incomplete measurements, because only linear relations between the components of the whole stress or strain tensors are known.

2.2.1. Manifolds instead of points: generalized mean and variance

Next, some mathematical generalizations of the mean concept will be derived. Let us assume that we have a set of N weighted points $\{(w_j, x_j)|\, , 0 \leq w_j,\, x_j \in \mathbb{R}^n, j = 1, \cdots, N\}$. A possible interpretation of this mathematical structure is a set of data points with different reliability. Weights can then be associated with the measurement accuracy, physical reliability (explored later in Section 2.3) or other reliability criteria as clustering or outlier filtering. We define the *mean squared error* (mse) function associated to a given point x (represented by its coordinates $\mathbf{x}$) as:

$$\text{mse}(x) = \sum_{j=1}^{N} w_j d^2(x, x_j) = \sum_{j=1}^{N} w_j ||\mathbf{x}_j - \mathbf{x}||^2 \tag{1}$$

A classical result from probability theory [40] states the following:

The function mse *is minimized when x has coordinates $\mathbf{x} = \bar{\mathbf{x}}$ and the value of this minimum is* $\text{Tr}(\mathbf{S})$ *where $\bar{\mathbf{x}} = \sum_{j=1}^{N} w_j \mathbf{x}_j$ is the weighted mean value of the data and $\mathbf{S} = \sum_{j=1}^{N} w_j (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T$ its covariance matrix.*

Let us suppose now that we have N manifolds $\mathcal{M}_1, \cdots, \mathcal{M}_N$. Given a set of N weights $\{0 \leq w_j,\ j = 1, \cdots, N\}$, we call **generalized mean value** to the value x^* that minimizes the weighted mean squared error function (1) except for the fact that now we consider manifolds instead of points (and therefore the distance from a point to a manifold, not between points). We can define a new unconstrained minimization problem, which is the natural generalization of the former, as:

$$\min_{x \in \mathbb{R}^n} \quad \text{mse}(x) = \sum_{j=1}^{N} w_j d^2(x, \mathcal{M}_j) \tag{2}$$

The value $V_G = \text{mse}(x^*)$ is called the *generalized variance*.

Computational solution.. Let us derive a computational solution to the problem (2) for linear manifolds. Let us consider the linear manifolds defined in terms of their vector director subspaces $\mathcal{M}_j = p_j + M_j$, where $p_j \in \mathbb{R}^n$ and M_j is the generator vector space associated to $\mathcal{M}_j$, that can be defined with an orthonormal basis $M_j = \langle \mathbf{u}_{j1}, \mathbf{u}_{j2}, \cdots, \mathbf{u}_{jm_j} \rangle$. Here m_j is the dimension of $\mathcal{M}_j$. Let $\mathbf{A}_j$ be the matrix with column vectors $\mathbf{u}_{ji}$, $A_{ji} = u_{ji}$, $j = 1, \cdots, N$, $i = 1, \cdots, m_j$, and $\mathbf{p}_j$ the vector associated with the point p_j, then we have:

Proposition 2.1 (Computational characterization). *The solution of the problem (2) is obtained by solving the linear system*

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3}$$

where:

$$\mathbf{A} = (\sum_{j=1}^{N} w_j)\mathbb{I} - \sum_{j=1}^{N} w_j \mathbf{A}_j \mathbf{A}_j^T \tag{4}$$

and

$$\mathbf{b} = (\sum_{j=1}^{N} w_j \mathbf{p}_j) - \sum_{j=1}^{N} w_j \mathbf{A}_j \mathbf{A}_j^T \mathbf{p}_j \tag{5}$$

Proof:. The distance from the point x to a linear manifold $\mathcal{M}_j = p_j + M_j$, d_j, is given by

$$d_j^2(x, \mathcal{M}_j) = ||\mathbf{x} - (\pi_{M_j}(\mathbf{x} - \mathbf{p}_j) + \mathbf{p}_j)||^2 \tag{6}$$

with, as always, $\mathbf{p}_j, \mathbf{x}$ refers to the coordinates, in a reference system, of points p_j, x, respectively, π_{M_j} is the (vectorial) orthogonal projection over the vector subspace M_j.

Using the matrix expression in coordinates of the orthogonal projection:

$$d_j^2 = ||\mathbf{x} - (\mathbf{A}_j\mathbf{A}_j^T(\mathbf{x} - \mathbf{p}_j) + \mathbf{p}_j)||^2 \tag{7}$$

with $\mathbf{A}_j$ the matrix associated to M_j.

To minimize $D^2 = \sum_{j=1}^N w_j d_j^2$, the function to minimize yields:

$$D^2(\mathbf{x}) = \sum_{j=1}^N w_j ||\mathbf{x} - (\mathbf{A}_j\mathbf{A}_j^T(\mathbf{x} - \mathbf{p}_j) + \mathbf{p}_j)||^2 \tag{8}$$

We can compute the gradient of D^2 as:

$$\frac{\partial(D^2)}{\partial \mathbf{x}} = 2\sum_{j=1}^N w_j \left(\mathbb{I} - \mathbf{A}_j\mathbf{A}_j^T\right)\left(\mathbf{x} - \left(\mathbf{A}_j\mathbf{A}_j^T(\mathbf{x} - \mathbf{p}_j) + \mathbf{p}_j\right)\right)$$

$$\frac{\partial(D^2)}{\partial \mathbf{x}} = 2\sum_{j=1}^N \left(\left[w_j\mathbb{I} - w_j\mathbf{A}_j\mathbf{A}_j^T\right]\mathbf{x} - \left[w_j\mathbb{I} - w_j\mathbf{A}_j\mathbf{A}_j^T\right]\mathbf{p}_j\right)$$

Solving for $\frac{\partial(D^2)}{\partial \mathbf{x}} = \mathbf{0}$, we obtain

$$\left[\left(\sum_{j=1}^N w_j\right)\mathbb{I} - \sum_{j=1}^N w_j\mathbf{A}_j\mathbf{A}_j^T\right]\mathbf{x} = \left[\sum_{j=1}^N w_j\mathbf{p}_j - \sum_{j=1}^N w_j\mathbf{A}_j\mathbf{A}_j^T\mathbf{p}_j\right] \tag{9}$$

□

We observe that if $w_j = \frac{1}{N}$ and the linear manifolds have 0 dimension, that is, they are points, $\mathcal{M}_j = \{p_j\}$ and $\mathbf{A}_j = \mathbf{0}$, then

$$\mathbf{A} = \mathbb{I} \tag{10}$$

$$\mathbf{b} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{p}_j = \bar{\mathbf{p}} \tag{11}$$

obtaining, therefore, the mean of the points p_j, being this the reason for the denomination of generalized mean.

2.2.2. Manifolds instead of the whole space: consistent mean and variance

Let us suppose now that we have a manifold $\mathcal{M}$ (that is, an incomplete measurement) and N points x_j of $\mathbb{R}^n$, $j = 1, \cdots, N$, described in terms of N coordinate vectors $\mathbf{x}_1, \cdots, \mathbf{x}_N$.

Given a set of N weights $\{0 \leq w_j,\, j = 1, \cdots, N\}$, the value $x^* \in \mathcal{M}$ that minimizes the weighted mean squared error function (1) is called the *consistent mean value* with respect to $\mathcal{M}$, which is the solution of the constrained minimization problem:

$$\min_{x \in \mathcal{M}} \quad \text{mse}(x) = \sum_{j=1}^{N} w_j d^2(x, x_j) \tag{12}$$

The value $V_{\mathcal{M}} = \text{mse}(x^*)$ is called the *consistent variance* with respect to $\mathcal{M}$. Note that now the manifold $\mathcal{M}$ acts as a constraint of the problem.

Computational solution. Now, to derive a computational solution to the problem (12) when $\mathcal{M}$ is a linear manifold, we solve an equivalent unconstrained minimization problem.

Proposition 2.2. *Let $\mathbf{x}_j$, $j = 1, \cdots, N$ coordinate vectors associated to points, $x_j \in \mathbb{R}^n$, such as $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the mean and covariance matrix of the vectors. Let $\mathcal{M}$ be a linear manifold of dimension $m \leq n$, then, the solution to the constrained minimization problem*

$$\min_{x \in \mathcal{M}} \quad \text{mse}(x) \tag{13}$$

is given by:

$$\mathbf{x}^* = \pi_{\mathcal{M}}(\bar{\mathbf{x}}) \tag{14}$$

Moreover, if $s^2 = \text{mse}(\mathbf{x}^*)$*, then:*

$$s^2 = \sum_{j=1}^{N} w_j(\mathbf{x}_j - \pi_{\mathcal{M}}(\mathbf{x}_j))^2 + \sum_{j=1}^{N} w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x}^*)^2 \tag{15}$$

Proof:. The proof is based on Pythagoras theorem. So, we can get:

$$\begin{aligned}
\text{mse}(\mathbf{x}) &= \sum_{j=1}^{N} w_j(\mathbf{x}_j - \mathbf{x})^2 \\
&= \sum_{j=1}^{N} w_j \left[(\mathbf{x}_j - \pi_{\mathcal{M}}(\mathbf{x}_j))^2 + (\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x})^2\right] \\
&= \sum_{j=1}^{N} w_j(\mathbf{x}_j - \pi_{\mathcal{M}}(\mathbf{x}_j))^2 + \sum_{j=1}^{N} w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x})^2
\end{aligned} \tag{16}$$

If we analyze this last expression, the only term depending on $\mathbf{x}$, is $\sum_{j=1}^{N} w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x})^2$, then the minimum is achieved for $\mathbf{x}^* = \overline{\pi_{\mathcal{M}}(\mathbf{x}_j)} = \pi_{\mathcal{M}}(\bar{\mathbf{x}})$, where in the last expression we use linearity of the projection operator.

Additionally, using the equation (16), and $s^2 = \text{mse}(\mathbf{x}^*)$, we obtain:

$$s^2 = \text{mse}(\mathbf{x}^*) = \sum_{j=1}^{N} w_j(\mathbf{x}_j - \pi_{\mathcal{M}}(\mathbf{x}_j))^2 + \sum_{j=1}^{N} w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x}^*)^2 \tag{17}$$

□

We have then obtained an orthogonal decomposition of the quadratic spread of vectors $\mathbf{x}_j$. The term $\sum_{j=1}^{N} w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \mathbf{x}^*)^2$ is denoted as $s^2_{\mathcal{M}}$ since it represents the spread of the points $\mathbf{x}_j$ projected on the manifold $\mathcal{M}$.

Therefore, the next result is straightforward.

Corollary 2.1 (Computational characterization). *Using the same hypothesis as in the previous result and if the linear manifold* $\mathcal{M}$ *is described using an orthonormal basis* $\mathcal{R} = \{p; \mathbf{u}_1, \cdots, \mathbf{u}_m\}$*, we have:*

$$\mathbf{x}^* = \mathbf{A}\mathbf{A}^T\bar{\mathbf{x}}$$
$$s^2_{\mathcal{M}} = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T\mathbf{S})$$

with $\mathbf{A}$ *the matrix with column vectors* $\mathbf{u}_i$*,* $\bar{\mathbf{x}}$ *the coordinates in the reference frame of the mean of points* x_j*, that is,* $\bar{\mathbf{x}} = \sum_{j=1}^N w_j\mathbf{x}_j$ *and* $\mathbf{S}$ *the covariance matrix of points* x_j*, that is* $\mathbf{S} = \sum_{j=1}^N w_j(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T$*.*

Proof:. Let us define $\mathbf{T} = \sum_{j=1}^N w_j(\pi_{\mathcal{M}}(\mathbf{x}_j) - \pi_{\mathcal{M}}(\bar{\mathbf{x}})(\pi_{\mathcal{M}}(\mathbf{x}_j) - \pi_{\mathcal{M}}(\bar{\mathbf{x}}))^T$, therefore:

$$\begin{aligned}
\mathbf{T} &= \sum_{j=1}^N w_j(\mathbf{A}\mathbf{A}^T\mathbf{x}_j - \mathbf{A}\mathbf{A}^T\bar{\mathbf{x}})(\mathbf{A}\mathbf{A}^T\mathbf{x}_j - \mathbf{A}\mathbf{A}^T\bar{\mathbf{x}})^T \\
\mathbf{T} &= \mathbf{A}\mathbf{A}^T\left[\sum_{j=1}^N w_j(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T\right]\mathbf{A}\mathbf{A}^T \\
\mathbf{T} &= \mathbf{A}\mathbf{A}^T\mathbf{S}\mathbf{A}\mathbf{A}^T
\end{aligned} \tag{18}$$

Therefore $s^2_{\mathcal{M}} = \mathrm{Tr}(\mathbf{T}) = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T\mathbf{S}\mathbf{A}\mathbf{A}^T) = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T\mathbf{S})$ where we have used the fact that Tr is a cyclic operator and $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ as $\mathbf{u}_i$ are orthonormal vectors.

□

We observe that, if $w_j = \frac{1}{N}$ and $\mathcal{M} = \mathbb{R}^n$, that is, $\mathcal{M}$ is the whole space, $\mathbf{A} = \mathbf{I}_n$ and

$$\mathbf{x}^* = \bar{\mathbf{x}} \tag{19}$$

$$s^2_{\mathcal{M}} = \mathrm{Tr}(\mathbf{S}) \tag{20}$$

obtaining directly the mean value for the points x_j and the whole uncertainty, being this the reason for the denomination of consistent mean.

2.2.3. Filling data using the consistent generalized mean and variance: an unsupervised learning technique

The idea of using both generalizations at the same time for missing data techniques is natural. Let us assume that we have a set of partially incomplete data $\mathcal{D}$ of size N, that is a set of manifolds $\{\mathcal{M}_j, j = 1, \cdots, N\}$ with $\mathcal{M}_j \subset \mathbb{R}^n$ with respective weights $\{w_j, j = 1, \cdots, N\}$. Here, $\mathbb{R}^n$ is the embedding space and $\mathcal{M}_j$ represents the (incomplete) measurements related to different states.

One strategy for data completion of missing data is using the tools presented in section 2.2.1. When those measurements are defined in terms of linear manifolds, we have derived, also, a closed linear expression (Proposition 2.1 and Corollary 2.1). The idea is, therefore, to solve the next minimization problem for each measurement $\mathcal{M}_i$, $i = 1, \cdots, N$:

$$\min_{x \in \mathcal{M}_i} \quad \text{mse}(x) = \sum_{j=1}^{N} w_j d^2(x, \mathcal{M}_j) \tag{21}$$

The solution x_i of this problem is the completed data associated to the incomplete data $\mathcal{M}_i$. To summarize, and for linear manifolds, a strategy is assumed for finding a minimum candidate, which the natural and strongly reduces the time required to obtain the actual solution of the minimization problem 21 directly using minimization algorithms. This strategy follows two steps:

1. **Global computation step.** Computation of the solution point for the unconstrained minimization problem using the expression given in Proposition 2.1.
2. **Projection step.** Computation of the solution point for the constrained minimization problem using the projection of the mean value and uncertainty defined in Corollary 2.1.

The geometric interpretation of this method is provided in Figure 1.

Remark. This two-step method may be used when working with non-linear manifolds as an iterative tangent based algorithm. At each step, the tangent space to the nonlinear manifold is computed at the current point and the linear

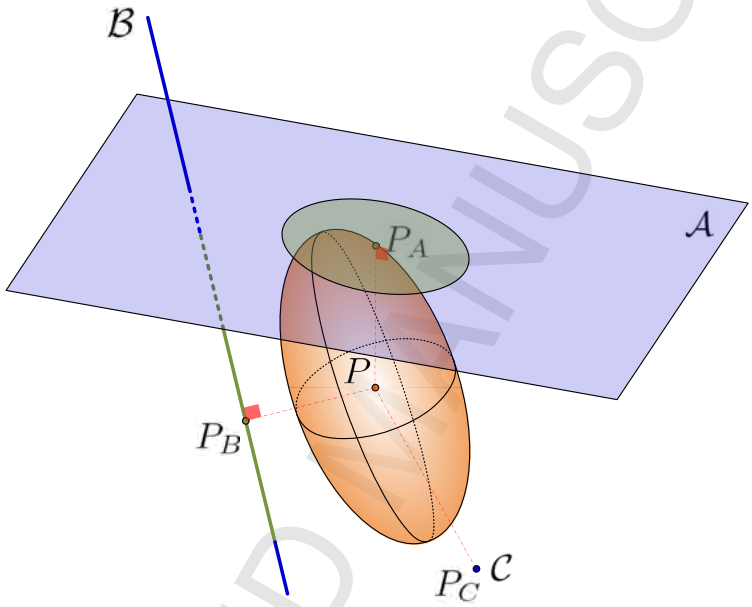


Figure 1: Geometric representation of the method. $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ represent three linear manifolds of dimension 2, 1 and 0 respectively, associated to three measurements in a space of dimension 3, the last one complete. Point P is the generalized mean in the sense defined in Equation 2, that is, the point minimizing the sum of (eventually weighted) squared distances to the manifolds. In orange, the uncertainty ellipsoid, related to the generalized variance, represents the spread of (eventually incomplete) measurements. Points P_A, P_B and P_C are the consistent generalized means associated to each of the manifolds in the sense defined in Equation 21, that is, the projection of the generalized mean on each manifold. In green, the associated uncertainty ellipsoid, related to the consistent generalized variance, for each manifold is depicted.

problem is solved. Thus, a point belonging to the tangent space is obtained. Using the exponential map [41], it is possible to obtain an associated point belonging to the manifold, which is used as the starting point for the next iteration. This strategy is usual in nonlinear computational mechanics [42, 43]. The iteration scheme stops when the distance between the subsequent global solutions is lower than a given tolerance. This construction is illustrated by the schematic diagram in Figure 2. This algorithm may be computationally expensive and is very dependent on the manifold smoothness and convexity, being this type of problems out of the scope of this paper. □

In any case, when the algorithm achieves convergence, N values of X are obtained, one for each manifold, x_i, $i = 1, \cdots, N$, representing the expected value associated with the manifold $\mathcal{M}_i$. Besides, for linear problems, we obtain for each manifold a value $s^2_{\mathcal{M}_i}$ that characterizes the uncertainty related to the manifold $\mathcal{M}_i$ and the matrix $\mathbf{S}_{\mathcal{M}_i}$ characterizing the uncertainty spread on this manifold. This *uncertainty ellipsoid* can be seen in Figure 1 for linear manifolds of dimension 0, 1 and 2, respectively.

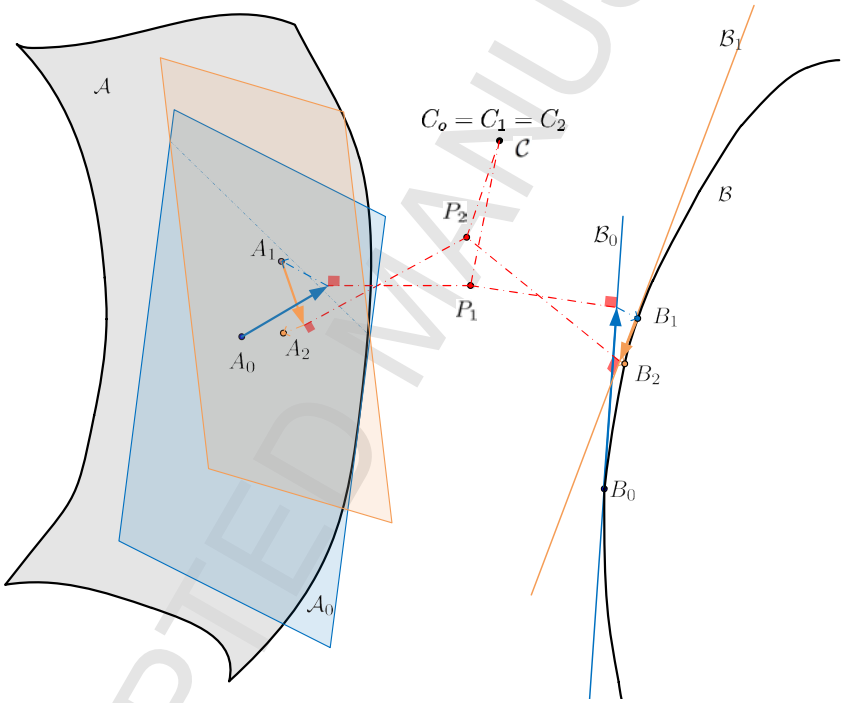


Figure 2: Extension of the method to nonlinear problems: Initially, for each manifold ($\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$) an initial point is selected (A_0, B_0 and $\mathcal{C}_0 = \mathcal{C}$) and the tangent spaces are computed. Then, the optimal completion, i.e., the generalized consistent mean for each linear manifold (in red), is computed. Using the exponential map, these projections on the linear manifolds are translated to the respective manifolds ($\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$) obtaining a new point in each manifold (A_1, B_1 and $C_1 = C_0$). The process is repeated until convergence.

2.3. Introducing physical laws: weighting strategy

Here we introduce the weighting strategy. Let us consider a system defined in terms of a physical manifold $\mathcal{M}$ and a set of measure manifolds $\mathcal{N}_j, j = 1, \cdots, N$. The starting point from which we want to derive the state of the system is another measure, that is, another manifold $\mathcal{N}$ that could be, for instance, boundary conditions for a given problem. The fundamental idea is to compute the manifolds $\mathcal{M}_j = \mathcal{N}_j \cap \mathcal{N} \cap \mathcal{M} = \mathcal{N}_j \cap \mathcal{P}$ where $\mathcal{P} = \mathcal{N} \cap \mathcal{M}$ and to perform the unsupervised learning strategy for this reduced space. This strategy has two direct consequences:

- **Physical consequence**: Since we are only learning second order statistics of physically admissible manifolds, the result will have a more physical sense.
- **Numerical consequence**: Projections are performed in a smaller space so the computational cost will be lower.

In that case, we are looking for a physically admissible (incomplete) data point measure that has the lowest uncertainty. However, this strategy can dramatically fail for few data with non-negligible uncertainty. For example, a measure manifold $\mathcal{N}_j$ may be close to the manifold $\mathcal{P}$ but $\mathcal{P} \cap \mathcal{N}_j = \emptyset$ so this measure will not be used for the system learning, even though it is very close to the real state.

An intermediate solution is using an activation function in the learning step, depending on the distance to the manifold $\mathcal{P}$, $d = d(\mathcal{P}, \mathcal{N}_j)$. That is, a function $\phi : \mathbb{R}^+ \to [0; 1]$ so that if $z = 0$, $\phi(z) = 1$ and eventually if $z \to +\infty$, $\phi(z) \to 0$. Given $\mathcal{N}_j \subset \mathbb{R}^n$, defining $u_j = u(\mathcal{N}_j) = d(\mathcal{N}_j, \mathcal{P})$, it is possible to define $w_j = \phi(u_j)$ in the learning process. In other words, the nearer the considered data set to $\mathcal{P}$, the higher the weight should be in the minimization of the optimal distance. Some possible activation functions are the step function $\phi(u) = \chi_{[0;a]}(u)$, where χ_A is the characteristic function of the set A and $a \geq 0$, radial basis functions (RBF), $\phi(u) = \exp\left(-\frac{u^2}{2\varsigma^2}\right)$ with $\varsigma > 0$, homographic

functions, $\phi(u) = \left(\frac{a}{a+bu}\right)^k$, with $a, b, k > 0$, or generalized ramp functions $\phi(u) = \left(1 - \left(\frac{u}{a}\right)^k\right) \chi_{[0;a]}(u)$, with $a, k > 0$. Figure 3 illustrates the geometric idea under the presented filling method when combined with the physics of the problem.

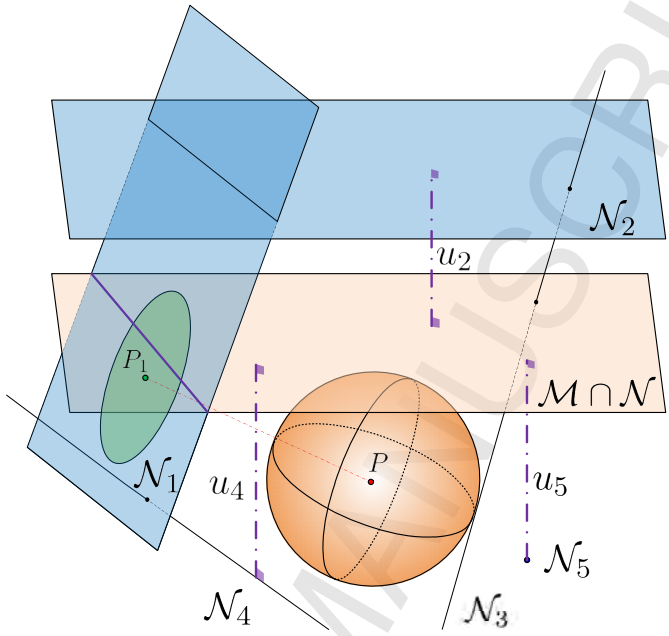


Figure 3: Geometric idea of the learning phase for data completion of the incomplete measure $\mathcal{N}_1$ and its corresponding uncertainty. Blue manifolds $\mathcal{N}_j$, $j = 1, \cdots, 5$. represent measurements (all incomplete except $\mathcal{N}_5$, which is a point). These measurements may correspond to states far from the current state of the system, that should belong to the orange physical manifold $\mathcal{P} = \mathcal{M} \cap \mathcal{N}$ that corresponds to the intersection of the points satisfying the governing equations of the problem (equilibrium, thermodynamics, Maxwell's equations etc.) and knowledge about the system state, e.g. boundary conditions or measured control variables. Each of the measure manifolds is at a certain distance u_j from the physical manifold. In particular, in the figure case, $u_1 = 0$ and $u_3 = 0$ because $\mathcal{N}_1 \cap \mathcal{M} \neq \emptyset$ and $\mathcal{N}_3 \cap \mathcal{M} \neq \emptyset$. Note that this situation is frequent in a three-dimensional state-space, but it is less and less probable when the dimension of the total space becomes very high. From these distances, the weights are calculated by $w_j = \phi(u_j)$, so that if $u = 0$, $\phi(u) = 1$ and if $u \to \infty$, $\phi(u) \to 0$. Thus, if $x \in \mathbb{R}^3$, the distance $D^2(x) = \sum_{j=1}^{5} w_j d_j^2(x) = \sum_{j=1}^{5} \phi(u_j) d^2(x, \mathcal{N}_j)$ is minimized (generalized mean) and the solution point P (in red) is projected onto the measure manifolds (consistent mean) obtaining the measure completion, as illustrated in Figure 1. Likewise, depending on the "spread" of the sets, an ellipsoid of three-dimensional uncertainty (in red), is obtained, related to the generalized variance. This ellipsoid is projected in each of the manifolds, obtaining ellipsoids related to the associated consistent variance (in green). Note that if the red ellipsoid is very slender in the direction orthogonal to the measure manifold that we are completing, this would have no impact on the projected green ellipsoid. The point P_1, is then the generalized consistent mean associated to incomplete measure $\mathcal{N}_1$ weighted by the neighborhood to the problem physics.

2.4. Looking for the nearest measure to a given point: solving the data-driven problem

2.4.1. Preliminary mathematical results

Let $\mathcal{M} \subset \mathbb{R}^n$ be an embedded manifold of the Euclidean space of dimension n with associated probability distribution ρ, that accounts for the probability distribution of a given random point x belonging to $\mathcal{M}$. For instance, in the case of linear manifolds, $\mathcal{M}$ can be described as $\mathcal{M} = p + \langle \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m \rangle$, where $\langle \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m \rangle$ is the linear span generated by $\mathbf{v}_1, \cdots, \mathbf{v}_m$. That is $\mathcal{R} = \{p; \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m\}$ is a basis of $\mathcal{M}$ and $m = \dim(\mathcal{M})$.

Let $x \in \mathcal{M}$ and $\rho : \mathcal{M} \to \mathbb{R}^+$, be the probability distribution describing the position of x. Then we define the square distance random variable $D^2 = d^2(p, X)$, with p (deterministic) and X (random) defined by their coordinates $\mathbf{p}$, $\mathbf{X}$ in a (global) reference frame, as

$$D^2 = ||\mathbf{p} - \mathbf{X}||^2 \tag{22}$$

It is possible to define (under some integrability conditions) the expected value $\mathbb{E}[D^2] = \int_{\mathcal{M}} d^2(p, X)\rho \, dV$, or, using the point coordinates $\mathbf{x}$ (note that $\mathbf{X}$ is a random vector while $\mathbf{x}$ the vector point coordinates), $\mathbb{E}[D^2] = \int_{\mathcal{M}} ||\mathbf{p} - \mathbf{x}||^2 \rho(\mathbf{x}) \, dV(\mathbf{x})$. More important than the explicit computation of the expected value of D^2 in terms of a given parametrization are the following results for linear manifolds, proven in Appendix A, and relating the moments of D^2 to the moments of X. As usual, the random variable X and the point p are identified with their coordinates description $\mathbf{X}$, and $\mathbf{p}$. If $\boldsymbol{\mu}$ is the expected value of the random vector $\mathbf{X}$ ($\mu_i = \mathbb{E}[X_i]$), $\boldsymbol{\Sigma}$ its variance-covariance matrix ($\Sigma_{ij} = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$) and $\boldsymbol{\Upsilon}$ its fourth order moment tensor ($\Upsilon_{ijkl} = \mathbb{E}[X_i X_j X_k X_l]$), then:

$$\mathbb{E}[D^2] = ||\mathbf{p} - \pi_{\mathcal{M}}(\mathbf{p})||^2 + ||\pi_{\mathcal{M}}(\mathbf{p}) - \boldsymbol{\mu}||^2 + \mathrm{Tr}(\boldsymbol{\Sigma}) \tag{23}$$

$$\mathrm{Var}[D^2] = \mathbf{I} : \boldsymbol{\Upsilon} : \mathbf{I} - \big(\mathrm{Tr}(\boldsymbol{\Sigma}) + (\pi_{\mathcal{M}}(\mathbf{p}) - \boldsymbol{\mu})^T(\pi_{\mathcal{M}}(\mathbf{p}) - \boldsymbol{\mu})\big) \tag{24}$$

Here, $\mathbf{I}$ is the second order identity tensor. Finally, under normality conditions, D^2 follows a noncentral χ^2 distribution with $m = \dim(\mathcal{M})$ degrees of freedom and non-centrality parameter $\lambda = (\pi_{\mathcal{M}}(\mathbf{p}) - \boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\pi_{\mathcal{M}}(\mathbf{p}) - \boldsymbol{\mu})$.

2.4.2. The closest point in a stochastic sense

Let us go back now to the methodology and tools introduced in section 2.2. Once the constrained minimization problem is solved for each incomplete measurement $\mathcal{M}_i$ and all filled data points x_i are derived, it is possible to compute how far is a state $p \in \mathbb{R}^n$ of the system from a given data-point i. Moreover, we can define for each $p \in \mathbb{R}^n$ which is the closest measure, and from this to define a tessellation of the state space in terms of the measurements.

One could consider the deterministic distance $d_i = d(p, x_i)$ but this distance would not have into consideration the accuracy of the filling step and the effect on physical weights on uncertainty. It is more natural to consider a stochastic distance. Indeed, considering again the random variable D_i^2 defined at Section 2.4.1 associated with the manifold $\mathcal{M}_i$, we may define, denoting $s^2_{\mathcal{M}_i}$ by s_i^2 and $\pi_{\mathcal{M}_i}$ by π_i:

$$d_i^2 = \mathbb{E}[D_i^2] = d^2(p, \pi_i(p)) + d^2(\pi_i(p), x_i) + s_i^2 \tag{25}$$

The manifold $\mathcal{M}_i$ verifying that d_i^2 is minimal is the **closest manifold in the statistical sense** to the point p. Besides, each term in d_i^2 has its own interpretation:

- $T_{i,1} = d^2(p, \pi_i(p))$ is the statistical error due to finite measurements of the sample. It is related to the lack of knowledge about the system, since the information is obtained by means of a given finite data-set. The more measurements are added, the lower the error usually is.
- $T_{i,2} = d^2(\pi_i(p), x_i)$ is inherent to the manifold and depends on the manifold selection. It is unavoidable to some extent.
- $T_{i,3} = s_i^2$ is the term associated with the uncertainty and is characteristic of the self-learning process: the worse the manifold learning, the higher

this term. Locating properly the *real* point in a manifold, even though it actually belongs to that manifold, is less accurate when this term increases.

Moreover, if measurement uncertainty is taken into consideration, it is possible to state $T_{i,3} = s_i^2 + s_i'^2$ where $s_i'^2$ is the quadratic uncertainty of the i-th measurement and, therefore, *orthogonal* to the uncertainty associated to the filling procedure. This uncertainty, nevertheless, is not being considered in the applications presented in this work.

The geometric idea behind these considerations is illustrated in Figure 4.

Once the learning step is finished, the data-driven problem may be solved as usual [15, 17], provided the uncertainty of the completion is considered, as just explained. Figure 5 illustrates the geometric idea.

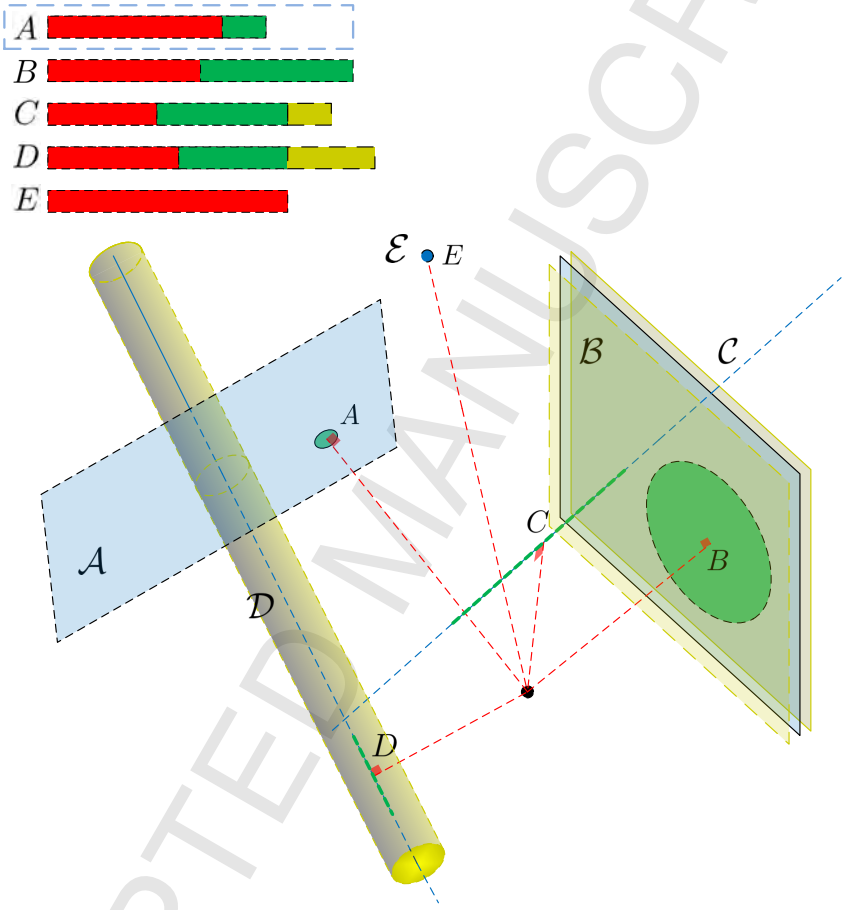


Figure 4: Stochastic distances to different measurements. Manifolds $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{E}$, associated to incomplete measurements, have been completed using the procedure described above. Completed measurements are represented by the points A, B, C, D and E as well as their associated uncertainty ellipsoids (including terms $T_{j,2}$ and $T_{j,3}$). Measurement uncertainty is illustrated in yellow and is taken into account in distance computations. Even if the completed measure associated to the manifold $\mathcal{C}$ is the closest in a deterministic sense, the one associated to the manifold $\mathcal{A}$ is the closest in stochastic sense and the one associated to the manifold $\mathcal{D}$ the farthest

.

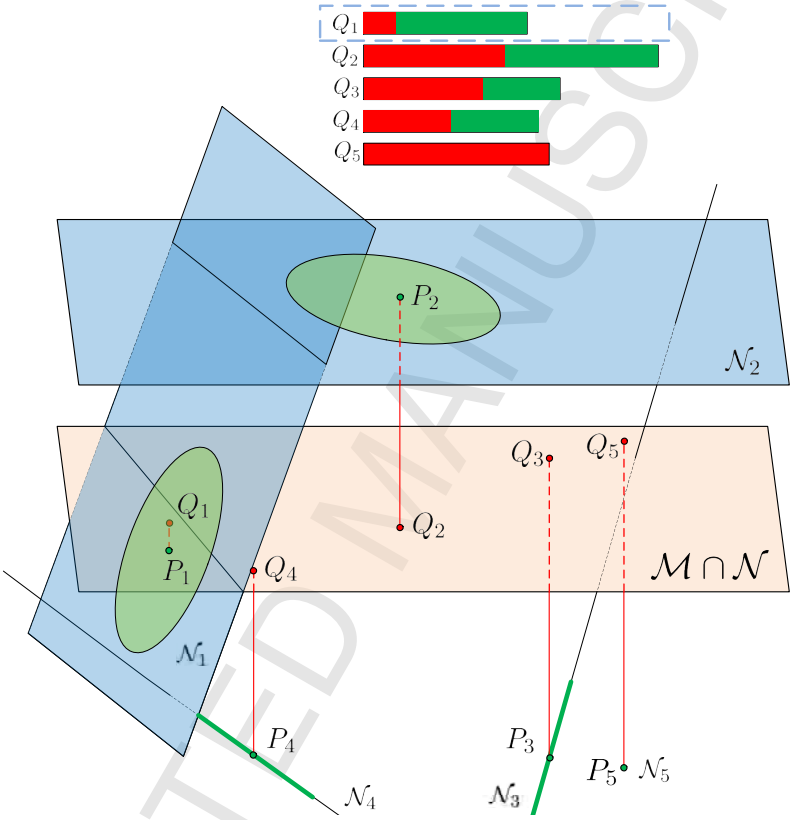


Figure 5: Geometric idea of the problem solving stage starting from complete measures given by points P_i, $i = 1, \cdots, 5$, and associated quadratic uncertainties (consistent generalized variances) s_i^2, $i = 1, \cdots, 5$. Suppose that the previous process has been repeated for all the measure manifolds and that we have associated complete expected points with their associated uncertainty. Now, the algorithm looks for the complete point closest to the physical manifold, including both the deterministic and the stochastic parts of the squared distance (as shown in Figure 4). Then the point associated to the complete measure closest to the physical manifold (in this case it would be the measure associated with the manifold $\mathcal{N}_1$) is selected and its projection over the physical manifold $\mathcal{M}$ is the solution to the problem. In this case, Q_1 is the solution point.

3. Applications

Next, we analyze three applications of the presented method. The first corresponds to a standard regression problem. Several model-free missing data techniques will be compared with the one proposed in this work. The second is a physically based example illustrating how the method can be seen as a physically-based mean generalization, including constraints based on the problem discretization. Finally, the third one illustrates how the described methodology is particularly suitable for general (eventually time-dependent) problems based on a physical frame, where some physical underlying knowledge is specified explicitly by means of specific governing equations but some physical knowledge (such as empirical constitutive equations) is not known.

3.1. Standard data-science problem

Let us consider different concrete material specimens. Each of them is characterized in terms of the mass fraction of their constituents: cement, slag, fly ash, water, superplasticizer, coarse aggregate and fine aggregate (in $\mathrm{kg/m^3}$). For each sample, the compression strength at the 28th day is tested. Assuming a linear relationship between the compression strength, that is the response variable Y, and the water content, that is the explanatory variable X, we set-up a linear regression model, $Y = aX + b$. The goal is to obtain an estimate of the strength for $X = 100$. This can be easily obtained using the standard least squares technique.

One the full data analysis is performed, we define the following data loss process from the complete data-set, depending on a threshold parameter $0 \leq p \leq 1$

- For data having a water content lower than the $1 - p/2$ quantile and higher than the $p/2$, the water content is removed. This represents a loss of the $100p\%$ of the data due to, for example, experimental difficulties for characterizing high and low water contents.

- For data having a cement content higher than the $1 - p$ quantile, the strength is removed. This represents a loss of the $100p\%$ of the data due to, for example, loss of the data for a given batch of experimental trials.

Note that the described loss process is MNAR so that we are in a context where the filling data method should be fine enough to not include bias and then, error in the predicted value.

As the presented method is non-parametric, it is compared to other non-parametric standard methods: Listwise Deletion, and four interpolation techniques (linear interpolation, nearest point interpolation, piece-wise cubic spline interpolation and shape-preserving piecewise cubic interpolation). The error of the method is defined as

$$\epsilon = \frac{|Y - Y_c|}{Y_c} \tag{26}$$

where Y is the prediction of the incomplete data, following the filling data procedure described before and performing linear regression as if it was the complete data-set and Y_c is the target value.

In that case, incomplete measurements are the canonical manifolds defined as follows. If $\mathbf{X}$ is the matrix of data where each row represents a specimen and each column a variable (cement, slag, fly ash, water, superplasticizer, coarse aggregate and fine aggregate content) a missing data value is described by some specimen i where the j field value is lost. We may then define a missing value matrix $\mathbf{M}$ where $M_{ij} = 1$ if the data at $i - j$ slot is missed, and $M_{ij} = 0$ otherwise. Suppose that we have N specimens, $N - K$ of them have the $n = 7$ values fully reported while the rest K have incomplete data vectors. For each of the $I = 1, \cdots, K$, incomplete data, the missing value matrix is completed such that $i = I$ and for each row, some j values are removed, so $M_{ij} = 1$. Let us suppose we have for the first incomplete vector $I = 1$, $i = 24$ and $j = 1, 4, 6$. This incomplete data point is then associated with the manifold that may be

described using a parametric equation:

$$\mathcal{M}_1 = \{(\lambda, X_{24,2}, X_{24,3}, \mu, X_{24,5}, \nu, X_{24,7}) | (\lambda, \mu, \nu) \in \mathbb{R}^3\} \tag{27}$$

Note that our method could define missing values in a much more sophisticated framework (oblique linear manifolds or even nonlinear manifolds, where we know a relationship between some variables but not the variable itself) using the general expression.

$$\mathcal{M}_I = \{(x_1, x_2, x_3, x_4, x_5, x_6, x_7) | \Phi^I(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = 0\} \tag{28}$$

Since we are interested in the estimate of Y for $X = 100$ kg/m^3, we introduce here the *physical* or *target* manifold:

$$\begin{aligned} \mathcal{M} &= \{(\mu_1, \mu_2, \mu_3, 100, \mu_4, \mu_5, \mu_6) | (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) \in \mathbb{R}^6\} \\ &= \{(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \in \mathbb{R}^7 | x_4 = 100\} \end{aligned} \tag{29}$$

The presented methodology is then used with the manifolds $\mathcal{M}^I$, $I = 1, \cdots, K$ and $\mathcal{M}$ defined by the equations (27) and (29). The RBF function $\phi(u) = \exp\left(-\frac{u^2}{2\varsigma^2}\right)$ with $\varsigma = 20$ kg/m^3 is selected as weighting function.

The number of incomplete data, K, is dependent on the parameter p. For the present example, we deal with $N = 103$ specimens and our method is compared to the other ones for different values of p. The results are shown in Figure 6 in terms of the fraction of missing data with respect to the complete data-set $F = \frac{K}{N}$.

It is clear from Figure 6 that the presented methodology yields better results than the rest of standard filling methods. This is due to the fact that the bias induced by the missing process is here corrected since the filling procedure takes into account how far the data points used are from the target manifold. This local counterpart of the presented methodology makes the method more robust with respect to the missing data fraction in comparison with other interpolation

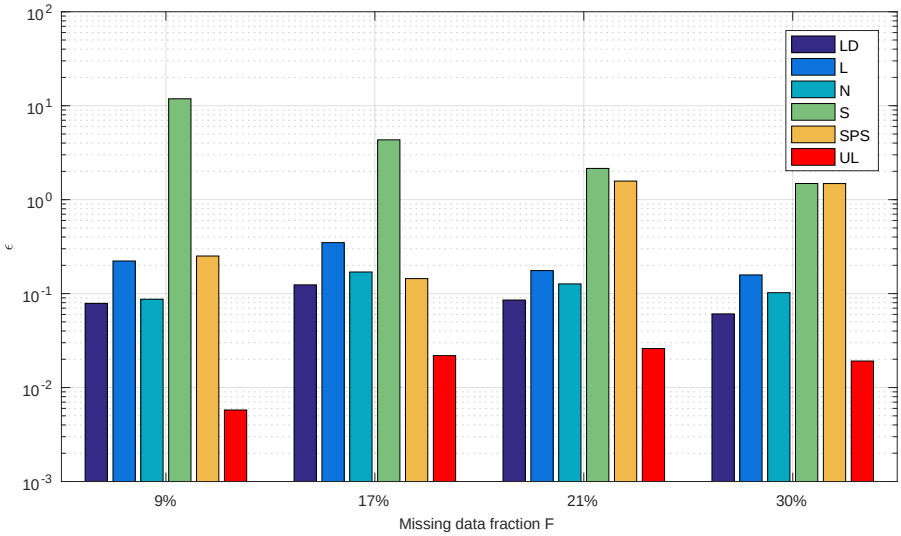


Figure 6: Error of the different model-free filling data procedures for the estimation of Y for $X = 100$ and different missing data fractions F (LD: Listwise deletion, L: Linear interpolation, N: Nearest neighbor interpolation, S: Piecewise cubic spline interpolation, SPS: Shape preserving cubic spline interpolation, UL: Unsupervised learning).

methods. Note that the error of the presented method (UL) does not increase with the amount of lost data. This is due to the fact that the performance of the method depends on how far are the missed data points and not on their number, i.e. data quality and not data quantity. On the other hand, standard interpolation techniques are rather dependent on the distance of the missing data from the true solution than in the volume of missing data itself.

Moreover, low order interpolation techniques (nearest or linear interpolation) are sometimes unable to reproduce the underlying data structure, whereas high order interpolation techniques (cubic splines) performance is strongly dependent on data sampling [44]. Shape preserving interpolation, for example, was conceived as a compromise solution to these problems, but is still strongly dependent on the missing data process as has been demonstrated [45]. Listwise deletion is the most robust method with respect to missing data fraction, but has statistical power and the bias problems for MNAR data as reported in literature [27], [23]. The proposed method shows a more robust behavior with respect to the missing data fraction.

3.2. Model-based data-driven problem

The performance of the method is now illustrated in a classical problem of strength of materials. Let us consider a two-end clamped beam of length L under bending by a linearly distributed load $q = q(x)$, $0 \leq x \leq L$. In the Euler-Bernoulli framework, and supposing the beam composed of a linear elastic material with Young modulus $E = E(x)$, $0 \leq x \leq L$, and a section with moment of inertia $I = I(x)$, the vertical beam displacement $u = u(x)$, $0 \leq x \leq L$, may be computed solving the linear differential equation:

$$\frac{d^2}{dx^2}\left(EI\frac{d^2u}{dx^2}\right) = q \tag{30}$$

with boundary conditions:

$$\begin{aligned} u(0) &= \left.\frac{du}{dx}\right|_{x=0} = 0 \\ u(L) &= \left.\frac{du}{dx}\right|_{x=L} = 0 \end{aligned} \tag{31}$$

Equation (30) with boundary conditions (31) may be solved numerically using any standard numerical procedure (e.g. Finite Elements or Finite Differences). Once the problem is discretized using a mesh of characteristic size h and the boundary conditions are applied, the nodal displacements $\mathbf{u}^h$ are obtained solving the linear system (it is important to note that this equation is characteristic of a broad family of linear discretized problems in Physics and Engineering, not only the Euler-Bernoulli bending beam):

$$\mathbf{K}^h\mathbf{u}^h = \mathbf{f}^h \tag{32}$$

In order to test our method, we may proceed by considering the equation (32) as a physical constraint to a data-set of measurements $\mathcal{E} = \{\mathbf{u}_i^{h'}, i = 1, \cdots, N\}$ equally spaced h'. Note that this approach makes sense when $h' << h$ (this may be the case when equation (32) is computationally expensive to solve with

very fine meshes while measurements are easy to obtain). In this case, we deal with complete measurements, but they are subjected to error (bias and noise for example due to experimental reasons). The presented method is able to detect how far a given measure is from the physics of the problem in terms of the distance to the manifold defined by equation (32). The different measurements will be weighted differently depending on their distance to the manifold. Recall that a standard procedure of averaging all measurements may induce an error if there is a systematic bias in the measurements, which is a well-known problem of mean imputation [25]. The weighting strategy considered associates the bias in the estimation to data quality in a physical sense.

In order to illustrate the application of the methodology, let us solve the defined problem for $q(x) = 10$ kN/m, $L = 10$ m, $E(x)I(x) = 1 \cdot 10^6$ kN $\cdot$ m^2. In that case, the analytical solution is given by $u(x) = -\frac{qL^4}{24EI}\left(\frac{x}{L} - \frac{x^2}{L^2}\right)^2$.

The measurements are randomly generated from the analytical solution sampled in a mesh of size $h' = \frac{1}{m-1}$, $u^k \sim \mathcal{N}(u(x = k\frac{L}{h'}) + b, \sigma)$, where $b = \alpha\bar{u}$ is a bias, $\alpha \sim \mathcal{N}(0.03, 0.03)$, $\sigma = \beta\bar{u}$, $\beta = 0.002$ and $\bar{u} = \frac{1}{30}\frac{qL^4}{EI}$ is the mean value of the analytical solution. For this example, $m = 100$ points are considered along the beam length and $N = 6$ samples are evaluated. The six families of points are shown in Figure 7. Also, and in the same figure, the results computed for different spacings h used to establish the physical constraint are shown. In particular, the mean of the samples is represented by the bold orange color line and the true analytical solution is represented by the continuous bold blue line. The RBF function $\phi(u) = \exp\left(-\frac{u^2}{2\varsigma^2}\right)$ with $\varsigma = 5 \cdot 10^{-4}$ m was selected as the weighting function.

As pointed out above, the method here described corrects partially the bias of the measurements. This correction is done automatically by computing the distance of each measure sample to the physical manifold defined by the discretized equilibrium equation and boundary conditions, and transforming this distance using the RBF function. A finer mesh in the discretized physical problem (lower h) takes into account more points to compute the distance from the measure to the physics of the problem, while this physics is more accurate. It

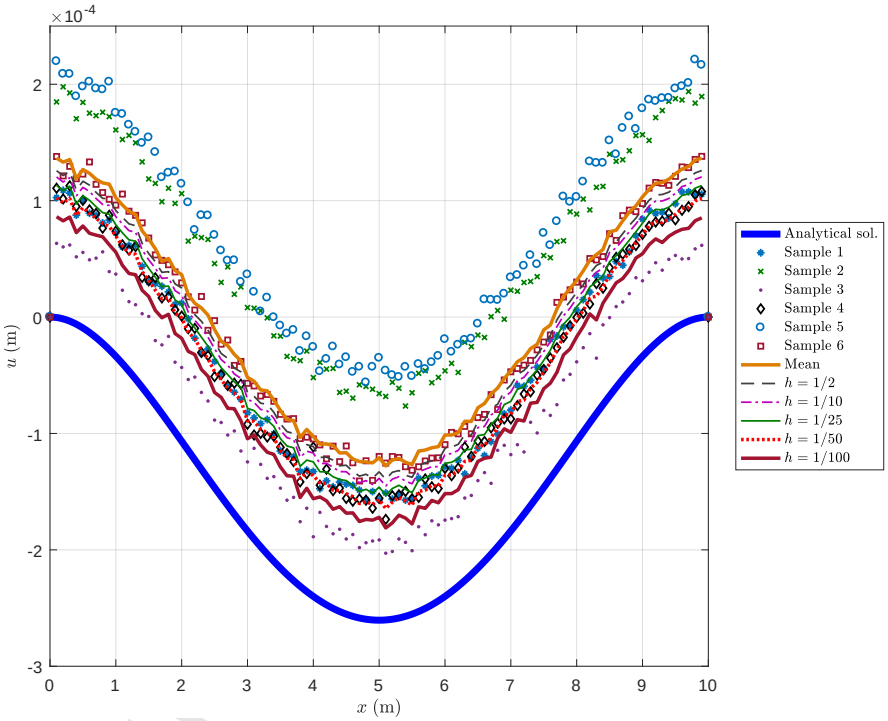


Figure 7: Comparison of the solutions obtained using the presented method for different mesh sizes h to define the physical constraint.

is clear, however, that the accuracy of the estimation in cases as this one with a systematic positive bias will never be better than the best measure.

In order to have a deeper understanding of the method, it is worth to make a physical interpretation of this example. For a fixed h, equation (32) is a constraint relating $n = 1/h + 1$ variables of the m-dimensional space. The manifold $\mathcal{P}^h \simeq \mathbb{R}^n$ of dimension $n \ll m$ defined by these n coordinates is the manifold where the relevant physics of the problem is evaluated. It is easy to figure up what the method does by considering only the projections of the points in $\mathbb{R}^m$ on $\mathcal{P}^h$. In $\mathcal{P}^h$, the physical manifold (that is, the constraint) is given by a single point p: the numerical solution obtained solving the equation (32). Therefore, the distances of the different samples to the physical manifold may be interpreted as Euclidean distances in $\mathcal{P}^h \simeq \mathbb{R}^n$ between sample projections on $\mathcal{P}^h$ (i.e. the consideration of the n coordinates related to the mesh with h spacing) and point P, $d_i = ||\pi_{\mathcal{P}^h}(\mathbf{u}_i) - \mathbf{u}^h||$. We could have defined the distance in a more general framework, such as Hilbert spaces, using the Finite Element approximation, but a simpler and more interpretable norm was selected for illustration purposes. The weighted mean is then computed by using the solution $\mathbf{u}^h$ as the reference point in the weights computations, $w_i = \phi(d_i)$. The physics of the problem is inferred outside the manifold $\mathcal{P}^h$, that is, in $(\mathcal{P}^h)^\perp$. In Figure 8 all described geometric elements are shown for $h = 1/10$: projections over manifold $\mathcal{P}^h$ of m-dimensional samples, numerical solution $\mathbf{u}^h$ and associated generalized mean and variance. The analytical solution is also plotted even though it does not appear in the computation.

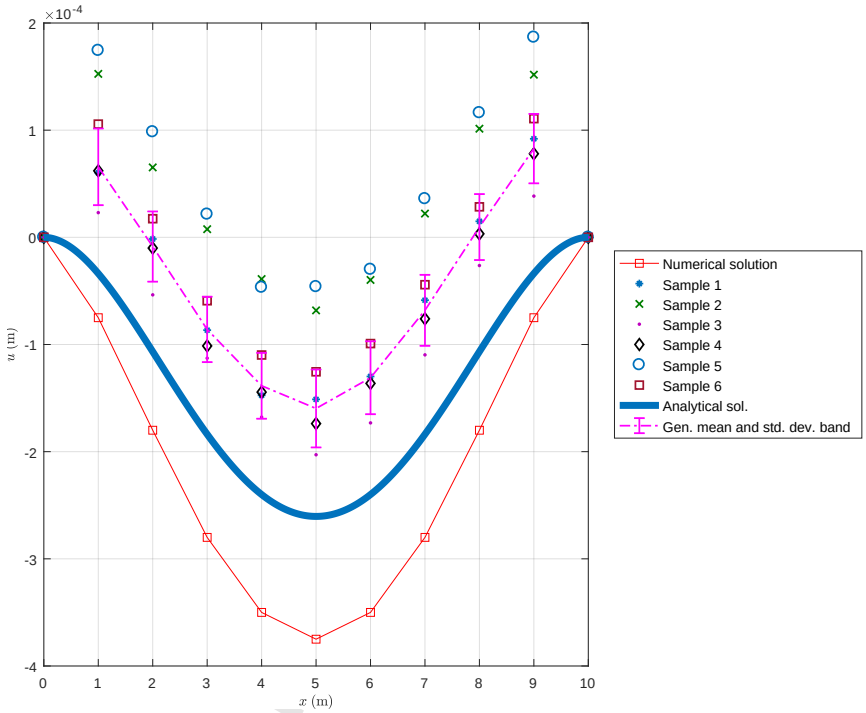


Figure 8: Projections over the manifold $\mathcal{P}^h$ of all geometric elements used in computations for $h = 1/10$. The dotted magenta line represents the generalized mean (in this case it is equivalent to the weighted mean with complete measurements). The error bands correspond to the projections over the different coordinates lines (manifolds of dimension 1) of the generalized variance ellipsoid (in orange in Figure 1 or Figure 3). Green ellipsoids represented in these figures collapse in the data samples points because samples are complete so the measure manifolds are points.

3.3. Model-free data-driven problem

The final application lays within the framework in which the method has been conceived: model-free data-driven problems. The objective is now to solve a real physically-based problem, formulated in terms of a set of governing equations encoding the physical information, and some empirical knowledge, formulated in terms of a given data-set.

The main objective of this example, despite its simplicity, is to illustrate the performance and some of the properties and capabilities of the technique proposed in DDSBES rather than comparing it with other filling data methods, as in the previous examples. With this aim, the problem is first formulated in a classical framework approach, where no data-set is considered and the entire physics of the problem is supposed to be known and parametrized. Secondly, the problem is reformulated in the data-driven framework, where only the sound physics is postulated and the rest of the physical structure is built from data.

Let us suppose two reservoirs connected by a channel with section S, hydraulic diameter D and length L as illustrated in Figure 9. A fluid flows from one reservoir to the other depending on the water level in each of them, y_1 for the reservoir 1 and y_2 for the reservoir 2. The section of each reservoir is defined for each height y_i by a function $S_i = S_i(y_i)$, $i = 1, 2$, being clear that the volume occupied by the fluid at reservoir i for a height y_i is $V_i(y_i) = \int_0^{y_i} S_i(u)\, du$.

When considering the physics of the problem, two sound laws are invoked: conservation of mass (1) and conservation of energy (2). The first one is equivalent to impose (under the assumption of fluid incompressibility) zero net flow, $Q_{net} = 0$. That writes:

$$\frac{dV_1}{dt} + \frac{dV_2}{dt} = 0 \tag{33}$$

Using the expression of V_i in terms of y_i we get:

$$S_1(y_1)\dot{y_1} + S_2(y_2)\dot{y_2} = 0 \tag{34}$$

The second considers the energy loss due to viscous dissipation, unless we

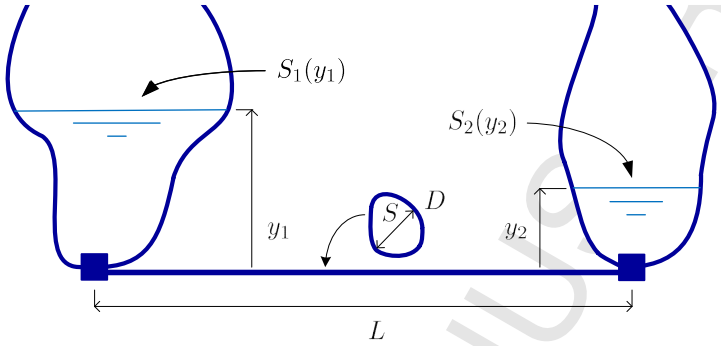


Figure 9: Schematic diagram of the fluidic device used to validate the presented methodology.

consider an inviscid fluid. Using Bernoulli energy conservation statement in terms of water height, conservation of energy writes

$$y_1 - y_2 = IL \tag{35}$$

where I is the hydraulic head loss slope, $I = I(Q) = f_D \frac{1}{2g} \frac{Q^2}{S^2 D}$ and $Q = S_1(y_1)\dot{y_1}$. In order to close the system of equations (34) and (35), it is necessary to define the Darcy friction factor f_D or at least to express it in terms of the state space variables y_1, y_2, $\dot{y_1}$ and $\dot{y_2}$, which is the critical step in classical approaches. In general, either one additional hypothesis is assumed (for example laminar regime), carry out simulations with complex fluid flow models or use a semi-empirical equation such as Kármán-Prandtl resistance equation for the smooth turbulent regime, [46], [47], the well-known Colebrook-White equation [48] or other more recent equations for the transition from a smooth pipe to a rough pipe flow [49], [50], [51]. One way or another, these approaches complete the physics based on the particular hypothesis stated a priori.

Assuming a laminar regime, $f_D = \frac{64}{\mathrm{Re}} = \frac{64 S \nu}{QD}$, with ν the kinematic viscosity and defining the initials conditions, the problem may be solved numerically integrating the system of equations. For instance, we may fix $y_1(t=0) = H$, $y_2(t=0) = 0$, $\dot{y_1}(t=0) = 0$ and $\dot{y_2}(t=0) = 0$ (reservoir 1 at level H and

reservoir 2 empty, both at rest).

The data-driven approach is based on the use of a data-set that will complete implicitly the physics of the problem *a posteriori*. Using this approach, only the conservation of mass (equation (34)) is taken into account while the energy equation is replaced by a data-set sampled from the state-space $\mathcal{M} = \{(y_1, y_2, \dot{y_1}, \dot{y_2}) |\, S_1(y_1)\dot{y_1} + S_2(y_2)\dot{y_2} = 0\}$. We have therefore a data-set $\mathcal{S} = \{(y_1^i, y_2^i, \dot{y_1}^i, \dot{y_2}^i) |\, i = 1, \cdots, N\}$. As we deal again with a missing data problem, we define the missing process as follows: due to experimental limitations, it is impossible to measure the velocity of the free surface level when the free surface is higher than a defined threshold h^*. This condition tries to reproduce a realistic missing data process related to the experimental setup. We are clearly in a MNAR situation. The presented method, which is local in a certain sense due to the weighting process, should be insensitive to the missing process.

In order to illustrate the methodology, let us solve the problem in a particular case using both the classical approach assuming laminar regime and the data-driven framework, using as data-set the solution in the laminar regime with noise: for each variable $x_{\text{noise}} = x_{\text{exact}} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, p\sigma)$, being $p = 1/10$ and σ the standard deviation of the exact values x_{exact}. A truncated cone geometry is assumed for both reservoirs with lower radii $r = 1.33$ mm and slope $s = 0.035$. The density and dynamic viscosity of water are $\rho = 1000$ kg/m^3 and $\mu = 1.006 \cdot 10^{-3}$ Pa $\cdot$ s, $\nu = \frac{\mu}{\rho}$ and $g = 9.81$ m/s^2, respectively. The initial level of the fluid is $H = 5$ cm and the channel has length $L = 10$ cm and a rectangular section of width $w = 750$ μm and height $h = 200$ μm. The laminar solution in terms of the two heights and the flow is shown in Figure 10.

Our filling data strategy is tested by solving the data-driven problem with a loss fraction of data defined by fixing the threshold h^* as described before. The components y_1 and y_2 of the complete data-set are shown in Figure 11a. Our aim is solving the problem when we observe a value of $y_1 = m$, that is, to obtain the other state variables y_2, $\dot{y_1}$ and $\dot{y_2}$. In other words, to find the closest point to our incomplete data-set satisfying mass conservation (physical manifold $\mathcal{M}$) and using as reference our initial set of observations (observation

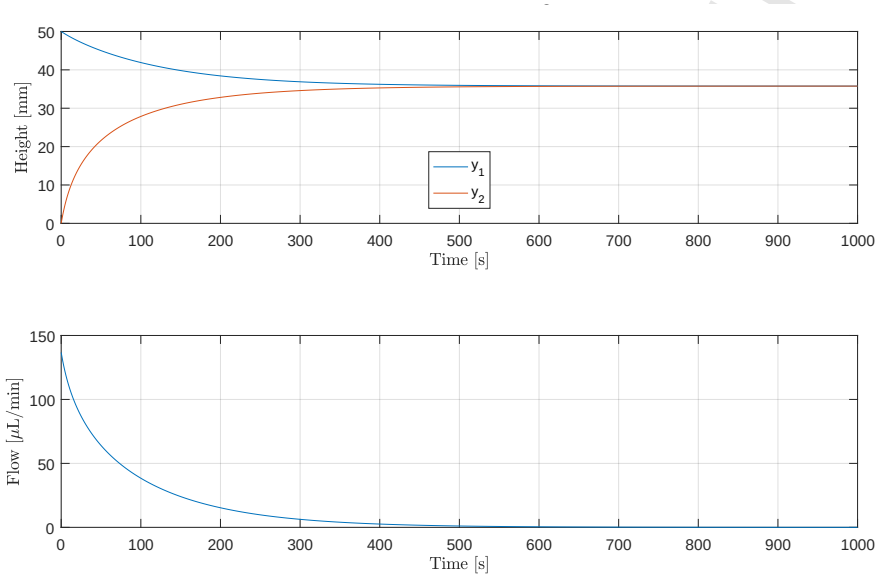
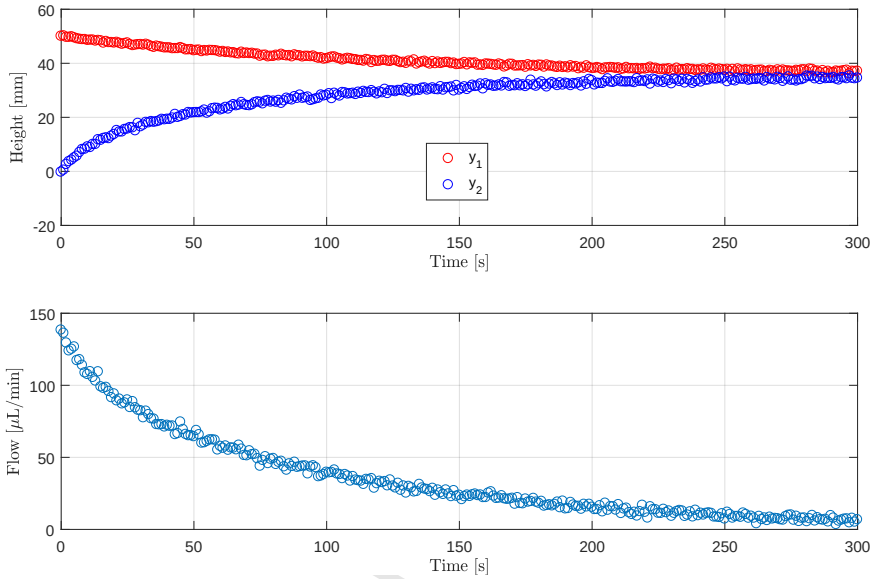


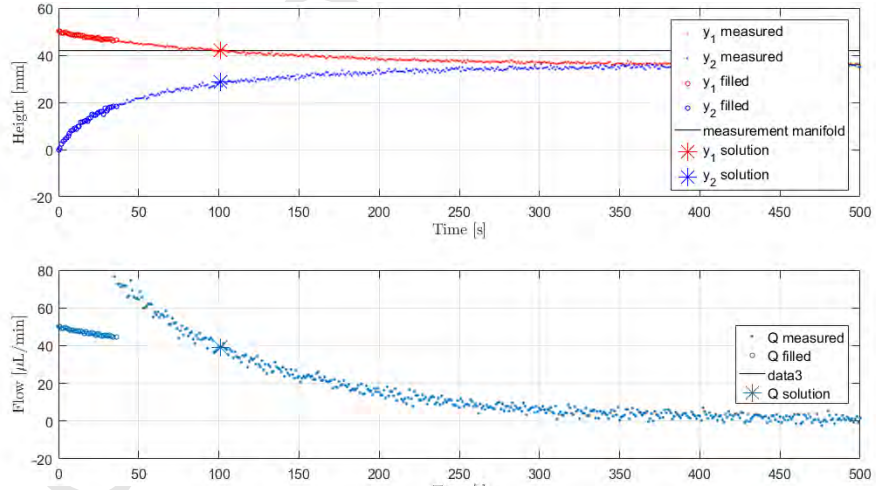
Figure 10: Classical solution of the problem solving the system of ordinary differential equations.

manifold $\mathcal{N}$). The manifold $\mathcal{M}$ is defined by the mass conservation equation (34) while the manifold $\mathcal{N}$ is defined by equation $y_1 - m = 0$. Incomplete data are filled using the methodology presented in this work. Completed data are shown in Figure 11b for $m = 42\,\text{mm}$ and $h^* = 1.1 \cdot m$. As stated before, an RBF function is chosen as weighting function with $\varsigma = 0.2\,\text{mm}$ while, for illustration purposes, weights are computed only in terms of the distance to the manifold $\mathcal{N}$ instead of $\mathcal{N} \cap \mathcal{M}$, which is nonlinear. Finally, the solution point is chosen by solving the data-driven minimization problem suggested in [16, 17], that is, by minimizing the distance to the (filled) data-set, provided the physical constraint. The difference here is the fact that a stochastic distance is computed, including the deterministic term ($T_{i,1}$ and $T_{i,2}$, related to the generalized mean) and the quadratic uncertainty term ($T_{i,3}$, related to the generalized variance), both resulting from the filling procedure, as it is described in section 2.4.2.

Figure 11b depicts the solution of the problem in terms of intuitive plots or state variables but does not illustrate the geometry behind. In order to



(a) Data samples corresponding to state-space variables y_1 and y_2.



(b) Completion of data using the described filling data methodology

Figure 11: Original complete data-set and data-set constructed from incomplete data using unsupervised learning.

illustrate the geometric idea of the method, Figure 12 represents the same as Figure 11b but including the complete measurements, the physical manifold and the solution point in a three dimensional projection of the state space. Note that the physical manifold is $\mathcal{P} = \{(y_1, y_2, \dot{y_1}, \dot{y_2}) \in \mathbb{R}^4 | \, S_1(y_1)\dot{y_1} + S_2(y_2)\dot{y_2} = 0, y_1 = m\}$ that lives in a four-dimensional space, whose projection in the three-dimensional space $\dot{y}_1, \dot{y}_2$ and y_2 is shown. The fourth dimension is illustrated using colors representing the distance to the manifold $\mathcal{N}$. It is observed now that the complete measurements own a more complex geometry in the state space, being therefore this figure a clearer representation of the geometry of the problem.

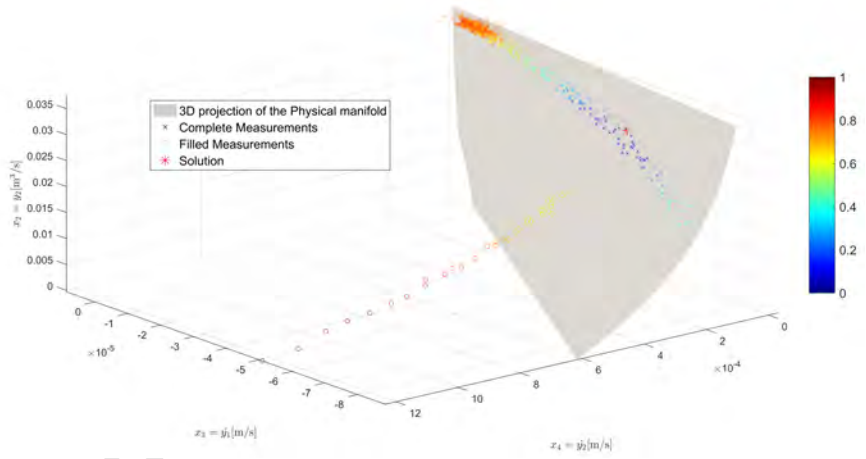


Figure 12: Representation of different geometric elements in the space $(\dot{y_1}, \dot{y_2}, y_2) \in \mathbb{R}^3$. Colors represent the normalized distance (between 0 and maximum) of each of the plotted points to the linear manifold $\mathcal{N}$ defined by $y_1 = m$, that is not represented in this 3D representation. As the weights have been computed using distances to the manifold $\mathcal{N}$, all completed points are close to $\mathcal{N}$ even if they do not belong to the manifold $\mathcal{M}$. The best measure should be then a point close (in the described stochastic sense) to the surface represented in this 3D plot and to the manifold $\mathcal{N}$. The solution point is then computed by projecting the closest measure into the physical manifold or in the nonlinear case, looking for the point of the physical manifold closest to this measure.

The accuracy of the solution obtained with some fraction of data with respect

to the solution using the complete data is again evaluated by means of the error defined in (26). Figure 13 shows the accuracy of the solution with respect to the missing data fraction, that depends on the selected threshold h^*, indicated on the figure. Note that the missing data fraction has only a statistical sense, but the selected threshold has a physical meaning that can be related to the problem. As shown in 13, the accuracy of the method depends again primarily on h^*, that is, on how far the missing data are from the observation manifold than in the amount of missing data. Actually, when $h^* < m$, the error of the method increases: it is clear that there is no data sample close to the observation manifold or the physical manifold, neither the presented method nor any other in a data-driven context could reconstruct the data structure at this region. The presented method, however, link the data-driven solution to the physical insights concerning the missing process, according to considerations pointed out in [38]. In the same figure, in the dashed blue line, the error considering the complete noisy data with respect to the laminar solution is highlighted. The accuracy of the presented method is of the order of this error, revealing that for $h^* \geq m$, the error of our method is not a consequence of the filling process but of the data-driven nature.

4. Conclusions

In this work, a new completion data method has been established, adapted to the data-driven simulation-based engineering and sciences framework. The method can be seen as a generalization of the classical mean imputation. Indeed, the presented method works when each of the data points is constrained to an oblique or even nonlinear manifold, whereas the mean imputation considers the points in the canonical coordinate manifolds. The presented method is based on the definition of a generalized weighted mean as the solution of the constrained minimization problem:

$$\min_{x \in \mathcal{M}} \quad \sum_{j=1}^{N} w(\mathcal{M}_j) d^2(x, \mathcal{M}_j) \tag{36}$$

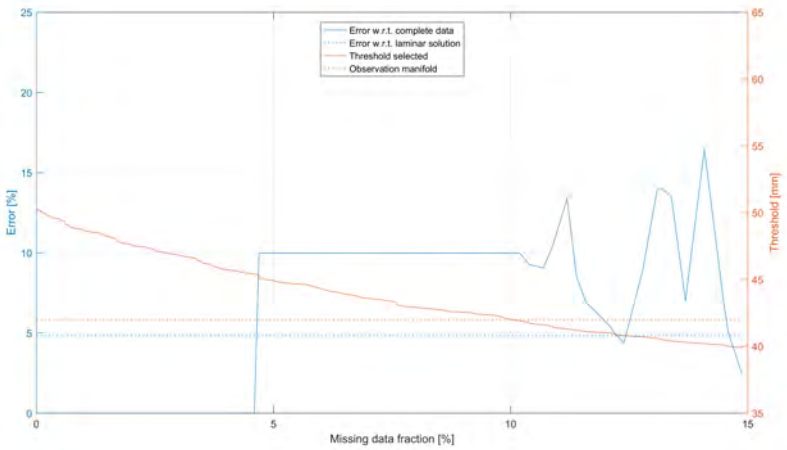


Figure 13: Evaluation of the presented method for different missing data fractions and thresholds.

Here, $\mathcal{M}$ is the manifold of points fulfilling a certain underlying (physical) structure and $\mathcal{M}_j$, $j = 1, \cdots, N$ represent the incomplete data-sets. When $\mathcal{M} = \mathbb{R}^n$ and $\mathcal{M}_j = \{p_j\}$, that is, points, we recover the usual definition of the mean. With this generalization, the mean can be defined in non-Euclidean frameworks and the imputation of the missing values can be consistent with any structure of the problem of interest to maintain.

However, the presented method is more than a simple imputation method. The abstract framework in which it is formulated facilitates its use in combination with a data-driven approach for the resolution of simulation-based problems when the data used to feed the data-driven algorithm is incomplete. The physical interpretation of the data in terms of state variables belonging to a given local structure (physical manifold) is compatible with the generalized defined imputation. Moreover, if some global physical conditions must be fulfilled, that is, state variables are embedded in a more manageable space and/or the data involve uncertainty, a weighting strategy is proposed in order to take all these considerations into account.

It has been shown that the presented imputation method, though it is used

in the usual framework (with the canonical coordinate manifolds), improves the classical imputation approaches when the desired prediction can be stated in a framework including some constraints. The first example, of a pure data-science nature, illustrates how the weighting strategy can be used to quantify the admissibility of a point in the imputation method.

The second example illustrates how the presented methodology incorporates the physics of the problem to computations. It is pointed out that the method may be used for any physical problem where some fundamental physical constraints are invoked in combination with experimental measurements. Here, the method provides an alternative to highly demanding computational solutions based on numerical procedures, when experimental measurements can be easily obtained. The presented algorithm takes into account the physical quality of the data and, therefore, is more robust in problems with experimental bias.

The last example, much richer, illustrates the full power of the presented method. Here, all features of the methodology are taken into account for the solution of a model-free approach to a typical engineering problem involving fluid mechanics where the formulation of the problem should guarantee the fulfillment of some fundamental physical laws (mass conservation) and the operation condition. The results show a good agreement with a model-based approach to the problem (using a laminar assumption for the flow) and demonstrate that the accuracy does not depend on the missing data fraction, but on how far the missing data are from the operating conditions. This fact links the performance of the model to physical considerations, more than statistical ones (that is, in data quality more than in data quantity), as it would be desirable in data-driven engineering problems.

The filling data methodology presented is conceptually simple, because it is based on the minimization problem (36). However, this constrained minimization problem is in general nonlinear and not always smooth. In the present work, a computational expression for the solution of (36) is derived in the linear case and an iterative algorithm is presented for nonlinear problems, based on tangent linearization, the application of the linear solution and a standard

strategy for returning to the manifold. No mathematical results are presented in this work about the convergence of the presented algorithm, which is crucial, and depends on the geometry of the problem, particularly the convexity of the data manifolds. Fortunately, the existence of extensive software for solving constrained optimization problems can save this inconvenience in many problems, but the selected solution method would be, also, context dependent.

To conclude, the presented imputation method is a starting point for a new domain in Engineering, which responds to the need of data-driven simulation-based engineering and sciences, that is, the adaptation of the classical statistical tools to engineering problems where some physics defines the geometric structure of the problem and has to be fulfilled. This domain, that could be named as data-driven simulation-based statistics is no more than the meeting point between Mathematical Physics, whose mathematical language is differential geometry and Statistics, whose mathematical language is measure theory.

Acknowledgements

The authors gratefully acknowledge the financial support from the Spanish Ministry of Economy and Competitiveness (MINECO MAT2016-76039-C4-4-R, AEI/FEDER, UE), the Government of Aragon (DGA-T24_17R) and the Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN). CIBER-BBN is financed by the Instituto de Salud Carlos III with assistance from the European Regional Development Fund.

Appendix A. Mathematical proofs

Proposition Appendix A.1 (Expected value of the squared distance). *Let $\mathcal{V}$ a linear manifold and let $\pi_{\mathcal{V}}$ the orthogonal projection on $\mathcal{V}$, $P \in \mathbb{R}^n$, $X \in \mathcal{V}$ a random vector and $D = d(P, X)$. We identify X with random vector $\mathbf{X}$ and point P with its coordinates $\mathbf{p}$ in a given reference frame. Let us suppose that $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ and $\boldsymbol{\Sigma} = \mathrm{COV}(\mathbf{X})$ are finite. Therefore:*

$$\mathbb{E}[D^2] = ||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + ||\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu}||^2 + \mathrm{Tr}(\boldsymbol{\Sigma}) \qquad \text{(Appendix A.1)}$$

Proof:. Using Pythagoras theorem we have:

$$\begin{aligned} D^2 &= ||\mathbf{p} - \mathbf{X}||^2 \\ &= ||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + ||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2 \end{aligned} \qquad \text{(Appendix A.2)}$$

Therefore

$$\begin{aligned} \mathbb{E}[D^2] &= \mathbb{E}\left[||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + ||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right] \\ &= ||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + \mathbb{E}\left[||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right] \end{aligned} \qquad \text{(Appendix A.3)}$$

Now, we have

$$\begin{aligned} \mathbb{E}\left[||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right] &= \mathrm{Tr}(\boldsymbol{\Sigma}) + ||\mathbb{E}[\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}]||^2 \\ &= \mathrm{Tr}(\boldsymbol{\Sigma}) + ||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbb{E}[\mathbf{X}]||^2 \end{aligned} \qquad \text{(Appendix A.4)}$$

Combining equations Appendix A.3 and Appendix A.4 we obtain the result. □

We may observe that when using a reference frame in $\mathcal{V}$, $\mathrm{rang}(\boldsymbol{\Sigma}) = \dim(\mathcal{V})$ and then $\mathrm{Tr}(\boldsymbol{\Sigma})$ has as many terms as the dimension of $\mathcal{V}$.

An analogous result for the variance of D^2 can be derived.

Proposition Appendix A.2 (Variance of the squared distance). *Let $\mathcal{V}$ a linear manifold and let $\pi_{\mathcal{V}}$ the orthogonal projection on $\mathcal{V}$, $P \in \mathbb{R}^n$, $X \in \mathcal{V}$ a random vector and $D = d(P, X)$. We identify X with random vector $\mathbf{X}$ and point P with its coordinates $\mathbf{p}$ in a given reference frame. Let us suppose that $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$, $\boldsymbol{\Sigma} = \mathrm{COV}(\mathbf{X})$ and $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}(\mathbf{X})$ the centered tensor moment of order 4 of $\mathbf{X}$ are finite. Therefore:*

$$\mathrm{Var}[D^2] = \mathbb{I} : \boldsymbol{\Upsilon} : \mathbb{I} - \left[\mathrm{Tr}(\boldsymbol{\Sigma}) + (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})^T (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})\right] \quad \text{(Appendix A.5)}$$

Proof:. We have seen in previous proof that:

$$D^2 = ||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + ||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2 \quad \text{(Appendix A.6)}$$

Therefore:

$$\begin{aligned} \mathrm{Var}[D^2] &= \mathrm{Var}\left(||\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})||^2 + ||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right) \\ &= \mathrm{Var}\left(||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right) \end{aligned} \quad \text{(Appendix A.7)}$$

But we have,

$$\begin{aligned} \mathrm{Var}\left(||\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}||^2\right) &= \mathbb{I} : \boldsymbol{\Upsilon} : \mathbb{I} - \left(\mathrm{Tr}(\boldsymbol{\Sigma}) + \mathbb{E}[\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}]^T \mathbb{E}[\pi_{\mathcal{V}}(\mathbf{p}) - \mathbf{X}]\right) \\ &= \mathbb{I} : \boldsymbol{\Upsilon} : \mathbb{I} - \left(\mathrm{Tr}(\boldsymbol{\Sigma}) + (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})^T (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})\right) \end{aligned} \quad \text{(Appendix A.8)}$$

Combining equations Appendix A.7 and Appendix A.8 we obtain the result.

□

Under normality conditions, we have the following result:

Proposition Appendix A.3 (Squared distance distributional properties). *Let $\mathcal{V}$ a linear manifold and let $\pi_{\mathcal{V}}$ the orthogonal projection on $\mathcal{V}$, $P \in \mathbb{R}^n$, $X \in \mathcal{V}$ a random vector and $D = d(P, X)$. We identify X with random vector*

$\mathbf{X}$ *and point* P *with its coordinates* $\mathbf{p}$ *in a given reference frame. Let us assume that* $\mathbf{X}$ *follows a multivariate normal distribution with mean* $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ *and covariance matrix* $\boldsymbol{\Sigma} = \mathrm{COV}(\mathbf{X})$. *Let* $\chi^2 = (\mathbf{p} - \mathbf{X})^T(\boldsymbol{\Sigma})^{-1}(\mathbf{p} - \mathbf{X})$. *Then* χ^2 *follows a non-central* χ^2 *distribution with* $k = \dim(\mathcal{V})$ *degrees of freedom and non-centrality parameter* $\lambda = (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})$.

Proof:. Let $\mathbf{U} = \mathbf{p} - \mathbf{X}$. Therefore, $\mathbf{U}$ follow a multivariate normal distribution with mean $\mathbf{p}-\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, $\boldsymbol{\Sigma}^{-1/2}\mathbf{U} \sim \mathcal{N}(\boldsymbol{\Sigma}^{-1/2}(\mathbf{p}-\boldsymbol{\mu}), \mathbb{I})$. Using non-central χ^2 distribution definition:

$$\chi^2 = \mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U} = (\boldsymbol{\Sigma}^{-1/2}\mathbf{U})^T(\boldsymbol{\Sigma}^{-1/2}\mathbf{U}) \sim \chi^2(n, \lambda) \qquad \text{(Appendix A.9)}$$

Where $n = \mathrm{rang}(\boldsymbol{\Sigma}) = \dim(\mathcal{V})$ and $\lambda = (\boldsymbol{\Sigma}^{-1/2}(\mathbf{p} - \boldsymbol{\mu}))^T(\boldsymbol{\Sigma}^{-1/2}(\mathbf{p} - \boldsymbol{\mu})) = (\mathbf{p} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{p} - \boldsymbol{\mu}))$.

But we have, $\mathbf{p} - \boldsymbol{\mu} = (\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p})) + (\pi_{\mathcal{V}}(\mathbf{p})) - \boldsymbol{\mu})$. As $\mathbf{p} - \pi_{\mathcal{V}}(\mathbf{p}) \in \ker(\boldsymbol{\Sigma}^{-1})$,then $(\mathbf{p} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{p} - \boldsymbol{\mu}) = (\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\pi_{\mathcal{V}}(\mathbf{p}) - \boldsymbol{\mu})$. □

References

[1] L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Computer networks 54 (15) (2010) 2787–2805.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, Big data: The next frontier for innovation, competition, and productivity.

[3] D. T. Larose, Discovering knowledge in data: an introduction to data mining, John Wiley & Sons, 2014.

[4] T. M. Mitchell, Machine learning. 1997, Burr Ridge, IL: McGraw Hill 45 (37) (1997) 870–877.

[5] S. Khan, Introduction to machine learning (adaptive computation and machine learning series), Natural Language Engineering 14 (01) (2008) 133–137.

[6] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.

[7] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain., Psychological review 65 (6) (1958) 386.

[8] A. Krenker, J. Bester, A. Kos, Introduction to the artificial neural networks, in: Artificial neural networks-methodological advances and biomedical applications, InTech, 2011.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

[10] F. Darema, Dynamic data driven applications systems: A new paradigm for application simulations and measurements, in: International Conference on Computational Science, Springer, 2004, pp. 662–669.

[11] R. E. Kalman, A new approach to linear filtering and prediction problems, Journal of basic Engineering 82 (1) (1960) 35–45.

[12] B. Peherstorfer, K. Willcox, Dynamic data-driven reduced-order models, Computer Methods in Applied Mechanics and Engineering 291 (2015) 21–41.

[13] R. Ibanez, E. Abisset-Chavanne, J. V. Aguado, D. Gonzalez, E. Cueto, F. Chinesta, A manifold learning approach to data-driven computational elasticity and inelasticity, Archives of Computational Methods in Engineering (2016) 1–11.

[14] P. Ladevèze, The large time increment method for the analysis of structures with non-linear behavior described by internal variables, COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE II 309 (11) (1989) 1095–1099.

[15] T. Kirchdoerfer, M. Ortiz, Data-driven computational mechanics, Computer Methods in Applied Mechanics and Engineering 304 (2016) 81–101.

[16] T. Kirchdoerfer, M. Ortiz, Data driven computing with noisy material data sets, arXiv preprint arXiv:1702.01574.

[17] J. Ayensa-Jiménez, M. H. Doweidar, J. A. Sanz-Herrera, M. Doblaré, A new reliability-based data-driven approach for noisy experimental data with physical constraints, Computer Methods in Applied Mechanics and Engineering 328 (2018) 752–774.

[18] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[19] J. L. Schafer, J. W. Graham, Missing data: our view of the state of the art., Psychological methods 7 (2) (2002) 147.

[20] P. D. Allison, Missing data: Quantitative applications in the social sciences, British Journal of Mathematical and Statistical Psychology 55 (1) (2002) 193–196.

[21] P. D. Allison, Missing data techniques for structural equation modeling., Journal of abnormal psychology 112 (4) (2003) 545.

[22] J. W. Graham, Missing data analysis: Making it work in the real world, Annual review of psychology 60 (2009) 549–576.

[23] M. Nakai, W. Ke, Review of the methods for handling missing data in longitudinal data analysis, International Journal of Mathematical Analysis 5 (1) (2011) 1–13.

[24] C.-Y. J. Peng, M. Harwell, S.-M. Liou, L. H. Ehman, et al., Advances in missing data methods and implications for educational research, Real data analysis 3178.

[25] J. L. Schafer, Analysis of incomplete multivariate data, CRC press, 1997.

[26] P. D. Allison, Missing data: Sage university papers series on quantitative applications in the social sciences (07–136), Thousand Oaks, CA.

[27] A. Briggs, T. Clark, J. Wolstenholme, P. Clarke, Missing.... presumed at random: cost-analysis of incomplete data, Health economics 12 (5) (2003) 377–392.

[28] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmospheric Environment 38 (18) (2004) 2895–2907.

[29] R. J. Little, D. B. Rubin, Bayes and multiple imputation, Statistical Analysis with Missing Data, Second Edition (2002) 200–220.

[30] D. B. Rubin, Multiple imputation after 18+ years, Journal of the American statistical Association 91 (434) (1996) 473–489.

[31] N. Schenker, J. M. Taylor, Partially parametric techniques for multiple imputation, Computational statistics & data analysis 22 (4) (1996) 425–446.

[32] J. L. Schafer, Multiple imputation: a primer, Statistical methods in medical research 8 (1) (1999) 3–15.

[33] H. Hartley, R. Hocking, The analysis of incomplete data, Biometrics (1971) 783–823.

[34] C. K. Enders, A primer on maximum likelihood algorithms available for use with missing data, Structural Equation Modeling 8 (1) (2001) 128–141.

[35] C. K. Enders, D. L. Bandalos, The relative performance of full information maximum likelihood estimation for missing data in structural equation models, Structural equation modeling 8 (3) (2001) 430–457.

[36] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society. Series B (methodological) (1977) 1–38.

[37] C. K. Enders, Applied missing data analysis, Guilford Press, 2010.

[38] B. G. Tabachnick, L. S. Fidell, S. J. Osterlind, Using multivariate statistics.

[39] G. J. Sussman, J. Wisdom, Structure and interpretation of classical mechanics, Mit Press, 2015.

[40] A. Papoulis, Probability & statistics, Vol. 2, Prentice-Hall Englewood Cliffs, 1990.

[41] M. P. Do Carmo, J. Flaherty Francis, Riemannian geometry, Vol. 115, Birkhäuser Boston, 1992.

[42] J. Simo, N. Tarnow, M. Doblare, Non-linear dynamics of three-dimensional rods: Exact energy and momentum conserving algorithms, International Journal for Numerical Methods in Engineering 38 (9) (1995) 1431–1473.

[43] S. Niroomandi, I. Alfaro Ruiz, E. Cueto Prendes, Real-time simulation of surgery by model reduction and x-fem techniques.

[44] D. Kahaner, C. Moler, S. Nash, Numerical methods and software, Englewood Cliffs: Prentice Hall, 1989.

[45] F. N. Fritsch, R. E. Carlson, Monotone piecewise cubic interpolation, SIAM Journal on Numerical Analysis 17 (2) (1980) 238–246.

[46] H. J. Tracy, C. Lester, Resistance coefficients and velocity distribution smooth rectangular channel, US Government Printing Office, 1961.

[47] B. J. McKEON, M. V. ZAGAROLA, A. J. SMITS, A new friction factor relationship for fully developed pipe flow, Journal of Fluid Mechanics 538 (2005) 429443. doi:10.1017/S0022112005005501.

[48] C. F. Colebrook, T. Blench, H. Chatley, E. Essex, J. Finniecome, G. Lacey, J. Williamson, G. Macdonald, Correspondence. turbulent flow in pipes, with particular reference to the transition region between the smooth and rough pipe laws.(includes plates)., Journal of the Institution of Civil engineers 12 (8) (1939) 393–422.

[49] N. Afzal, Friction factor directly from transitional roughness in a turbulent pipe flow, Journal of Fluids Engineering 129 (10) (2007) 1255–1267.

[50] N. Afzal, Erratum:" friction factor directly from transitional roughness in a turbulent pipe flow"[asme trans. j. fluid eng., 2007, 129, pp. 1255-1267], Transactions of the ASME-I-Journal of Fluids Engineering 133 (10) (2011) 107001.

[51] N. Afzal, A. Seena, A. Bushra, Turbulent flow in a machine honed rough pipe for large reynolds numbers: General roughness scaling laws, Journal of hydro-environment research 7 (1) (2013) 81–90.