Barycentres de séries temporelles : une nouvelle approche basée sur la méthode de la signature

Raphaël MIGNOT[1], Konstantin USEVICH[2], Marianne CLAUSEL[1], Georges OPPENHEIM[3], Laure COUTIN[4], Antoine LEJAY[1]

[1]Université de Lorraine, CNRS, IECL, Nancy, France

[2]Université de Lorraine, CNRS, CRAN, Nancy, France.

[3]Université Paris-Est-Marne la Vallée, Département de Mathématiques, Marne-la-Vallée, France

[4]Université Paul Sabatier, IMT, Toulouse, France

raphael.mignot@univ-lorraine.fr

Résumé – Le but de ce travail est de moyenner des séries temporelles multivariées en utilisant une représentation des séries temporelles à base d'intégrales de moments d'ordres différents, constituant sa signature. La contribution de cet article est de proposer d'exploiter la structure de groupe de Lie pour moyenner des signatures en utilisant les opérations exp et log du groupe à l'image de la géométrie Riemannienne.

Abstract – The aim of this work is to average multidimensional time series through the representation of time series with integrals of various moment orders, constituting its signature. The contribution of this article is to suggest to take advantage of the Lie group structure to average signatures, using the exp and log operation of the group.

1 Introduction

L'analyse statistique de séries temporelles multivariées est un problème difficile, avec de nombreuses applications dans des domaines divers tels la finance, l'environnement ou le domaine médical. Une des principales difficultés est d'encoder de manière pertinente la dépendance temporelle inhérente à chaque composante ainsi que les dépendances possiblement non linéaires entre les composantes. Une approche prometteuse introduite tout récemment vise à utiliser la méthode dite de la signature [4], introduite initialement dans les années 1950 par K-T. Chen [3] dans le cadre de la théorie du contrôle et qui a connu de nombreux développements liés à la théorie des trajectoires rugueuses [6].

La méthode de la signature a souvent été couplée avec les méthodes usuelles d'analyse de séries temporelles multivariées, comme une première étape d'encodage des dépendances inter et intra composantes. Cette approche s'est révélée être redoutablement efficace pour de nombreuses applications, comme la reconnaissance d'idéogrammes manuscrits [12], la détection du trouble bipolaire [8] ou encore en océanographie [10]. Ici, nous nous intéressons à la résolution d'un problème important dans le domaine de l'apprentissage : celui du calcul de barycentres de séries temporelles. Nous montrons que notre approche basée sur le calcul de la signature a de bonnes propriétés sur le plan computationnel et statistique.

2 Processus multivariés et méthode de la signature

Soit $X : [0,1] \to \mathbb{R}^d$ un processus multivarié continu à variations bornées. Pour tout $0 \leq t_1 < t_2 \leq 1$ et m un entier naturel, on définit sa signature d'ordre m sur le segment $[t_1, t_2]$ de la manière suivante :

$$S^{(m)}_{[t_1,t_2]}(X) \overset{\text{déf}}{=} \int\cdots\int_{t_1<u_1<\cdots<u_m<t_2} dX_{u_1} \otimes \cdots \otimes dX_{u_m} \quad (1)$$

où $\otimes$ est le produit tensoriel. La signature d'ordre m est un tenseur de $(\mathbb{R}^d)^{\otimes m}$ dont chaque coefficient est intuitivement une mesure d'association entre m composantes du processus multivarié considéré. La signature du processus multivarié X est alors la collection infinie des signatures de tous les ordres :

$$\mathbf{S}_{[t_1,t_2]}(X) = \left(1, S^{(1)}_{[t_1,t_2]}(X), S^{(2)}_{[t_1,t_2]}(X), S^{(3)}_{[t_1,t_2]}(X), \dots\right) \quad (2)$$

où par convention on a fixé $S^{(0)}_{[t_1,t_2]}(X) = 1$. On a $S^{(1)}_{[t_1,t_2]}(X)$ qui est un vecteur, puis $S^{(2)}_{[t_1,t_2]}(X)$ une matrice, $S^{(3)}_{[t_1,t_2]}(X)$ un tenseur tri-dimensionnel, etc.

Par exemple, supposons X linéaire. Par un calcul simple, on trouve que la valeur du tenseur $S^{(m)}_{[0,1]}(X)$ pour tout jeu d'indice $(i_1, \dots, i_m)$ avec i_j compris entre 1 et d est :

$$\left[S^{(m)}_{[0,1]}(X)\right]_{(i_1,\dots,i_m)} = \frac{1}{m!}\prod_{j=1}^{m}(X_1^{(i_j)} - X_0^{(i_j)}) \in \mathbb{R}. \quad (3)$$

La signature vérifie deux propriétés fondamentales :

— c'est une description *intrinsèque*, caractérisant sous certaines conditions un signal multivarié aux translations et aux reparamétrisations près.
— l'espace de la signature est un groupe de Lie non compact pour l'opération $\otimes$. Ici l'opération $\otimes$ est une extension du produit tensoriel [4]. Par abus de notation, on ne distinguera pas les deux. Cette opération est intimement liée à la concaténation $\star$ de deux processus par la relation de Chen

$$\mathbf{S}_{[t_1,t_3]}(X \star Y) = \mathbf{S}(X)_{[t_1,t_2]} \otimes \mathbf{S}_{[t_2,t_3]}(Y). \quad (4)$$

L'espace des signatures tronqué à l'ordre m est un sous-espace de $T^m(\mathbb{R}^d) := \bigoplus_{k=0}^m (\mathbb{R}^d)^{\otimes k}$. Cet espace $(T^m(\mathbb{R}^d), +, \cdot, \otimes)$ est une algèbre de Lie non-commutative, dont le crochet est $[u, v] = u \otimes v - v \otimes u$. Par ailleurs, on définit l'exponentielle, le logarithme et l'inverse sur $T^m(\mathbb{R}^d)$ comme :

$$\exp(g) := \sum_{i\geq 0} \frac{g^{\otimes i}}{i!} \quad \text{et} \quad \log(1+g) := \sum_{i\geq 1} (-1)^{i-1} \frac{g^{\otimes i}}{i} \quad (5)$$

$$(1+g)^{-1} := \sum_{i=0}^{m} (-1)^i g^{\otimes i}. \quad (6)$$

Les intégrales itérées (1) sont définies pour des processus multivariés continus. Lorsqu'on analyse des séries temporelles multivariées discrètes, on peut leur associer par interpolation un processus multivarié continu. Ici, nous effectuerons une interpolation linéaire et on note $\tilde{X}$ le processus obtenu. Les intégrales itérées de $\tilde{X}$ sont alors calculées numériquement jusqu'à un certain ordre de troncature m à choisir.

3 Barycentre de séries temporelles multivariées et signature

Soit $(X_i)_{1\leq i\leq n}$ un jeu de données de n séries temporelles multivariées à d composantes discrètes. Notre objectif est de définir une notion de barycentre de ces séries temporelles, et pour cela une étape intermédiaire va être de trouver le barycentre de leurs signatures respectives.

Pour se fixer les idées, nous supposerons que chaque série temporelle est de taille $d \times l$ avec d le nombre de composantes et l le nombre de points d'échantillonage dans le domaine temporel. On cherche à définir un barycentre des n signatures $\{\mathbf{S}(X_1), \ldots, \mathbf{S}(X_n)\}$ de poids $(w_i)_{1\leq i\leq n}$. Un avantage de la signature est que l'on peut la comparer pour deux séries de longueurs l différentes, tant que leur dimension d est la même : en effet, dans ce cas les deux signatures appartiennent à $T^m(\mathbb{R}^d)$. Voici trois stratégies qui ont toutes pour point de départ la nature de groupe de Lie G de l'espace des signatures. On notera $\mathfrak{g}$ son algèbre de Lie, exp et log ses applications exponentielles et logarithmes.

3.1 Barycentre log-euclidien

La méthode naïve consistant à calculer la moyenne euclidienne $\bar{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^n \mathbf{S}(X_i)$ n'est pas pertinente sur le groupe de Lie des signatures car $\bar{\mathbf{S}}$ peut ne pas être une signature. Pour obtenir une moyenne qui reste dans l'espace des signatures, une première stratégie consiste à effectuer une moyenne euclidienne dans l'algèbre de Lie $\mathfrak{g}$ puis de repasser dans le groupe de Lie G :

$$\bar{\mathbf{S}}_{LE} = \exp\left(\frac{1}{n}\sum_{i=1}^n w_i \log \mathbf{S}(X_i)\right). \quad (7)$$

Cependant, ce barycentre n'est pas invariant par translation à gauche ou à droite : pour tout chemin Y, nous avons

$$\exp\left(\frac{1}{n}\sum_{i=1}^n w_i \log\left[\mathbf{S}(X_i) \otimes \mathbf{S}(Y)\right]\right) \neq \exp\left(\frac{1}{n}\sum_{i=1}^n w_i \log \mathbf{S}(X_i)\right) \otimes \mathbf{S}(Y). \quad (8)$$

Par l'Equation (4), cette invariance peut s'illustrer dans l'espace des chemins comme en Figure 1. Cette notion de barycentre amène donc à des phénomènes non intuitifs que nous souhaitons éviter. Ceci motive ainsi la stratégie suivante que nous proposons.

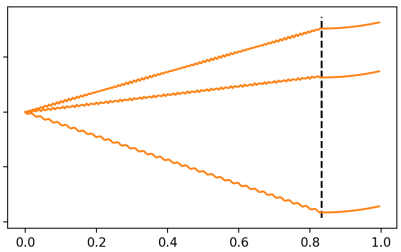

FIGURE 1 – Translation à droite : trois courbes X_1, X_2, X_3 auxquelles on concatène la même courbe Y (partie à droite de la barre verticale pointillée). Une moyenne m est invariante à droite si $m(\{X_i \star Y\}_{1\leq i\leq n}) = m(\{X_i\}_{1\leq i\leq n}) \star Y$.

3.2 Barycentre exponentiel de groupe

Cette stratégie vise à adapter [7, Algorithme 1] au contexte de la signature afin de définir de manière itérative un barycentre sur groupe de Lie, invariant par translation à droite. Il convient de noter ici que le groupe de Lie des signatures ne possède pas de métrique Riemannienne bi-invariante (à droite et à gauche) [1].

Après initialisation de $\bar{\mathbf{S}}_{(0)}$ première valeur du barycentre [2], l'étape de mise à jour de l'algorithme est la suivante. Tant que le critère d'arrêt n'est pas atteint, la valeur du barycentre est modifiée :

$$\bar{\mathbf{S}}_{(k+1)} = \bar{\mathbf{S}}_{(k)} \otimes \exp\left(\sum_{i=1}^n w_i \log\left((\bar{\mathbf{S}}_{(k)})^{-1} \otimes \mathbf{S}(X_i)\right)\right) \quad (9)$$

1. L'invariance à gauche n'est pas nécessaire car on peut choisir de translater toutes les séries afin que $X_i(0) = 0$ pour tout $1 \leq i \leq n$.
2. La procédure d'initialisation est à choisir. Par exemple $\bar{\mathbf{S}}_{(0)} = S(X_i)$ pour un i tiré aléatoirement.

où k est le nombre d'itérations déjà effectué. Le critère d'arrêt peut être par exemple un nombre d'itérations maximal ou lorsque la distance $d(\bar{\mathbf{S}}_{(k)}, \bar{\mathbf{S}}_{(k+1)})$ est inférieur à un seuil fixé. Cette procédure converge vers une solution, à condition que les données d'entrée $(\mathbf{S}(X_i))_{1\leq i\leq n}$ soient suffisamment proches de l'élément neutre (faible dispersion), et que l'initialisation $\bar{\mathbf{S}}_{(0)}$ soit suffisamment proche des données [7, Corollaire 5]. Dans notre contexte, on observe que lorsque l'algorithme converge, la solution est atteinte en moins de 10 itérations.

3.3 Optimisation sur l'espace des trajectoires

Avec les deux méthodes précédentes, le barycentre obtenu est une signature. Si l'on souhaite obtenir un chemin correspondant à $\bar{\mathbf{S}}$, c'est-à-dire un chemin X tel que $\mathbf{S}(X) = \bar{\mathbf{S}}$, une reconstruction est nécessaire. Cette reconstruction est de mieux en mieux maîtrisée mais reste loin d'être exacte et nécessite un ordre élevé de la signature [2].

Dans la méthode suivante, le barycentre obtenu est directement une trajectoire discrète. Celui-ci est obtenu par optimisation sur l'espace des matrices. Soit $X \in \mathbb{R}^{d\times \tilde{l}}$ une trajectoire discrète avec $1 \leq \tilde{l} \leq l$ et soit m un ordre de troncature pour la signature. On définit notre fonction objectif à minimiser :

$$f(X) = \sum_{i=1}^{n}\sum_{p=1}^{m} d(S^{(p)}(X), S^{(p)}(X_i)) \tag{10}$$

avec d une distance à choisir, par exemple $d(g,h) = ||g-h||_F^2$, avec $||.||_F$ la norme de Frobenius ou bien $d(g,h) = ||g^{-1}\otimes h||_{CC}$ avec $||.||_{CC}$ la norme de Carnot-Carathéodory. La fonction objectif peut être minimisée par exemple par descente de gradient. Numériquement, le gradient peut facilement être calculé par différentiation automatique.

Afin d'accroître les performances et la robustesse de cette méthode, on peut jouer sur plusieurs paramètres. La longueur $\tilde{l}$ de la matrice X est un hyper-paramètre à régler afin d'éviter un sur ou sous-apprentissage. On peut jouer sur le nombre d'initialisations. En choisissant un unique point de départ pour l'optimisation, on peut tomber dans un minimum local, surtout si le point choisi se trouve dans une région éloignée du minimum global. Effectuer p initialisations accroît les chances de se trouver proche du minimum global au départ. On choisit ensuite le plus petit minimum obtenu parmi les p solutions.

3.4 Comparaison des trois stratégies

Les trois méthodes précédentes ont des cas d'utilisations préférentiels : selon que l'on souhaite obtenir une signature ou une trajectoire. Seules les deux premières méthodes tirent vraiment parti de la structure intrinsèque de l'espace des signatures (groupe de Lie). En pratique, le temps de calcul de la première méthode est quasi-instantané. Alors que celui des deux autres varie selon les paramètres choisis et la convergence n'est pas assurée.

4 Expériences numériques

4.1 Méthode des K-moyennes

La définition de notions de barycentre nous permet maintenant de coupler la signature aux algorithmes classiques d'apprentissage de données. Par exemple, pour le partitionnement de données, la méthode omniprésente est le K-moyennes. On souhaite séparer nos données $(X_i)_{1\leq i\leq n}$ en K groupes. La procédure est la suivante.

1. Initialisation de $\mu_1, \ldots, \mu_K$ centres de chacun des K groupes.
2. Tant que le critère d'arrêt n'est pas satisfait :
   — Assignation : pour tout $1 \leq i \leq n$, X_i est assigné au groupe dont le centre est le plus proche de X_i selon une distance choisie.
   — Mise à jour : chaque centre $\mu_1, \ldots, \mu_K$ est le barycentre des points de son groupe (après l'étape d'assignation).

Ainsi, deux paramètres sont à définir pour effectuer les K-moyennes : une métrique et une notion de barycentre. En suivant l'algorithme décrit ci-dessus, on minimise l'inertie :

$$\text{inertie} = \arg\min_{\mathbf{C}} \sum_{k=1}^{K}\sum_{j:X_j\in C_k} d(X_j, \mu_k)^2 \tag{11}$$

où $\mathbf{C} = \{C_1, \ldots, C_K\}$ est le partitionnement et d la distance.

Pour évaluer les performances d'un algorithme de partitionnement, plusieurs métriques existent. Une des métriques les plus utilisée est le *rand index* (RI). On suppose que l'on dispose du *vrai* partitionnement $\mathbf{C} = \{C_1, \ldots, C_K\}$. On note $\hat{\mathbf{C}} = \{\hat{C}_1, \ldots, \hat{C}_K\}$ le partitionnement obtenu. Soit a le nombre de paires d'observations (X_i, X_j) qui sont dans un même groupe pour $\mathbf{C}$ et dans un même groupe pour $\hat{\mathbf{C}}$. Soit b le nombre de paires d'observations qui sont dans deux groupes différents pour $\mathbf{C}$ et dans deux groupes différents pour $\hat{\mathbf{C}}$. On définit le *rand index* comme $\mathrm{RI} = \frac{a+b}{\binom{n}{2}}$. On remarque que si $\mathbf{C}$ et $\hat{\mathbf{C}}$ sont égaux, alors $\mathrm{RI} = 1$. Plus les deux partitionnements sont différents, plus cette métrique décroît vers zéro.

4.2 Application à la signature

On utilise le jeu de données PenDigit [1] où chaque observation est un chiffre manuscrit. On possède $n = 10992$ séries temporelles bi-dimensionnelle de longueur $l = 8$. On possède les n vrais étiquettes. On cherche à faire un partitionnement des n séries. A noter que le bon nombre de groupes à choisir n'est pas forcément 10 : un même chiffre peut être représenté de plusieurs manières. On effectuera le K-means avec quatre valeurs pour K : 8, 10, 12 et 14 groupes. Le critère d'arrêt du K-means est un nombre maximum d'itérations à effectuer, fixé à 10.

Cinq stratégies sont comparées : voir Table 1. La méthode DTW (*Dynamic Time Warping*) *Barycenter Averaging* [9] est calculée avec la bibliothèque `tslearn` [11]. La signature est calculée avec `signatory` [5] jusqu'à l'ordre 8.

TABLE 1 – Paramètres pour les K-moyennes. Les deux premières méthodes n'utilisent pas la signature.

Type de barycentre	Distance
moyenne euclidienne	euclidienne
DTW *Barycenter Averaging* [9]	DTW
log-euclidien (Sec. 3.1)	euclidienne
barycentre exponentiel de groupe (Sec. 3.2)	euclidienne
optimisation espace trajectoires (Sec. 3.3)	euclidienne

Les résultats sont présentés en Figures 2. On voit que les trois méthodes utilisant la signature ont de moins bon scores sur ce problème de partitionnement. Les méthodes avec signature utilisent la métrique euclidienne afin de réduire le temps de calcul. Des distances plus adaptées pour les signatures existent et permettraient peut-être d'obtenir de meilleurs performances mais sont beaucoup plus coûteuses en temps. Par ailleurs, on observe des résultats quasiment confondus pour les méthodes barycentre log-euclidien et barycentre exponentiel de groupe. Ce constat n'est pas général : en effet, les barycentres selon le vrai groupe sont différents pour ces deux méthodes. Par ailleurs, la méthode des K-moyennes est inductive, adaptée aux données isotropes ce qui ne correspond peut-être pas à l'espace des signatures et pourrait expliquer ces résultats. A noter que l'on a aussi calculé la *Adjusted Mutual Information* et que l'on obtient des résultats similaires au *Rand Index*.

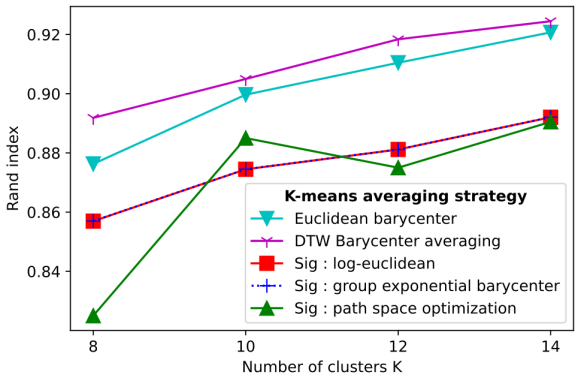

FIGURE 2 – *Rand Index* pour chacune des méthodes de la Table 1. La signature est calculée jusqu'à l'ordre 8. Les trois méthodes utilisant la signature sont précédées par "Sig".

5 Conclusion

Le développement d'une notion de barycentre pour les intégrales itérées ouvre la voie à l'extension d'algorithmes omniprésents en apprentissage statistique. Nous l'avons illustré avec l'exemple des K-moyennes. D'autres seront explorés dans de futurs travaux comme le partitionnement hiérarchique ou l'analyse en composante principale (ACP). D'autres jeux de données de séries temporelles multivariées seront aussi utilisés.

Références

[1] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv :1811.00075*, 2018.

[2] Jiawei Chang and Terry Lyons. Insertion algorithm for inverting the signature of a path. *arXiv preprint arXiv :1907.08423*, 2019.

[3] Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, pages 163–178, 1957.

[4] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning, 2016.

[5] Patrick Kidger and Terry Lyons. Signatory : differentiable computations of the signature and logsignature transforms, on both cpu and gpu. *arXiv preprint arXiv :2001.00706*, 2020.

[6] Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2) :215–310, 1998.

[7] Xavier Pennec and Vincent Arsigny. *Exponential Barycenters of the Canonical Cartan Connection and Invariant Means on Lie Groups*, page 123–166. Springer Berlin Heidelberg, 2013.

[8] Imanol Perez Arribas, Guy M Goodwin, John R Geddes, Terry Lyons, and Kate EA Saunders. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1) :1–7, 2018.

[9] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3) :678–693, Mar 2011.

[10] Nozomi Sugiura and Shigeki Hosoda. Machine learning technique using the signature method for automated quality control of argo profiles. *Earth and Space Science*, 7(9) :e2019EA001019, 2020.

[11] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res.*, 21(118) :1–6, 2020.

[12] Weixin Yang, Lianwen Jin, and Manfei Liu. Deepwriterid : An end-to-end online text-independent writer identification system. *IEEE Intelligent Systems*, 31(2) :45–53, 2016.