Projection-based curve clustering

Benjamin Auder, Aurélie Fischer

Projection-based curve clustering

Benjamin AUDER

DEN/DER/SESI/LCFR

CEA Cadarache

13108 SAINT PAUL LEZ DURANCE Cedex, FRANCE

Aurélie FISCHER*

Laboratoire de Statistique Théorique et Appliquée

Université Pierre et Marie Curie – Paris VI

Boîte 158, couloir 15–16, 2e étage

4 place Jussieu, 75252 PARIS Cedex 05, FRANCE

Abstract – This paper focuses on unsupervised curve classification in the context of nuclear industry. At the Commissariat à l'Energie Atomique (CEA), Cadarache (France), the thermal-hydraulic computer code CATHARE is used to study the reliability of reactor vessels. The code inputs are physical parameters and the outputs are time evolution curves of a few other physical quantities. As the CATHARE code is quite complex and CPU-time consuming, it has to be approximated by a regression model. This regression process involves a clustering step. In the present paper, CATHARE output curves are clustered using a k-means scheme, with a projection onto a lower dimensional space. We study the properties of the empirically optimal cluster centers found by the clustering method based on projections, compared to the "true" ones. The choice of the projection basis is discussed, and an algorithm is implemented to select the best projection basis among a library of orthonormal bases. The approach is illustrated on a simulated example and then applied to the industrial problem.

Index Terms — Clustering, basis expansion, projections, k-means, wavelet packets.

*Corresponding author. E-mail: `aurelie.fischer@upmc.fr`

1 Introduction

1.1 The CATHARE code

A major concern in nuclear industry is the life span of reactor vessels. To go on using the current nuclear reactors, their reliability has to be proved. For this purpose, complex computer codes are developed to simulate the behavior of the vessel under different sequences of accidents. At the Commissariat à l'Energie Atomique (CEA), Cadarache (France), one of the main types of accident under study is the pressurized thermal shock. This is a problem due to the combined stresses from a rapid temperature and pressure change. More specifically, as a reactor vessel gets older, the potential for failure by cracking when it is cooled rapidly at high pressure increases greatly. The analysis of pressurized thermal shock is made of two main steps. First, a thermal-hydraulic analysis is done to determine the temporal evolutions of temperature, pressure and thermal exchange coefficient in the vessel annular space, since these features have an influence on the mechanical and thermal charge on the vessel inner surface. Some evolution curves $x_1(t), \ldots, x_n(t)$ corresponding to the thermal exchange coefficient are depicted in Figure 1 ($n = 66$). Each curve $x_i(t)$ is obtained as the simulation result for a certain vector of input physical parameters. The curves of temperature, pressure and thermal exchange coefficient obtained during this first step are then used as limit conditions in the second part of the analysis, which is a mechanical investigation aiming at checking if some defects on the vessel annular space could propagate and gain importance to such an extent that this would cause a break of the vessel inner surface. For further details on the reliability of reactor vessels, we refer the reader to Auder, De Crecy, Iooss, and Marquès [5].

The simulation step relies on a computer code called CATHARE (Code Avancé de THermohydraulique pour les Accidents des Réacteurs à Eau, in English Code for Analysis of THermalhylaudrics during an Accident of Reactor and safety Evaluation). The CATHARE code is a system code for pressurized water reactors safety analysis, accident management, definition of plant operating procedures and for research and development. The project is a result of a joint effort of the reactor vendor AREVA, the CEA, EDF (Electricité de France) and the IRSN (Institut de Radioprotection et de Sûreté Nucléaire). The first delivered version V1.3L was available in 1997. The CEA team CATHARE located in Grenoble (France) is in charge of the development, the assessment and the maintenance of the code. (See `http://www-cathare.cea.fr`.)

The CATHARE code allows to simulate the evolution of temperature, pressure and thermal exchange coefficient, given the physical parameters as inputs. However, this code is so slow (about 6 to 10 hours for one run) that it cannot be used directly for reliability calculations. To bypass this obstacle, the strategy drawn up by the CEA is to build a so-called metamodel which is a fast approximation of the original code, precise enough to carry out statistical computations. The term "metamodel" indicates that a computer code approximating a physical process has already been developed, and now this code is modeled in turn. Here, the purpose is the construction of a regression model based on a few hundreds CATHARE code outputs,
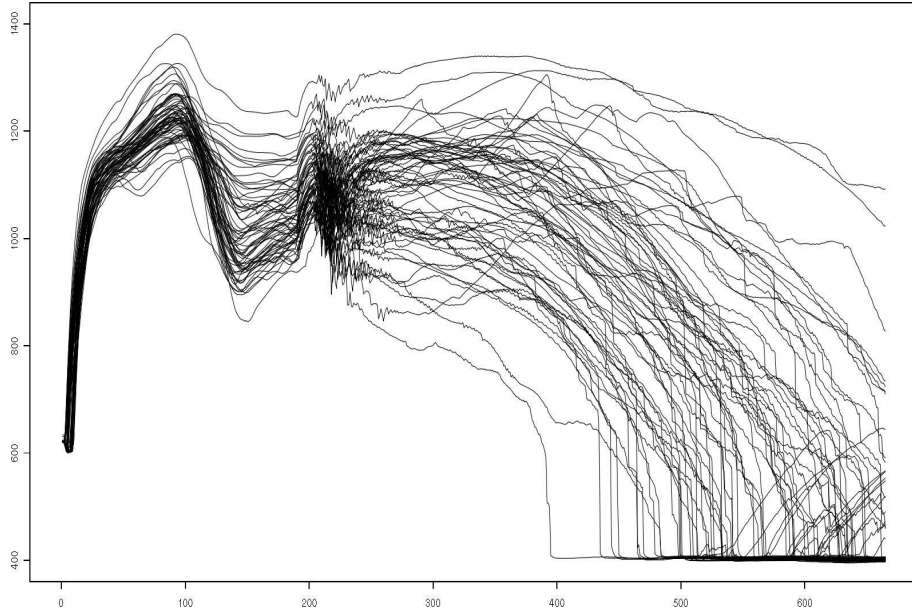
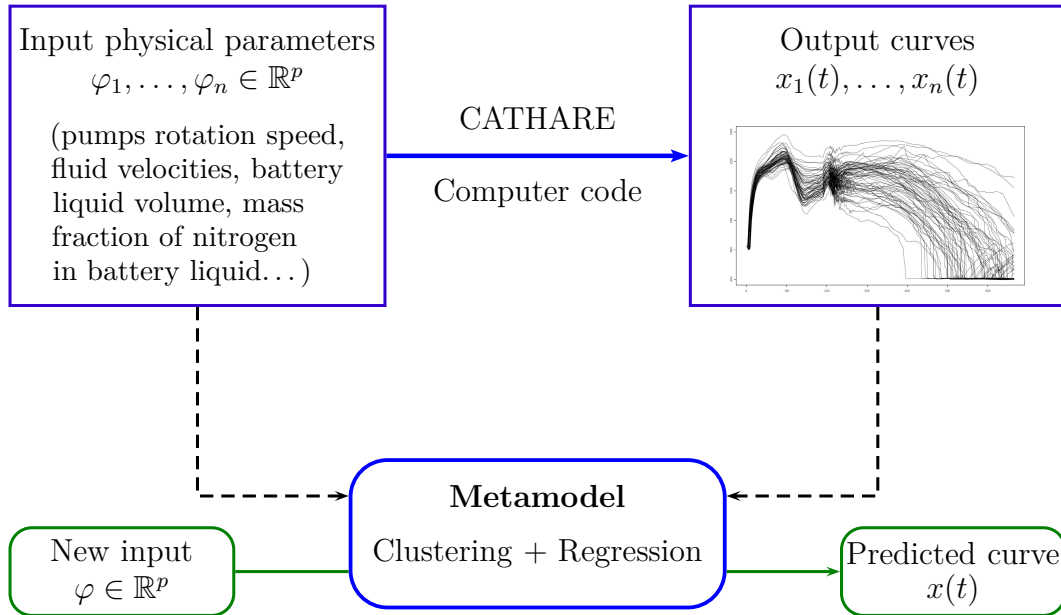Figure 1: 66 evolution curves of thermal exchange coefficient in a nuclear vessel.

Figure 2: Flowchart of the CATHARE code.

obtained during one week of computation on a supercomputer in 2007. The inputs were sampled randomly by latin hypercube methods (see, e.g., McKay, Conover, and Beckman [28] and Loh [26]), so that we have no control over the inputs in the learning sample. As different kinds of behavior for temperature, pressure and thermal exchange coefficient may be observed depending on the physical parameters, a preliminary unsupervised classification of CATHARE code output curves is essential. Once the curves have been clustered in meaningful classes, the regression model can be adjusted for each group of outputs separately. The clustering step is the object of the present paper.

1.2 Clustering

Clustering is the problem of partitioning data into a finite number of groups (denoted hereafter by k), or clusters, so that the data items inside each of them are very similar among themselves and as different as possible from the elements of the other clusters (Duda, Hart, and Stork [14, Chapter 10]). In our industrial context, the data is made of evolution curves of temperature, pressure or thermal exchange coefficient. Using a probabilistic point of view, these curves can be seen as independent draws $X_1(t), \ldots, X_n(t)$ with the same distribution as a generic random variable $X(t)$ taking values in a functional space $(E, \| \cdot \|)$ — typically, the Hilbert space of square integrable functions.

A widely used clustering method is the so-called k-means clustering, which consists in partitioning the random observations $X_1, \ldots, X_n \in E$ into k classes by minimizing the empirical distortion

$$W_{\infty,n}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{\ell=1,\ldots,k} \|X_i - c_\ell\|^2,$$

over all possible cluster centers $\mathbf{c} = (c_1, \ldots, c_k) \in E^k$. Here, μ_n denotes the empirical measure associated with the sample $X_1, \ldots, X_n$, i.e.,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}}$$

for every Borel subset A of E. In other words, we look for a Voronoi partition of E. The Voronoi partition $C_1, \ldots, C_k$ associated with $\mathbf{c} = (c_1, \ldots, c_k) \in E^k$ is defined by letting an element $x \in E$ belong to C_ℓ if it is closer (with respect to the norm $\| \cdot \|$) to c_ℓ than to any other c_j (ties are broken arbitrarily). The ℓ-th cluster is made of the observations X_i assigned to c_ℓ, or equivalently, falling in the Voronoi cell C_ℓ. In this framework, the accuracy of the clustering scheme is assessed by the distortion or mean squared error

$$W_\infty(\mathbf{c}) = \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|X - c_\ell\|^2 \right],$$

where $\mathbb{E}$ stands for expectation with respect to the distribution of X. This clustering method is in line with the more general theory of quantization. More specifically, it

corresponds to the empirical version of nearest neighbor quantization (Linder [24], Gersho and Gray [18], Graf and Luschgy [21]). However, the problem of finding a minimizer of the criterion $W_{\infty,n}(\mathbf{c})$ is in general NP-hard, and there is no efficient algorithm to find the optimal solution in reasonable time. That is why several iterative algorithms have been developed to give approximate solutions. One of the first to be described historically is Lloyd's algorithm (Lloyd [25]).

The challenge is to adapt the k-means clustering method to our setting. The main difficulty here is the high dimensionality of the data, which casts the problem into the general class of functional statistics. For a comprehensive introduction to this topic, see Ramsay and Silverman [32] and Ferraty and Vieu [15]. A possible approach to reduce the infinite dimension of the observations $X_1, \ldots, X_n$ consists in projecting them onto a lower-dimensional subspace. In this context, Abraham, Cornillon, Matzner-Løber, and Molinari [1] project the curves on a B-spline basis, and get clusters with the k-means algorithm applied to the coefficients. These authors argue that projecting onto a smooth spline basis plays the role of a denoising procedure, given that the observed curves could contain measurement errors, and also allows to deal with curves which were not measured at the same time. James and Sugar [23] use a B-spline basis to model the centers of the clusters and write each curve in cluster ℓ as a main effect defined by spline coefficients plus an error term. This allows for some deviations around a model curve specific to cluster ℓ. The authors add a Gaussian error term to model the individual variations among one cluster. This way, the main effect is enriched, and the model can take into account more complex behaviors. The method in Gaffney [17] is similar, also with a B-spline basis. Another option is to use a Self-Organizing Map algorithm on the coefficients (Rossi, Conan-Guez, and El Golli [33]), again obtained by projecting the functions onto a B-spline basis. These bases are often used because they are easy to implement, and require a relatively minimal number of parametric assumptions. Besides, Biau, Devroye, and Lugosi [6] examine the theoretical performance of clustering with random projections based on the Johnson-Lindenstrauss Lemma, which represent a sound alternative to orthonormal projections thanks to their distance-preserving properties. Chiou and Li [7] propose a method which generalizes the k-means algorithm to some extent, by considering covariance structures via functional principal component analysis. In the approach of these authors, each curve is decomposed on an adaptive local basis (valid for the elements in the cluster), and the clusters are determined according to the full approximation onto each basis. In the wavelet-based method for functional data clustering developed in Antoniadis, Brossat, Cugliari, and Poggi [3], a smooth curve is reduced to a finite number of representative features, by considering the contribution of each wavelet coefficient to the global energy of the curve.

In the present contribution, we propose to investigate the problem of clustering output curves $X_1, \ldots, X_n$ of the CATHARE code, assuming that they arise from a random variable X taking its values in some subset of the space of square integrable functions. As a general strategy, we reduce the infinite dimension of X by considering only the first d coefficients of the expansion on a Hilbertian basis, and then perform clustering in $\mathbb{R}^d$. We study the theoretical properties of this clustering method with projection. A bound expressing what is lost when replacing the empirically optimal

cluster centers by the centers obtained by projection is offered (Section 2). Since the result may depend on the basis choice, several projection bases are used in practice, and we look for the best one minimizing a given criterion. To this end, an algorithm based on Coifman and Wickerhauser [9] is implemented, searching for an optimal basis among a library of wavelet packet bases available in the **R** package *wmtsa*, and this "optimal basis" is compared with the Fourier basis, the Haar basis, and the functional principal component basis (Section 3). Finally, this algorithm is applied to a simulated example and to our industrial problem (Section 4). Proofs are postponed to Section 5.

2 Finite-dimensional projection for clustering

As mentioned earlier, we are concerned with square integrable functions. Since all results can be adapted to $L^2([a, b])$ by an appropriate rescaling, we consider for the sake of simplicity the space $L^2([0, 1])$. As an infinite-dimensional separable Hilbert space, $L^2([0, 1])$ is isomorphic via the choice of a Hilbertian basis to the space ℓ^2 of square-summable sequences. We focus more particularly on functions in $L^2([0, 1])$ whose coefficients in the expansion on a given Hilbertian basis belong to the subset $\mathcal{S}$ of ℓ^2 given by

$$\mathcal{S} = \left\{ \mathbf{x} = (x_j)_{j \geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\}, \tag{1}$$

where $R > 0$ and $(\varphi_j)_{j \geq 1}$ is a nonnegative increasing sequence such that

$$\lim_{j \to +\infty} \varphi_j = +\infty.$$

It is worth pointing out that $\mathcal{S}$ is closely linked with the basis choice, even if the basis does not appear explicitly in the definition. To illustrate this important fact, three examples are discussed below.

Example 2.1 (Sobolev ellipsoids). *For $\beta \in \mathbb{N}^*$ and $L > 0$, the periodic Sobolev class $W^{per}(\beta, L)$ is the space of all functions $f \in [0, 1] \to \mathbb{R}$ such that $f^{(\beta-1)}$ is absolutely continuous, $\int_0^1 (f^{(\beta)}(t))^2 dt \leq L^2$ and $f^{(\ell)}(0) = f^{(\ell)}(1)$ for $\ell = 0, \ldots, \beta - 1$. Let $(\psi_j)_{j \geq 1}$ denote the trigonometric basis. Then a function $f = \sum_{j=1}^{+\infty} x_j \psi_j$ is in $W^{per}(\beta, L)$ if and only if the sequence $\mathbf{x} = (x_j)_{j \geq 1}$ of its Fourier coefficients belongs to*

$$\mathcal{S} = \left\{ \mathbf{x} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\},$$

where

$$\varphi_j = \begin{cases} j^{2\beta} & \text{for even } j \\ (j-1)^{2\beta} & \text{for odd } j \end{cases}$$

and $R = \dfrac{L}{\pi^\beta}$.

For the proof of this result and further details about Sobolev classes, we refer the reader to the book of Tsybakov [36]. Note that the set $\mathcal{S}$ could also be defined by $\varphi_j = j^r e^{\alpha j}$ with $\alpha > 0$ and $r \geq -\alpha$ (Tsybakov [35]).

Example 2.2 (Reproducing Kernel Hilbert Spaces). *Let $K : [0,1] \times [0,1] \to \mathbb{R}$ be a Mercer kernel, i.e., K is continuous, symmetric and positive definite. Recall that a kernel K is said to be positive definite if for all finite sets $\{x_1, \ldots, x_m\}$, the matrix A defined by $a_{ij} = K(x_i, x_j)$ for $1 \leq i, j \leq m$ is positive definite. For example, the Gaussian kernel $K(x,y) = \exp(-\frac{(x-y)^2}{\sigma^2})$ and the kernel $K(x,y) = (c^2 + (x-y)^2)^{-a}$ with $a > 0$ are Mercer kernels. For $x \in [0,1]$, let $K_x : y \mapsto K(x,y)$. Then, Moore-Aronszajn's Theorem (Aronszajn [4]) states that there exists a unique Hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle)$ of functions on $[0,1]$ such that:*

1. *For all $x \in [0,1]$, $K_x \in \mathcal{H}_K$.*

2. *The span of the set $\{K_x, x \in [0,1]\}$ is dense in $\mathcal{H}_K$.*

3. *For all $f \in \mathcal{H}_K$ and $x \in [0,1]$, $f(x) = \langle K_x, f \rangle$.*

The Hilbert space $\mathcal{H}_K$ is said to be the reproducing kernel Hilbert space (for short, RKHS) associated with the kernel K. Next, the operator $\mathcal{K}$ defined by

$$\mathcal{K}f : y \mapsto \int_0^1 K(x,y)f(x)dx$$

is self-adjoint, positive and compact. Consequently, there exists a complete orthonormal system $(\psi_j)_{j \geq 1}$ of $L^2([0,1])$ such that $\mathcal{K}\psi_j = \lambda_j \psi_j$, where the set of eigenvalues $\{\lambda_j, j \geq 1\}$ is either finite or a sequence tending to 0 at infinity. Moreover, the λ_j are nonnegative. Suppose that $\mathcal{K}$ is not of finite rank — so that $\{\lambda_j, j \geq 1\}$ is infinite — and that the eigenvalues are sorted in decreasing order, that is $\lambda_j \geq \lambda_j + 1$ for all $j \geq 1$. Clearly, there is no loss of generality in assuming that $\lambda_j > 0$ for all $j \geq 1$. Indeed, if not, $L^2([0,1])$ is replaced by the linear subspace spanned by the eigenvectors corresponding to non-zero eigenvalues.

According to Mercer's theorem, K has the representation

$$K(x,y) = \sum_{j=1}^{+\infty} \lambda_j \psi_j(x)\psi_j(y),$$

where the convergence is absolute and uniform (Cucker and Smale [10, Chapter III, Theorem 1]). Moreover, $\mathcal{H}_K$ may be characterized through the eigenvalues of the operator $\mathcal{K}$ by

$$\mathcal{H}_K = \left\{ f \in L^2([0,1]) : f = \sum_{j=1}^{+\infty} x_j \psi_j, \sum_{j=1}^{+\infty} \frac{x_j^2}{\lambda_j} < \infty \right\},$$

with the inner product

$$\left\langle \sum_{j=1}^{+\infty} x_j \psi_j, \sum_{j=1}^{+\infty} y_j \psi_j \right\rangle = \sum_{j=1}^{+\infty} \frac{x_j y_j}{\lambda_j}$$

(Cucker and Smale [10, Chapter III, Theorem 4]). Then, letting

$$\mathcal{S} = \left\{ \mathbf{x} \in \ell^2, \sum_{j=1}^{+\infty} \frac{x_j^2}{\lambda_j} \leq R^2 \right\},$$

the set $\mathcal{S}$ is of the desired form (1), with $\varphi_j = 1/\lambda_j$.

Example 2.3 (Besov ellipsoids and wavelets). *Let $\alpha > 0$. For $f \in L^2([0,1])$, the Besov semi-norm $|f|_{B_2^\alpha}(L^2)$ is defined by*

$$|f|_{B_2^\alpha}(L^2) = \left(\sum_{j=0}^{+\infty} [2^{j\alpha} \omega_r(f, 2^{-j}, [0,1])_2]^2 \right)^{1/2}$$

where $\omega_r(f, t, [0,1])_2$ denotes the modulus of smoothness of f, as defined for instance in DeVore and Lorentz [12], and $r = \lfloor \alpha \rfloor + 1$. Let $\Lambda(j)$ be an index set at resolution level j and $(x_{j,\ell})_{j\geq 0, \ell \in \Lambda(j)}$ the coefficients of the expansion of f in a suitable wavelet basis. Then, for f such that $|f|_{B_2^\alpha}(L^2) \leq \rho$, the coefficients $x_{j,\ell}$ satisfy

$$\sum_{j=0}^{+\infty} \sum_{\ell \in \Lambda(j)} 2^{2j\alpha} x_{j,\ell}^2 \leq \rho^2 C^2,$$

where $C > 0$ depends only on the basis. We refer to Donoho and Johnstone [13] for more details.

Let us now come back to the general setting

$$\mathcal{S} = \left\{ \mathbf{x} = (x_j)_{j\geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\},$$

and consider the problem of clustering the sample $X_1, \ldots, X_n$ with values in $\mathcal{S}$. Some notation and assumptions are in order. First, we will suppose that $\mathbb{P}\{\|X\| \leq R\} = 1$. Notice that the fact that X takes its values in $\mathcal{S}$ is in general not enough to imply $\mathbb{P}\{\|X\| \leq R\} = 1$. Secondly, let j_0 be the smallest integer j such that $\varphi_j > 0$. To avoid technical difficulties, we require in the sequel $d \geq j_0$. For all $d \geq 1$, we will denote by Π_d the orthogonal projection on $\mathbb{R}^d$ and let $\mathcal{S}_d = \Pi_d(\mathcal{S})$. Lastly, observe that $\mathcal{S}_d$ identifies with the ellipsoid

$$\left\{ \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : \sum_{j=1}^{d} \varphi_j x_j^2 \leq R^2 \right\}.$$

As explained in the introduction, the criterion to minimize is

$$W_{\infty,n}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{\ell=1,\ldots,k} \|X_i - c_\ell\|^2,$$

and the performance of the clustering obtained with the centers $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{S}^k$ is measured by the distortion

$$W_\infty(\mathbf{c}) = \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|X - c_\ell\|^2 \right].$$

The quantity

$$W_\infty^* = \inf_{\mathbf{c} \in \mathcal{S}^k} W_\infty(\mathbf{c})$$

8

represents the optimal risk we can achieve. With the intention of performing clustering in the projection space $\mathcal{S}^d$, we also introduce the "finite-dimensional" distortion

$$W_d(\mathbf{c}) = \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\right]$$

and its empirical counterpart

$$W_{d,n}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{\ell=1,\ldots,k} \|\Pi_d(X_i) - \Pi_d(c_\ell)\|^2,$$

as well as

$$W_d^* = \inf_{\mathbf{c}\in\mathcal{S}^k} W_d(\mathbf{c}).$$

Let us observe that, as the support of the empirical measure μ_n contains at most n points, there exists an element $\hat{\mathbf{c}}_{d,n}$ which is a minimizer of $W_{d,n}(\mathbf{c})$ on $\mathcal{S}^k$. Moreover, in view of its definition, $W_{d,n}(\mathbf{c})$ only depends on the centers projection $\Pi_d(\mathbf{c})$ (one has $W_{d,n}(\mathbf{c}) = W_{d,n}(\Pi_d(\mathbf{c}))$ for all $\mathbf{c}$) and we can thus assume that $\hat{\mathbf{c}}_{d,n} \in (\mathcal{S}_d)^k$. Notice also that for all $c \in \mathcal{S}$,

$$\|\Pi_d(X) - \Pi_d(c)\|^2 \le \|X - c\|^2$$

(the projection Π_d is 1-Lipschitz), which implies that

$$W_d(\mathbf{c}) \le W_\infty(\mathbf{c})$$

for all $\mathbf{c}$.

The following lemma provides an upper bound for the maximal deviation

$$\sup_{\mathbf{c}\in\mathcal{S}^k}[W_\infty(\mathbf{c}) - W_d(\mathbf{c})].$$

Lemma 2.1. *We have*

$$\sup_{\mathbf{c}\in\mathcal{S}^k}[W_\infty(\mathbf{c}) - W_d(\mathbf{c})] \le \frac{4R^2}{\varphi_d}.$$

We are now in a position to state the main result of this section.

Theorem 2.1. *Let $\hat{\mathbf{c}}_{d,n} \in (\mathcal{S}_d)^k$ be a minimizer of $W_{d,n}(\mathbf{c})$. Then,*

$$\mathbb{E}[W_\infty(\hat{\mathbf{c}}_{d,n})] - W_\infty^* \le \mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^* + \frac{8R^2}{\varphi_d}. \tag{2}$$

Theorem 2.1 expresses the fact that the expected excess clustering risk in the infinite dimensional space is bounded by the corresponding "finite-dimensional risk" plus an additional term representing the price to pay when projecting onto $\mathcal{S}_d$. Yet, the first term in the right-hand side of inequality (2) above is known to tend to 0 when n goes to infinity. More precisely, as $\mathbb{P}\{\|\Pi_d(X)\| \le R\} = 1$, we have

$$\mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^* \le \frac{Ck}{\sqrt{n}},$$

where $C = 12R^2$ (Biau, Devroye, and Lugosi [6]). In our setting, to keep the same rate of convergence $O(1/\sqrt{n})$ in spite of the extra term $8R^2/\varphi_d$, φ_d must be of the order $\sqrt{n}$. For Sobolev ellipsoids (Example 2.1), where $\varphi_j \geq (j-1)^{2\beta}$, this means a dimension d of the order $n^{1/4\beta}$. When $\varphi_j = j^r e^{\alpha j}$, the rate of convergence is $O(1/\sqrt{n})$ as long as d is chosen of the order $\ln n/(2\alpha)$. In the RKHS context (Example 2.2), consider the case of eigenvalues $\{\lambda_j, j \geq 1\}$ with polynomial or exponential-polynomial decay, which covers a broad range of kernels (Williamson, Smola, and Schölkopf [38]). If $\lambda_j = O(j^{-(\alpha+1)}), \alpha > 0$, then $1/\varphi_d = O(d^{-(\alpha+1)})$, and d must be of the order $n^{1/(2\alpha+2)}$, whereas $\lambda_j = O(e^{-\alpha j^p}), \alpha, p > 0$, leads to a projection dimension d of the order $(\ln n/(2\alpha))^{1/p}$. Obviously, the upper bound (2) is better for large φ_d, and consequently large d. Nevertheless, from a computational point of view, the projection dimension should not be chosen too large.

Remark 2.1. Throughout, we assumed that $\mathbb{P}\{\|X\| \leq R\} = 1$. This requirement, called the peak power constraint, is standard in the clustering and signal processing literature. We do not consider in this paper the case where this assumption is not satisfied, which is feasible but leads to technical complications (see Merhav and Ziv [29], Biau, Devroye, and Lugosi [6] for results in this direction). Besides, the number of clusters is assumed to be fixed throughout the paper. Several methods for estimating k have been proposed in the literature (see, e.g., Milligan and Cooper [31] and Gordon [20]).

As already mentioned, the subset of coefficients $\mathcal{S}$ is intimately connected to the underlining Hilbertian basis. As a consequence, all the results presented strongly depend on the orthonormal system considered. Therefore, the choice of a proper basis is crucial and is discussed in the next section.

3 Basis selection

Wavelet packet best basis algorithm In this section, we describe an algorithm searching for the best projection basis among a "library". If $\{\psi_\alpha, \alpha \in I\} \subset L^2([0,1])$ is a collection of elements in $L^2([0,1])$ which span $L^2([0,1])$ and allow to build several different bases by choosing various subsets $\{\psi_\alpha, \alpha \in I_\beta\} \subset L^2([0,1])$, the collection of bases built this way is called a library of bases. Here, I is some index set, and β runs over some other index set.

More specifically, we focus on the best basis algorithm of Coifman and Wickerhauser [9] (see also Wickerhauser [37]), which yields an optimal basis among a library of wavelet packets. Wavelets are functions which cut up a signal into different frequency components to study each component with a resolution matched to its scale. Unlike the Fourier basis, wavelets are localized both in time and frequency. Hence, they have advantages over traditional Fourier methods when the signal contains discontinuities as well as noise. For detailed expositions of the mathematical aspects of wavelets, see the books of Daubechies [11], Mallat [27] and Meyer [30].

Let the sequence of functions $(\psi_\nu)_{\nu \geq 0}$ be defined by

$$\psi_0(t) = H\psi_0(t), \quad \int_{\mathbb{R}} \psi_0(t)dt = 1,$$

$$\psi_{2\nu}(t) = H\psi_\nu(t) = \sqrt{2} \sum_{p \in \mathbb{Z}} h(p)\psi_\nu(2t - p),$$

$$\psi_{2\nu+1}(t) = G\psi_\nu(t) = \sqrt{2} \sum_{p \in \mathbb{Z}} g(p)\psi_\nu(2t - p),$$

where H and G are orthogonal quadrature filters, i.e., convolution-decimation operators satisfying some algebraic properties (see, e.g., Wickerhauser [37]). Let Λ_ν denote the closed linear span of the translates $\psi_\nu(\cdot - p), p \in \mathbb{Z}$, of ψ_ν, and

$$\sigma^s \Lambda_\nu = \{2^{-s/2} x(2^{-s}t), x \in \Lambda_\nu\}.$$

To every such subspace of $L^2(\mathbb{R})$ corresponds a dyadic interval

$$I_{s\nu} = \left[\frac{\nu}{2^s}, \frac{\nu+1}{2^s} \right[.$$

For all (s, ν), these subspaces give an orthogonal decomposition

$$\sigma^s \Lambda_\nu = \sigma_{s+1} \Lambda_{2\nu} \oplus \sigma_{s+1} \Lambda_{2\nu+1}.$$

Observe that for $\nu = 0, \ldots, 2^s - 1$, the $I_{s\nu}$ are dyadic subintervals of $[0, 1[$.

The next proposition provides a library of orthonormal bases built with functions of the form $\psi_{s\nu p} = 2^{-s/2}\psi_\nu(2^{-s}t - p)$, called wavelet packets of scale index s, frequency index ν and position index p.

Proposition 3.1 (Wickerhauser [37]). *If $s \leq L$ for some finite maximum L, H and G are orthogonal quadrature filters and $\mathcal{I}$ is a collection of disjoint dyadic intervals whose union is $\mathbb{R}^+$, then $\mathcal{B}_{\mathcal{I}} = \{\psi_{s\nu p}, p \in \mathbb{Z}, I_{s\nu} \in \mathcal{I}\}$ is an orthonormal basis for $L^2(\mathbb{R})$. Moreover, if $\mathcal{I}$ is a disjoint dyadic cover of $[0, 1[$, then $\mathcal{B}_{\mathcal{I}}$ is an orthonormal basis of Λ_0.*

This construction yields orthonormal bases of $L^2(\mathbb{R})$. Some changes must be made to obtain bases of $L^2([0, 1])$. Roughly, they consist in considering not all scales and shifts, and adapting the wavelets which overlap the boundary of $[0, 1]$ (see for instance Cohen, Daubechies, and Vial [8]).

The library can be seen as a binary tree whose nodes are the spaces $\sigma^s \Lambda_\nu$ (Figure 3 and 4). An orthonormal basis is given by the leaves of some subtree. Figure 5 and 6 show two examples of bases which can be obtained in this way.

To define an optimal basis, a notion of information cost is needed. Coifman and Wickerhauser [9] propose to use Shannon entropy. In our context, the basis choice will be done with respect to some reference curve x_0 which has to be representative of the data. We compute, for each basis in the library, the Shannon entropy of the coefficients of x_0 in this basis, and select the basis minimizing this entropy. The

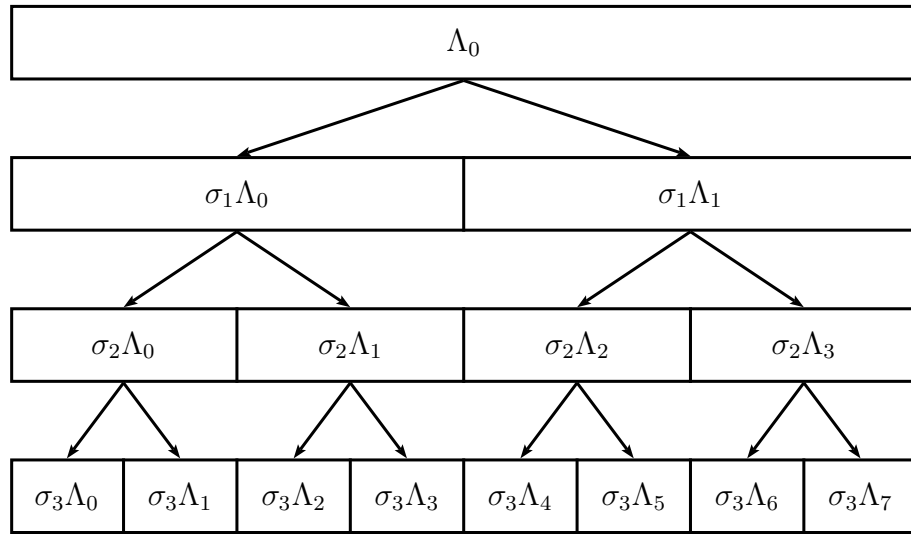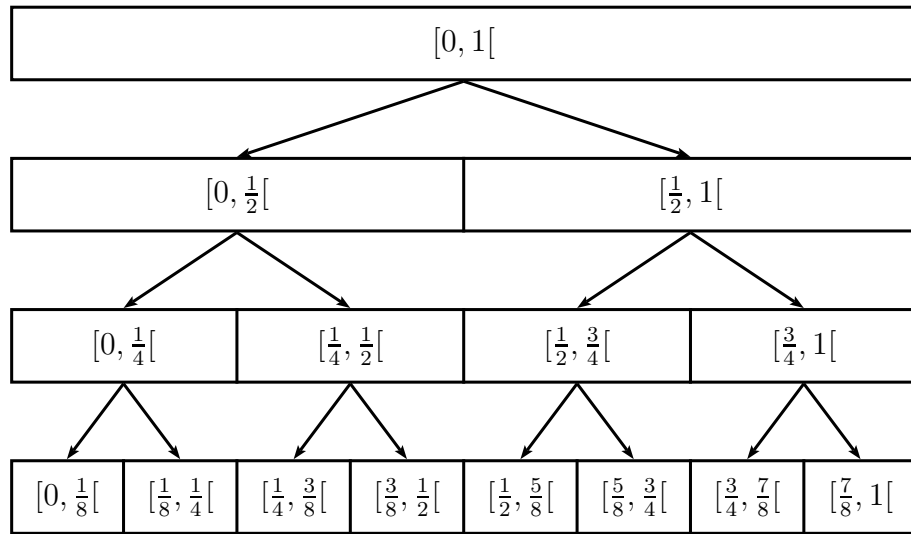Figure 3: Tree structure of wavelet packet bases.

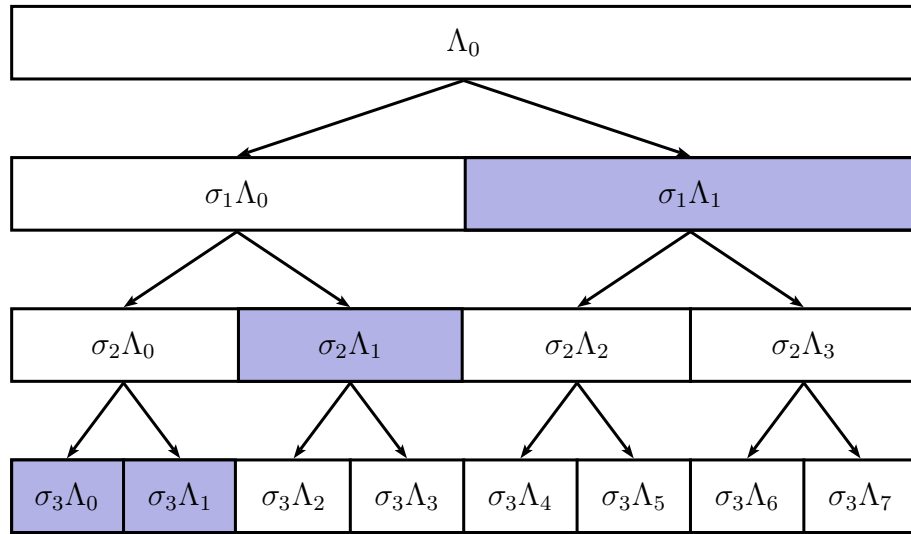Figure 4: Correspondence with dyadic covers of [0,1[.
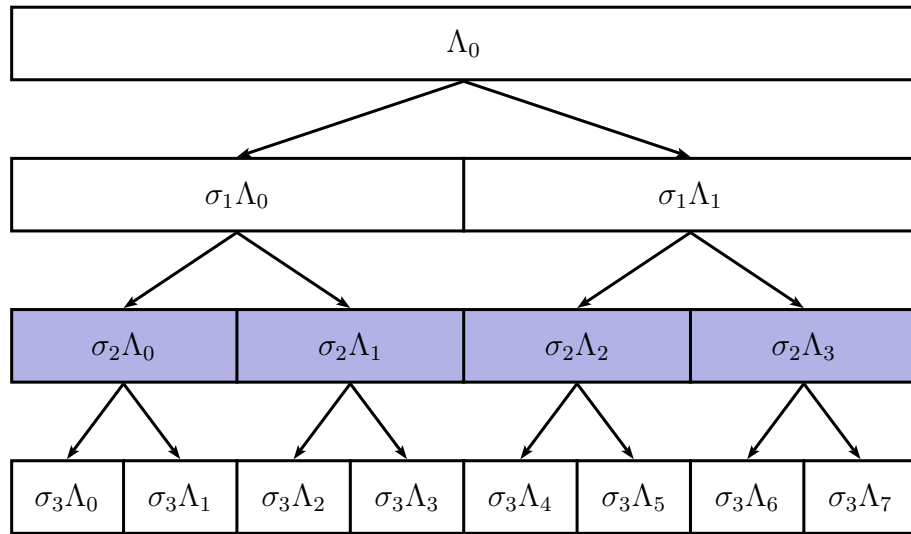
Figure 5: The wavelet basis.

Figure 6: An example of fixed level wavelet packet basis.

construction of this best basis, relying on the binary tree structure of the library, is achieved by comparing, at each node, starting from the bottom of the tree, parents with their two children. We decided to take for x_0 the median curve in the sample in the L^2 norm sense. Indeed, it is more likely to present characteristic behaviors of the other curves than the mean, because the mean curve is a smooth average representative, which is probably too easy to approximate with a few basis functions. However, we observed that these two functions surprisingly give rise to almost the same basis. Both choices are thus possible. The mean curve is useful if we know that some noise has to be removed, whereas the median curve seems a better choice to reflect the small-scale irregularities.

We have implemented the algorithm in **R**, and it has been run through all filters available in the **R** package *wmtsa*. These filters belong to four families, extremal phase family (Daubechies wavelets), including Haar basis, least asymmetric family (Symmlets), "best localized" wavelets and Coiflets. For example, the least asymmetric family contains ten different filters "s2", "s4", "s6", "s8", "s10", "s12", "s14", "s16", "s18", "s20". Finally, we keep the clustering result obtained with the basis minimizing the distortion among the various filters. In the sequel, this basis will be called Best-Entropy basis.

In the applications, the performance of the Best-Entropy basis will be compared with the Haar wavelet basis, the Fourier basis and the functional principal component analysis basis. For the sake of completeness, we recall here the definition of these bases.

The Haar wavelet basis   Let $\phi(t) = \mathbf{1}_{[0,1]}(t)$ and $\psi(t) = \mathbf{1}_{[0,1/2[}(t) - \mathbf{1}_{[1/2,1]}(t)$. Then, the family $\{\phi, \psi_{j,\ell}\}$, where

$$\psi_{j,\ell}(t) = 2^{j/2}\psi(2^j t - \ell), j \geq 0, 0 \leq \ell \leq 2^j - 1,$$

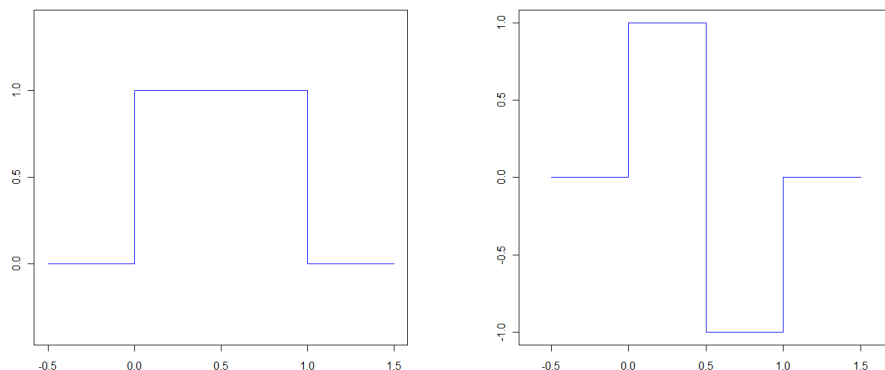constitutes a Hilbertian basis of $L^2([0,1])$, called Haar basis.

Figure 7: Haar scaling function ϕ and mother wavelet function ψ.

The Fourier basis   The Fourier basis on $[0,1]$ is the complete orthonormal system of $L^2([0,1])$ built with the trigonometric functions

$$\psi_1(t) = 1, \quad \psi_{2j}(t) = \sqrt{2}\cos(2\pi jt), \quad \psi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt), \quad j \geq 1.$$

Functional principal component analysis   Principal component analysis for functional data (for short, functional PCA) is the generalization of the usual principal component analysis for vector data. The Euclidean inner product is replaced by the inner product in $L^2([0,1])$. More precisely, functional PCA consists in writing $X_i(t)$ under the form

$$X_i(t) = \mathbb{E}X(t) + \sum_{j=1}^{+\infty} x_{ij}\psi_j(t),$$

where the $(x_{ij})_{j\geq 1}$ and the functions $(\psi_j)_{j\geq 1}$ are defined as follows. At the first step, the function ψ_1 is chosen to maximize

$$\frac{1}{n}\sum_{i=1}^{n} x_{i1}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\int \psi_1(t)X_i(t)dt\right]^2$$

subject to

$$\int \psi_1(t)^2 dt = 1.$$

Then, each ψ_j is computed by maximizing

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2$$

subject to

$$\int \psi_j(t)^2 dt = 1$$

and to the orthogonality constraints

$$\int \psi_\ell(t)\psi_j(t)dt = 0, \ 1 \leq \ell \leq j-1.$$

Functional PCA can be characterized in terms of the eigenanalysis of covariance operators. If $(\lambda_j)_{j\geq 1}$ denotes the eigenvalues and $(\psi_j)_{j\geq 1}$ the eigenfunctions of the operator $\mathcal{C}$ defined by $\mathcal{C}(f)(s) = \int_0^1 C(s,t)f(t)dt$, where $C(s,t) = \text{cov}(X(s),X(t))$, then

$$X_i(t) = \mathbb{E}X(t) + \sum_{j=1}^{+\infty} x_{ij}\psi_j(t),$$

where the x_j are uncorrelated centered random variables with variance $\mathbb{E}[x_{ij}^2] = \lambda_j$. There are similarities with the context of Example 2.2, but here the kernel depends on X. In practice, the decomposition can easily be computed with discrete matrix operations, replacing $C(s,t)$ by the covariance matrix of the X_i. This basis has some nice properties. In particular, considering a fixed number of coefficients, it minimizes among all orthogonal bases the average squared L^2 distance between the original curve and its linear representation (see, e.g., the book of Ghanem and Spanos [19]).

15

For more details on functional PCA, we refer the reader to Ramsay and Silverman [32].

Observe that since the functional PCA basis is a stochastic basis and the Best-Entropy basis algorithm also uses the data, rigorously the sample should be divided into two subsamples, one to build the basis, and the other for clustering.

4 Experimental results and analysis

We evaluated the performance of the clustering method with projection, using the various bases described in the previous section, for two different kinds of curves. First, a simulated example where the right clusters are known is discussed, to illustrate the efficiency of the method. Then, we focus on our industrial problem and cluster output curves of a "black box" computer code.

Observe that, although the curves considered in Section 2 and Section 3 were true functions, in practice, we have to deal with curves sampled on a finite number of discretization points. Therefore, a preprocessing step based on spline interpolation is necessary.

4.1 Synthetic control chart time series

Control chart time series are used for monitoring process environments, to achieve appropriate control and to produce high quality products. Different types of series can be encountered, but only one, a kind of white noise, indicates a normal working. All the other types of series must be detected, because they correspond to abnormal behavior of the process.

The data set contains a few hundreds to a few thousands curves generated by the process described in Alcock and Manolopoulos [2], discretized on 128 time points. There are six types of curves: normal, cyclic, increasing trend, decreasing trend, upward shift and downward shift, which are represented in Figure 8. The equations which generated the data are indicated below.

(A) *Normal pattern*: $y(t) = m + rs$ where $m = 30$, $s = 2$ and $r \sim \mathcal{U}(-3, 3)$.

(B) *Cyclic pattern*: $y(t) = m + rs + a \sin \frac{2\pi t}{T}$ where $a, T \sim \mathcal{U}(10, 15)$.

(C) *Increasing shift*: $y(t) = m + rs + gt$ with $g \sim \mathcal{U}(0.2, 0.5)$.

(D) *Decreasing shift*: $y(t) = m + rs - gt$.

(E) *Upward shift*: $y(t) = m + rs + hx$ where $x \sim \mathcal{U}(7.5, 20)$, $h = \mathbf{1}_{[t_0, D]}$, $t_0 \sim \mathcal{U}\left(\frac{D}{3}, \frac{2D}{3}\right)$, and D is the number of discretization points.
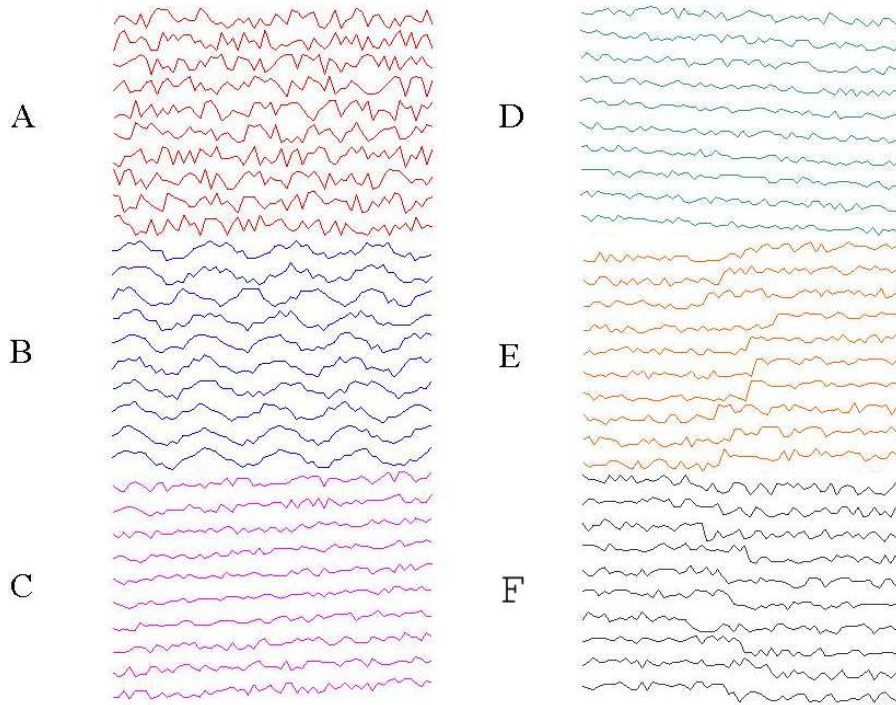
(F) *Downward shift*: $y(t) = m + rs - hx$.

Figure 8: 10 example curves for each of the 6 types of control chart.

The two main advantages using this synthetic data set is that we can simulate as many curves as we wish and we know the right clusters.

The k-means algorithm on the projected coefficients has been run for all four bases (Best-Entropy, Haar, functional PCA and Fourier basis), with varying sample size n and projection dimension d. Since the result of a k-means algorithm may depend on the choice of the initial centers, the algorithm is restarted 100 times. The maximum number of iterations per run is set to 500. This program tries to globally minimize the projected empirical distortion $W_{d,n}(\mathbf{c})$. To evaluate its performance, we compute an approximation $W(d, n)$ of the distortion $W_\infty(\hat{\mathbf{c}}_{d,n})$ using a set of 18000 sample curves. This is possible in this simulated example, since we can generate as many curves as we want. The distortion $W(d, n)$ is computed for d varying from 2 to 50 and n ranging from 100 to 3100. We restricted ourselves to the case $d \leq 50$, since there are only 128 discretization points. Moreover, as pointed out earlier, d must not be too large for computational complexity reasons. Indeed, a projection dimension $d = 50$ is already high for a practical use.

Figures 9 and 10 show the contours plots corresponding to the evolution of $W(d, n)$ as a function of d and n, for the functional PCA and the Haar basis. We remark that the norm of the gradient of $W(d, n)$ vanishes when d and n are close to their maximal values. Hence, as expected according to Theorem 2.1, $W(d, n)$ is decreasing in d and n. When d or n is too small (for instance $d = 2$ or $n = 100$), the clustering results are inaccurate. Besides, they are not stable with respect to the n observations chosen. However, for larger values of these parameters, the partitions obtained quickly become satisfactory. The choice $n \geq 300$ together with $d \geq 6$ generally
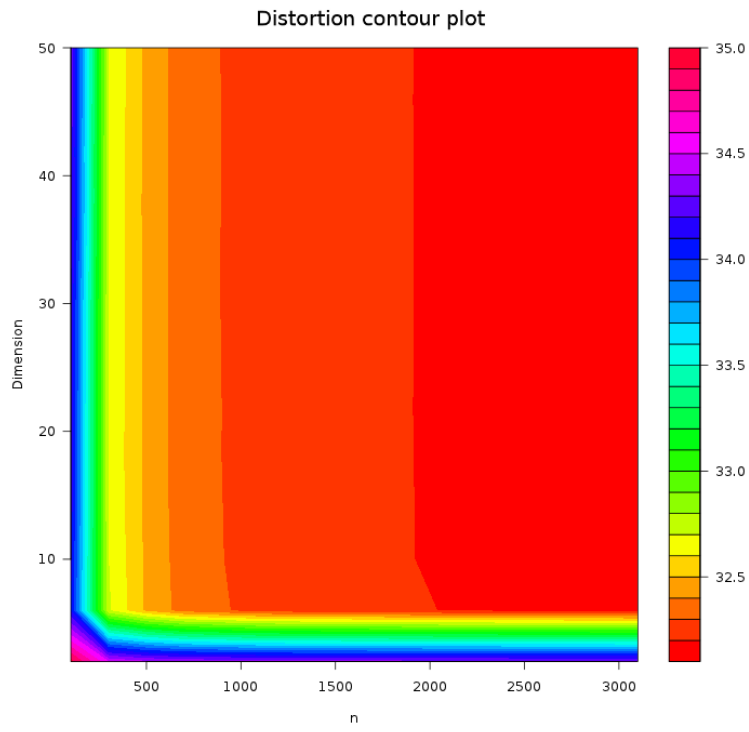
17

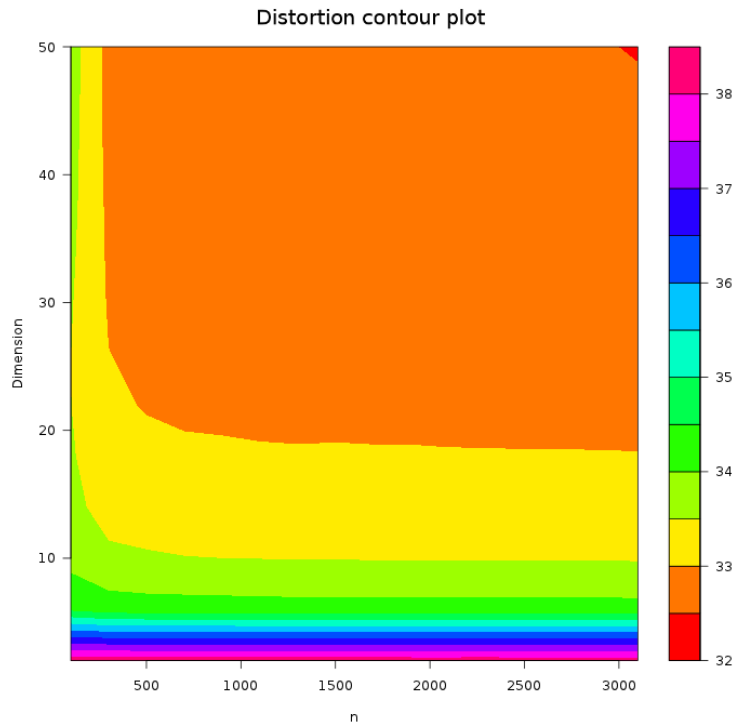Figure 9: Contour plot of $W(d, n)$ for the functional PCA basis.

Figure 10: Contour plot of $W(d, n)$ for the Haar basis.

provides good results and is reasonably low for many applications.

Figure 11 shows the curve corresponding to the evolution of $W(d, n)$ as a function of d for $n = 500$, for all four bases, whereas Figure 12 represents the evolution of $W(d, n)$ versus n for $d = 10$. According to Section 2, φ_d must be of the order $\sqrt{n}$. For $n = 500$, $\sqrt{n}$ is about 22. Considering that φ_d and d are approximately of the same order, a projection dimension close to 22 should thus be suitable. Indeed, we see via Figure 11 that $W(d, n)$ does not decrease much more after this value.
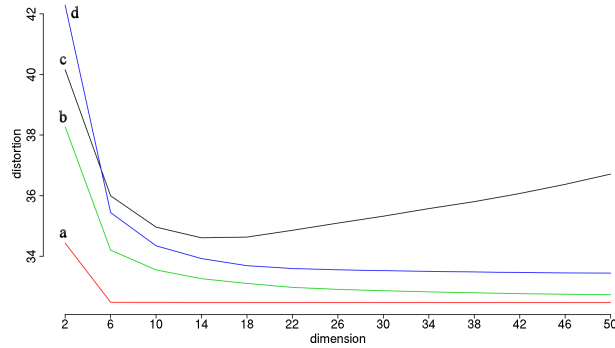

Figure 11: Evolution of $W(d, n)$ for $n = 500$ and d ranging from 2 to 50, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.
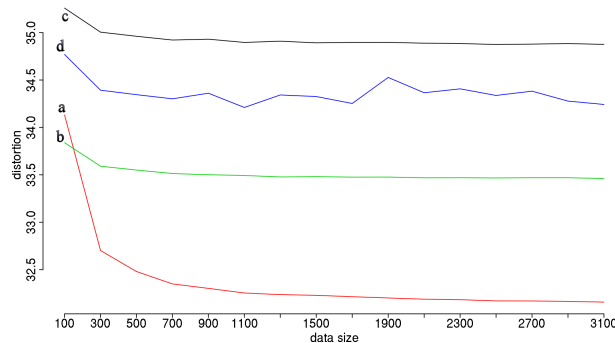

Figure 12: Evolution of $W(d, n)$ for $d = 10$ and n ranging from 100 to 3100, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

The evolution of the distortion for the Fourier basis looks quite odd: it shows a first decreasing step before increasing again. However, this increasing can be explained in the following way. The centers chosen first are wrong, but seem to give a better distortion than the "real" clustering. When the dimension grows large enough, these first wrong centers no longer represent a local minimum, and the k-means algorithm moves slowly toward the "right" clustering, although losing a bit in distortion. This interpretation is confirmed if we look at the clusters corresponding to each distortion. Furthermore, when the dimension is high relatively to the number of discretization

points, some basis functions which oscillate a lot may not be sampled correctly. As a result, the coefficients estimated by approximating the inner products can become very inaccurate as d increases. For very high d, these computed coefficients confound with some noise. Consequently, data becoming more noisy without adding any information, the distortion will increase. The other bases tend to oscillate too, so that they would probably show the same behavior if d were increased above 50. Besides, the small fluctuations observed for the Best-Entropy basis indicate that this basis is not suitable for clustering of control chart time series. The functional PCA basis always gives the lowest distortion. However, the distortions obtained for the three other bases are quite similar, with a preference for the Haar basis over the Best-Entropy wavelet basis, the Fourier basis being the worst choice. As an example, Table 1 gives the values of $W(d,n)$ for $n = 1100$ and $d = 30$.

Basis	Fourier	Functional PCA	Haar	Best-Entropy
Distortion	35.3	32.3	32.8	33.4

Table 1: Distortion $W(d,n)$ for $n = 1100$ and $d = 30$.

Figure 13 represents the 6 clusters for the Fourier basis, for $n = 300$ and $d = 10$, whereas Figure 14 shows them for $d = 30$. The classes obtained with the algorithm are shown in colors, and the real clusters are indicated in the caption. For relatively small values of d, the normal and cyclic patterns are merged into one big cluster, and one cluster corresponding to increasing (or decreasing) shift pattern is split in two. For large enough d, the normal and cyclic designs are well detected, and the overall clustering is correct despite some mixing increasing-upward shift or decreasing-downward shift. We also tested the algorithm for smaller values of the number k of classes. As expected, for the particular choice $k = 3$, clusters A and B are merged into one single group, and the same occurs for the cluster pairs $\{C, E\}$ and $\{D, F\}$.
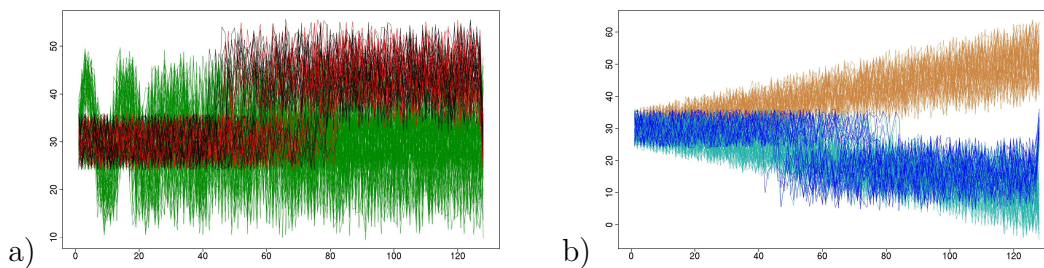

Figure 13: (a) Clusters A and B in green, cluster E in red and black. (b) Clusters C, D and F in brown, light blue and blue respectively. (Fourier basis, $n = 300$, $d = 10$.)
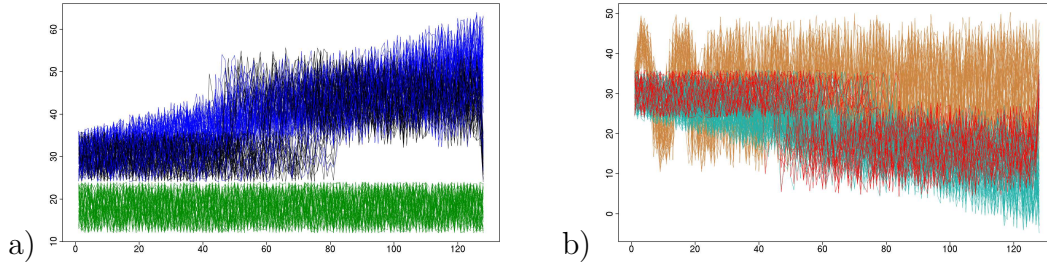
20

Figure 14: (a) Clusters A, C and E in green, blue, black. (b) Clusters B, D and F in brown, light blue and red. (Fourier basis, $n = 300$, $d = 30$.)

4.2 Industrial code examples

Let us now turn to the industrial issue which motivated our study. As explained in the introduction, the computer codes used in nuclear engineering have become very complex and costly in CPU-time consumption. That is why we try to approximate them with a cheap function substituted to the code. In order to build a regression model, a preliminary analysis of the different types of outputs is essential. This leads to data clustering, applied here to a computer code with functional outputs. Two different kinds of outputs are presented, the temperature evolution with a data set containing 100 curves, and the thermal exchange coefficient evolution with a data set of 200 curves.

Temperature curves   The data is made of 100 CATHARE code outputs representing the evolution of the temperature in the vessel annular space (Figure 15). Here, the sample size is fixed to $n_0 = 100$. However, the discretization can be controlled to some extent with spline interpolation. In this case, 256 discretization points are used.
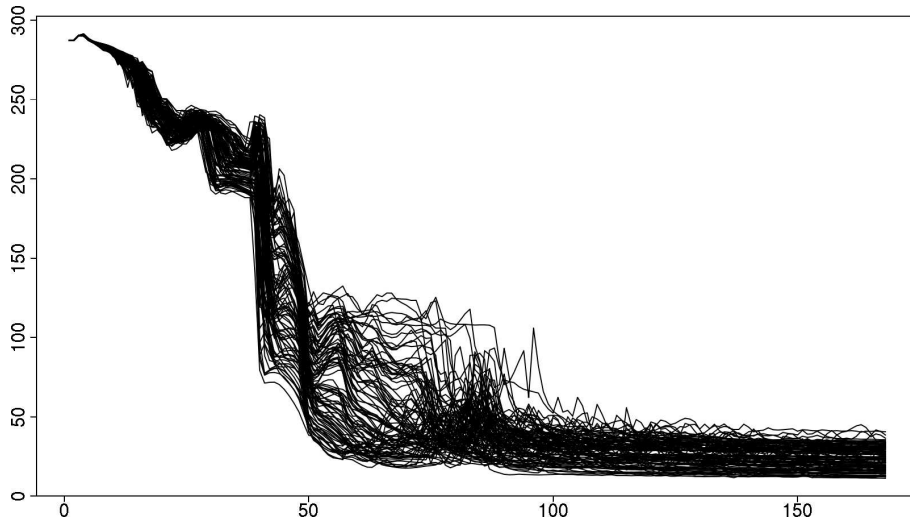

Figure 15: The 100 temperature curves.

Observe that all curves converge in the long-time limit to the same value, corresponding to the temperature of the cold water injected. These curves have been clustered, for physical reasons pertaining to nuclear engineering, in two groups. More precisely, there is a critical set of physical parameters beyond which the thermal shock is more violent and the temperature changes more rapidly (see Auder, De Crecy, Iooss, and Marquès [5] for more details). The algorithm on the projected coefficients has been run for the Best-Entropy, Haar, functional PCA and Fourier bases, with varying dimension, with the same settings as before. Since we consider real-life data, it is not possible to compute an approximation of $W_\infty(\hat{\mathbf{c}}_{d,n})$ as in the simulated example. Hence, the distortion $W(d, n_0)$ is simply the output $W_{d,n_0}(\hat{\mathbf{c}}_{d,n_0})$ of the clustering algorithm, with fixed $n_0 = 100$. This distortion is computed for d varying from 2 to 50. Figure 16 shows the curve corresponding to the evolution of $W(d, n)$ as a function of d. As expected, it is decreasing in d.
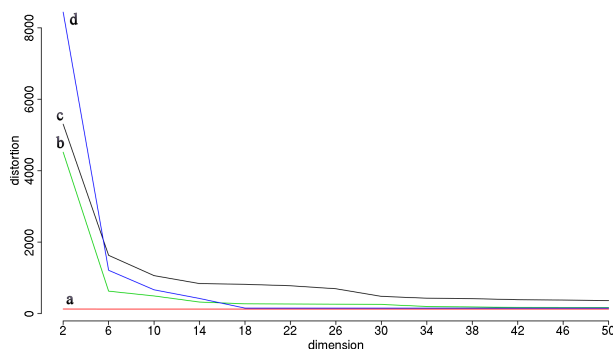

Figure 16: Evolution of $W(d, n_0)$ for the 100 temperature curves, d ranging from 2 to 50, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

Basis	Fourier	Functional PCA	Haar	Best-Entropy
Distortion	480.8	125.3	254.9	151.5

Table 2: Distortion values for $d = 30$.

We note that until $d = 16$, the Haar basis provides lower distortion, but for larger values of d, the Best-Entropy basis is better. As before, the Fourier basis is the worst and the functional PCA basis is the best. This can also be checked from Table 2, which presents the distortion obtained for each basis. Although the functional PCA basis gives the best result in terms of distortion, we see that using any of the other three bases is not that bad. Indeed, the same partitioning is found every time (Figure 17).
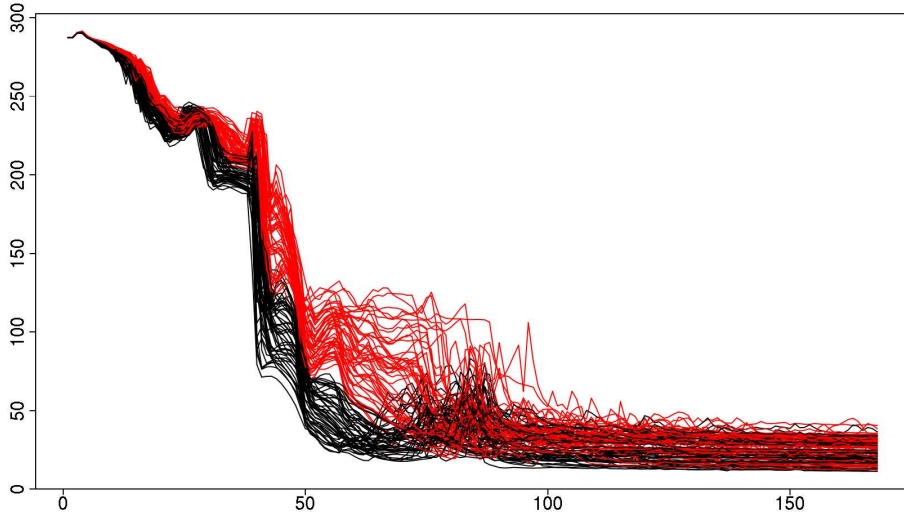
Figure 17: The temperature curves divided in two groups.

Finally, Figure 18 shows the two centers representing the classes obtained for $d = 30$ with the Fourier basis, functional PCA basis, Haar basis and Best-Entropy basis. The two curves obtained with the functional PCA basis characterize with an especially good accuracy the shape of the data items in the corresponding clusters.
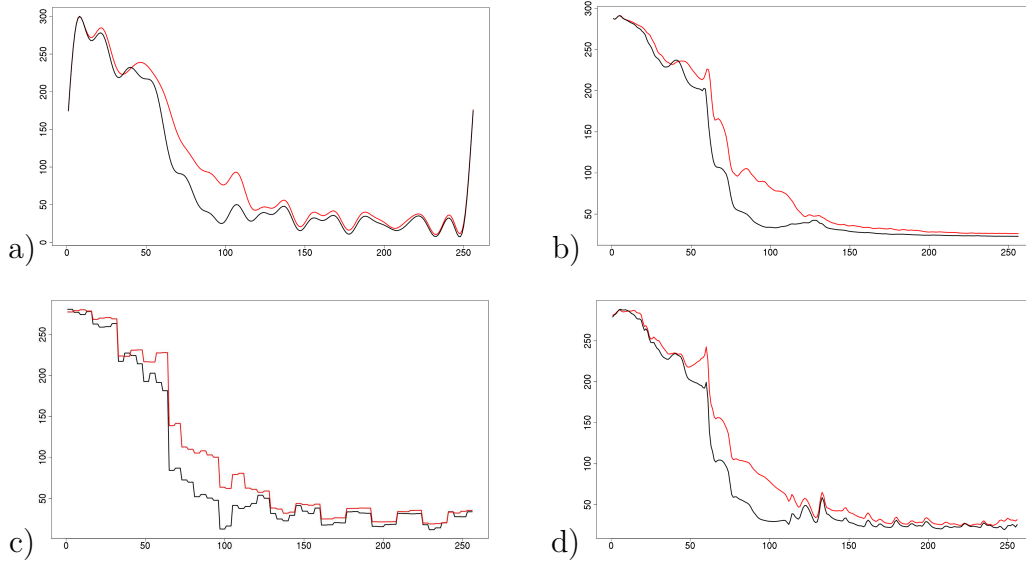

Figure 18: The two centers for $d = 30$ for (a) Fourier, (b) functional PCA, (c) Haar and (d) Best-Entropy basis.

Thermal exchange coefficient curves   Figure 19 shows all 200 CATHARE code outputs. Here, the number of discretization points is set to 1024. The data has been partitioned in three groups. As for the temperature curves, $W(d, n_0)$ is computed for d varying from 2 to 50 ($n_0 = 200$).
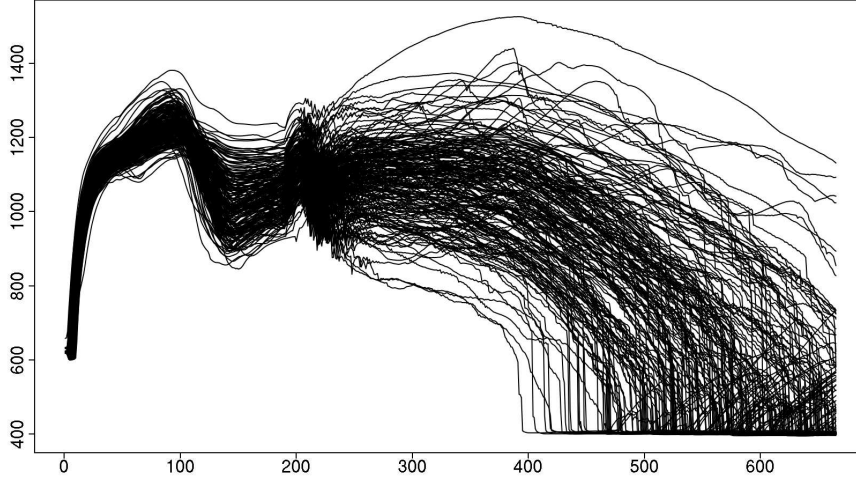
Figure 19: The 200 thermal exchange coefficient curves.

We can see via Figure 20 that $W(d, n_0)$ is decreasing in d again. The functional PCA basis is still the best choice, with a fast convergence (stabilization from $d = 10$). Interestingly, the Fourier basis shows smaller distortion values than the two wavelets basis in this case, as Table 3 indicates.
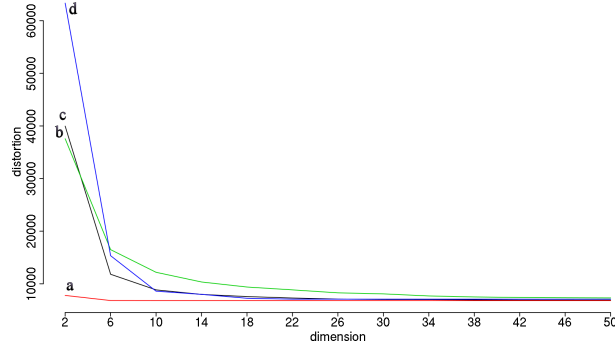

Figure 20: Evolution of $W(d, n_0)$ for the 200 thermal exchange coefficient curves, d ranging from 2 to 50 for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

Basis	$d = 6$	$d = 10$	$d = 18$	$d = 26$	$d = 34$	$d = 42$	$d = 50$
Fourier	11810	8843.8	7570.0	7097.0	7013.2	6893.3	6881.2
Functional PCA	6815.8	6802.0	6801.4	6801.2	6801.1	6801.1	6801.0
Haar	16491	12185	9385.2	8279.3	7688.0	7390.6	7313.9
Best-Entropy	15338	8596.2	7244.5	7082.1	7082.1	7082.1	6801.0

Table 3: Distortion values for the thermal exchange coefficient curves.

All the partitions obtained are very similar. A typical example is given by Figure 21. However, as for the temperature curves, it is interesting to look at the curves

selected as centers. Figure 22 shows the three centers obtained with the Fourier basis, functional PCA basis, Haar basis and Best-Entropy basis, for $d = 30$.
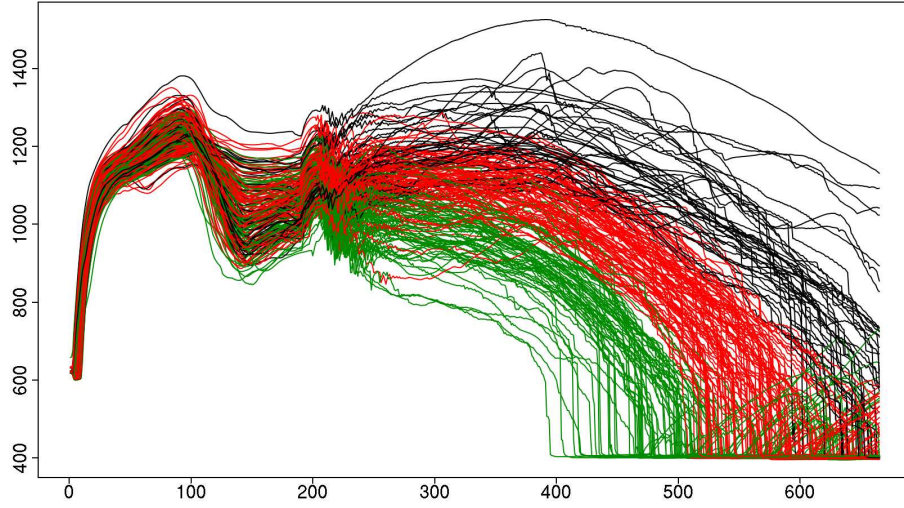

Figure 21: Three clusters obtained with $d = 14$, functional PCA basis.
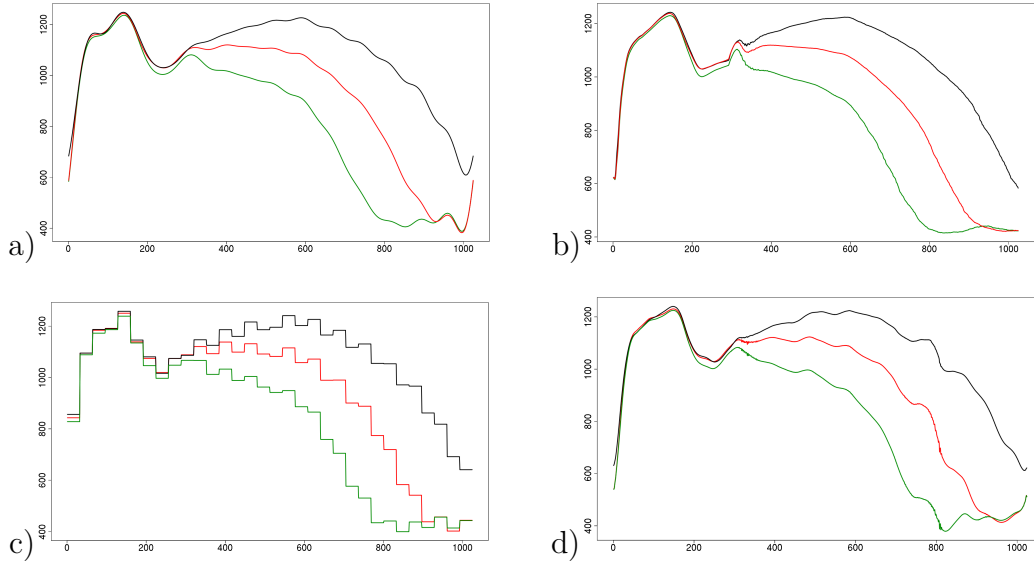

Figure 22: The three centers with (a) Fourier, (b) functional PCA, (c) Haar and (d) Best-Entropy basis, for $d = 30$.

5 Conclusion

These clusters allow to build accurate models for the industrial application. The partitioning method presented in this article has been integrated in our metamodel written in **R**. More specifically, given an array of n input vectors corresponding to n output curves, the purpose is to learn a function $\phi : z \mapsto x$ mapping an input vector to a continuous curve. In order to improve the accuracy of this task, we begin with

25

a clustering step and then look for a regression model in each cluster separately. The metamodel lets the user choose between several clustering techniques, either assuming some clusters shapes (like this projected k-means) or trying to discover them in data (like the ascendant hierarchical clustering). The latter are attractive as they do not make assumptions about the results, but they generally need a relatively good sampling of the data. Consequently, the k-means-like techniques are useful in many of our industrial applications, where only a few samples are available. Moreover, these methods provide easily interpretable clusters. In each cluster, after a dimension reduction step, which can either be achieved through the decomposition on an orthonormal basis (linear), or any manifold learning algorithm (nonlinear, with the assumption that the outputs lie on a functional manifold), a statistical learning method is applied to predict representation within this cluster. The mainly used method at this stage is the Projection Pursuit Regression algorithm (see Friedman, Jacobson, and Stuetzle [16]). Finally, a simple k-nearest neighbors classifier gives the most probable cluster for a new input, the corresponding regression function is applied, and the curve can be reconstruct from its predicted representation.

For the moment, our metamodel with the clustering method presented have successfully been used on two different scenarios involving the CATHARE code (minor or major break, for each we get temperature, pressure and thermal exchange coefficient curves).

Another research track could consist in considering other types of distances between curves. Distances involving derivatives might be hard to estimate on the thermal exchange coefficient dataset, because several curves are varying rapidly over short period of time, contrasting for instance with the Tecator dataset (`http://lib.stat.cmu.edu/datasets/tecator`), on which such distances proved successful (Ferraty and Vieu [15, Chapter 8], Rossi and Villa [34]). However, further investigations are needed to know if a smoothing step before clustering based on m-order derivatives would lead to improved results. As the "true" classes are unknown, such a procedure can only be validated within a cross validation framework involving the full metamodel. Experiments with the L^1 distance or some mixed distances related to functions shapes (Heckman and Zamar [22]) could also be studied in future research.

6 Proofs

6.1 Proof of Lemma 2.1

If we define the remainder R_d by $R_d(\mathbf{x}) = \mathbf{x} - \Pi_d(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$, then for $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$,

$$
\begin{aligned}
\|R_d(\mathbf{x} - \mathbf{y})\|^2 &\leq 2\|R_d(\mathbf{x})\|^2 + 2\|R_d(\mathbf{y})\|^2 \\
&= 2\sum_{j=d}^{+\infty} x_j^2 + 2\sum_{j=d}^{+\infty} y_j^2 \\
&= 2\sum_{j=d}^{+\infty} \frac{\varphi_j x_j^2}{\varphi_j} + 2\sum_{j=d}^{+\infty} \frac{\varphi_j y_j^2}{\varphi_j} \\
&\qquad (\varphi_j > 0 \text{ for all } j \geq d, \text{ since } d > j_0) \\
&\leq 2\sum_{j=d}^{+\infty} \frac{\varphi_j x_j^2}{\varphi_d} + 2\sum_{j=d}^{+\infty} \frac{\varphi_j y_j^2}{\varphi_d} \\
&\leq \frac{4R^2}{\varphi_d}.
\end{aligned}
$$

Thus, for $\mathbf{c} \in \mathcal{S}^k$,

$$
\begin{aligned}
W_\infty(\mathbf{c}) - W_d(\mathbf{c}) &= \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|X - c_\ell\|^2 - \min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2 \right] \\
&= \mathbb{E}\left[\min_{\ell=1,\ldots,k} \|\Pi_d(X) + R_d(X) - \Pi_d(c_\ell) - R_d(c_\ell)\|^2 \right. \\
&\qquad \left. - \min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2 \right] \\
&= \mathbb{E}\left[\min_{\ell=1,\ldots,k} \left(\|\Pi_d(X) - \Pi_d(c_\ell)\|^2 + \|R_d(X) - R_d(c_\ell)\|^2 \right) \right. \\
&\qquad \left. - \min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2 \right] \\
&\qquad (\text{since } \Pi_d \text{ is the orthogonal projection on } \mathbb{R}^d) \\
&\leq \mathbb{E}\left[\max_{\ell=1,\ldots,k} \|R_d(X) - R_d(c_\ell)\|^2 \right] \\
&\leq \frac{4R^2}{\varphi_d}.
\end{aligned}
$$

Hence,

$$
\sup_{\mathbf{c} \in \mathcal{S}^k} \left[W_\infty(\mathbf{c}) - W_d(\mathbf{c}) \right] \leq \frac{4R^2}{\varphi_d},
$$

as desired.

6.2 Proof of Theorem 2.1

We have

$$
W_\infty(\hat{\mathbf{c}}_{d,n}) - W_\infty^* = W_\infty(\hat{\mathbf{c}}_{d,n}) - W_d(\hat{\mathbf{c}}_{d,n}) + W_d(\hat{\mathbf{c}}_{d,n}) - W_d^* + W_d^* - W_\infty^*.
$$

27

According to Lemma 2.1, on the one hand,

$$W_\infty(\hat{\mathbf{c}}_{d,n}) - W_d(\hat{\mathbf{c}}_{d,n}) \leq \sup_{\mathbf{c} \in \mathcal{S}^k}[W_\infty(\mathbf{c}) - W_d(\mathbf{c})]$$
$$\leq \frac{4R^2}{\varphi_d},$$

and on the other hand,

$$W_d^* - W_\infty^* = \inf_{\mathbf{c} \in \mathcal{S}^k} W_d(\mathbf{c}) - \inf_{\mathbf{c} \in \mathcal{S}^k} W_\infty(\mathbf{c})$$
$$\leq \sup_{\mathbf{c} \in \mathcal{S}^k}[W_\infty(\mathbf{c}) - W_d(\mathbf{c})]$$
$$\leq \frac{4R^2}{\varphi_d},$$

and the theorem is proved.

References

[1] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised Curve Clustering using B-Splines. *Scandinavian Journal of Statistics*, 30:581–595, 2003.

[2] R. J. Alcock and Y. Manolopoulos. Time-series similarity queries employing a feature-based approach. In *7th Hellenic Conference on Informatics*, Ioannina, Greece, 1999.

[3] A. Antoniadis, X. Brossat, J. Cugliari, and J-M. Poggi. Clustering functional data with wavelets. In Y. Lechevallier and G. Saporta, editors, *e-Book COMP-STAT 2010*, pages 697–704. Springer, 2010.

[4] N. Aronszajn. Theory of reproducing kernel. *Transactions of American Mathematical Society*, 68:337–404, 1950.

[5] B. Auder, A. De Crecy, B. Iooss, and M. Marquès. Screening and metamodeling of computer experiments with functional outputs. Application to thermal-hydraulic computations. `http://hal.archives-ouvertes.fr/docs/00/52/54/91/PDF/ress_samo10_BA.pdf`.

[6] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 2008.

[7] J. M. Chiou and P. L. Li. Functional Clustering of Longitudinal Data. In S. Dabo-Niang and F. Ferraty, editors, *Functional and Operatorial Statistics*, Contributions to Statistics, chapter 17, pages 103–107. Physica-Verlag, Heidelberg, 2008.

[8] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1:54–81, 1993.

[9] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

[10] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

[11] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[12] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer-Verlag, Berlin, 1993.

[13] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26:879–921, 1998.

[14] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, second edition, 2000.

[15] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer-Verlag, New York, 2006.

[16] J. H. Friedman, M. Jacobson, and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–846, 1981.

[17] S. Gaffney. *Probabilistic Curve-Aligned Clustering and Prediction with Mixture Models*. PhD thesis, Department of Computer Science, University of California, Irvine, 2004. http://www.ics.uci.edu/~sgaffney/papers/sgaffney_thesis.pdf.

[18] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, 1992.

[19] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.

[20] A. D. Gordon. *Classification*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, London, second edition, 1999.

[21] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2000.

[22] N.E. Heckman and R.H. Zamar. Comparing the shapes of regression functions. *Biometrika*, 87:135–144, 2000.

[23] G. M. James and C. A. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98:397–408, 2003.

[24] T. Linder. Learning-Theoretic Methods in Vector Quantization. In L. Györfi, editor, *Principles of Nonparametric Learning*, CISM lecture Notes. Springer, Wien, New York, 2002.

[25] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[26] W. L. Loh. On latin hypercube sampling. *The Annals of Statistics*, 24:2058–2080, 1996.

[27] S. Mallat. *A wavelet tour of signal processing, The Sparse Way.* Academic Press, Burlington, third edition, 2009.

[28] M. D. McKay, W. J. Conover, and R. J. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

[29] N. Merhav and J. Ziv. On the amount of statistical side information required for lossy data compression. *IEEE Transactions on Information Theory*, 43, 1997.

[30] Y. Meyer. *Wavelet and Operators.* Cambridge University Press, Cambridge, 1992.

[31] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–79, 1985.

[32] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, New York, second edition, 2006.

[33] F. Rossi, B. Conan-Guez, and A. El Golli. Clustering functional data with the som algorithm. In *ESANN'2004 proceedings - European Symposium on Artificial Neural Networks*, pages 305–312, Bruges (Belgium), 28-30 April 2004.

[34] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69:730–742, 2006.

[35] A. B. Tsybakov. On the best rate of adaptive estimation in some inverse problems. *Comptes-rendus de l'Académie des Sciences, Paris*, 2000.

[36] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York, 2009.

[37] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software.* A. K. Peters/CRC Press, Wellesley, 1996.

[38] R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47:2516–2532, 2001.