Signature Estimation and Signal Recovery Using Median of Means

Stéphane Chrétien, Rémi Vaucher

Signature estimation and signal recovery using Median of Means

Stéphane Chrétien[1] and Rémi Vaucher[1,2]

[1] Laboratoire ERIC, Bron 69676, France stephane.chretien@univ-lyon2.fr
Website https://sites.google.com/site/stephanegchretien
[2] Halias Technologies, Meylan 38240, France, remi.vaucher@halias.fr

Abstract. The theory of Signatures [1,7] is a fast growing field which has demonstrated wide applicability to a large range of applications, from finance to health monitoring[2,6,10]. Computing signatures often relies on the assumptions that the signal under study is not corrupted by noise, which is rarely the case in practice. In the present paper, we study the influence of noise on the computation of signature via the theory of anti-concentration. We then propose a median of means (MoM) approach to the estimation problem and give a bound on the estimation error using Rademacher complexity.

Keywords: Signature Theory · Median of Means · Anticoncentration.

1 Introduction: Using the signature of a function dictionary to get the signature of a path

Signatures, a transform that applies to time dependent signals, have become a tool of choice for the analysis of multidimensional dynamical phenomena which are pervasive in many applications of machine learning. Computing Signatures allows to extract meaningful features about the various time dependencies of the components of the signal in a natural way, even when sampling may possibly be irregular and at different time stamps for different components.

1.1 Goal of the paper

In practice, signals are often corrupted by additive observation noise, and we can write

$$X = X^* + \epsilon. \tag{1}$$

X is a discrete path, as observations occurs at a finite number of time $\{t_1, .., t_\nu\}$, but can be viewed as continuous by linearly interpolating between two observations.
In the present paper, we will assume for the sake of simplicity that ϵ is a Gaussian white noise fonction.

The main problems studied in the present paper are:

- estimating the Signature tensor of X^*
- recovering the original signal X^* in an appropriate basis, based on the Signature tensor of a subsampled noisy version of X^*.

Under observation noise, the computed signature may not be an accurate estimation of the signature of the true signal and one needs to resort to an appropriate regularisation procedure in order to recover the seeked signature tensor.

In the first part of this paper, we study anticoncentration of the components of the noisy signature tensor, a result which raises the questions of how reliable estimation problems for Signature-based machine learning can be. In the second part, we show how to use a new technique from the field of Robust Statistics, named Median of Means, in order to estimate a robust version of the signature tensor.

2 Anti-concentration for the 3-signature coefficients

One standard approach to estimating $S^{(3)}(\mathbb{E}[X])$, i.e. the signature of the original signal is to solve the following least-square regression problem:

$$\min_{A\in\mathbb{R}^{n\times m}} \left\| S^{(3)}(X) - [\![C; A, A, A]\!] \right\|_F^2 . \tag{2}$$

Using that, after [4], for an orthogonal basis ψ,

$$\epsilon = E\psi \tag{3}$$

where E is a i.i.d. Gaussian matrix, and using the expansion

$$X^* = \mathbb{E}[X] = A^*\psi \tag{4}$$

of X^* in the basis ψ, we obtain

$$\begin{aligned} S^{(3)}(X^* + \epsilon) &= S^{(3)}(A^*\psi + E\psi) \\ &= S^{(3)}((A^* + E)\psi) \\ &= [\![C; A^* + E, A^* + E, A^* + E]\!]. \end{aligned}$$

For the sake of making the problem finite dimensional, we will further approximate $S^{(3)}(X^* + \epsilon)$ with $S^{(3)}(XT_\nu{}^* + \epsilon T_\nu)$.

2.1 Expanding the expression of the Signature tensors

To simplify, we will note

- $[\![C]\!]_{A+E} = [\![C; A + E, A + E, A + E]\!]$
- $[\![C]\!]_{A+E,\alpha,\beta,\gamma} = [\![C; A + E, A + E, A + E]\!]_{\alpha,\beta,\gamma}$

First, it is easy to see that

$$\begin{aligned}[\![C]\!]_{A+E} = [\![C;A,A,A]\!] + [\![C;E,E,E]\!] \\ + \sum_{i=0}^{2} \left([\![C;\sigma^i(A,A,E)]\!] + [\![C;\sigma^i(A,E,E)]\!]\right)\end{aligned} \tag{5}$$

with $\sigma = (1,2,3) \in S_3$. So each coefficients from $[\![C]\!]_{A+E}$ takes the form of a polynomial of coefficients $e_{\alpha,i}$ with $\alpha \in [\![1;n]\!], i \in [\![1,m]\!]$.

We deduce from (**??**) and (5):

$$[\![C]\!]_{A+E,\alpha,\beta,\gamma} = P_1(E) + P_2(E) + P_3(E) + R = P(E) \tag{6}$$

where:

- $P_1(E) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} c_{ijk} e_{\alpha,i} e_{\beta,j} e_{\gamma,k}$
- $P_2(E) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} c_{ijk}(a_{\alpha,i}a_{\beta,j}e_{\gamma,k} + a_{\alpha,i}e_{\beta,j}a_{\gamma,k} + e_{\alpha,i}a_{\beta,j}a_{\gamma,k})$
- $P_3(E) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} c_{ijk}(a_{\alpha,i}e_{\beta,j}e_{\gamma,k} + e_{\alpha,i}e_{\beta,j}a_{\gamma,k} + e_{\alpha,i}a_{\beta,j}e_{\gamma,k})$
- $R = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} c_{ijk} a_{\alpha,i} a_{\beta,j} a_{\gamma,k}$

2.2 Anti-concentration of coefficients

We use here Theorem 1.8 from [8] or Theorem 1.2 from [5] depending on multilinearity (or not) of P (i.e. depending on $\alpha \neq \beta \neq \gamma$ or not):

The case $\alpha \neq \beta \neq \gamma$ (P is multilinear): Theorem 1. *There is an absolute constant B such that the following holds. Let $\xi_1, \ldots, \xi_n$ be independent (but not necessarily iid) random variables. Let P be a polynomial of degree d with the form*

$$P(\xi_1, \ldots, \xi_n) = \sum_{S \subset 1,\ldots,n, |S| \leq d} a_S \prod_{i \in S} \xi_i$$

whose rank $r \geq 2$. Assume that there are positive numbers p and ε such that for each $1 \leq i \leq n$, there is a number y_i such that $\min\{\mathbf{P}(\xi_i \leq y_i), \mathbf{P}(\xi_i > y_i)\} = p$ and $\mathbf{P}(|\xi_i - y_i| \geq 1) \geq \varepsilon$. Assume furthermore that $\tilde{r} := (p\varepsilon)^d r \geq 3$. Then for any interval I of length 1

$$\mathbf{P}(P(\xi_1, \ldots, \xi_n) \in I) \leq \min\left(\frac{Bd^{4/3}(\log \tilde{r})^{1/2}}{(\tilde{r})^{1/(4d+1)}}, \frac{\exp\left(Bd^2\left(\log\log(\tilde{r})^2\right)\right)}{\sqrt{\tilde{r}}}\right) \tag{7}$$

The application here is simple. P is here a multilinear polynomial of degree $d = 3$, and all $e_{\alpha,i}$ are independants. The existence of p and ε such that, for all $e_{\alpha,i}$ (with $(\alpha,i) \in [\![1,n]\!] \times [\![1,m]\!]$) there exist $y_{\alpha,i}$ verifying:

$$\min\{\mathbb{P}(e_{\alpha,i} \leq y_{\alpha,i}), \mathbb{P}(e_{\alpha,i} \geq y_{\alpha,i})\} = p \tag{8}$$

and

$$\mathbb{P}(|e_{\alpha,i} - y_{\alpha,i}| \geq 1) = \varepsilon \tag{9}$$

depends on the $e_{\alpha,i}$ distributions.
Finally, the rank r depends on the dictionary ψ: $r = \#\{c_{ijk}, \quad |c_{ijk}| > 1\}$. For the last hypothesis ($\tilde{r} \geq 3$), we need $r \geq \frac{3}{(p\varepsilon)^3}$. Under these assumptions on ψ and $a = \{a_{\alpha,i}\}$, the equation (7) applies.

<u>The two other cases:</u> For the two other cases, as the polynomials are no more multilinear, we need an anti-concentration inequality for multivariate polynomials. With this in mind, we use Theorem 2. First, we need to define PSD anti-concentration:

A distribution $\mathcal{D}$ has PSD anti-concentration if there exist $C, c > 0$ such that the following holds. Let A be an $n \times n$ positive semi-definite matrix with $\mathrm{Tr}(A) = 1$. Then for any $\varepsilon > 0$,

$$\mathbb{P}_{\mathbf{x} \in \mathcal{D}^n} \left[\mathbf{x}^t A \mathbf{x} \leq \varepsilon \right] \leq C \cdot \varepsilon^c.$$

We now fix a notation for sum and subtraction of independent variables for the same distribution:
Define $d\mathcal{D} := \mathcal{D} + \ldots + \mathcal{D}$ to be the distribution of the sum of d independent elements sampled from D, and $\mathcal{D} - \mathcal{D}$ to be the distribution of the difference of two independent elements sampled from D.
And now the theorem:

Theorem 2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ such that $\mathcal{D} - \mathcal{D}$ has PSD anti-concentration. Then there exist $C_d, c_d > 0$ such that the following holds. Let $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ be a degree d polynomial, normalized to have $\mathrm{Var}_{(d\mathcal{D})^n}[f] = 1$. Then for any $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathbb{P}_{\mathbf{x} \sim (d\mathcal{D})^n} [|f(x) - t| \leq \varepsilon] \leq C_d \cdot \varepsilon^{1/c_d}, \tag{10}$$

where $c_d = O\left(d \cdot 2^{O(d)}\right)$.

Under this assumption, which is clearly satisfied for i.i.d. Gaussian matrices, on

$$E = \{e_{\alpha,i}, e_{\beta,j}, e_{\gamma,k}, \quad \alpha, \beta, \gamma \in [\![1, n]\!], \quad i, j, k[\![1, m]\!]\}, \tag{11}$$

we can apply (10).

3 Estimation of the signal decomposition using Median of Means (MoM)

We now turn to the problem of estimating the Signature coefficients.

For this purpose, we will first consider the general problem of estimating the expectation $\mu_P = P[X]$ of a distribution P from the observation of an i.i.d. sample $\mathcal{D}_N = (X_1, \ldots, X_N)$ of real valued random variables with common distribution P.

In this part, we will note ϵ for the noise of an observation (so $X = X^* + \epsilon$) to avoid confusion with Rademacher variables.

3.1 The MoM principle

Let K and b such that $N = Kb$ and let $B_1, \ldots, B_K$ denote a partition of $\{1, \ldots, N\}$ into subsets of cardinality b. For any $k \in \{1, \ldots, K\}$, let $P_{B_k} X = b^{-1} \sum_{i \in B_k} X_i$. The MOM estimators of μ_P are defined by

$$\mathrm{MOM}_K[X] \in \mathrm{median}\left\{P_{B_k} X, k \in \{1, \ldots, K\}\right\}.$$

Recall that the Rademacher complexity of a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$D(\mathcal{F}) = \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i f\left(X_i\right)\right\}\right]\right)^2. \tag{12}$$

where ξ_i, $i = 1, \ldots, N$ are independant ± 1 Rademacher random variables.

Theorem 1. *(Concentration for suprema of MOM processes). Let $\mathcal{F}$ denote a separable set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\sup_{f \in \mathcal{F}} \sigma^2(f) = \sigma^2 < \infty$, where $\sigma^2(f) = \mathrm{Var}(f(X))$. Then, for any $K \in \{1, \ldots, N/2\}$,*

$$\mathbb{P}\left(\sup_{f \in F} |\mathrm{MOM}_K[f] - Pf| \geq 128\sqrt{\frac{D(\mathcal{F})}{N}} \vee 4\sigma\sqrt{\frac{2K}{N}}\right) \leq e^{-K/32}. \tag{13}$$

3.2 Application to signal decomposition

Let $X = X^* + \epsilon$ be a continuous noisy observation of X^*, i.e. corrupted by a continuous noise ϵ on the interval $[0, T]$. and C_Ψ be the third degree signature of an orthogonal basis ψ on $[0, T]$. We will assume that C_Ψ is computable without subsampling; otherwise, we can approximate C_Ψ by subsampling the basis functions in Ψ as well. Given an positive integer ν, and a random set of timestamps T_ν, our goal is to estimate the quantity

$$\mathbb{E}_{T_n, \epsilon}\left[\|S^{(3)}(X^*_{T_\nu} + \epsilon_{T_\nu}) - [\![C_\psi; A, A, A]\!]\|_F^2\right] \tag{14}$$

In order to put the MoM principle to work, we need n samples of the variable

$$Y = \left\|S^{(3)}(X^*_{T_\nu} + \epsilon_{T_\nu}) - [\![C_\psi; A, A, A]\!]\right\|_F^2. \tag{15}$$

Let $\{T_\nu^{(1)}, T_\nu^{(2)}, \ldots, T_\nu^{(N)}\}$ be a set of N sets of timestamps with same distribution as T_ν. For all $i = 1, \ldots, N$, define

$$Y^{(i)} = \left\|S^{(3)}(X^*_{T_\nu^{(i)}} + \epsilon_{T_\nu^{(i)}}) - [\![C_\psi; A, A, A]\!]\right\|_F^2. \tag{16}$$

Notice that the vectors $\epsilon^{(i)}$, $i = 1, \ldots, N$ with

$$\epsilon^{(i)} = \begin{bmatrix} \epsilon^{(i)}_{t_1^{(i)}} \\ \vdots \\ \epsilon^{(i)}_{t_\nu^{(i)}} \end{bmatrix} \tag{17}$$

are i.i.d. and therefore, the signals

$$X^*_{T_\nu^{(i)}} + \epsilon_{T_\nu^{(i)}}, \tag{18}$$

$i = 1, \ldots, N$ are i.i.d. as well. Based on these assumptions, we can use the MoM approach to estimate the expectation (15) Instead of estimating (15) using the mean of $Y^{(i)}$, $i = 1, \ldots, N$, we will turn to the Median of Means technique. Let $N = Kb$, Writing $P_{B_k} Y = \frac{1}{b} \sum\limits_{i \in B_k} Y^{(i)}$:

$$\begin{aligned} MOM(Y) &= \text{median}\left\{P_{B_k} Y, k \in \{1, \ldots, K\}\right\} \\ &= \text{median}\left\{\frac{1}{b} \sum_{i \in B_k} \left\| S^{(3)}(X^*_{T_\nu^{(i)}} + \epsilon_{T_\nu^{(i)}}) - [\![C_\Psi; A, A, A]\!] \right\|_F^2, k \in \{1, \ldots, K\}\right\} \end{aligned} \tag{19}$$

In order to apply Theorem 1, we need to compute $D(\mathcal{F})$, where

$$\begin{aligned} \mathcal{F} = \Bigg\{ (\epsilon, T) \mapsto \|S^{(3)}(X^*_T + \epsilon)\|_F^2 - 2 \left\langle S^{(3)}(X^*_T + \epsilon), [\![C_\Psi; A, A, A]\!] \right\rangle \\ + \|[\![C_\Psi; A, A, A]\!]\|_F^2 \Bigg\}. \end{aligned} \tag{20}$$

Thus, we have

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{\frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i f(X_i)\right\}\right] \leq \mathbb{E}\Bigg[\sup_{A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n} \Bigg\{\frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i \Bigg(\|S^{(3)}(b + \epsilon)\|_F^2 \\ - 2 \left\langle S^{(3)}(b + \epsilon), [\![C_\Psi; A, A, A]\!] \right\rangle + \|[\![C_\Psi; A, A, A]\!]\|_F^2 \Bigg)\Bigg\}\Bigg] \end{aligned} \tag{21}$$

conditioning on ϵ gives

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{\frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i f(X_i)\right\}\right] \leq \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sup_{A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n} \Bigg\{\frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i \Bigg(\|S^{(3)}(b + \epsilon)\|_F^2 \\ - 2 \left\langle S^{(3)}(b + \epsilon), [\![C_\Psi; A, A, A]\!] \right\rangle + \|[\![C_\Psi; A, A, A]\!]\|_F^2 \Bigg)\Bigg\}\Bigg] \mid \epsilon\Bigg] \\ = \mathbb{E}[\mathcal{R}_\epsilon(\mathcal{F}) \mid \epsilon] \end{aligned} \tag{22}$$

In (22), $\mathcal{R}_\epsilon(\mathcal{F})$ is the empirical Rademacher complexity. In order to bound this quantity, it is important to note that (15) is the squared Frobenius norm of a matrix where every coefficient is a degree 3 polynomial of ϵ. Hence, the final quantity is a degree 6 polynomial in the variable ϵ. Rademacher complexity of polynomials has been addressed in earlier sources such as, e.g. [9].

At this point, we are refering to [3] to "convert" a polynomial function to a polynomial network and [11] for the Rademacher complexity.

Consider $\mathcal{E} = (\epsilon_1, \ldots, \epsilon_n)$ a set of n i.i.d. samples from the same distribution as $\epsilon \in \mathbb{R}^d$. Let $\mathcal{C}$ denote the event

$$\mathcal{C} = \{\|\epsilon_k\|_\infty \leq 1, k \in \{1, \ldots, n\}\} \tag{23}$$

Then, it is proved in [11] that there exist two constants μ and λ such that

$$\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \leq 2\mu\lambda\sqrt{\frac{12\log(d)}{n}} \tag{24}$$

on $\mathcal{C}$. It follows from (22) that

$$\begin{aligned} D(\mathcal{F}) &\leq \mathbb{E}\left[\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \mathcal{C}\right]\mathbb{P}(\mathcal{C}) + \mathbb{E}\left[\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \overline{\mathcal{C}}\right]\mathbb{P}(\overline{\mathcal{C}}) \\ &\leq 2\mu\lambda\sqrt{\frac{12\log(d)}{n}}\mathbb{P}(\mathcal{C}) + \frac{c}{\sqrt{n}} \end{aligned} \tag{25}$$

where we used that $\sqrt{n}\mathbb{E}\left[\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \overline{\mathcal{C}}\right]\mathbb{P}(\overline{\mathcal{C}})$ can be shown to be bounded by a constant using a peeling argument. Combining (1), (25), we obtain the following result

Theorem 2. *Let Y be defined by* (15)*, $Y^{(I)}$, $i = 1, \ldots, N$ be defined* (16)*, and the MoM estimator be defined by* (19)*. Then, we have*

$$\begin{aligned} &\mathbb{P}\left(\sup_{Y\in\mathcal{F}} |\mathrm{MOM}_K[Y] - PY| \geq 128\sqrt{\frac{2\mu\lambda\sqrt{\frac{12\log(d)}{n}}\mathbb{P}(\mathcal{C}) + \frac{c}{\sqrt{n}}}{N}} \vee 4\sigma\sqrt{\frac{2K}{N}}\right) \\ &\qquad \leq e^{-K/32}. \end{aligned} \tag{26}$$

Complete proof details will be provided in an extended version of the paper.

References

1. Chen, K.T.: Iterated integrals and exponential homomorphisms. Proceedings of the London Mathematical Society **3**(1), 502–512 (1954)
2. Chevyrev, I., Kormilitzin, A.: A primer on the signature method in machine learning. arXiv preprint arXiv:1603.03788 (2016)

3. Chrysos, G.G., Moschoglou, S., Bouritsas, G., Panagakis, Y., Deng, J., Zafeiriou, S.: P-nets: Deep polynomial neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7325–7335 (2020)
4. Johnstone, I.M.: Function estimation and gaussian sequence models. Unpublished manuscript **2**(5.3), 2 (2002)
5. Lovett, S.: An elementary proof of anti-concentration of polynomials in gaussian variables. In: Electron. Colloquium Comput. Complex. vol. 17, p. 182 (2010)
6. Lyons, T., McLeod, A.D.: Signature methods in machine learning. arXiv preprint arXiv:2206.14674 (2022)
7. Lyons, T.J.: Differential equations driven by rough signals. Revista Matemática Iberoamericana **14**(2), 215–310 (1998)
8. Meka, R., Nguyen, O., Vu, V.: Anti-concentration for polynomials of independent random variables. arXiv preprint arXiv:1507.00829 (2015)
9. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018)
10. Schell, A., Oberhauser, H.: Nonlinear independent component analysis for discrete-time and continuous-time signals. Annals of Statistics (2023)
11. Zhu, Z., Latorre, F., Chrysos, G.G., Cevher, V.: Controlling the complexity and lipschitz constant improves polynomial nets. arXiv preprint arXiv:2202.05068 (2022)