

# CLUSTERING "OPTIMAL" DANS DES ESPACES FONCTIONNELS

Benjamin Auder<sup>[1]</sup>, Bertrand Iooss<sup>[2]</sup>

[1] : CEA, DEN, Centre de Cadarache, DER/SESI/LCFR,13108 Saint Paul lez Durance - UPMC Paris 6

[2] : CEA, DEN, Centre de Cadarache, DER/SESI/LCFR,13108 Saint Paul lez Durance

## Résumé

Dans l'industrie nucléaire, les codes de calcul simulant des phénomènes physiques sont de plus en plus complexes, et donc coûteux en temps de calcul. Nous nous intéressons au cas de codes dont les sorties sont des évolutions de paramètres physiques dans le temps. Ces dernières évolutions sont des courbes assimilables à des fonctions réelles d'un intervalle  $I \subset \mathbb{R}$  dans  $\mathbb{R}$ . Nous effectuons dans ce travail un prétraitement avant une analyse plus fine du code : le regroupement des courbes par similarités. Ce regroupement est également appelé clustering, ou classification non supervisée. Nous obtenons une estimation de l'erreur de généralisation attendue pour un clustering effectué sur  $N$  courbes. Une application avec six clusters distincts illustrera les résultats sur la vitesse de convergence de la qualité optimale théorique du clustering.

## Abstract

Computer codes used in support of nuclear industry are more and more complex, and consequently more and more CPU time consuming. We are here interested in such code, in the special case of functional output : the code output represents the evolutions of some physical parameters in time. Those last curves are functions from an interval  $I \subset \mathbb{R}$  to  $\mathbb{R}$ , which will be preprocessed in order to cluster them in a few meaningful groups (clustering, or unsupervised classification). The aim of our work is the estimation of the convergence speed of clustering error estimates. After finding bounds on convergence speeds, we will illustrate this on an example with six distinct groups of curves.

## 1 Introduction

Les courbes en sortie d'un code de calcul physique sont analysées par l'utilisateur du code, à des fins diverses : analyse de sensibilité, propagation d'incertitudes, calibration de code par exemple. L'un des objectifs de ce dernier utilisateur peut être de construire un modèle du code ("métamodèle"), simulant à moindre coût un grand nombre d'échantillons de courbes, facilitant donc beaucoup les analyses précitées. Notre travail tente d'apporter une aide directe pour ce dernier objectif, ainsi qu'une aide indirecte pour les autres analyses. En effet, regrouper les courbes suivant un certain critère de similarité permet d'isoler différents types de réponses du code, permettant d'établir un modèle plus fin.

Deux possibilités existent pour classer un jeu de  $N$  courbes  $y_1, \dots, y_N$  :

- Regrouper les courbes directement dans leur espace de départ ;

- Projeter les courbes sur une base de fonctions, puis déduire le clustering du regroupement effectué sur les coefficients de projection.

Nous étudions la seconde option, qui est attrayante car nettement moins coûteuse en nombre de calculs. Notre objectif est de déterminer la qualité de l'approximation réalisée alors. La mesure retenue pour évaluer la qualité du clustering en  $k$  groupes  $C_1, \dots, C_k$  distincts ( $k \in \mathbb{N}^*$ , supposé constant) est la *distorsion*  $W$ . Pour calculer  $W$ , on détermine d'abord un point de référence  $c_i$  appelé *centre* ou *centroïde* pour chaque  $C_i$ . Nous choisissons pour  $c_i$  le centre de gravité de  $C_i$  dans nos applications.  $W$  est alors égale à la moyenne des carrés des distances des échantillons  $y_i$  au centre le plus proche. Nous distinguons quatre cas, suivant que le clustering est effectué dans l'espace de départ ou en dimension finie  $d$ , et si nous considérons la distribution de probabilités des  $y_i$  ou seulement l'échantillon de  $N$  courbes. Nous montrons que sous certaines conditions les bornes inférieures de toutes ces quantités convergent assez rapidement vers la même valeur.

Dans un premier temps nous définissons toutes les quantités nécessaires, avant de donner les résultats théoriques de convergence, ainsi qu'une majoration des vitesses associées. Ensuite, un cas d'application artificiel sera présenté pour visualiser quelques résultats.

## 2 Notations et définitions

Nous supposons l'existence d'un générateur aléatoire  $\mathcal{G}$  de sorties de code notées  $y_i$ , suivant une loi de probabilité de densité  $g$ . La variable aléatoire correspondante à  $\mathcal{G}$  est noté  $Y$ . Les sorties  $y_i$  (déterministes) appartiennent a priori à l'ensemble des fonctions de  $I$  dans  $\mathbb{R}$   $\mathcal{F}(I, \mathbb{R})$ , mais ce cadre est beaucoup trop général. Nous supposons en fait l'hypothèse suivante  $H$  vérifiée dans toute la suite.

Fixons une base orthonormée  $\beta = (\beta_1, \dots, \beta_m, \dots)$  de  $L_2(I, \mathbb{R})$ ,  $(\varphi_j)_{j \in \mathbb{N}}$  une suite de réels strictement positifs divergente vers  $+\infty$ , et  $R$  un réel positif. Considérons alors le sous ensemble  $\Gamma^{(2)}(\varphi, R) \subset L_2(I, \mathbb{R})$  de fonctions  $f$  qui vérifient

$$\sum_{j=0}^{+\infty} \varphi_j \langle f, \beta_j \rangle^2 \leq R.$$

### **Hypothèse H :**

Les fonctions  $y_i$  ainsi que le support de la densité  $g$  sont contenus dans l'espace  $\Gamma^{(2)}(\varphi, R)$ ,  $\beta, \varphi$  et  $R$  étant fixés.

Cette dernière hypothèse revient à imposer plus de régularité à notre espace. C'est une contrainte peu restrictive en pratique, car les courbes observées s'inscrivent en général naturellement dans un tel espace moyennant un bon choix de la base  $\beta$  (dont  $\Gamma$  dépend principalement). Afin d'alléger les notations nous écrirons simplement  $\Gamma$  au lieu de  $\Gamma^{(2)}(\varphi, R)$ , et  $\Gamma_d = \Gamma \cap \text{Vect}(\beta_1, \dots, \beta_d)$ .

Notant  $c = (c_1, \dots, c_k) \in \Gamma^k$  et  $\pi_d$  la projection orthogonale sur  $\text{Vect}(\beta_1, \dots, \beta_d)$  dans  $\Gamma$ , nous sommes maintenant prêts pour définir les quatre types de distorsion mentionnés en introduction :

$$W_\infty(c) = \mathbb{E}_g \left[ \min_{j=1..k} \|Y - c_j\|^2 \right], \quad W_N(c) = \frac{1}{N} \sum_{i=1}^N \min_{j=1..k} \|y_i - c_j\|^2, \quad \text{et}$$

$$W_d(c) = \mathbb{E}_g \left[ \min_{j=1..k} \|\pi_d(Y) - \pi_d(c_j)\|^2 \right], \quad W_{d,N}(c) = \frac{1}{N} \sum_{i=1}^N \min_{j=1..k} \|\pi_d(y_i) - \pi_d(c_j)\|^2,$$

représentant respectivement la distorsion moyenne dans  $\Gamma$ , la distorsion empirique dans  $\Gamma$  puis les distorsions moyennes et empiriques dans l'espace projeté  $\Gamma_d$ , tout ceci en fonction du clustering défini par  $c \in \Gamma^k$ . Pour  $a \in \{d, N, \infty\}$ , on note la borne inférieure de la distorsion comme suit :

$$W_a^* = \inf_{c \in \Gamma^k} W_a(c).$$

En pratique, on ne peut qu'approcher algorithmiquement  $W_{d,N}^*$  voire  $W_{\infty,N}^*$  en cherchant des valeurs de  $c$  adéquates, compte-tenu des limitations informatiques.

Faisons une observation avant de poursuivre :  $\Gamma$  est fermé. On le montre par un simple théorème de prolongement des inégalités, appliqué deux fois. On peut donc adopter les notations suivantes pour les centres optimaux :

$$c_{d,N}^* = \arg \min_{c \in \Gamma^k} W_{d,N}(c), \quad c_d^* = \arg \min_{c \in \Gamma^k} W_{d,\infty}(c), \quad \text{et}$$

$$c_N^* = \arg \min_{c \in \Gamma^k} W_{\infty,N}(c), \quad c_\infty^* = \arg \min_{c \in \Gamma^k} W_{\infty,\infty}(c).$$

Le but final de cette étude est de montrer que l'écart entre la distorsion moyenne utilisant les centroïdes approchés et celle utilisant les centroïdes optimaux est rapidement aussi petit que l'on veut.

### 3 Convergence de la distorsion empirique

En décomposant la variable aléatoire  $Y$  sur la base  $\beta$ , on vérifie le lemme suivant en quelques lignes. Celui-ci borne  $Y$  presque sûrement, ce qui sera utile dans toute la suite.

**Lemme**

$$\mathbb{P} \left( \|Y\|^2 \leq \frac{R}{\varphi_{j_0}} \right) = 1,$$

la norme étant celle héritée de l'espace  $L_2(I, \mathbb{R})$ .

Cette constatation permet d'obtenir une majoration de l'erreur réalisée, lorsqu'on utilise les centres optimaux pour calculer la distorsion empirique à  $N$  courbes, résumée dans la propriété 1. On obtient une estimation de la valeur de  $N$  "minimale" pour approcher

la moyenne de la distorsion (au sens de la mesure  $g$ ), en supposant connaître les centres optimaux (non empiriques) dans ce dernier. La preuve se trouve par exemple dans l'article de Linder et al. (1994), utilisant une inégalité de Hoeffding (1963).

**Proposition 1** *Pour tout  $\alpha > 0$ ,*

$$\mathbb{P}(W_{d,N}(c_{d,N}^*) - W_d^* > \alpha) \leq e^{-\frac{2N\alpha^2}{8B^2}} \quad , \quad \text{et} \quad \mathbb{P}(W_N(c_\infty^*) - W_\infty^* > \alpha) \leq e^{-\frac{2N\alpha^2}{8B^2}} .$$

Nous disposons également d'un résultat sur l'erreur réalisée en utilisant la distorsion empirique dans l'espace projeté à  $d$  dimensions, au lieu de l'espace fonctionnel de dimension infinie. La borne est cette fois non probabiliste, et on obtient une estimation de la valeur de  $d$  "minimale" pour bien retranscrire l'espace de départ  $\Gamma$ .

**Proposition 2** *Posant  $\varphi_{min}^d = \min_{j \in ]d, +\infty[} \varphi_j$ ,*

$$\sup_{c \in \Gamma^k} \{W_N(c) - W_{d,N}(c)\} \leq \frac{2R}{\varphi_{min}^d} \quad , \quad \text{et} \quad \sup_{c \in \Gamma^k} \{W_\infty(c) - W_d(c)\} \leq \frac{2R}{\varphi_{min}^d} .$$

La preuve de ce dernier résultat s'obtient facilement à l'aide de l'hypothèse effectuée en première partie. Nous examinons à présent une dernière quantité, représentant l'erreur moyenne réalisée en approximant  $c_\infty^*$  par  $c_{d,N}^*$ . C'est celle-ci qui indique la capacité de généralisation du clustering empirique effectué en dimension finie :

$$\Delta_{d,N} = W_\infty(c_{d,N}^*) - W_\infty^* .$$

Pour  $\alpha > 0$ , posons  $D_\alpha = \min \left\{ j_0 \in \mathbb{N} / \forall j \geq j_0, \frac{2R}{\varphi_j} < \alpha \right\}$ , plus petit indice au-delà duquel  $\frac{2R}{\varphi_j}$  est strictement inférieur à  $\alpha$ .

**Théorème :**

Soit  $\alpha > 0$ , et  $d \geq D_\alpha$ . Alors :

$$\mathbb{P}(\Delta_{d,N} > \alpha) \leq 4(2N)^{k(d+1)} \exp \left( -N \left( \alpha - \frac{4R}{\varphi_{min}^d} \right)^2 / 512B^2 \right)$$

*Étapes de la preuve (donnée dans Linder et al. (1994)) :*

Nous commençons par nous ramener en dimension finie en montrant le résultat suivant

$$\Delta_{d,N} \leq W_d(c_{d,N}^*) - W_d^* + \frac{4R}{\varphi_{min}^d} .$$

Concernant  $W_d(c_{d,N}^*) - W_d^*$ , on se ramène ensuite à une borne supérieure faisant intervenir la distorsion empirique :

$$W_d(c_{d,N}^*) - W_d^* \leq 2 \sup_{c \in \Gamma^k} \{W_{d,N}(c) - W_d(c)\} .$$

Enfin, une inégalité due à Vapnik et Chervonenkis (1971) avec des paramètres bien choisis permet de montrer le résultat suivant, après quelques développements :

$$W_d(c_{d,N}^*) - W_d^* \leq 4(2N)^{k(d+1)} e^{\frac{-N\alpha^2}{512B^2}}.$$

On en déduit le résultat en choisissant  $d \geq D_\alpha$ .  $\diamond$ .

**Bilan :** Le théorème final indique l'erreur de généralisation attendue dans le pire cas, c'est-à-dire quand le clustering se fait en dimension finie empiriquement. La propriété 1 nous renseigne quant à elle sur les valeurs pratiques pour  $N$  et  $d$ , mais est difficilement utilisable car elle suppose le regroupement optimal connu. Nous remarquons que la borne obtenue dans le dernier théorème est elle aussi difficilement exploitable, compte-tenu de la dimension présente en exposant.

Un meilleur résultat existe sur la majoration de  $\mathbb{P}(\Delta_{d,N} > \alpha)$ , en supposant la dimension réduite par le biais de projections aléatoires comme décrit dans l'article de Biau et al. (2005). Posant  $d = \lceil 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log(N/\sqrt{\delta}) \rceil$  pour  $\varepsilon > 0, \delta > 0$ , les auteurs montrent l'inégalité suivante, valide au moins avec la probabilité  $(1 - \delta)(1 - e^{-x}), x > 0$  :

$$\Delta_{d,N} \leq \frac{24kB^2 + 12B^2\sqrt{2x}}{\sqrt{N}} + 4\varepsilon B^2.$$

## 4 Application

Nous choisissons la fonction paramétrique suivante, définie sur  $]0, +\infty[$  :

$$f_{a,b,c,w}(x) = x(w \cos(x + a) - b \log(x)) + c.$$

Supposons que  $a$  et  $c$  suivent des lois uniformes respectivement dans  $[0, \pi/2]$  et  $[0, 100]$ .  $b$  et  $w$  suivent quant à eux les lois discrètes suivantes :  $b = \frac{1}{3}\delta_1 + \frac{1}{3}\delta_3 + \frac{1}{3}\delta_6$ ,  $w = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_4$ . Par construction, à  $k = 6$  il y a exactement six clusters suivant les valeurs de  $b$  et  $w$ . Nous représentons leurs courbes principales sur la figure 1 pour le domaine d'apprentissage ( $x \in [1, 71]$ ),  $b$  faisant varier la "pente" de la courbe, et  $w$  son amplitude.

La figure 2 montre le clustering obtenu avec 1000 courbes aléatoires (en imposant  $k = 6$ ) : celui-ci est cohérent avec nos prédictions. Choisisant la majoration de la proposition 1 avec  $\alpha = 0.5$  et  $B \simeq 1000$  (valeur minimale constatée), il faudrait plusieurs dizaines de millions de courbes aléatoires pour voir une vraie convergence, soit plusieurs jours de calcul. Ainsi, nous nous contentons d'indiquer les résultats obtenus (moyennés sur 10 exécutions) jusqu'à des valeurs assez élevées de  $N$ , à l'aide d'un algorithme de complexité  $O(dN)$  écrit dans le langage *R*.

$N$	500	1000	10000	50000	100000
$(W_N(c_d^*))^{1/2}$	723.3	722.0	721.2	719.7	719.2

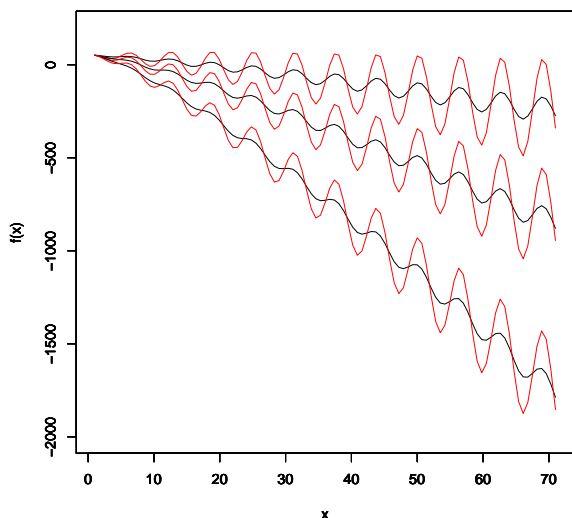


FIG. 1 – Les six courbes principales de  $f_{a,b,c,w}$ . En noir pour  $w = 1$ , en rouge pour  $w = 4$ .

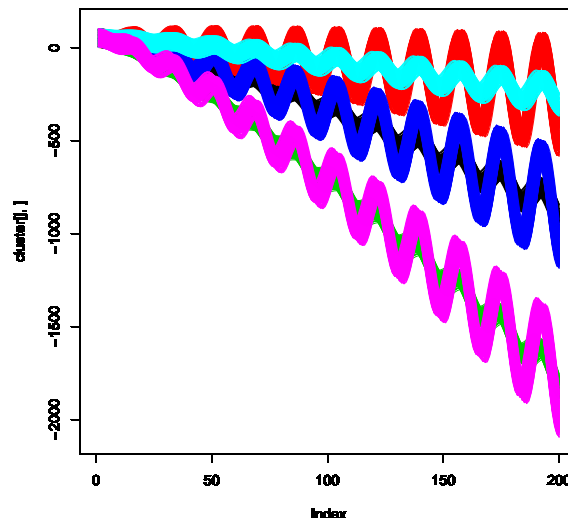


FIG. 2 – Clustering obtenu pour 1000 courbes aléatoires  $f_{a,b,c,w}$ .

## 5 Conclusion

Nous avons à présent des résultats de convergence théoriques sur le groupement de courbes en  $k$  clusters en dimension infinie, dont un est vérifié en pratique. Un problème crucial qui n'a pas été abordé est celui du choix de la bonne valeur de  $k$ . Il existe des heuristiques en déterminant une valeur approchée, mais comme l'évaluation du nombre de clusters est finalement un problème subjectif de visualisation, des études à ce sujet sont encore en cours. Une prochaine étape consistera à classer automatiquement les entrées du code selon les groupes de leurs sorties correspondantes (classification supervisée).

## Bibliographie

- [1] Linder, T., Lugosi, G. et Zeger, K. (1994) Rates of Convergence in the Source Coding Theorem, in Empirical Quantizer Design, and in Universal Lossy Source Coding, *IEEE Transactions on Information Theory*, 40, 1728–1740.
- [2] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables, *American Statistical Association*, 58, 13–30.
- [3] Vapnik, V. N. et Chervonenkis, A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities, *Applied Probability*, 16, 264–280.
- [4] Biau, G., Bunea, F., Wegkamp, M. H. (2005) Functional Classification in Hilbert Spaces, *IEEE Transactions on Information Theory*, 51, 2163–2172.