Signature, chemins rugueux et apprentissage

Adeline Fermanian [1]

[1] *Sorbonne université, LPSM, Campus Pierre et Marie Curie, Case courrier 158, 4 place Jussieu, 75252 Paris Cedex 05, adeline.fermanian@upmc.fr*

Résumé. Les applications modernes de la statistique et de l'apprentissage automatique ont mené à une explosion de données temporelles. On peut par exemple penser à la finance quantitative, aux enregistrements d'appareils médicaux ou à des trajectoires d'écriture manuscrite. De tels flux de données sont classiquement considérés comme des réalisations de processus stochastiques échantillonnés. Afin d'utiliser des algorithmes d'apprentissage classiques, il est nécessaire de représenter ces processus sous la forme de vecteurs de dimension finie. Nous présentons ici la transformation d'un flux de données multi-dimensionnel en sa signature, qui encode des propriétés géométriques du processus associé. La signature a été introduite dans les année 60 quand Chen (1958) a remarqué qu'un chemin peut être représenté par ses intégrales itérées, et a ensuite été au centre de la théorie des chemins rugueux de Lyons dans les année 90. La transformation en signature combinée avec un algorithme d'apprentissage a obtenu des résultats de pointe pour plusieurs applications, comme par exemple Yang (2016), ce qui soulève la question de ses propriétés statistiques. Nous allons donc présenter les principales propriétés de la signature puis étudier ses applications en apprentissage. Nous nous intéresserons en particulier à l'utilisation de la signature en régression, présentée dans Levin (2013). Compte tenu des résultats prometteurs dans la littérature, nous mènerons plusieurs tests empiriques sur les performances de cette transformation en comparaison d'autres algorithmes classiques.

Mots-clés. Apprentissage statistique, données séquentielles, chemins rugueux.

Abstract. Modern applications of statistics and machine learning have lead to a tremendous amount of temporal data. Think for example of quantitative finance, signals from medical devices or handwriting trajectories. Such data flows are traditionally considered as realisations of sampled stochastic processes. In order to use classical learning algorithms, it is necessary to represent these processes as vectors of finite dimension. We present in the following the signature transformation of a multidimensional data flow, which encodes geometric properties of its associated process. The signature dates back from the 60s when Chen (1958) noticed that a path can be represented by its iterated integrals and it has been at the center of Lyons' rough paths theory in the 90s. The signature transformation combined with a learning algorithm has achieved state of the art results for several applications, see e.g. Yang (2016), which raises the issue of its statistical properties. Therefore, we review the main properties of the signature and we investigate its applications in statistical learning. We look more closely at the signature transformation in a regression framework, as presented in Levin (2013). In light of promising results in

the literature, we undertake some empirical tests on the signature transformation and its performance compared to other classical algorithms.

Keywords. Statistical learning, sequential data, rough paths.

1 Signature d'un chemin: définition et propriétés

Soit $X : [0,1] \to \mathbb{R}^d \in BV(\mathbb{R}^d)$ un chemin à variation bornée. La signature de X est la série infinie définie par

$$S(X) = (1, \mathbf{X^1}, \mathbf{X^2}, ..., \mathbf{X^k}, ...)$$

où pour tout entier $k \in \mathbb{N}$

$$\mathbf{X^k} = \int\limits_{0<u_1<\cdots<u_k<1} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_k} \in (\mathbb{R}^d)^{\otimes k}.$$

$\otimes$ est ici le produit tensoriel de $\mathbb{R}^d$, c'est à dire que si $(e_i)_{1\leq i\leq d}$ est une base de $\mathbb{R}^d$, alors $(e_{i_1} \otimes \cdots \otimes e_{i_k})_{(i_1,...,i_k)\subset\{1,...d\}^k}$ est une base de $\left(\mathbb{R}^d\right)^{\otimes k}$. On peut donc écrire $\mathbf{X^k}$ sur cette base comme

$$\mathbf{X^k} = \sum_{(i_1,...,i_k)\subset\{1,...d\}^k} S^{i_1,...,i_k}(X) e_{i_1} \otimes \cdots \otimes e_{i_k}$$

avec

$$S^{i_1,...,i_k}(X) = \int\limits_{0<u_1<\cdots<u_k<1} \cdots \int dX^{i_1}_{u_1} \dots dX^{i_k}_{u_k}.$$

Les coefficients de la signature correspondent à des propriétés géométriques de X. Par exemple, les coefficients du premier ordre $S^i(X) = X^i_1 - X^i_0$ correspondent au déplacement de X dans la direction i. De même, les coefficients du deuxième ordre ont un lien direct avec l'aire parcourue par X. Par exemple, pour les deux premières dimensions de X, l'aire de Levy est

$$A = \frac{1}{2}\left(S^{1,2}(X) - S^{2,1}(X)\right) = A^+ - A^-$$

où A^+ et A^- sont représentés sur la Figure 1.

De plus, la signature possède un certain nombre de propriétés désirables pour bien représenter un chemin. Tout d'abord, si une des coordonnées de X est monotone, la signature est unique, voir Hambly (2010). D'autre part, la signature vérifie l'identité de Chen : soit $X : [s,t] \to \mathbb{R}^d$, $Y : [t,u] \to \mathbb{R}^d$ et $Z : [s,u] \to \mathbb{R}^d$ la concaténation de X et Y, alors

$$S(Z) = S(X) \otimes S(Y).$$

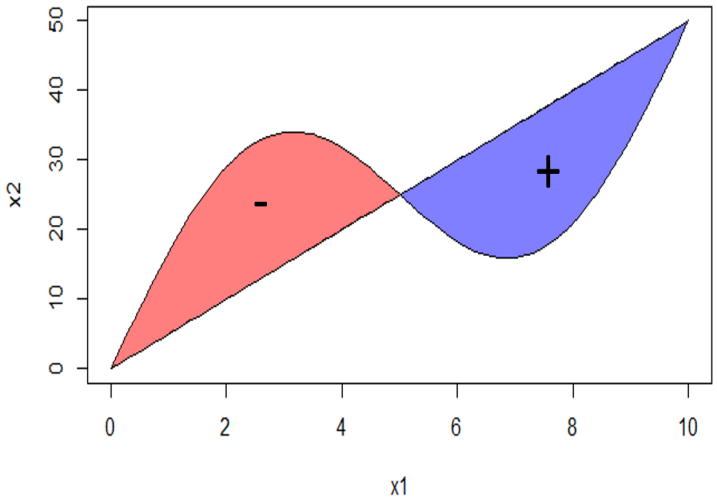

Figure 1: Aire de Levy. Rouge: A^-, bleu: A^+.

Cette identité permet de doter l'espace des formes linéaires sur la signature d'un produit nommé "shuffle product", de telle sorte que cet espace soit une algèbre. Cela permet d'utiliser le théorème de Stone-Weierstrass pour prouver que l'espace des formes linéaires sur la signature est dense dans l'espace des fonctions continues d'un compact de $BV(\mathbb{R}^d)$ dans $\mathbb{R}$. Ainsi, les formes linéaires sur la signature peuvent approcher arbitrairement bien des fonctions complexes du chemin X, ce qui est extrêmement utile dans les applications statistiques.

2 La signature en apprentissage statistique

On veut prédire une réponse Y en fonction d'une séquence $x = (x_{t_1}, \dots x_{t_k})$. Afin d'utiliser la signature, il est nécessaire dans un premier temps de plonger cette séquence dans un espace de processus. Par exemple, on va interpoler linéairement x, ou bien lui appliquer une transformation lead-lag, voir Chevyrev (2016). Ensuite, on peut calculer la signature de ce processus, la tronquer et utiliser un algorithme classique, par exemple une régression ridge ou un réseau de neurone. On peut résumer ce mécanisme par le schéma suivant :

$$\text{Données} \longrightarrow \text{Plongement} \longrightarrow \text{Signature} \longrightarrow \text{Algorithme statistique.}$$

On a vu que les formes linéaires sur la signature peuvent approcher des fonctions non linéaires de X, ce qui nous amène à nous intéresser en particulier au modèle linéaire de

la signature. Ainsi, on suppose le modèle

$$Y = \left\langle \beta^*, S^{m^*}(X) \right\rangle + \varepsilon$$

avec $\mathbb{E}[\varepsilon|X] = 0$ et $\mathbb{E}[\varepsilon^2|X] = \sigma^2 < \infty$ et on va s'intéresser à l'estimation de m^* et β^*. C'est un problème de sélection de modèle, car on doit à la fois estimer le coefficient de régression β^* et la taille du modèle m^*. Une première approche est de faire une régression ridge sur la signature tronquée à l'ordre k pour différents k, puis de sélectionner par validation croisée le meilleur k. On étudiera expérimentalement cette approche et on observera qu'elle donne des résultats compétitifs avec des méthodes plus intenses en calcul comme des forêts aléatoires ou l'algorithme XGBoost.

Bibliographie

Chen, K. (1958), Integration of paths—a faithful representation of paths by non-commutative formal power series, *Transactions of the American Mathematical Society*, 89 (2), pp. 395–407.

Chevyrev I. et Kormilitzin A. (2016), A primer on the signature method in machine learning, *arXiv preprint arXiv:1603.03788*.

Hambly B. et Lyons T. (2010), Uniqueness for the signature of a path of bounded variation and the reduced path group, *Annals of Mathematics*, pp. 109–167.

Levin, D., Lyons T. et Ni H. (2013), Learning from the past, predicting the statistics for the future, learning an evolving system, *arXiv preprint arXiv:1309.0260*.

Yang, W., Lyons T., Ni H., Schmid C., Jin L., et Chang J. (2017), Leveraging the Path Signature for Skeleton-based Human Action Recognition, *arXiv preprint arXiv:1707.03993*.