Design of Computer Experiments using Competing Distances between Set-valued Inputs

David Ginsbourger, Jean Baccou, Clément Chevalier and Frédéric Perales

Abstract In many numerical simulation experiments from natural sciences and engineering, inputs depart from the classical moderate-dimensional vector set-up and include more complex objects such as parameter fields or maps. In this case, and when inputs are generated using stochastic methods or taken from a pre-existing large set of candidates, one often needs to choose a subset of "representative" elements because of practical restrictions. Here we tackle the design of experiments based on distances or dissimilarity measures between input maps, and more specifically between inputs of set-valued nature. We consider the problem of choosing experiments given dissimilarities such as the Hausdorff or Wasserstein distances but also of eliciting adequate dissimilarities not only based on practitioners' expertise but also on quantitative and graphical diagnostics including nearest neighbour cross-validation and non-Euclidean structural analysis. The proposed approaches are illustrated on an original uncertainty quantification case study from mechanical engineering, where using partitioning around medoids with ad hoc distances gives promising results in terms of stratified sampling.

David Ginsbourger
Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, CH-1920 Martigny and IMSV, Department of Mathematics and Statistics, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland. e-mail: ginsbourger@idiap.ch and e-mail: ginsbourger@stat.unibe.ch

Jean Baccou and Frédéric Perales
Institut de Radioprotection et de Sûreté Nucléaire, PSN-RES, SEMIA, Centre de Cadarache and Laboratoire de Micromécanique et d'Intégrité des Structures, IRSN-CNRS-UMII, 13115 Saint-Paul-lès-Durance, France. e-mail: jean.baccou@irsn.fr and e-mail: frederic.perales@irsn.fr

Clément Chevalier
Institut de Statistique, Université de Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel and Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. e-mail: clement.chevalier@unine.ch

1 Introduction

Here we consider a function $f : \mathbf{x} \in D \longrightarrow f(\mathbf{x}) \in \mathbb{R}$ stemming from some expensive deterministic computer experiment, where the input space D can possibly be a subset of $\mathbb{R}^q$ or some more complicated set of structured objects such as curves, maps, or trees. A main prerequisite is that D is endowed with a distance d, or more generally with a *dissimilarity* $\delta : (\mathbf{x},\mathbf{y}) \in D \times D \longrightarrow \mathbb{R}$, a function satisfying $\forall \mathbf{x},\mathbf{y} \in D, \delta(\mathbf{x},\mathbf{y}) \geq 0, \delta(\mathbf{x},\mathbf{x}) = 0$, and that reflects to some extent how different the outcome is supposed to be for any given couple of inputs. Here we also require the symmetry condition $\forall \mathbf{x},\mathbf{y} \in D, \delta(\mathbf{x},\mathbf{y}) = \delta(\mathbf{y},\mathbf{x})$ [11]. Depending on the context, δ may either be prescribed by expert knowledge, learnt based on data, or a combination of both. Here we are interested in the use of such dissimilarities for the design of experiments, with a mechanical engineering case study serving as motivating example, and a main focus on two interrelated questions. The first question, assuming that several candidate dissimilarity functions are available, is how to choose the most relevant one in order to model and predict the response of interest. Addressing it will lead us to consider some quantitative and graphical diagnostics. The second one, assuming δ is given, is how to extract a sub-sample from a sample of inputs, with the aim to suitably represent the distribution of responses. Our proposed approach consists in appealing to a clustering algorithm such as *Partitioning Around Medoids* (PAM) with respect to the chosen δ. In the considered test case, inputs are modelled as point sets, and several candidate distances between sets (Hausdorff, Wasserstein, and variants thereof) are considered. We appeal to non-Euclidean structural analysis concepts from spatial statistics [4] and also to nearest-neighbour approaches for studying the adequacy of dissimilarities and diagnosing what may reasonably be expected when appealing to "distance methods" based on a reference sample and candidate dissimilarities. The paper is organized as follows. In Section 2, we describe the motivating case study. In Section 3, we review a few selected distance methods and present the specific dissimilarities (distances between point sets) considered for the case study. In Section 4, we first consider a dissimilarity-based variography framework and display experimental results obtained on test case data. Then a nearest-neighbour approach serves to produce additional diagnostics. The candidate distances are used to perform stratified subsampling relying on PAM. Conclusions and perspectives as well as a list of references are presented in Section 5.

2 Motivating case study

In the framework of a research program at the French Institute for Radiological Protection and Nuclear Safety, mechanical simulations are performed with the CASTEM code [2] in order to calculate equivalent stresses on biphasic materials subjected to uni-axial traction. The system, an elastoplastic matrix containing elastic inclusions, is modelled as a unit square containing p circular inclusions, all with the same radius R. Macroscopic calculations are done by averaging over a large

number of squares with random inclusion locations. Two examples of inclusion patterns with response values of equivalent von Mises stress are sketched on Figure 1.

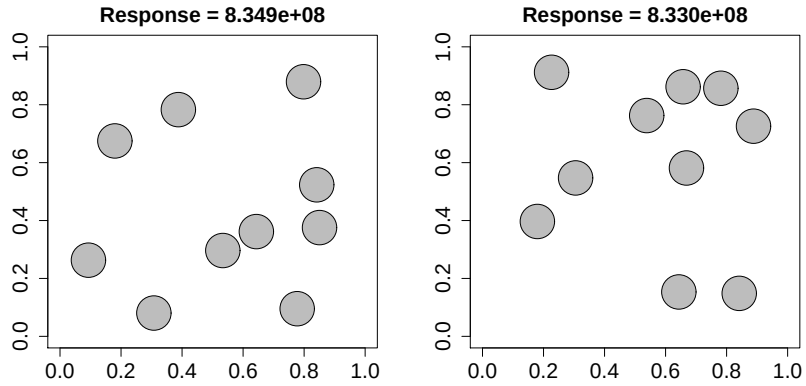

Fig. 1 Two randomly drawn CASTEM inputs with associated equivalent stress responses

Simulations were all performed with input media containing $p = 10$ randomly distributed inclusions with constant radii ($R = 0.056419$). In such a context, an input configuration $\mathbf{x} = \{\mathbf{c}_1, \ldots, \mathbf{c}_p\}$ is parametrized by a set of 2-dimensional points $\mathbf{c}_i \in [0,1]^2$ ($1 \leq i \leq p$). Given two arbitrary configurations $\mathbf{x}$ and $\mathbf{y}$, we aim at finding a relevant distance between them and making predictions of equivalent stress at inputs for which the simulation outcome is unknown, based on available simulation results. Of course, there are many ways of defining distances between such $\mathbf{x}$ and $\mathbf{y}$; and this richness can turn into a drawback when facing a choice between numerous candidates. In the next section, we principally focus on four candidate distances for the design of CASTEM simulations. The main application-driven objectives pursued here concern parsimoniously choosing sub-samples of inputs by relying on distance methods in order to get an empirical distribution of responses in the sub-sample close to the empirical distribution of responses in the larger sample.

3 Distance methods: basics and considered distances

Before focusing on the specifics of our test case, let us give a brief overview of distance methods from machine learning to spatial statistics. One of the most simple and yet efficient distance-methods for classification and prediction is the *k*-nearest neighbors algorithm (*k*NN). Assuming that the response of interest was observed for N instances $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the input, predicting the response at an arbitrary $\mathbf{x} \in D$ by *k*NN consists in averaging the responses of $\mathbf{x}$'s k nearest neighbors (among $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and in the sense of the chosen distance d), where $k \geq 1$ is a parame-

ter of the algorithm. Beyond kNN, further distance methods have been proposed for data analysis, including multi-dimensional scaling, phylogenetic trees, spectral clustering and more (see, e.g., [6] and references therein for an overview).

Distances are also key in spatial statistics [3, 5]. A spatially varying measurement of interest, say $z : \mathbf{x} \in D \longrightarrow \mathbb{R}$, is modelled as one realization of a random field $Z = \{Z_{\mathbf{x}}(\omega)\}_{\mathbf{x}\in D}$ ($\omega \in \Omega$, Ω standing for the underlying probability space). Assuming that Z's increments are squared integrable, $\gamma : (\mathbf{x},\mathbf{y}) \in D^2 \longrightarrow \gamma(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathrm{Var}[Z_{\mathbf{x}} - Z_{\mathbf{y}}] \in \mathbb{R}$ exists and is called *semi-variogram* of Z. $\sqrt{2\gamma}$ then defines over D the so-called *canonical distance* associated with Z [1]. Given an arbitrary fixed distance d over D, when Z's mean is constant and $\gamma(\mathbf{x},\mathbf{y})$ depends on $(\mathbf{x},\mathbf{y})$ through $d(\mathbf{x},\mathbf{y})$, Z and γ are called *isotropic*. Estimating how γ depends on d based on data is a difficult step, often referred to as *empirical variography*. In Section 4, we investigate distance methods and empirical variography in the case where $\mathbf{x}$ is point set ("configuration") and the underlying distances are chosen accordingly, as discussed now.

In our motivating case study we consider four candidate distances between input configurations, namely the Hausdorff [8] and Wasserstein [10] distances, together with ad hoc symmetrizations of them. In all cases, the configurations are represented by the locations of their respective centers. Denoting by $\mathbf{c}'_1,\ldots,\mathbf{c}'_p$ the centers of the configuration $\mathbf{y}$, the Hausdorff distance between $\mathbf{x}$ and $\mathbf{y}$ is defined in our case by

$$d_H(\mathbf{x},\mathbf{y}) = \max\left(\max_{1\le i\le p}\min_{1\le j\le p} d_{\mathbb{R}^2}(\mathbf{c}_i,\mathbf{c}'_j), \max_{1\le j\le p}\min_{1\le i\le p} d_{\mathbb{R}^2}(\mathbf{c}_i,\mathbf{c}'_j)\right) \tag{1}$$

where $d_{\mathbb{R}^2}$ denotes the Euclidean distance over $\mathbb{R}^2$. The Hausdorff distance is quite popular in probability theory as it comes with a number of seminal mathematical results, but also in applied fields such as computer vision. Beyond d_H, another family of distances considered here, related to optimal transportation, are the so-called Wasserstein distances. Unlike Hausdorff distances, Wasserstein (or "earth mover's") distances do not only quantify the closeness of points of configurations to those of the other one, but also incorporate some more "physical" information on the cost to transform one configuration into the other one. Configurations (e.g., $\mathbf{x}$ and $\mathbf{y}$) are seen as "patterns" (i.e. finite support measures), and the distance between two patterns is calculated by pairing their respective points in such a way that each point of be in correspondence with a unique point of the other pattern. The distance is then defined based on the minimal value, over all possible pairings, of the average (be it in the L^2 sense or other) or maximal "ambient" (Euclidean, here) distances between paired points. We consider the following instance of this family of distances:

$$d_W(\mathbf{x},\mathbf{y}) = \min_{\sigma\in\mathscr{S}_p}\sqrt{\frac{1}{p}\sum_{i=1}^{p} d^2_{\mathbb{R}^2}\left(\mathbf{c}_i,\mathbf{c}'_{\sigma(i)}\right)}, \tag{2}$$

where $\mathscr{S}_p$ is the set of permutations of $\{1,\ldots,p\}$. Additionally, some simple physical knowledge was used to design symmetrized distances dedicated to the test case. As the vertical force applied to the micro-structure is symmetrical with respect to the $(x_1 = 0.5)$-axis, f is known to be invariant under the axial flip $v : (x_1,x_2) \in$

$[0,1]^2 \longrightarrow (1-x_1,x_2) \in [0,1]^2$. So we designed the corresponding symmetrized (or *quotient*) versions of d_L ($L \in \{W,H\}$): $d_{Lsym}(\mathbf{x},\mathbf{y}) = \min\{d_L(\mathbf{x},\mathbf{y}), d_L(\mathbf{x},v(\mathbf{y}))\}$.

4 Application results

Data sets and preliminary results. We consider two data sets, a preliminary 900-element one (*data set A*) and the main one (*data set B*), consisting of 404 instances of numerical simulation input/output tuples. Data set A was generated by sequentially drawing inclusion centers uniformly over $[0,1]^2$ with rejection in case of overlap between new and already included inclusions. While A was judged satisfactory in terms of space-fillingness, results (not displayed here by space limitation) from the targeted distance-based methods happened to be disappointing precisely because configurations turn out to be too "far" from each other to produce exploitable results, a consequence of the *curse of dimensionality* (see [9] and references therein). In order to get around this and produce a more informative data set (especially with respect to empirical variograms presented next), we conducted a new batch of simulations (data set B) with a more exploitable design. Data set B was generated by first sampling 116 random configurations following the same procedure as for data set A. Then the remaining configurations were generated by applying local perturbations (scrambling of randomly selected centers, with small to medium –i.e. 10^{-3} to 10^{-2}– orders of magnitude). While there is some arbitrariness in the way B was designed, it has to be kept in mind that identifying modes of relevance (resp. of failure) of the proposed approaches fully falls in the objectives of this study. Besides this, we are interested in investigating which distance works best and how to uncover and/or quantify it based on empirical diagnostics. Diagnostics obtained on data set B via empirical variography and cross-validated kNN predictions are presented and discussed next. The section then presents some further results in stratification.

Diagnostics based on structural analysis vs kNN cross-validated predictions. The first graphical diagnostic considered here is a set of four non-Euclidean empirical variograms based on the respective candidate distances. Denoting by d any arbitrary distance from $d_H, d_W, d_{Hsym}, d_{Wsym}$, we plot in Figure 2 the corresponding empirical semi-variogram (*Matheron estimator*, see [3]) values:

$$\widehat{\gamma}(h_\ell) = \frac{1}{2\sharp N(h_\ell)} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in N(h_\ell)} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)^2, \tag{3}$$

where $N(h_\ell) = \left\{(\mathbf{x}_i,\mathbf{x}_j) : d(\mathbf{x}_i,\mathbf{x}_j) \in [\max(0,h_\ell-\Delta), h_\ell+\Delta]\right\}$ with a given tolerance parameter $\Delta > 0$, and $\sharp N(h_\ell)$ is the number of couples in $N(h_\ell)$. In Figure 2, empirical semi-variograms are plotted with common values of h_ℓ (here $1 \leq \ell \leq 10$) for the four candidate distances; the coloured numbers near the points stand for the corresponding $\sharp N(h_\ell)$ values. A first comment regarding the left panel (Hausdorff distances) is that taking the symmetry into account does not seem to affect things at short distances, while a slight departure is observed at the penultimate bin center.

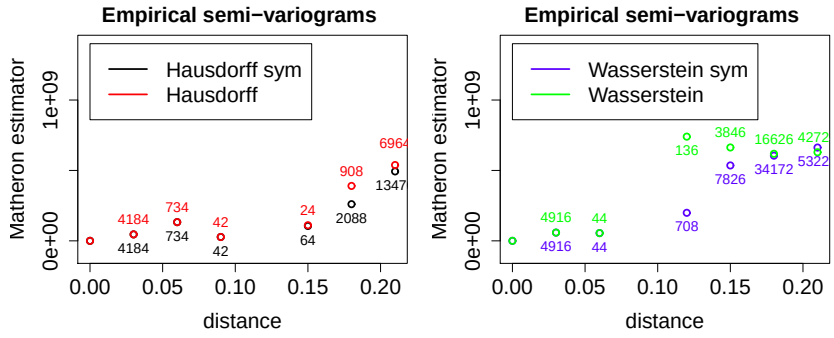

Fig. 2 Empirical variograms with respect to the four candidate distances (Hausdorff vs Wasserstein, with possible symmetrization) of the equivalent constraint depending on inclusion centers.

Focusing on the number of elements in the bins, the central regions of the empirical semi-variograms appear to be based on little information, and thus more faith should be put on their ends. As for Wasserstein distances (right panel), even if the central region creates the impression of a chaotic behaviour –especially for the d_W green curve–, it appears that the number of points at low distance is greater than in the Hausdorff case, a worthwhile piece of information for forthcoming considerations.

A second approach, investigated next for quantifying and graphically diagnosing how the candidate distances absolutely and comparatively perform, is based on cross-validated kNN predictions. For the four distances and a number of neighbours k ranging from 1 to 20, the following is implemented: for all $i \in 1,\ldots,N$ leave $\mathbf{x}_i$ out and predict it both by kNN and by taking the arithmetic average of the $N-1$ remaining responses. Finally, evaluate kNN's relative performance by taking either the ratio of the mean kNN error over the mean error when predicting by the mean, or the ratio of the median kNN error over the median error when predicting by the mean. Results are represented in Figure 3. For the first criterion, symmetrizing the Hausdorff distance appears to lead to improved performances. In median, however, all distances perform similarly, and well compared to the prediction by the mean.

Distance-based stratification. We finally appeal to a clustering method, *Partitioning Around Medoids* (PAM), in order to select "representative" sub-samples of $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ and approximate the cumulative distribution function of associated responses. This time, the letter k refers to the subsample size, which is in turn here the number of medoids in the PAM method. We rely on the Kolmogorov-Smirnov (K-S) statistic in order to assess and compare the performances of the different considered approaches. In addition to the PAM approach with the four candidate distances, we also consider uniform random sampling of k among N configurations. Values of the K-S statistic for the four candidate strategies versus 500 replications of the random strategy are plotted as a function of k in the left panel of Figure 4. It appears that for most values of k, the K-S statistic is smaller than the median K-S under the random scenario, whatever the candidate distance. Stratification with PAM hence

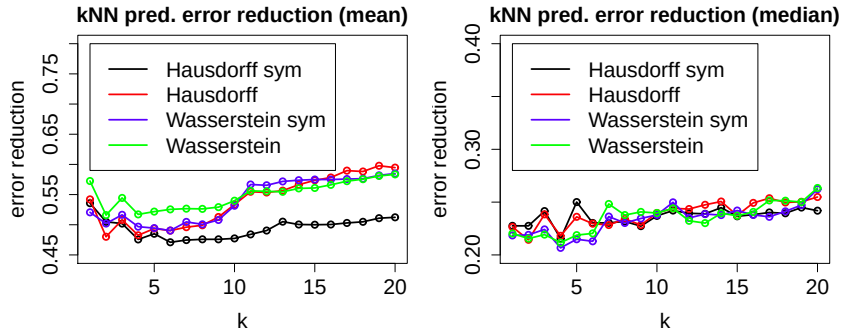

Fig. 3 Leave-one-out compared performances of kNN prediction versus naive prediction by the mean, both in terms of ratios of average errors and of median errors.

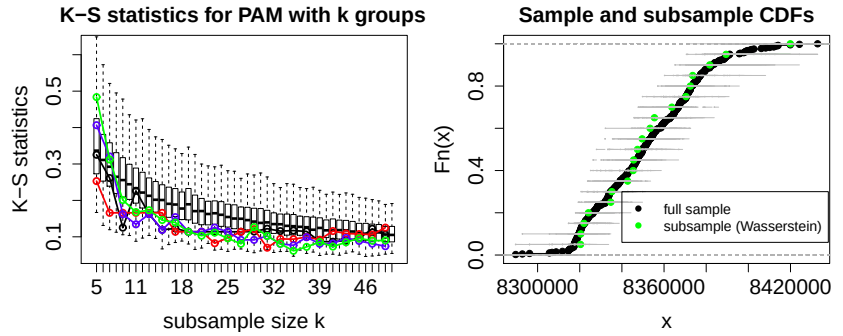

Fig. 4 Left: Performance assessment of random vs stratified sampling using $d_H, d_{Hsym}, d_W, d_{Wsym}$ (left). Right: full sample CDF vs subsample CDF obtained with the Wasserstein distance ($k = 20$). Gray points stand for quantiles obtained by drawing 50 subsamples uniformly at random.

appears as an appealing alternative to random sampling, as it avoids its variability. As an illustration, the empirical CDF obtained by PAM with the Wasserstein distance and $k = 20$ –approx. 5% of the total budget– is represented (in green) against the empirical CDF of the full sample of responses. To finish with, a number of other approaches based on dissimilarity matrices may be beneficially applied to this problem; in particular, maximin and minimax distance designs [7] appear as credible alternatives for choosing representative sub-samples out of a reference sample.

5 Conclusion and Perspectives

We have investigated distance methods for the design of computer experiments with a view toward subsampling. A case study in mechanical engineering, where inputs

are parametrized by point sets and four distances are in competition, has served as our basis. While non-Euclidean variography appeared appealing, results obtained on the test case did not shed much light on which distance was best suited for distance-based modelling. On the other hand, one of the two diagnostics based on kNN highlighted the potential interest of symmetrizing the Hausdorff distance. Nevertheless, all distances gave decent stratification results. While not suffering from the variability inherent to random subsampling, using the PAM algorithm with any of these distances and for most considered subsample sizes also delivered subsample distributions closer to the full sample distribution than obtained in median via random subsampling. However, these results have to be tempered for several reasons. First, all tests rely on a single data set with a specific structure, so that more studies with alternative data sets are necessary for validating and refining the conclusions by exploring their degree of generality. From the point of view of variogram design, it would be worthwhile to search for admissible models with respect to the Hausdorff and/or Wasserstein distances considered here (Gaussian and exponential models are not, according to numerical tests), but also to question the intrinsic stationarity assumption which was quietly made here. Besides this, the distances considered were almost off-the-shelf, and it would probably be beneficial to investigate further candidate distances for this or other test cases, taking available expertise into account [5]. Further perspectives include using this framework to adaptively refine the subsample, e.g., in order to foster the exploration of distributional tails.

Acknowledgements: Part of this work has been conducted within the frame of the ReDice Consortium, which gathered industrial (CEA, EDF, IFPEN, IRSN, Renault) and academic (Ecole des Mines de Saint-Etienne, INRIA, and the University of Bern) partners around advanced methods for Computer Experiments.

References

1. Adler, R.J., Taylor, J.E., Random Fields and Geometry, Springer (2007)
2. Cast3M software, http://www-cast3m.cea.fr
3. Cressie, N.A.C., Statistics for Spatial Data, Wiley (1993)
4. Curriero, F.C., On the use of non-Euclidean distance measures in geostatistics, Mathematical Geology, **38** (8), 907–926 (2006)
5. Ginsbourger, D., Rosspopoff, B., Pirot, G., Durrande, N., Renard, P., Distance-based kriging relying on proxy simulations for inverse conditioning, Adv Water Res **52**, 275–291 (2013)
6. Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag New York (2009)
7. Johnson, M.E., Moore, L.M., Ylvisaker, D., Minimax and maximin distance designs, Journal of Statistical Planning and Inference **26**, 131–148 (1990)
8. Molchanov, I., Theory of Random Sets, Springer (2005)
9. Radovanović, M., Nanopoulos, A., Ivanović, M., Hubs in space: popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487-2531 (2010)
10. Schuhmacher, D., Xia, A., A new metric between distributions of point processes. Advances in Applied Probability 40, 651–672 (2008)
11. von Luxburg, U., Statistical Learning with Similarity and Dissimilarity Functions. PhD thesis, Technische Universität Berlin (2004)