UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

EL MÈTODE DE LA SIGNATURA I APLICACIONS A DADES DE TIPUS FINANCER

Autor: David Cots Mañas

Director: Dr. José Manuel Corcuera Valverde
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 13 de juny de 2023

Abstract

The signature captures a non-parametric characteristic of a data stream that has been previously transformed into a path through an embedding algorithm.

In this work, we will introduce the signature method, which consists of embedding a data stream into a path to hen use its signature as a feature in machine learning problems.

First, we will define the signature of a path and present its most important properties.

Then, we will explore different ways of embedding data streams into a continuous path and examine the application of this method to financial time series data.

Finally, we will see how we can implement a supervised machine learning model based on signatures.

Resum

La signatura extrau una característica no paramètrica d'un flux de dades que ha sigut previament transformat en un camí mitjançant un algorisme d'incrustació. A aquest treball introduïrem el mètode de la signatura, aquest consisteix en incrustar un flux de dades a un camí i utilitzar la signatura d'aquest com a característica en problemes d'aprenentatge automàtic. Primer definirem la signatura d'un camí en donarem les seves propietats més importants.

Després explorarem diferents formes d'incrustar un flux de dades en un camí continu i veurem l'aplicació d'aquest mètode a dades de sèrie temporal de tipus financer.

Finalment veurem com podem implementar un model d'aprenentatge automàtic supervisat basat en signatures.

2020 Mathematics Subject Classification. 60L10

Agraïments

En primer lloc vull agraïr al Dr. José Manuel Corcuera per tutoritzar i proposar el tema d'aquest treball.
M'agradaria també agraïr als meus pares per haver-me donat la possibilitat de cursar aquests estudis i per donar-me tot el seu suport.
M'agradaria fer també un agraïment general a tots els docents del grau que a través de la seva passió i dedicació inspiren a molts alumnes.
Finalment, voldria agraïr també a tots els amics i companys, així com a la meva germana, que m'han acompanyat i donat suport durant aquest llarg camí.

Índex

1 Introducció

La signatura d'un camí és una col·lecció infinita de nombres reals que captura certes propietats geomètriques d'aquest. Tot i que un camí no està totalment determinat per la seva signatura, la signatura determina completament les propietats geomètriques d'un camí sota certes condicions de no degeneració. Això la fa útil per a caracterizar fluxos de dades, incrustats en camins prèviament. En això últim consisteix precisament el mètode de la signatura: donat un flux de dades, l'incrustem a un camí continu i posteriorment usem la signatura com a característica d'aquest dins del marc de l'aprenentatge automàtic.

Estructura de la Memòria

En el primer capítol donem la definició de signautra d'un camí, així com las dels conceptes previs necessaris. A la mateixa secció hi trobem certes propietats importants de la signatura, aixií com exemples i la intuïcuó geomètrica darrera de la signatura. En el segon capítol expliquem primer diferents maneres d'incrustar un flux de dades e un camí. Al quart capítol explorem la signatura aplicada a dades de caire financer, o més en general, dades en forma de sèrie temporal. Finalment a l'últim capítol discutim un model d'aprenentatge automàtic basat en la signatura.

2 La signatura d'un camí

2.1 Conceptes previs

A continuació revisem una sèrie de nocions bàsiques d'anàlisi matemàtic.

Definició 2.1. *Un un espai normat* $(V, \|\cdot\|)$ *és un parell format per un espai vectorial* V *sobre un cos* $\mathbb{K}$ *i una norma* $\|\cdot\| : V \to \mathbb{K}$.

Les normes indueixen una funció distància invariant per traslacions anomenada la *distància canòninca* o *distància induïda per la norma*, definida $\forall u, v \in V$ com:

$$d(u, v) := \|x - y\| = \|y - x\|.$$

Podem formar, per tant, l'espai mètric (V, d).

Definició 2.2. *Una successió* $\{v_i\}_{i\in\mathbb{N}}$ *en un espai mètric* (V, d) *és diu de Cauchy si per a cada* $r > 0$ *real,* $\exists N \in \mathbb{N} \cup \{0\}$ *tal que* $\forall m, n > N$

$$d(v_n, v_m) < r$$

Definició 2.3. *Un espai mètric* (V, d) *es diu complet si es compleix alguna de les següents condicions equivalents:*

1. *Tota succesió de Cauchy de punts en* V *té un límit en* V*-*
2. *Tota succesió de Cauchy en* V *convergeix a* V*.*
3. *Tota succesió decreixent de subespais no buits de* V*, amb diàmetres que tendeixen a 0, té una intersecció no buida: si* F_n *es tancat i no buit,* $F_{n+1} \subseteq F_n$ *per a cada* n*, i* $diam(F_n) \to 0$*, aleshores hi ha un punt* $v \in V$ *comú a tots els conjunts* F_n*.*

Definició 2.4. *Un espai de Banach* V *és un espai normat i complet* $(V, \|\cdot\|)$ *en la mètrica induïda per la seva norma.*

Exemple 2.5. L'espai euclidià $(\mathbb{R}^d, \|\cdot\|_2)$ és un espai de Banach.

En la següent secció treballarem sobre l'espai euclidià $\mathbb{R}^d$, per major simplicitat, però notem que tot el que s'exposa a continuació es pot extendre a un espai de Banach V qualsevol.

Definició 2.6. *Un camí en un espai euclidià* $\mathbb{R}^d$*, és una aplicació contínua* $X : [a, b] \to \mathbb{R}^d$*, on* $[a, b] \subset \mathbb{R}$.

Notarem $X_t = X(t)$ per a denotar la dependència del temps $t \in [a, b]$. Per a $X_t \in \mathbb{R}^d$, representarem la parametrització del camí de la forma:

$$X_t = \{X_t^1, ..., X_t^d\}$$

D'ara en endavant asumirem que tots el camins que considerem són diferenciables a trossos. Quan parlem de camí *suau* voldrem dir que té derivades de tots els ordres.

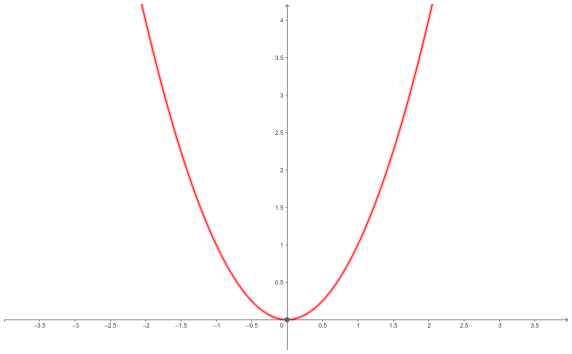


Figura 1: Exemple d'un camí bidimensional suau

Exemple 2.7 (Figura 1)**.** Un exemple de camí suau a $\mathbb{R}^2$ és el següent:

$$X_t = \{X_t^1, X_t^2\} = \{t, t^2\},\ t \in [-4, 4]$$

Exemple 2.8. Un exemple d'un camí diferenciable a trossos pot ser el que observem a la figura 2

$$X_t = \{X_t^1, X_t^2\} = \{t, f(t)\},\ t \in [0, 1]$$

on f és una funció lineal a trossos. Per exemple, f podria ser el preu d'un actiu a temps t. Aquest tipus de camins no suaus poden representar dades seqüencials, com per exemple, sèries temporals. A la Figura 2 observem un exemple de camí estocàstic bidimensional no suau.

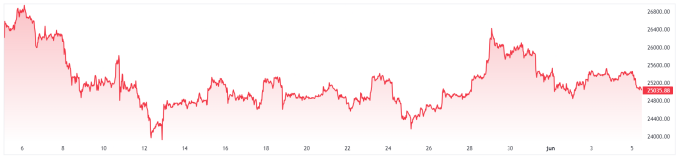


Figura 2: BTC/EUR període: 05/05/2023-05/06/2023.

Definició 2.9. *Siguin $X : [a, b] \to \mathbb{R}$ un camí 1-dimensional i una funció $f : \mathbb{R} \to \mathbb{R}$, la integral de línia de X respecte f es defineix com*

$$\int_a^b f(X_t) dX_t = \int_a^b f(X_t) \dot{X}_t dt, \tag{2.1}$$

on la última integral és de Rieman d'una funció continua acotada i on $\dot{X}_t = dX_t/dt$.

Exemple 2.10. Per al camí constant $Y_t = 1, \forall t \in [a, b]$. La integral de línia de Y_t respecte qualsevol camí $X : [a, b] \to \mathbb{R}$ és l'increment de X :

$$\int_a^b dX_t = \int_a^b \dot{X}_t dt = X_b - X_a$$

Exemple 2.11. Per al camí constant $X_t = t, \forall t \in [a, b]$, $\dot{X}_t = 1$ i la integral de línia per a qualsevol $Y : [a, b] \mapsto \mathbb{R}$ és simplement la integral de Riemann de Y:

$$\int_a^b Y_t dX_t = \int_a^b Y_t dt.$$

Definició 2.12. *Sigui $X : [a,b] \to \mathbb{R}^d$ un camí d-dimensional. Diem que X té una p-variació finita o acotada per cert $p \geq 1$ si la p-variació de X es defineix com*

$$\|X\|_{p,[a,b]} = \left(\sup_{D \subset [a,b]} \sum_l \|X_{t_l} - X_{t_{l-1}}\|^p \right)^{1/p} < \infty$$

on el suprem s'agafa sobre totes les possibles particions finites $D = t_{ll}$ de l'interval $[a,b]$. Denotarem $\mathcal{V}^p([a,b], \mathbb{R}^d)$ el conjunt de tots els camins continus $X : [a,b] \to \mathbb{R}^d$ de p-variació finita.

De vegades als camins de 1-variació finita els anomenarem simplement de variació acotada o finita.

Exemple 2.13. Un moviment brownià fins a temps T, indicat per $\mathcal{B}[0,T]$, té p-variació finita quasi segur si i només si $p > 2$.

A continuació presentem alguns conceptes previs d'ànalisi estocàstic.

Definició 2.14 (Procés estocàstic)**.** *Donat un espai de probabilitats $(\Omega, \mathcal{F}, \mathbb{P})$. Un procés estocàstic X és una aplicació:*

$$X : (\omega, t) \in \Omega \times \mathbb{T} \to X_t(\omega) \in \mathbb{R}$$

mesurable, és a dir, tal que

$$X^{-1}(B) \in \mathcal{F} \otimes \mathcal{B}(\mathbb{T})$$

per a tot $B \in \mathcal{B}(\mathbb{R})$. On $\mathbb{T}$ és un conjunt que mesura el temps. En temps continu, és habitual tenir $\mathbb{T} = [0,T]$ o $\mathbb{T} = [0,\infty)$. En temps discret, $\mathbb{T} := \{0,1,\ldots,N\}$ o $\mathbb{T} = \mathbb{N}$.

Definició 2.15 (Reticulat o norma d'una partició)**.** *Donada una partició d'un interval $[a,b]$*

$$a = x_0 < x_1 < ... < x_{n-1} < x_n = b$$

definim la norma de la partició el reticulat com la longitud

$$màx\{|x_i - x_{i-1}| : i = 1, ..., n\}.$$

És a dir, la longitud més gran de les longituds dels subintervals $[x_{i-1}, x_i]$.

Definició 2.16 (variació quadràtica)**.** *Sigui X_t un procés estocàstic real sobre un espai de probabilitats $(\Omega, \mathcal{F}, \mathbb{P})$ i amb l'índex real $t \in \mathbb{R}_{\geq 0}$. La seva variació quadràtica, denotada $[X]_t$ es defineix com*

$$[X]_t = \lim_{\|P\| \to 0} \sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})^2$$

on P recorre sobre particions de l'interval $[0,t]$ i la norma de la partició P és el reticulat. Aquest límit, d'existir, és la noció de límit en probabilitats.

La variació quadràtica és un cas particular de la covariació variaciáncia creuada.

Definició 2.17 (Covariació)**.** *La covariació o variaciància creuada de dos processos X i Y es defineix com*

$$[X,Y]_t = \lim_{\|P\| \to 0} \sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})(Y_{t_k} - Y_{t_{k-1}})$$

Observació 2.18. La covariació es pot escriure en funció de la variació degut a la identitat de polarització:

$$[X,Y]_t = \frac{1}{2}([X+Y]_t - [X]_t - [Y]_t)$$

La variacó quadràtica també es pot escriure $\langle X\rangle_t$ o $\langle X,X\rangle_T$, així com $QV(X)$

2.2 Definició de signatura

Sigui $X : [a,b] \to \mathbb{R}^d$ un camí, $\forall i \in \{1,...,d\}$, definim la quantitat

$$S(X)^i_{a,t} = \int_{a<s<t} dX^i_s = X^i_t - X^i_0 \tag{2.2}$$

que és l'increment de la i-èssima coordenada de camí a temps $t \in [a,b]$. Observem que $S(X)^i_{a,\cdot} : [a,b] \to \mathbb{R}$ és en si mateixa un camí 1-dimensional.

Definició 2.19. *Sigui $X : [a,b] \to \mathbb{R}^d$ un camí, $\forall i,j \in \{1,...,d\}$, definim la integral doble iterada com*

$$S(X)^{i,j}_{a,t} = \int_{a<s<t} S(X)^i_{a,s} dX^j_s = \int_{a<r<s<t} dX^i_r dX^j_s \tag{2.3}$$

on $S(X)^i_{a,s}$ es donat per (2.2) *i el límits d'integració són:*

$$a<r<s<t = \begin{cases} a<r<s \\ a<s<t. \end{cases} \tag{2.4}$$

Anàlogament, $\forall i,j,k \in \{1,...,d\}$ podem definir la integral triple iterada com

$$S(X)^{i,j,k}_{a,t} = \int_{a<s<t} S(X)^{i,j}_{a,s} dX^k_s = \int_{a<q<r<s<t} dX^i_q dX^j_r dX^k_s \tag{2.5}$$

Podem continuar recursivament:

Per a qualsevol $k \geq 1$ i $i_1,...,i_k \in \{1,...,d\}$ definim la integral k vegades iterada de X recorrent els índexos $i_1,...,i_k$ com

$$S(X)^{i_1,...,i_k}_{a,t} = \int_{a<s<t} S(X)^{i_1,...,i_{k-1}}_{a,s} dX^{i_k}_s = \int_{a<t_k<t} ... \int_{a<t_1<t_2} dX^{i_1}_{t_1} ... dX^{i_k}_{t_k} \tag{2.6}$$

Definició 2.20 (Signatura)**.** *La signatura d'un camí $X : [a,b] \to \mathbb{R}^d$ és la col·lecció (sèrie infinita) de totes les integrals iterades de X. Formalment, és la succesió de nombres reals*

$$S(X)_{a,b} = \left(1, S(X)^1_{a,b}, ..., S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, ...\right) \tag{2.7}$$

on els superíndexos recorren el conjunt de tots els multi-índexos

$$W = \{(i_1,...,i_k) \mid k \geq 1, i_1,...,i_k \in \{1,...,d\}\}. \tag{2.8}$$

El conjunt W es sol anomenar el conjunt de *paraules* de l'*alfabet* $A = \{1,...,d\}$ de d *lletres*.
De vegades denotarem simplement $S(X) = (1, S(X)^1, S(X)^2, ...)$, per a la signatura d'un camí $X : [a,b] \mapsto \mathbb{R}^d$, ometent el subíndex per tal d'aulleugerar notació.

Exemple 2.21. Si considerem l'alfabet de tres lletres $\{1,2,3\}$, tenim un nombre infinit de paraules que podem que podem fer amb aquest alfabet: $(1,2,3,11,12,13,21,22,23,31,32,33,111,112,113,121,...)$.

Observació 2.22. Observem que les integrals iterades d'un camí són independents del punt inicial de X. És a dir, si per algun $x \in \mathbb{R}^d$, definim el camí $\tilde{X}_t = X_t + x$, aleshores $S(\tilde{X})_{a,b}^{i_1,...,i_k} = S(X)_{a,b}^{i_1,...,i_k}$.

A continuació explorem una definició més completa de signatura.

Definició 2.23 (Sèrie de Potències Formal)**.** *Siguin $e_1,...,e_d$ d indeterminats formals. L'àlgebra de sèries de potències formals (no commutatives) en d indeterminats és l'espai vectorial de totes les sèries de la forma*

$$\sum_{k=0}^{\infty} \sum_{i_1,...,i_k \in \{1,...,d\}} \lambda_{i_1,...,i_K} e_{i_1}...e_{i_k}, \tag{2.9}$$

Definició 2.24. *Un polinomi formal (no commutatiu) és una sèrie de potències formal per a la que només un nombre finit de coeficients $\lambda_{i_1,...,i_k}$ és no nul. Els termes $e_{i_1},...e_{i_k}$ s'anomenen monomis.*

El terme que correpson a $k=0$ és un nombre real λ_0. L'espai de les sèries formals es sol anomenar l'àlgebra tensorial de $\mathbb{R}^d$, i el podem escriure $T(\mathbb{R}^d)$.
Observem que si definim la suma i el producte escalar de la següent manera

$$\begin{aligned}\left(\sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} \lambda_{i_1,...,i_K} e_{i_1}...e_{i_k}\right) + \left(\sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} \mu_{i_1,...,i_K} e_{i_1}...e_{i_k}\right) = \\ \sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} (\lambda_{i_1,...,i_K} + \mu_{i_1,...,i_K}) e_{i_1}...e_{i_k})\end{aligned} \tag{2.10}$$

i

$$c\left(\sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} \lambda_{i_1,...,i_K} e_{i_1}...e_{i_k}\right) = \sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} c\lambda_{i_1,...,i_K} e_{i_1}...e_{i_k}) \tag{2.11}$$

Podem definir, a més, el producte $\otimes$ entre monomis ajuntant multi-índexos.

$$e_{i_1}...e_{i_k} \otimes e_{j_1}...e_{j_k} = e_{i_1}...e_{i_k} e_{j_1}...e_{j_k} \tag{2.12}$$

El producte $\otimes$ s'exten llavors de forma única i lineal a totes les sèries de potències formals. Observem-ne els primers termes:

$$\begin{aligned}\left(\sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,...,d\}} \lambda_{i_1,...,i_K} e_{i_1}...e_{i_k}\right) \otimes \left(\sum_{k=1}^{\infty} \sum_{i_1,...,i_k \in \{1,,,d\}} \mu_{i_1,...,i_K} e_{i_1}...e_{i_k}\right) = \\ \lambda_0\mu_0 + \sum_{i=1}^{d} (\lambda_0\mu_i + \lambda_i\mu_0) e_i + \sum_{i=1}^{d} (\lambda_0\mu_{i,j} + \lambda_i\mu_j + \lambda_{i,j}\mu_0) e_i e_j + ...\end{aligned} \tag{2.13}$$

L'espai de sèries formals forma un àlgebra en l'espai vectorial que queda definit junt amb $\otimes$.

Hom pot observar que el conjunt d'índexos dels monomis $e_{i_1}, ..., e_{i_K}$, coincideix amb el conjunt d'índexos dels termes de la signatura d'un camí $X : [a, b] \to \mathbb{R}^d$, més en particular la coll·llecció de tots els multi-índexos $(i_1, ..., i_K)$ on $i_1, ..., i_k \in \{1, ..., d\}$. Podem observar doncs que una forma d'expressar la signatura del camí X és amb una sèrie de potències formal en que el coeficient de cada monomi $e_{i_1}, ..., e_{i_k}$ és defineix com $S(X)_{a,b}^{i_1,...i_k}$, és a dir:

$$S(X)_{a,b} = \sum_{k=0}^{\infty} \sum_{i_1,...,i_k \in \{1,...d\}} S(X)_{a,b}^{i_1,...i_k} e_{i_1} ... e_{i_K} \tag{2.14}$$

Per tant, podem definir alternativament la signatura d'una camí de la forma que s'exposa a continuació:

Definim la n-èssima potència tensorial de $\mathbb{R}^d$ com:

$$(\mathbb{R}^d)^{\otimes n} := \underbrace{\mathbb{R}^d \otimes \mathbb{R}^d \ldots \otimes \mathbb{R}^d}_{n}$$

on $\otimes$ és el producte tensorial. Definim l'*àlgebra tensorial* com

$$T(\mathbb{R}^d) := \{(a_0, a_1, a_2, \ldots) : a_n \in (\mathbb{R}^2)^{\otimes n} \forall n \geq 0\},$$

on prenem $(\mathbb{R}^d)^{\otimes 0} = \mathbb{R}$ per conveni. A més, definim l'*àlgebra tensorial truncada* com

$$T(\mathbb{R}^d) := \bigoplus_{i=0}^{n} (\mathbb{R}^d)^{\otimes i}.$$

Donats dos elements $a = (a_0, a_1, \ldots), b = (b_0, b_1, \ldots) \in T(\mathbb{R}^d)$ introduïm la suma i producte de a i b com

$$a + b := (a_0 + b_0, a_1 + b_1, \ldots)$$

$$a \otimes b = ab := \left(\sum_{i=0}^{n} a_i \otimes b_{n-i} \right)_{n \geq 0}.$$

Que es defineix de forma anànaloga en l'àlgebra tensorial truncada.
Definim doncs:

Definició 2.25 (Signatura)**.** *La signatura d'un camí $X : [a, b] \to \mathbb{R}^d$ és la seqüència*

$$S(X)_{a,b} = \left(1, X_{a,b}^1, X_{a,b}^2, \ldots\right) \in T(\mathbb{R}^d) \tag{2.15}$$

on

$$X_{a,b}^n := \int_{a<t_1<t_2<\ldots<t_n<b} \cdots \int dX_{t_1} \otimes \ldots \otimes dX_{t_n} \in (\mathbb{R}^d)^{\otimes n} \ \forall n \geq 0 \tag{2.16}$$

De nou, de vegades ometrem els subíndexos per a alleugerar notació.

Exemple 2.26. 1. Quan X té 1-variació finita, la integració es pot entendre en el sentit de la integral de Stieltjes.

2. Quan X és un moviment brownià, la seva signatura es pot definir en el sentit de la integral d'Itô o la integral de Stratonovich.

També ens interessa la següent definició:

Definició 2.27 (Nivell k-éssim de la signatura)**.** *Definim el nivell k-èssim de la signatura o signatura truncada d'ordre k d'un camí* $X : [a, b] \to \mathbb{R}^d$*, com la col·lecció finita de tots els termes* $S(X)_{a,b}^{i_1,...,i_k}$*, on el multi-índex és de longitud k.*

Així, per exemple, el primer nivell de la signatura és la col·lecció de d nombres reals $S(X)_{a,b}^1, ..., S(X)_{a,b}^d$ i el segon nivell és la col·lecció de d^2 nombres reals

$$S(X)_{a,b}^{1,1}, ..., S(X)_{a,b}^{1,d}, S(X)a, b^{2,1}, ..., S(X)_{a,b}^{2,d}.$$

En termes de la segona definiciíque hem ofert de signatura la *signatura truncada* de nivell k és

$$S(X)_{a,b}^k = (1, X^1, ..., X^k).$$

Exemple 2.28. Considerem un camí unidimensional $X : [a, b] \to \mathbb{R}$, $X_t = t$, en aquest cas l'alfabet és $A = \{1\}$ i le conjunt de paraules és $W = \{(1, ..., 1) \mid k \geq 1\}$, on 1 apreix k vegades en $(1, ..., 1)$. La signatura és doncs

$$S(X)_{a,b}^1 = X_b - X_a, \tag{2.17}$$

$$S(X)_{a,b}^{1,1} = \frac{(X_b - X_a)^2}{2!},$$

$$S(X)_{a,b}^{1,1,1} = \frac{(X_b - X_a)^3}{3!},$$

$$\vdots$$

Observació 2.29. Es por demostrar que aquesta últijma expressió és certa per a qualsevol camí $X : [a, b] \to \mathbb{R}$. Per a camins unidimensionals la signatura només depen de l'increment $X_b - X_a$.

Exemple 2.30. Considerem un camí bidimensional. Tenim l'alfabet $A = \{1, 2\}$ com a conjunt de índexos, i el conjunt de paraules o multi-índexos $W = \{(i_1, ..., i_K) \mid k \geq 1,\ i_1, ..., i_k \in \{1, 2\}\}$. Considerem el següent camí a $\mathbb{R}^d$:

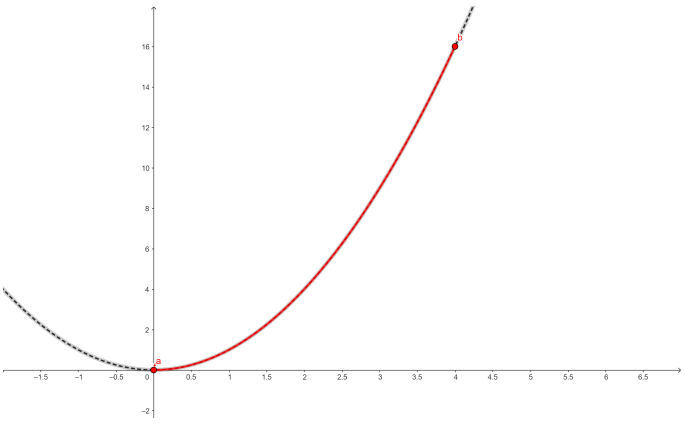


Figura 3: Camí bidimensional X_t, $t \in [0, 4]$.

$$\begin{aligned} X_t &= \{X_t^1, X_t^2\} = \{t, t^2\}, t \in [0, 4]. \\ dX_t &= \{dX_t^1, dX_t^2\} = \{dt, 2tdt\}. \end{aligned} \tag{2.18}$$

Computem la signatura

$$S(X)^1_{0,4} = \int 0 < t < 4 dX^1_t dt = \int_0^4 = dt = X^1_4 - X^1_0 = 4, \tag{2.19}$$

$$S(X)^2_{0,4} = \int_{0<t<4} dX^2_t dt = \int_0^4 2t dt = X^2_4 - X^2_0 = 16,$$

$$S(X)^{1,1}_{0,4} = \iint_{0<t_1<t_2<4} dX^1_{t_1} dX^1_{t_2} dt = \int_0^4 \left[\int_0^{t_2} dt_1\right] dt_2 = 8,$$

$$S(X)^{1,2}_{0,4} = \iint_{0<t_1<t_2<4} dX^1_{t_1} dX^2_{t_2} dt = \int_0^4 \left[\int_0^{t_2} dt_1\right] 2t_2 dt_2 = \frac{128}{3},$$

$$S(X)^{2,1}_{0,4} = \iint_{0<t_1<t_2<4} dX^2_{t_1} dX^1_{t_2} dt = \int_0^4 \left[\int_0^{t_2} 2t_1 dt_1\right] dt_2 = \frac{64}{3},$$

$$S(X)^{2,2}_{0,4} = \iint_{0<t_1<t_2<4} dX^2_{t_1} dX^2_{t_2} dt = \int_0^4 \left[\int_0^{t_2} 2t_1 dt_1\right] 2t_2 dt_2 = 128,$$

$$S(X)^{1,1,1}_{0,4} = \iiint_{0<t_1<t_2<t_3<4} dX^1_{t_1} dX^1_{t_2} dX^1_3 = \int_0^4 \left[\int_0^{t_3} \left[\int_0^{t_2} dt_1\right] dt_2\right] dt_3 = \frac{32}{3},$$

$$\vdots$$

Si seguim de la mateixa manera, podem calcular cada terme $S(X)^{i_1,...,i_K}_{0,5}$ de la signtatura per a cada multi-índex $(i_1, ..., i_k)$ amb $i_1, ..., i_k \in \{1, 2\}$.

2.3 Intuició geomètrica dels primers nivells

A continuació donem el significat geomètric dels dos primers nivells de la signatura.

2.3.1 Primer nivell

El primer nivell $(1, X^1, ..., X^d)$ és l'increment del camí $X : [a, b] \to \mathbb{R}^d$, és a dir que

$$X^{(i)} = X^i_b - X^i_a$$

2.3.2 Termes d'ordre superior

Les integrals iterades d'ordre superior poden ser interpretades com polinomis generalitzats de camins. La regressió lineal sobre aquests polinomis generalitzats es pot interpretar com a regressió polinomial directe sobre els camins.
No obstant, no tots els termes son necessaris per a la representació única d'un camí en termes de les seves integrals iterades doncs certes integrals iterades estan determinades per altres integrals iterades. Per exemple

$$\begin{aligned} X^{(i,i)} &= \frac{(X^{(i)})^2}{2} = \frac{(X^i_b - X^i_a)^2}{2},\ i \in \{1, ..., d\} \\ X^{(i,j)} &= X^{(i)} X^{(j)} - X^{(j,i)},\ i, j \in \{1, ..., d\} \end{aligned} \tag{2.20}$$

Aquestes últimes igualtats són conseqüència de la identitat del producte mescla (Teorema 2.8) que veiem més endavant a la subsecció 2.4.2.. La *log-signatura* que es presenta a la subsecció 2.4.5 presenta la mínima informació que determina únicament totes les integrals iterades d'un camí a través d'operacions no lineals.

2.3.3 Segon nivell (àrea)

Observem doncs que $X^{(i,i)}_{a,b}$ és $(X^i_b - X^i_a)^2/2$. Per al terme $X^{(i,j)}_{a,b}$ per $i \neq j$. Per $1 \leq i \leq j \leq d$, els termes

$$A^{i,j}_{a,b} := \frac{1}{2}\left(\int_{a<t_1<t_2<b} dX^1_{t_1} dX^2_{t_2} - \int_{a<t_1<t_2<b} dX^2_{t_1} dX^1_{t_2}\right) = \frac{1}{2}\left(X^{(i,j)}_{a,b} - X^{(j,i)}_{a,b}\right)$$

determinen l'àrea signada entre la corba $t \mapsto (X^i_t, X^j_t)$, per $t \in [a,b]$ i la corda que uneix els punts (X^i_a, X^j_a) i (X^i_b, X^j_b). Aquesta área es la que es coneix com *àrea de Lévy*.

En la figura 4, pel camí bidimensional $X_t = \{X^1_t, X^2_t\}$ les àrees signades A_- i A_+ són negativa i positiva respectivament i ΔX^1_t i ΔX^2_t són els increments en cada coordenada.

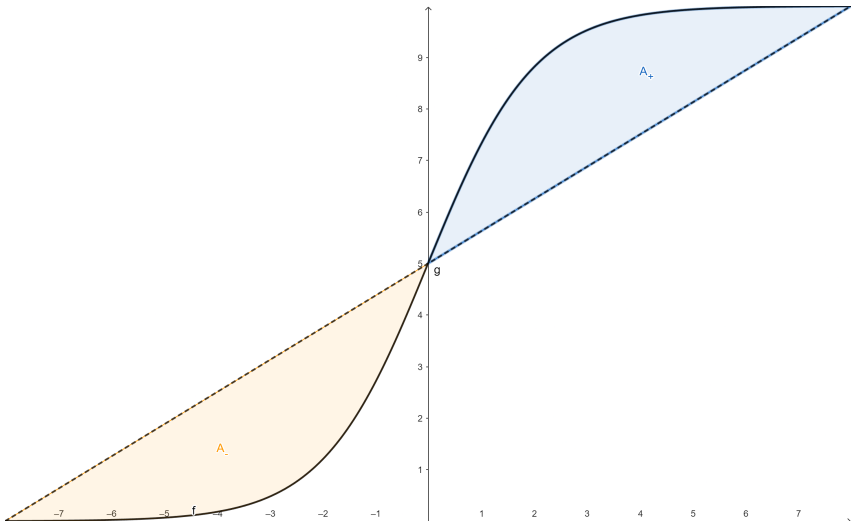


Figura 4: Camí bidimensional X_t, $t \in [0,4]$.

2.3.4 Tercer nivell (àrea de segon ordre)

La gràfica de la figura 4 es pot dividir en dues trajectòries en dues dimensions. Totes dues trajectòries tenen increments unitaris en ambdues components i àrees signades nulles. L'àrea de segon ordre es calcula mitjançant la fórmula

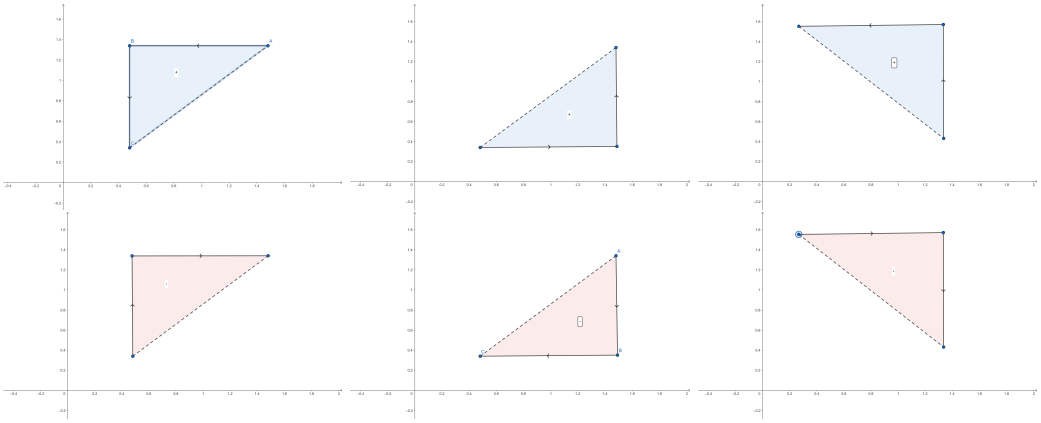
$$A^{a,b}_{1,(1,2)} := \frac{1}{2}\left(\int_{a<t_1<t_2<b} dX^1_{t_1} dA^{1,2}_{b,t_2} - \int_{a<t_1<t_2<b} dA^{1,2}_{b,t_1} dX^1_{t_2}\right) = \frac{1}{2}(X^{(1,1,2)}_{a,b} - X^{(1,2,1)}_{s,t}),$$

l'últim pas s'infereix de la propietat que es presenta a la subsecció 2.4.2 (una demostració d'això es por trovar a [13]).

Mitjançant arguments recursius similars, es poden identificar àrees d'ordre superior i representar-les en termes de combinacions lineals d'integrals iterades.

2.3.5 Signe the l'àrea signada

A la figura 5 observem les sis possibilitats en funció de la direcció del camí amb el que estem treballant: les tres de dalt son positives i les tres d'abaix negatives. Això té relació directe amb l'índex del camí (vegis [14]).



2.4 Propietats de la signatura

A continuació presentem algunes propietats de la signatura, en deixem diverses sense demostrar però les demostracions es poden trobar en la bibliografia citada

2.4.1 Invariància respecte a reparametritzacions del temps

Anomenem *reparametrització* a una aplicació $\psi : [a,b] \to [a,b]$ exhaustiva, contínua i no decreixent. D'ara en endavant considerarem reparametritzacions suaus només.

Considerem dos camins $X, Y : [a,b] \to \mathbb{R}$ i una reparametrització $\psi : [a,b] \to [a,b]$. Definim els camins $\tilde{X}, \tilde{Y} : [a,b] \to R$ com $\tilde{X}_t = X_{\psi(t)}$ i $\tilde{Y}_t = Y_{\psi(t)}$. Notem que $\dot{\tilde{X}}_t = \dot{X}_{\psi(t)}\dot{\psi}(t)$ del que es segueix que

$$\int_a^b \tilde{Y}_t d\tilde{X}_t = \int_a^b Y_{\psi(t)}\dot{X}_{\psi(t)}\dot{\psi}(t)dt = \int_a^b Y_u dX_u, \tag{2.21}$$

on a la última igualtat hem usat la sustitució $u = \psi(t)$. És a dir, que les integrals de línia són invariants sota reparametritzacions del temps dels dos camins.

Considerem ara un camí $X : [a,b] \to \mathbb{R}^d$ i una reparametrització $\psi : [a,b] \to [a,b]$. Com abans, $\tilde{X}_t = X_{\psi(t)}$. Com que cada terme de la signatura $S(X)_{a,b}^{i_1,...,i_k}$ és una integral de línia iterada de X, es dedueix de (2.19) que

$$S(\tilde{X})_{a,b}^{i_1,...,i_K} = S(X)_{a,b}^{i_1,...,i_k}, \forall k \geq 0,\ i_1,...,i_k \in \{1,...,d\}. \tag{2.22}$$

2.4.2 Linealitat: Producte mescla

Una permutació σ de $\{1,...,k+m\}$ s'anomena (k,m)-mescla si $\sigma^{-1}(1) < ... < \sigma^{-1}(k)$ i $\sigma^{-1}(k+1) < ... < \sigma^{-1}(k+m)$. La llista $(\sigma(1),...,\sigma(k+m))$ també s'anomena una mescla de $(1,...,k)$ i $(k+1,...,k+m)$. Denotarem $M(k,m)$ la col·lecció de totes les (k,m)-mescles.

Definició 2.31 (Producte Mescla)**.** *Considerem dos multi-índexos $I = (i_1,...,i_k)$ i $J = (j_1,...,j_m)$ amb $i_1,...,i_k,j_1,...,j_m \in \{1,...,d\}$. Considerem el multi-índex*

$$(r_1,...,r_k,r_{k+1},...,r_{k+m}) = (i_1,...,i_k,j_1,...,j_m). \tag{2.23}$$

El producte mescla de I i J, que denotem $I \sqcup\!\sqcup J$, és un conjunt finit de multi-índexos de longitud $k+m$ de la següent manera

$$I \sqcup\!\sqcup J = \{(r_{\sigma(1)},...,r_{\sigma(k+m)}) \mid \sigma \in M(k,m)\} \tag{2.24}$$

Teorema 2.32 (Identitat del producte mescla). *Donat un camí $X : [a,b] \to \mathbb{R}^d$ i dos multi-índexos $I = (i_1, ..., i_k)$ i $J = (j_1, ..., j_m)$ amb $i_1, ..., i_k, j_1, ..., j_m \in \{1, ..., d\}$ aleshores*

$$S(X)^I_{a,b} S(X)^J_{a,b} = \sum_{K \in I \sqcup\!\sqcup J} S(X)^K_{a,b}. \tag{2.25}$$

El producte mescla implica, en particular, que el producte de dos termes de la signatura es pot expressar com a una combinació lineal de termes d'ordre superior, cosa que serà útil de cara a les aplicacions pràciques.

2.4.3 Identitat de Chen

Definició 2.33 (Concatenació). *Per a dos camins $X : [a,b] \to \mathbb{R}^d$ i $Y : [b,c] \to \mathbb{R}^d$, definim la seva concatenació com el camí $X * Y : [a,c] \to \mathbb{R}^d$ de la següent forma*

$$(X * Y)_t = \begin{cases} X_t & t \in [a,b] \\ X_b + (Y_t - Y_b) & t \in [b,c] \end{cases}$$

La prova de la següent identitat es pot trobar a [13].

Teorema 2.34 (Identitat de Chen). *Sigui $p \in [1,2)$ $X : [a,b] \to \mathbb{R}^d$ i $Y : [a,b] \to \mathbb{R}^d$ camins de p-variació finita. Aleshores*

$$S(X * Y)_{a,c} = S(X)_{a,b} \otimes S(Y)_{b,c} \tag{2.26}$$

La identitat, ens diu que la signatura trasnforma el producte $*$, o concatenació en el producte $\otimes$.

Exemple 2.35. Sigui X un camí lineal a trossos de valor real, és a dir, X es la concatenació d'un nombre finit de camins lineals, o sigui que existeixen un enter positiu l i $X_1, X_2, ..., X_l$ camins lineals tals que

$$X = X_1 * X_2 * \cdots * X_l$$

La identitat de Chen ens dona un mètode per a calcular la signatura d'un camí lineal a trossos com

$$S(X) = \bigotimes_{i=1}^{l} exp(X_i)$$

2.4.4 Inversió en el temps

Definició 2.36 (Inversió en eltemps). *Sigui $X : [a,b] \to \mathbb{R}^d$, definim la seva inversió en el temps com el camí $\overleftarrow{X} : [a,b] \to \mathbb{R}^d$ de forma que $\overleftarrow{X}_t = X_{a+b-t}$, $\forall t \in [a,b]$.*

Tenim la següent propietat algebraica de la signatura:

Teorema 2.37 (Signatura invertida en el temps). *Donat un camí $X : [a,b] \to \mathbb{R}^d$, aleshores*

$$S(X)_{a,b} \otimes S(\overleftarrow{X})_{a,b} = 1. \tag{2.27}$$

2.4.5 Log-signatura

Donada una sèrie de potències formal,

$$x = \sum_{k=0}^{\infty} \sum_{i_1,...,i_k \in \{1,...,d\}} \lambda_{i_1,...,i_k} e_{i_1} ... e_{i_k} \tag{2.28}$$

amb $\lambda_0 > 0$, definim el seu logaritme com la sèrie de potències

$$\log x = log\lambda_0 + \sum_{n \geq 1} \frac{(-1)^n}{n} \left(1 - \frac{x}{\lambda_0}\right)^{\otimes n}, \tag{2.29}$$

on amb $\otimes n$ denotem la n-èssima potència respecte el prodcute $\otimes$.
D'aquesta manera, per exemple, per $\lambda \in \mathbb{R}$ i la sèrie:

$$x = 1 + \sum_{k \geq 1} \frac{\lambda^k}{k!} e_1^{\otimes k}, \tag{2.30}$$

aleshores, es pot verificar que

$$\log x = \lambda e_1. \tag{2.31}$$

Observació 2.38. En general, $\log x$ és una sèrie amb una quantitat infinita de termes, emperò per a cada multi-índex $(i_1, ..., i_k)$, el coeficient de $e_{i_1} ... e_{i_k}$ en $\log x$ només depèn dels coeficients de x de la forma $\lambda_{j_1,...,j_m}$ amb $m \leq k$, dels quals n'hi ha un nombre finit, de forma que $\log x$ és ben diefinit sense que calgui considerar la convergència de la sèrie infinita.

Definició 2.39 (Log-signatura)**.** *Sigui $X : [a, b] \to \mathbb{R}^d$ un camí, definim la log-signatura de X com la sèrie de potències* $\log S(X)_{a,b}$.

Definició 2.40 (Parèntesi de Lie)**.** *Donades x i y dues sèries de potències formals, definim el parèntesi de Lie per*

$$[x, y] = x \otimes y - y \otimes x \tag{2.32}$$

Els primers termes de la log-signatura s'expressen:

$$\log S(X)_{a,b} = \sum_{i=1}^{d} S(X)_{a,b}^{i} e_1 + \sum_{1 \leq i < j \leq d} \frac{1}{2} \left(S(X)_{a,b}^{i,j} - S(X)_{a,b}^{j,i} \right) [e_i, e_j] + ... \tag{2.33}$$

Observem, de fet, que el coeficient del segon terme és precisament, l'àrea de Lévy.
El següent teorema, demostrat per Chen [5], generaliza la fórmula de Camper-Baker-Hausdorff.

Teorema 2.41. *Sigui $X : [a, b] \to \mathbb{R}^d$ un camí. Existeixen nombres reals $\lambda_{i_1,...i_k}$ tal que*

$$\log S(X)_{a,b} = \sum_{k \geq 1} \sum_{i_1,...,i_k \in \{1,...,d\}} \lambda_{i_1,...,i_k} [e_{i_1}, [e_{i_2}, ...[e_{i_{k1}}, e_{i_k}]...]]. \tag{2.34}$$

Observació 2.42. Els coeficients $\lambda_{i_1,...,i_k}$ en general no són únics ja que els polinomis de la forma $[e_{i_1}, [e_{i_2}, ..., [e_{i_{k-1}}, e_{i_k}]...]]$ no són linealment independents.

2.4.6 Iteracions de Picard: origen de la signatura

El primer cop que es va estudiar la signtura d'un camí va ser a mans del geometre Kuo-Tsai Chen a l'any 1957 en l'estudi de la integració de camins ([5]). Les integral iterades de camins apareixen a l'hora d'aplicar iteracions de Picard per a resoldre equacions diferencials ordinaries controlades de la forma

$$dY_t = g(Y_t)dX_t, \quad Y_a = y_a \tag{2.35}$$

Considerem el camí $X \in \mathcal{V}^1([a,b],\mathbb{R}^d)$. Anomenem $\mathcal{L}(\mathbb{R}^d,\mathbb{R}^e)$ a l'espai vectorial de les aplicacions lineals de $\mathbb{R}^d$ a $\mathbb{R}^e$. Equivalentment, $\mathcal{L}(\mathbb{R}^d,\mathbb{R}^e)$ es pot considerar com l'espai vectorial de les matrius reals de mida $d \times e$. Donat un camí $Z : [a,b] \to \mathcal{L}(\mathbb{R}^d,\mathbb{R}^e)$, observem que podem definir la integral

$$\int_a^b Z_t dX_t \tag{2.36}$$

com un element de $\mathbb{R}^e$ de la mateixa manera que la integral de línia usual. Donada una funció $g : \mathbb{R}^e \to \mathcal{L}(\mathbb{R}^d,\mathbb{R}^e)$ i un camí $Y : [a,b] \to \mathbb{R}^e$, diem que Y resol l'equació diferencial controlada 2.35 quan $\forall t \in [a,b]$ resol l'equació de Volterra

$$Y_t = y_a + \int_a^t g(Y_s)dX_s. \tag{2.37}$$

La funció g en l'expressió anterior sovint s'anomena una col·lecció de camps vectorials conduïts, el camí X s'anomena control o conductor, i Y s'anomena solució o resposta.

Un procediment típic per a obtenir una solució de 2.37 és mitjançant iteracions de Picard. Donat un camí arbitrari $Y : [a,b] \to \mathbb{R}^e$, definim un nou camí $F(Y) : [a,b] \to \mathbb{R}^e$ mitjançant

$$F(Y)_t = y_a + \int_a^t g(Y_s)dX_s. \tag{2.38}$$

Y és una solució de 2.37 si i només si és un punt fix de F. Considerem la succesió de camins $Y_t^n = F(Y^{n-1})_t$ amb un camí arbitrari inicial Y_t^0 (sovint pres com el camí constant $Y_t^0 = y_a$). Sota certes condicions adequades, es pot demostrar que F té un únic punt fix Y i que $Y_t^n \to Y$ per $n \to \infty$ en la norma $\|\cdot\|_{1,[a,b]}$ (*teorema del Picard-Lindelöf*).

Suposem ara que $g : \mathbb{R}^e \to \mathcal{L}(\mathbb{R}^d,\mathbb{R}^e)$ és una aplicació lineal (també podem tractar g com a una aplicació lineal $\mathbb{R}^d \to \mathcal{L}(\mathbb{R}^e,\mathbb{R}^e)$, on $\mathcal{L}(\mathbb{R}^e,\mathbb{R}^e)$ és l'espai de totes les matrius reals $e \times e$). Comencem les iteracions de Picard amb el camí constant inicial $Y_t^0 = y_a$ per a tot $t \in [a,b]$. Denotem per I_e la matriu identitat a $\mathcal{L}(\mathbb{R}^e,\mathbb{R}^e)$, podem expressar les iteracions de F de la següent manera:

$$\begin{aligned}
Y_t^0 &= y_a, \\
Y_t^1 = y_a + \int_a^t g(Y_s^0)\, dX_s &= \left[\int_a^t \frac{dg(X_s)}{ds} + I_e\right](y_a), \\
Y_t^2 = y_a + \int_a^t g(Y_s^1)\, dX_s &= \left[\int_a^t \int_s^a \frac{dg(X_u)}{ds}\, dg(X_s) + \int_a^t \frac{dg(X_s)}{ds} + I_e\right](y_a), \\
&\vdots
\end{aligned} \tag{2.39}$$

$$Y_t^n = y_a + \int_a^t g(Y_s^{n-1})\, dX_s = \left[\sum_{k=1}^{n} \int_{a<t_1<\ldots<t_k<t} \frac{dg(X_{t_1})}{dt_1} \ldots \frac{dg(X_{t_k})}{dt_k} + I_e \right] (y_a),$$

$$\vdots$$

Com que $\mathcal{L}(\mathbb{R}^e, \mathbb{R}^e)$ és un àlgebra de matrius, cada quantitat

$$\int_{a<t_1<\ldots<t_k<t} \frac{dg(X_{t_1})}{dt_1} \ldots \frac{dg(X_{t_k})}{dt_k} \tag{2.40}$$

es pot definir naturalment com un element de $\mathcal{L}(\mathbb{R}^e, \mathbb{R}^e)$, que, com es pot comprovar, queda completament determinat (de manera lineal) la signatura truncada de nivell k $S(X)_{a,t}$ amb $t \in [a, b]$.

Pel que en podem extreure que la solució Y_t queda completament determinada per la signatura $S(X)_{a,t}$ per a cada $t \in [a, b]$. En particular, si les signatures de dos conductors X i X_e coincideixen en el temps $t \in [a, b]$, és a dir, $S(X)_{a,t} = S(X_e)_{a,t}$, llavors les solucions corresponents a 2.37 també coincideixen en el temps t per a qualsevol elecció dels camps vectorials lineals g.

Un resultat menys evident, és que el mateix és cert per als camps vectorials no lineals g. Aquest resultat va ser obtingut per primer cop per K.T. Chen per a una certa classe de camins suaus a trossos, i recentmentha estat ampliat per Hambly i Lyons [7] per a camins de variació finita, i per Boedihardjo, Geng, Lyons i Yang [14] per a camins geomètrics rugosos en un marc de camins completament no suaus en què la signatura encara està ben definida. Aquesta darrera classe de camins és d'un interès especial en l'anàlisi estocàstic.

2.4.7 Declivi Factorial

Els ordinadors només poden emmagatzemar un nombre finit de termes de la seqüència, de manera que només podem treballar amb la signatura truncada. Per tant, tot i que la signatura d'un camí descrgués bé aquest camí, la signatura truncada podria no fer-ho, ja que els termes que s'eliminen podrien contenir una gran quantitat d'informació. No obstant, el resultat següent mostra que aquest no és el cas, que els termes de la signatura disminueixen factorialment en tamany a mesura que l'ordre augmenta.

Aquest teorema el podem trobar demostrat com a lema a [13].

Teorema 2.43. *Sigui $X : [a, b] \to \mathbb{R}^d$ un camí continu de variació acotada. Aleshores, donats $1 \leq i_1 \leq \ldots \leq i_n \leq d$,*

$$\left\| \int_{a<t_1<\ldots<t_n<b} \cdots \int dX_{t_1}^{i_1} \ldots dX_{t_n}^{i_n} \right\| \leq \frac{\|X\|_1^n}{n!}$$

on

$$\|X\|_1 := \sup_{t_i} \subset [a, b] \sum_i |X_{t_{i+1}} - X_{t_i}|,$$

i el suprem es pren sobre totes les particions de $[a, b]$.

Per tant, si la signatura d'un camí el descriu bé, la signatura truncada també ho farà i pot ser usada com a característica.

2.4.8 Unicitat de la signatura

Una quëstió que ens podem plantejar és si un camí està totalment determinat per la seva signatura, es despren tant de les identitas de Chen i del producte mescla com de les propietats d'inversió en el temps i d'invariància sota reparametritzacions, que la resposta en general és no. Així, per exemple, per la propietat d'invariància respecte parametritzacions, no es pot recuperar de la signatura d'un camí, la velocitat a la que s'estava recorrent, tampoc podem, per exemple, diferenciar la signatura d'un camí constant trivial de la d'un camí concatenat amb el seu invers en el temps.

No obstant, aquesta és fonamentalment la única informació que es perd del camí. Per a un camí X que no es creua a si mateix, la signatura descriu completament les propietats geomètriques del camí, més concretament, els punts pels que passa i en quin ordre passa per ells. Això últim es formalitzat a [7] amb la noció de camí *Tree-Like*.

Teorema 2.44. *Sigui $X : [a, b] \to \mathbb{R}^d$ un camí continu de variació acotada. Aleshores $S(X)_{a,b}$ determina X fins a una equivalencia Tree-Like.*

Això, vol dir que la signatura determina camins fins a seccions en les que el camí es creua a si mateix. Per tant, la signtaura únicament determina camins amb almenys una component monotona, ja que aquesta impedeix que un camí es creui a si mateix.

Exemple 2.45. Un camí de la forma $Y : [a, b] \to \mathbb{R}^d$, $Y(t) = (t, X_t)$ queda úincament determinat per la seva signatura de la manera en que hem descrit anteriorment.

3 Transformació d'un flux de dades en camí

Suposem que tenim un conjunt de dades (o flux de dades), el que volem és transformar aquesta successió discreta en una aplicació contínua, és a dir, en un camí, incrustant els punts (dades) en ella. A continuació presentem diferents maneres d'incrustar dades en camins continus. A la pràctica, quina d'aquestes maneres s'emplea depèn del que ens interessa més en cada cas.

Primer donem una definició més formal dels següents conceptes.

Definició 3.1. *Un conjunt (o flux) de dades $\Omega \subset \mathbb{R}^d$ és una seqüència, és a dir, un element de*

$$\mathcal{S}(\mathbb{R}^d) := \{(x_1, ..., x_k) \mid k \in \mathbb{N},\ x_i \in \mathbb{R}^d,\ i \in \{1, ..., k\}\}. \tag{3.1}$$

Per a cada $i \in \{1, ..., k\}$, diem que x_i és una dada.

Definició 3.2. *La longitud d'una seqüència $\Sigma \in \mathcal{S}(\mathbb{R}^d)$ és la seva cardinalitat: $\#(\Sigma)$.*

Definició 3.3. *Un conjunt de dades etiquetat $\Omega \subset \mathbb{R}^d \times \mathbb{R}$ és una seqüència (o flux), és a dir, un element de*

$$\mathcal{S}(\mathbb{R}^d \times \mathbb{R}) := \{((x_1, y_1), ..., (x_k, y_k)) \mid k \in \mathbb{N},\ (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R},\ i \in \{1, ..., k\}\}. \tag{3.2}$$

Per a cada $i \in \{1, ..., k\}$, diem que x_i és una dada i y_i la seva etiqueta.

3.1 Incrustació mitjançant interpolació linial a trossos

En aquest cas, donat el conjunt de dades, unim les dades amb línies rectes, és a dir amb el camí amb distància euclidiana menor, en un determinat ordre.
Si les dades són ordenades, per exemple, sèries temporals, tenim $\Sigma = ((t_1, x_1), ..., (t_k, x_k))$, amb $(t_i, x_i) \in \mathbb{R} \times \mathbb{R}^{d-1}$, $1 \leq i \leq k$. Aleshores, $\forall t_i, t_{i+1} \in \{t_1, ..., t_k\}$ interpolem amb el polinomi de Newton

$$f(t|t_i; t_{i+1}) = f(t_i) + \frac{f(t_{i+1}) - f(t_i)}{t_{i+1} - t_i}(t - t_i), \tag{3.3}$$

on $f(t_i) = x_i$ i $f(t_{i+1}) = x_{i+1}$. És a dir el graf de l'aplicació és el camí (geomètric) que uneix cada punt (t_i, x_i) amb el punt (t_{i+1}, x_{i+1}) en línia recta.

Exemple 3.4. Sigui $\{(t, X)\} = \{(t_i, X_i)\}_{i=0}^4 = \{(0, 1), (1, 4), (2, 2), (3, 6)\}$ a la figura 5 podem observar el resultat d'incrustar aquest flux amb interpolació lineal a trossos

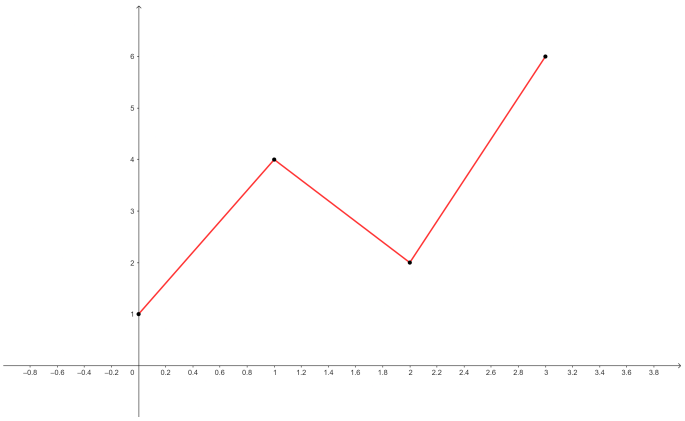


Figura 5: Interpolació lineal a trossos de $\{(t_i, X_i)\}_{i=0}^4$

3.2 Incrustació mitjançant interpolació rectilínia

En aquest cas unim cada parell de punts $\{(t_i, x_i), (t_{i+1}, x_{i+1})\}$ amb el camí més curt (de menor longitud) amb distància Manhattan, és a dir, la distància $d(\cdot)$ tal que $d(x, y) = d((x_1, ..., x_d), (y_1, ..., y_d)) = \sum_{i=1}^d |x_i - y_i|$, $\forall x, y \in \mathbb{R}^d$, recorrent sempre primer la recta inscrita a l'hiperplà $\{t = t_i\}$ (o a qualsevol altre, sempre i quan sigui el mateix durant tot el procés).

Exemple 3.5. Prenent el mateix flux que a l'últim exemple, a la figura 6 podem observar el resultat de la incrustació mitjançant interpolació rectilínia.

3.3 Suma cumulativa

Per a algunes aplicacions numèriques el flux de dades original, en ocasions, pot ser graficat de formes diferents. Una d'aquestes és la suma cumulativa o successió de sumes parcials.

Definició 3.6. *Donat $\{X_i\}_{i=0}^n$ definim la seva suma cumulativa o successió de sumes parcials com*

$$CS(\{X_i\}_{i=0}^n) = \{S_1, ..., S_n\}$$

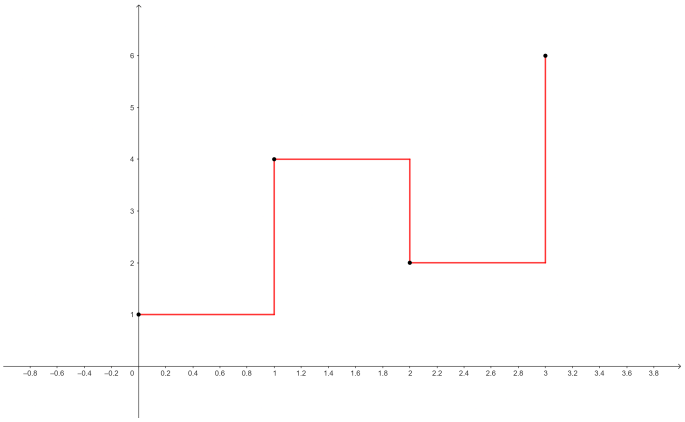


Figura 6: Interpolació rectilínia de $\{(t_i, X_i)\}_{i=0}^{4}$

on $S_k = \sum_{i=0}^{k} X_i$, *per a* $k \in \{1, .., n\}$.

Exemple 3.7. Tornem a prendre el flux dels dos últims exemples. Aleshores $CS(X) = \{1, 5, 7, 13\}$. A la figura 7 podem veure el resultat d'aplicar les dues interpolacions anteriors a $\{(t, CS(X)\}$

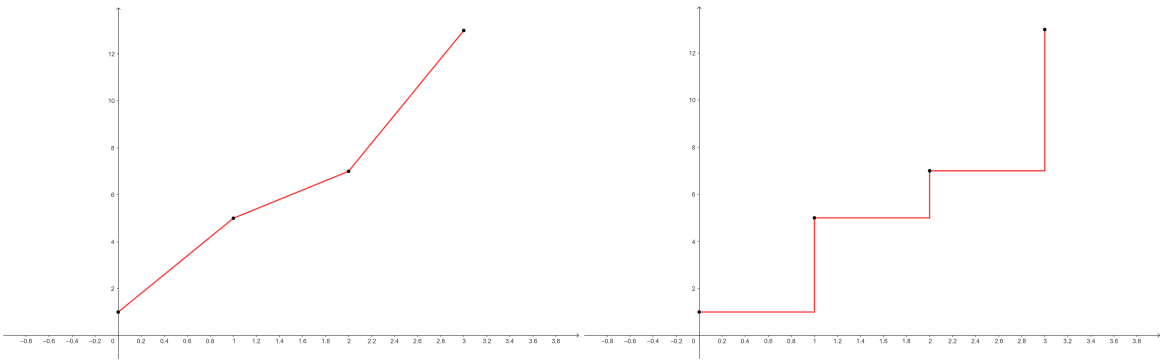


Figura 7: Interpolacions lineal a trossos i rectilinia de $\{(t, CS(X)\}$

3.4 Incrustació mitjançant augmentacions

Un altre tipus de transformacions més complexes passen per augmentar el flux $\Omega \subset \mathbb{R}^d$ abans d'interpolar entre els punts resultants per a incrustar-los en un camí. Més precisament, considerem una aplicació fixada $\theta : \mathbb{R}^d \to \mathbb{R}^e$, i interpolem el conjunt de dades augmentat $\theta(\Omega) \subset \mathbb{R}^e$. Algunes de les augmentacions explícites més comunment utilitzades són: l'augmentació *Lead-Lag*, l'augmentació *Time*, l'augmetanció de *No Future Pause* i l'augmentació *Invisibility Reset*.

Definició 3.8. *L'augmentació Lead-Lag és l'aplicació* $T : S(\mathbb{R}^d) \to S(\mathbb{R}^d \times \mathbb{R}^d)$ *definida per*

$$T(x_1, ..., x_d) := ((x_1, x_1), (x_2, x_1), (x_2, x_2), ..., (x_j, x_j), (x_{j+1}, x_j), ..., (x_d, x_d)) \tag{3.4}$$

L'augmentació *Lead-Lag* pren una seqüència de longitud d i la converteix en una seqüència de longitud $2d$, dividint la original en dues còpies etiquetades, *flux del futur* i *flux de passat*. Hi ha un retard entre quan s'actualitza el futur i quan s'actualitza el passat posteriorment. La longitud del retard pot variar, donant lloc a un Lead-Lag on el nombre de passos entre l'actualització del futur i la del passat és d'almenys dos. A més, es poden enregistrar més d'una seqüència del passat, sent cada una d'elles actualitzada amb un nombre diferent de passos de retard entre l'actualització de la seqüència del futur i la seva pròpia actualització.

Exemple 3.9. L'augmentació Lead-Lag, amb 2 passos de retard és l'aplicació $T : S(\mathbb{R}^d) \to S(\mathbb{R}^d \times \mathbb{R}^d)$ definida per

$$T(x_1, ..., x_d) := ((x_1, x_1), (x_2, x_1), (x_3, x_1), (x_3, x_2), (x_4, x_2), (x_5, x_2), ..., (0, x_d)) \tag{3.5}$$

L'augmentació *Lead-Lag* definida en primer lloc és l'agumentació *Lead-Lag* amb un pas de retard.

Definició 3.10. *Donats k fluxos del passat, definim l'augmentació No Future Pause com l'aplicació* $T : S(\mathbb{R}^d) \to S(\mathbb{R}^{d(k+1)})$ *que envia el flux* $(x_1, ..., x_d) \in S(\mathbb{R}^d)$ *a*

$$\left\{ \begin{pmatrix} x_1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, ..., \begin{pmatrix} x_{k+1} \\ x_k \\ x_{k-1} \\ \vdots \\ x_2 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_{k+2} \\ x_{k+1} \\ x_k \\ \vdots \\ x_3 \\ x_2 \end{pmatrix}, ..., \begin{pmatrix} x_d \\ x_{d-1} \\ x_{d-2} \\ \vdots \\ x_{d-k+1} \\ x_{d-k} \end{pmatrix}, \begin{pmatrix} 0 \\ x_d \\ x_{d-1} \\ \vdots \\ x_{d-k+2} \\ x_{d-k+1} \end{pmatrix}, ..., \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ x_d \\ x_{d-1} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ x_d \end{pmatrix} \right\} \tag{3.6}$$

a $S(\mathbb{R}^{d(k+1)})$.

Exemple 3.11. Considerem el flux $\Omega = \{1, 4, 2, 5, 3\}$, l'augmentació *Lead-Lag* envia a aquest flux al flux

$$\{(1,1), (4,1), (4,4), (2,4), (2,2), (5,2), (5,5), (3,5), (3,3)\} \in S(\mathbb{R}^2)$$

Si el retard és de dos passos, obtenim el flux

$$\{(1,1), (4,1), (2,1), (2,4), (5,4), (3,4), (3,2), (0,2), (0,2), (0,5), (0,5), (0,5), (0,3)\} \in S(\mathbb{R}^2)$$

Podem doncs, interpolar aquests fluxos en camins a $\mathbb{R}^2$.

També es poden ajuntar les dues augmentacions en una mateixa augmentació, enviant el flux $\Omega \subset S(\mathbb{R}^2)$ a flux de $S(\mathbb{R}^3)$

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 \\ 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \right\} \in S(\mathbb{R}^3)$$

Si apliquem la augmentació *No Future Past* amb $k = 2$, obtenim el flux

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix} \right\} \in S(\mathbb{R}^3)$$

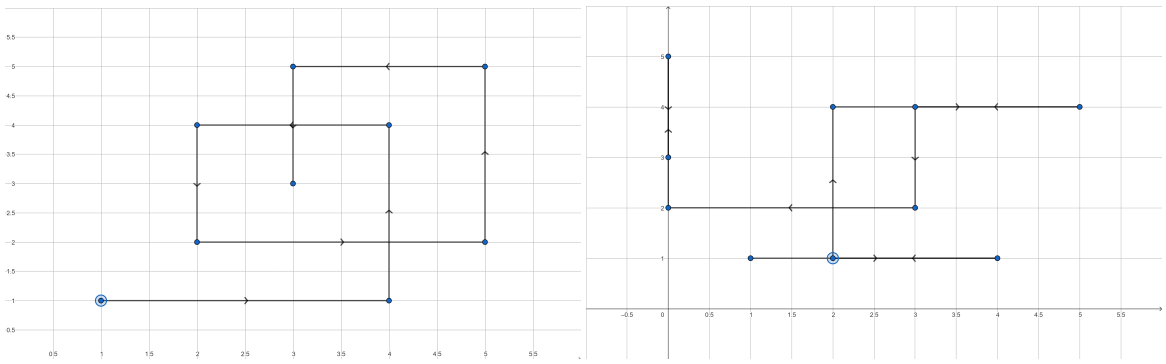


Figura 8: Transformacions Lead-Lag amb 1 i 2 passos de retard resp.

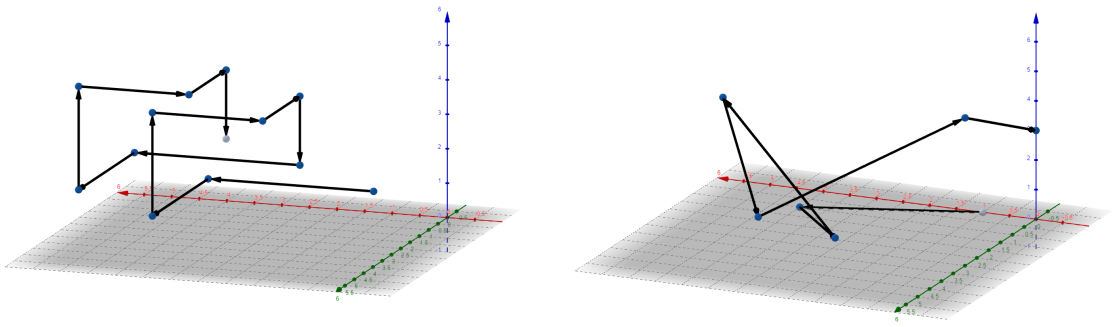


Figura 9:

Definició 3.12. *Definim l'augmentació Time com l'aplicació* $T : S(\mathbb{R}^d) \to S(\mathbb{R}^d \times \mathbb{R})$ *definida com*

$$T(x_1, ..., x_d) =:= ((x_1, t_0), ..., (x_j, t_j), ..., (x_d, t_d)) \tag{3.7}$$

per a una seqüència estrictament creixent de temps $0 \leq t_1 \leq t_2 \leq ... \leq t_k$.

Observació 3.13. La seqüència estrictament creixent de temps $0 \leq t_1 \leq t_2 \leq ... \leq t_k$ assegura que el camí resultant té una component estrictament monotona.

Una altra variant d'aquesta augmentació és l'augmentació *Time-difference*:

Definició 3.14. *L'augmentació Time-difference és l'aplicació* $T : S(\mathbb{R}^d) \to S(\mathbb{R}^d \times \mathbb{R})$ *definida com*

$$T(x_1, ..., x_d) =:= ((x_1, t_0), ..., (x_j, t_j - t_{j-1}), ..., (x_d, t_d - t_{d-1})) \tag{3.8}$$

per a una seqüència estrictament creixent de temps $0 \leq t_1 \leq t_2 \leq ... \leq t_k$.

Observació 3.15. Observem que tant l'agumentació *Time* com l'augmentació *Time-difference* envien fluxos de longitud k a fluxos de longituds k.

Exemple 3.16. Considerem el flux $\Omega = \{1, 5, 2, 9, 7, 6\} \in S(\mathbb{R})$ i la seqüència estríctament creixent 0, 1, 3, 6, 8, 12. Aleshores

$$T_{time}(\Omega) = \{(1, 0), (5, 1), (2, 3), (9, 6), (7, 8), (6, 12)\} \in S(\mathbb{R}^d) \tag{3.9}$$

i

$$T_{time-diff}(\Omega) = \{(1, 0), (5, 1), (2, 2), (9, 3), (7, 2), (6, 4)\} \in S(\mathbb{R}^d) \tag{3.10}$$

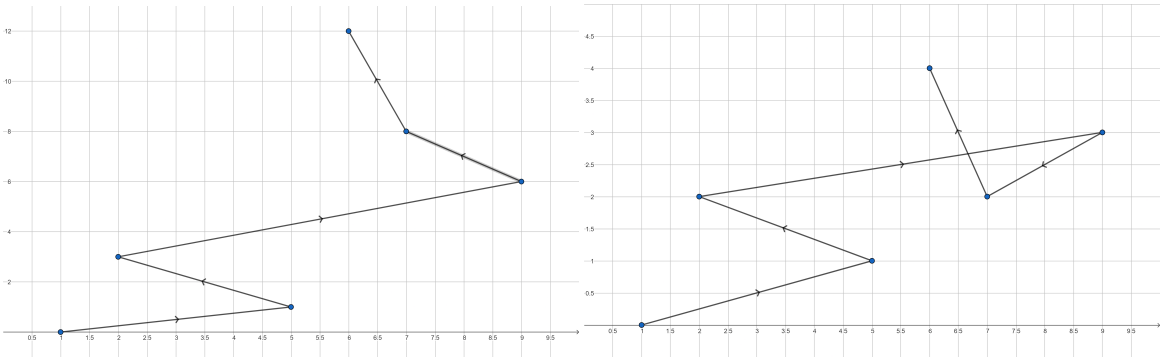


Figura 10: Interpolacions time i time-difference

La última augmentació que definirem és la següent:

Definició 3.17. *L'augmentació Invisibility Reset és l'aplicació definida com* $T : S(\mathbb{R}^d) \to S(\mathbb{R}^d \times \mathbb{R})$ *definida com*

$$T(x_1, ..., x_d) := ((x_1, 1), ..., (x_j, 1), ..., (x_d, 1), (x_d, 0), (0, 0)) \tag{3.11}$$

Observació 3.18. L'augmentació *Invisibility Reset* pren un flux de longitud d i l'envia a un de longitud $d+2$. El nou flux resultant conté la informació de la norma dels elements del flux original $x_1, ..., x_d$ de forma invariant per translació.

Exemple 3.19. Donats el fluxos $\Omega_1 = \{1, 3, 4, 8, 9\} \in S(\mathbb{R})$ i $\Omega_2 = \{(1, 2), (3, 4), (4, 6), (5, 9), (7, 10)\} \in S(\mathbb{R}^2)$. Aplicant l'augmentació *Invisibility Reset* obtenim:

$$T(\Omega_1) = \{(1, 1), (3, 1), (4, 1), (8, 1), (9, 1), (9, 0), (0, 0)\} \in S(\mathbb{R}^2) \tag{3.12}$$

$$T(\Omega_2) = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 9 \\ 1 \end{pmatrix}, \begin{pmatrix} 7 \\ 10 \\ 1 \end{pmatrix}, \begin{pmatrix} 7 \\ 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\} \in S(\mathbb{R}^3) \tag{3.13}$$

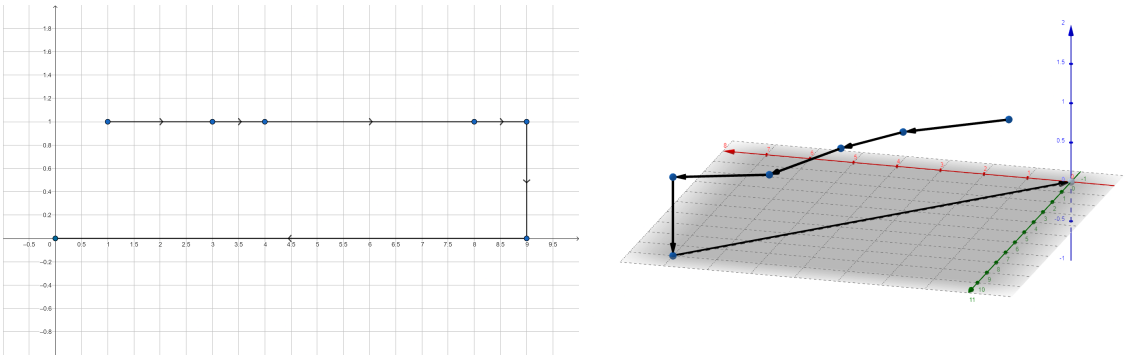


Figura 11: Interpolacions lineal a trossos i rectilinia de $\{(t, CS(X)\}$

3.5 Exemple explícit de càlcul

A continuació calculem algunes signatures basant-nos en la primera definició de signatura.
Considerem el camí de la figura 5, podem veure el seu increment i la seva àrea signada

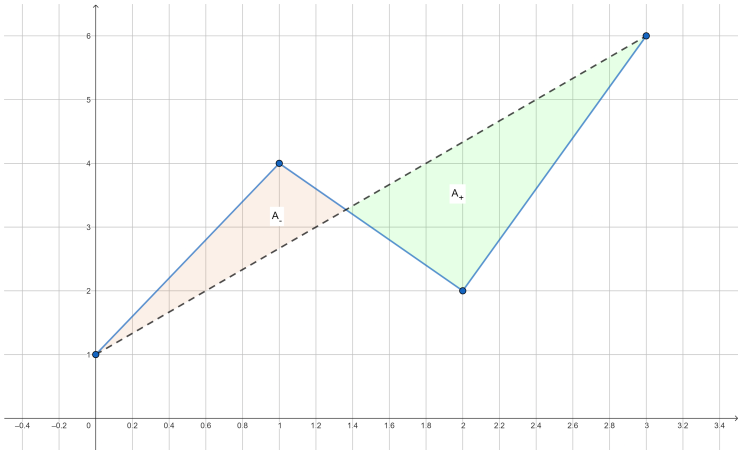


Figura 12: Camí amb l'àreea signada

a la figura ?. La seva signatura truncada de nivell 2 és:

$$S(X) = (1, 3, 5, 4.5, 8.5, 6.5, 12.5) = (1, S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)})$$

i

$$logS(X) = (3, 5, 1) = (S^{(1)}, S^{(2)}, S^{(1,2)})$$

l'últim terme de $logS(X)$ es calcula de $\frac{1}{2}(S^{(1,2)} - S^{(2,1)})$ i és l'àrea total entre els extrems. Observem que l'àrea ? és més gran que l'àrea ? doncs l'àrea total es positiva. A la figura ? podem entendre el significat geomètric dels termes de segon ordre $S^{(1,2)}$ i $S^{(2,1)}$.

Si canviem l'ordre d'integració sobre el camí en els termes $S^{(1,2)}$ i $S^{(2,1)}$ obtenim dues àrees que es completen l'una a l'altra i sumen l'àrea total d'un rectangle amb longituds laterals X1 i X2. Aquest senzill significat geomètric és la relació de producte mescla:

$$S^{(1)} \cdot S^{(2)} = S^{(1,2)} + S^{(2,1)}$$

$$3 \cdot 5 = 8.5 + 6.5.$$

La interpretació geomètrica dels termes d'ordre superior és menys intuitiva i els ometem.

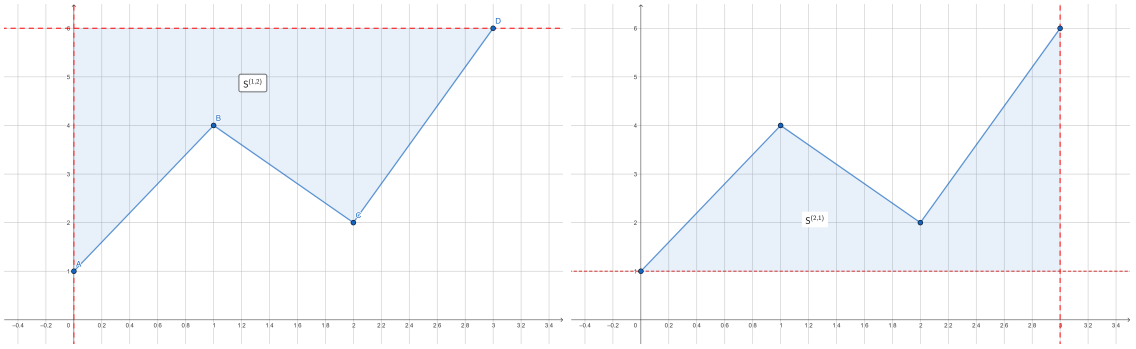


Figura 13: àrees $S^{(1,2)} = 8.5$ i $S^{(2,1)} = 6.5$

4 Signatura de fluxos de dades de caire financer

En el mon de les finances sovint es presenten dades en forma de sèries temporals, es a dir presenten trajectories indexades per un nombre finit de punts en el temps: $(\hat{X}_{t_i})_{i=0}^N$ a $\mathbb{R}^d$. Una forma de definir la signatura d'un flux dades d'aquest tipus es amb les integrals iterades del camí obtingut per interpolació lineal a trossos, com hem vist a l'apartat 3.1.. És a dir, introduïm el camí continu

$$X_u = \hat{X}_{t_i} + \frac{u - t_i}{t_{i+1} - t_i}(\hat{X}_{t_{i+1}} - \hat{X}_{t_i}),\ u \in [t_i, t_{i+1}],$$

i definim la signatura del flux $(\hat{X}_{t_i})_{i=0}^N$ com

$$S(\hat{X})_{t_0,t_N} := S(X)_{0,2N}.$$

4.1 La transformació *time-joined* i la signatura d'una sèrie temporal

Una altra manera de transformar un flux de dades $(\hat{X}_{t_i})_{i=0}^N = (t_i, \hat{X}_{t_i})_{i=0}^N$ en un camí continu és utilitzaar la transformació *time-joined* que es defineix com $R : [2, 2N] \to \mathbb{R}^+ \times \mathbb{R}talque$:

$$X_t^{TJ} = R(t) = \begin{cases} (t_0, \hat{X}_{t_0} t) & \text{per a } t \in [0,1], \\ (t_i + (t_{i+1} - t_i,)(t - 2i - 1), \hat{X}_{t_i}) & \text{per a } t \in [2i+1, 2i+2], \\ (t_{i+1}, \hat{X}_{t_i} + (\hat{X}_{t_{i+1}} - \hat{X}_{t_i})(t - 2i - 2)) & \text{per a } t \in [2i+2, 2i+3]. \end{cases}$$

per a $i = m, m+1, ..., n-1$. A l'annex B trobem un programa que computa aquesta transformació per a un flux de dades bidimensional donat.

Observació 4.1. La funció contínua R el que fa simplement és mantenint el valor $\hat{X}_{t_i}$ durant l'interval de temps $[t_i, t_{i+1})$. Quan les noves dades $\hat{X}_{t_{i+1}}$ arriben al temps t_{i+1}, hi ha un salt instantani de $\hat{X}_{t_i}$ a $\hat{X}_{t_{i+1}}$. Afegim un punt més, 0, al temps t_0, a la sèrie temporal $\{\hat{X}_i\}_{i=0}^N$ per tal de convertir-la en una nova sèrie temporal, de manera que la signatura de $\{\hat{X}_i\}_{i=0}^N$ pugui determinar unívocament $\{\hat{X}_i\}_{i=0}^N$.

La signatura de $\{(t_i, \hat{X}_i)\}_{i=0}^N$ es defineix com la signatura de la transformació *time-*textitjoined de $\{(t_i, \hat{X}_i)\}_{i=0}^N$, $N \in \mathbb{N}$. Es pot demostrar que la signatura d'una sèrie temporal la determina completament [8]

Definició 4.2. *Sigui $\{(t_i, \hat{X}_i)\}_{i=0}^N$ una sèrie temporal incrustada en el camí time-jonied R. La signatura de la sèrie temporal $\{(t_i, \hat{X}_{t_i})\}_{i=0}^N$ es defineix com la signatura del camí $\{R(s)\}_{s\in[0,2N]}$, i la denotem per $S(\{(t_i, \hat{X}_i)\}_{i=0}^N)$, on $1 \le m < n \le N$ i $m, n \in \mathbb{N}$.*

4.2 La signatura i els moments estadístics

El fet de combinar diferents fluxos de dades $X \equiv \{X_i\}$ en un sol ens permet calcular moments estadístics i correlacions entre els diferents fluxos.
A continuació veiem la relació entre els moments estadístics d'un conjunt de dades X en termes de la seva signatura.

4.3 Variació quadràtica

Tot i que els termes de la signatura mesuren diferents quantitats associades al camí, la variació quadràtica del procés no es captura directament. Aquesta quantitat és molt rellevant en aplicaions al camp de les finances, és per tant d'interés incorporarla en la signatura. Això últim es pot fer mitjançant la augmetnació *Lead-Lag* presentada a la secció 3.

Donat un flux $(\hat{X}_{t_i})_{i=0}^N \subset S(\mathbb{R}^d)$ él seu flux *Lead-transformat* $(\hat{X}_j^{lead})_{j=0}^{2N}$ és defineix com

$$\hat{X}_j^{lead} = \begin{cases} \hat{X}_{t_i}, \text{ si } j = 2i \\ \hat{X}_{t_i}, \text{ si } j = 2i-1 \end{cases}$$

Definim també el seu flux *Lag-transformat*, $(\hat{X}_j^{lag})_{j=0}^{2N}$, com

$$\hat{X}_j^{lag} = \begin{cases} \hat{X}_{t_i}, \text{ si } j = 2i \\ \hat{X}_{t_i}, \text{ si } j = 2i+1 \end{cases}$$

Definim doncs el flux *Lead-Lag-transformat*, que pren valors a $\mathbb{R}^{2d}$ com

$$(\hat{X}_j^{lead-lag})_{j=0}^{2N} = (\hat{X}_j^{lead}, \hat{X}_j^{lag})_{j=0}^{2N}$$

Observem que X^{lag} correspon a una reparametrització del temps de camí X:

$$S(\hat{X})_{t_0,t_N} = S(\hat{X}^{lag})_{0,2N}$$

Similarment

$$S(\hat{X})_{t_0,t_N} = S(\hat{X}^{lead})_{0,2N}$$

A més l'àrea signada entre la i-èssima component del camí *lead*-transformat i la component j-èssima del camí *lag*-transformat equival a la covariació de les trajectories $\hat{X}^i$ i $\hat{X}^j$. A la pràcticua s'utilitza una transoformació *Lead-Lag* parcial, prenent la tansformació *lead* del flux d'entrada i l'ajuntem amb la transformació *lag* d'aquells components pels quals la variació quadràtica és rellevant.

Una altra forma de definir més explicitament la transformació *Lead-Lag* d'un flux de dades $(\hat{X})_{i=0}^N = \{(t_i, \hat{X}_{t_i})_{i=0}^N\}$ és la següent:

$$X_t^{Lead-Lag} = \begin{cases} (X_{t_i}, X_{t_{i+1}}) & \text{per a } t \in [2i, 2i+1], \\ (X_{t_i}, X_{t_{i+1}} + 2(t-(2i+1))(X_{t_{i+2}} - X_{t_{i+1}}) & \text{per a } t \in [2i+1, 2i+\frac{3}{2}], \\ (X_{t_{i+2}} + 2(t-(2i+\frac{3}{2}))(X_{t_{i+1}} - X_{t_i}), X_{t_{i+2}}) & \text{per a } t \in [2i+\frac{3}{2}, 2i+2]. \end{cases}$$

per a $t \in [0, 2N]$. A l'annex A trobem programa que computa aquesta transformació per a un flux de dades bidimensional donat.

Exemple 4.3. Sigui

$$[X]_t = \sum_{i=0}^{N-1} (X_{t_{i+1}} - X_{t_i})^2$$

la variació quadràtica del camí construit amb el flux $(\hat{X}_{t_i})_{i=0}^N \subset S(\mathbb{R}^2)$. Aleshores, podem esciure

$$A_{lead-lag} = \frac{1}{2}[X]_t$$

4.4 La suma cumulativa d'una succesió

Veiem algunes propietats de camins obtinguts incrustant punts utilitzants sumes cumulatives.
La suma considerem una seqüencia unidimensional $X = \{\hat{X}_i\}_{i=0}^{N}$. La seva suma cumulativa és

$$\tilde{X} = S_1, ..., S_N$$

On $S_k = \sum_{i=0}^{k} \hat{X}_i$ amb $k \in \{1, ..., N\}$.

Ajuntant-la amb una component temporat $\{t_i\}_{i=0}^{N}$ obtenim un camí bidimensional X_t amb interpolació rectilínia.
Si augmentem la seqüència original $\{\hat{X}_i\}$ afegint el valor 0 al principi i trunquem la signtura de la nova seqüència $\{\tilde{X}\}$ al nivell L, podrem determinar els moments estadístics d'odre $k \leq L$ de la seqüència original.
És a dir passem de $\{\hat{X}_i\}_{i=1}^{N}$ a $\{0.\{\hat{X}_i\}_{i=0}^{N}\}$ a $CS(\{\hat{X}_i\}_{i=0}^{N}) = \{\tilde{X}\}$, li podem aplicar la augmentació *Lead-Lag* i calcular fins al segon nivell de la signatura, obtenint quantitats proporcionals als moments de primer i segon ordre, és a dir, la mitjana i la variància.

$$\Delta\tilde{X} = \sum_{i=0}^{N} \hat{X}_i$$

$$QV(\tilde{X}) = \sum_{i=0}^{N-1} (\tilde{X}_{i+1} - \tilde{X}_i)^2 = \sum_{i=0}^{N} (\hat{X}_i)^2$$

I podem observar fàcilment que la mitjana i la variància (per a dades uniformement distribuïdes) respectivament són:

$$\bar{X} = E[X] = \frac{1}{N} \sum_{i=0}^{N} \hat{X}_i = \frac{\Delta\tilde{X}}{N}$$

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 = \frac{1}{N} \left(QV(\tilde{X}) - \frac{1}{N} (\Delta\tilde{X})^2 \right)$$

Es pot demostrar també que els nivells d'ordre superior de la signatura determinen els moments d'ordre superior.

5 La signatura en l'aprenentatge automàtic

La idea principal d'utilitzar la transformació de la signatura per a problemes d'aprenentatge automàtic es dona per la seva capacitat d'extraure característiques distintives de les dades. Incrustar les dades en un camí i calcular la seva signatura ens proporciona informació important sobre les dades originals.
Les característiques extretes es poden utilitzar per a diversos tipus d'aplicacions a l'aprenentatge automàtic (o *machine learning*), incloent tant l'aprenentatge supervisat com l'aprenentatge no supervisat. Per exemple, es poden classificar sèries temporals o distingir clústers de dades. Un dels avantatges de l'extracció de característiques amb el mètode de la signatura és que aquesta és sensible a la forma geomètrica d'un camí, això últim,

per exemple, ha portat a una aplicació exitosa del mètode de la signatura en el problema de reconeixement de caràcters xinesos [9]. Una de les aplicacions més conegudes i naturals del mètode de la signatura és en finances quantitatives, concretament en l'anàlisi de dades de sèries temporals. Les dades de sèries temporals representen dades seqüencials ordenades, que són un candidat ideal per crear un camí a partir de les dades, seguit de calcular la signatura i aplicar algoritmes d'aprenentatge automàtic per a un anàlisi posterior. Qualsevol tipus de dades seqüencials ordenades en el temps encaixa naturalment en el marc de treball de la signatura. A més, si les dades d'entrada provenen de diverses fonts paral·leles, això donarà com a resultat un camí multidimensional. Un exemple d'aquest tipus de dades són les dades de panell (en econometria) o les dades longitudinals (en medicina, psicologia, bioestadística, etc.) que involucren observacions repetides de la mateixa quantitat durant períodes de temps.

5.1 Aprenentatge supervisat

Una classe molt àmplia de problemes en l'aprenentatge automàtic és l'aprenentatge supervisat. L'objectiu de l'aprenentatge supervisat és aprendre una funció que assigni una entrada (*input)* a una sortida (*output*). Això es fa utilitzant un conjunt de dades d'entrenament amb parells coneguts d'entrada-sortida (o *input-output*). És crucial fer una bona selecció d'entrades, conegudes com a *característiques*, per construir un model a partir d'aquestes característiques. Aquestes característiques haurien de descriure l'objecte que s'està estudiant de manera precisa, però al mateix temps, característiques amb una dimensió molt alta provoquen problemes computacionals i de sobreajust.

Suposem, per exemple, que volguessim construir un classificador que predigués si una persona és d'una a d'una altra ètnicitat determinada. La nacionalitat, pot ser una bona característica que podríem utilitzar, ja que els diferents països solen tenir alguna etnicitat majoritaria, però el sexe obviament no ho és, ja que per a una determinada etnicitat, hi ha aproximadament, el mateix nombre d'homes que de dones. Per tant, el sexe no és una característica que descrigui de manera precisa la etnicitat d'un grup d'individus.

5.2 Les signatures com a característiques

La pregunta que ens plantegem aquí és: Poden les signatures ser característiques efectives que es puguin utilitzar per a crear un model d'aprenentatge automàtic?
La resposta es troba en la secció 2 d'aquest treball.
Sabem que donat un camí de variació acotada podem construir la seqüència d'integrals iterades que defineixen la signatura. A la subsecció 2.4.7 hem vist que, tot i que donada una d'aquestes seqüències d'integrals iterades no existeix un únic camí amb variació acotada la signatura del qual sigui aquesta seqüència, si hi ha una certa noció d'unicitat sota certes condicions de regularitat del camí, concretament, per aquells camins que no es creuen a si mateixos. Per això, la signatura és una bona descripció del camí.

No obstant, donada la impossibilitat de computar la signatura (ja que és una sèrie infinita), cal assegurar que la signatura truncada proporciona també una bona descripició del camí, per la propietat de declivi factorial esmentada a la subsecció 2.4.6 deduïm que això és així, ja que la importància dels termes de la signatura decau factorialment respecte al seu ordre.

5.3 Model d'aprenentatge automàtic basat en signatures

Podem formular els efectes dels fluxos de dades com a variable dependent d'un problema de regressió, amb una variable explicativa que és un camí a $\mathcal{V}^p(J, \mathbb{R}^d)$, $J \subset \mathbb{R}$ compacte. Suposem que tenim un conjunt d'entrenament de parells d'entrada-sortida $\{(X_i, Y_i)\}_{i=0}^N$ on X_i és un camí de variació acotada i amb $Y_i \in \mathbb{R}$. Volem aprendre la funció que assigna cada entrada a la sortida corresponent. Per tant, assumirem que les entrades i les sortides estan relacionades mitjançant una funció desconeguda f de la forma

$$Y_i = f(X_i) + \varepsilon_i \tag{5.1}$$

On $X_i \in \mathcal{V}^p(J, \mathbb{R}^d)$, $Y_i \in \mathcal{V}^p([0,T], \mathbb{R}^e)$, $\mathbb{E}[\varepsilon_i \mid X_i] = 0$.

El següent teorema assegura que n'hi ha prou amb amb buscar funcions lineals.

Teorema 5.1. *Sigui $\mathcal{V}^1([0,T], \mathbb{R}^d)$ l'espai de camins d-dimensionals de variació acotada definits sobre l'interval $[0,T]$. Sigui $S(\mathcal{V}^1([0,T], \mathbb{R}^d)) := \{S(X) : X \in \mathcal{V}^1([0,T], \mathbb{R}^d)\}$ i $S_1 \subset S(\mathcal{V}^1([0,T], \mathbb{R}^d))$ un conjunt compacte. Aleshores, donat $\varepsilon > 0$ i una funció continua $g : S_1 \to \mathbb{R}$, existeix una funció lineal continua L tal que*

$$\|g(x) - L(x)\| < \varepsilon \ \forall x \in S_1$$

La demostració d'aquest teorema es pot trobar a [10].
Per tant, la funció f pot ser aproximada per funcions lineals. Com que la signatura truncada es pot interpretar com a un vector, podriem aleshores aplicar-hi regressió lineal per a resoldre aquest problema.
Usant tècniques clàssiques de regressió no paramètrica per al cas de dimensionalitat finita, cal identificar conjunts de característiques específiques de les dades d'entrada (sortida) observades per a linearitzar la relació funcional entre elles. Com hem vist a la secció 2.4.7 la signatura és en certa forma única i la usarem com a característica, per tant buscarem f tal que

$$S(Y_i) = f(S(X_i) + \varepsilon_i$$

Més en general, definim el següent model:

Definició 5.2 (Model de la Signatura Esperada)**.** *Siguin X i Y dos processos estocàstics que prenen valors a $\mathbb{R}^d$ i $\mathbb{R}^e$, respectivament. Suposem que les signatures de X i Y, denotades per $S(X)$ i $S(Y)$, estan ben definides q.s. Assumim que*

$$S(Y) = f(S(X)) + \varepsilon,$$

on $\mathbb{E}[\varepsilon|X] = 0$ i f és una funció lineal que fa una correspondència de $T(\mathbb{R}^d)$ a $T(\mathbb{R}^e)$.

Sota el model de la signatura esperada, donat un nombre de mostres $\{X_i, Y_i\}_{i=1}^N$, l'estimació de la signatura truncada esperada de Y de l'ordre m condicionada per $S(X)$, és a dir $\rho_m(\mathbb{E}[Y|X])$, o en altres paraules, la funció lineal $\rho_m \circ f$, resulta ser el problema estàndard de regressió lineal [8]; les integrals iterades coordenades de $S(Y)$ són regressants multidimensionals mentre que les integrals iterades coordinades de $S(X)$ són variables explicatives. Com hem discutit anteriorment, a la pràctica, hem de considerar la signatura truncada de X d'un ordre determinat en lloc de la signatura completa, ja que el nombre de variables explicatives ha de ser finit. Com que la regressió és lineal, hi ha molts métodes

estàndard de regressió lineal que es poden emplear. Podem utilitzar, a més tècniques de regularització o selecció de variables com ara LASSO o SVD per evitar la colinealitat de la matriu de disseny i el problema de sobreajustament. Anomenem aquest mètode de calibració l'aproximació de signatura esperada.

Per mesurar la bondat de l'ajustament del model, utilitzem l'error quadrat mitjà dels residus $\{a_i\}_{i=1}^N$, on

$$a_i = S(Y_i) - \hat{f}(S(X_i)), \quad \forall i = 1, \ldots, N.$$

Alternativament, podem utilitzar R^2 o R^2 ajustat com a indicador de la bondat de l'ajust.

5.4 Implementació del model a un cas pràctic

En aquesta secció, apliquem el model anteriorment descrit per a predir, a quin país pertany una empresa utilitzant l'evolució del preu de tancament diari de les seves accions i el volum negociat per a un període de temps donat.

Òbviament, la manera com el preu d'una acció canvia amb el temps varia d'empresa a empresa. Si una empresa és rendible i els inversors creuen que seguirà sent rendible en el futur, és probable que el preu de les seves accions augmenti. Per una altra banda, si una empresa està a prop de la fallida, és bastant probable que el preu baixi.

No obstant això, de manera intuitiva té sentit esperar que els factors externs que afecten al país en conjunt també tinguin un impacte en el preu d'una acció. Per tant, és raonable pensar que els preus de les accions que es negocien en el mateix país tindran certes similituds intrínseques que van més enllà del rendiment de cada empresa en particular.

Com que només volem mostrar la signatura en aquest context, limitarem el problema a tres països: els Estats Units, el Regne Unit i el Japó.

El que farem doncs, carregar la llista de tuples $(P_i, V_i)_{i=0}^N$ per a un període de N dies, on P_i i V_i son el preu de tancament i volum negociat de les seves accions el i-èssim dia, junt amb la categoria (país) a la que pertany l'empresa.

Per a això amb l'ajuda del portal web finance.yahoo.com hem exportat fulls de càlcul amb les primeres 200 empreses de cada un dels tres països ordenades per capitalització de mercat de major a menor (per tal de seleccionar aquelles accions més representatives de cada país),hem eliminat tota la informació que no siguin els *tickers* i afegit el país per a ser llegit després pel programa, despres usant el paquet de Python $yfinance$ podem obtenir les dades històriques d'una acció a través del seu *ticker*. Assignem també les tres categories a punts linealment independents.

Posteriorment, calcularem la signatura del camí (signatura truncada d'un cert nivell) incrustat de les evolucions dels preus de tancament i volum negociats del període inserit, per a fer-ho usem el paquet de Python *iisignature* que computa la signatura. Això ho farem per a diferents ordres de la signatura truncada.

Un cop tenim això barrejem les dades inserides i seleccionem el primer 70% de les dades per a entrenar el model i l'altre 30% com a test. Per a entrenar el model usarem

k	Exactitud
1	0.7278
2	0.7389
3	0.745
4	0.75
5	0.75

		Classe Real		
		E.E.U.U	R.U.	Japó
Classe predita	E.E.U.U	60	10	7
	R.U.	9	39	11
	Japó	4	8	36

regressió lineal amb LASSO, per a això utilitzem el paquet de Python *scikit learn* que és un paquet que conté diverses funcions per a l'aprenentatge automàtic. Cal incidir en que no pretenem aquí comparar aquest mètode amb un altre, sino mostrar la potència que la signatura pot tenir com a característica en problemes d'aprenentatge automàtic.

Després d'una execució obtenim que la exactitud (tasa d'encerts) en funció de l'ordre la signatura truncada (k) que podem observar a la taula superior. La matriu de confusió quan $k = 4$ és la que observem a dalt. Les dates usades són des de la data 01/08/2022 fins al 01/01/2023.

Hem provat a aplicar aquest model per a predir el sector al que pertany una empresa d'un mateix país (els Estats Units) basant-nos també en els preus de tancament diaris de les seves accions i el volum negociat però el model no ha donat bons resultats, és probable que el problema no estigui ben condicionat o que les accions dels diferents sectors estiguin massa correlacionades entre si, o bé que el preu de tancament i volum negociat diares no siguin per si sols una característica prou rellevant.

A l'annex C es un programa en Python que aplica aquest model.

6 Conclusions

Tot i que els resultats del model implementats a l'últim apartat estan lluny de ser perfectes, ens son prou per a poder observar que la signatura és de fet una característica acceptable capaç de predir la majoria de classes correctament. Dit això, els resultats són pitjors de l'esperat.

No obstant, això no significa que la signatura sigui una mala característica, revisant la literatura al respecte, part d'ella en la bibliografia citada a aquest treball, hi han numerosos exemples en que s'ha aconsseguit amb prou éxit utilitzar la signatura com a característica en problemes d'aprenentatge automàtic aplicats a dades de tipus financers, essent capaços amb aquest mètode de classificar correctament més del 90% de classes. És clar doncs, més enllà d'aquest experiment anecdòtic que la signatura és una eina molt potent per a classificar fluxos de dades en problemes d'aprenentatge automàtic.

Referències

[1] Ilya Chevyrev; Andre y Komilitzin: A Primer on the Signature Method, `arXiv:1603.03788`, març de 2016.

[2] Bourbaki Nicolas (1987): Topological Vector Spaces: Chapters 1−5. Éléments de mathématique. Translated by Eggleston, H.G.; Madan, S. Berlin New York: Springer-Verlag.

[3] Terry Lions; Andrew D. McLeod: Signature Methods in Machine Learning, `arXiv:2206.14674v1`, juny de 2022.

[4] Kuo-Tsai Chen: Iterated Integrals and Exponential Homomorphisms, *Prod. London Math. Soc.* 4 (1954).

[5] Kuo-Tsai Chen: Integrations of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Ann. of Math. (2)* 65:163-178, 1957

[6] Paul Wilmott, *Machine Learning: An Applied Mathematics Introduction*, Panda Ohana Publishing (2019).

[7] Ben Hambly; Terry Lyons: Uniqueness for the signature of a path of bounded variation and the reduced path group, *Ann. of Math., Volume 171*, pp.109-167, 2010.

[8] Jonathann Field; Lajos Gergely Gyurkó; Mark Kontkowski; Terry Lyons: Extracting information from the signature of a financial data stream. arXiv:1307.7244, July 2014. Preprint.

[9] Benjamin Graham. Sparse arrays of signatures for online character recognition. arXiv:1308.0371, December 2013. Preprint.

[10] Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. arXiv:1309.0260, September 2015. Preprint.

[11] Protter, Philip E. (2004), Stochastic Integration and Differential Equations (2nd ed.), Springer

[12] [2] Király, F. J. and Oberhauser, H. (2016) "Kernels for sequentially ordered data, arXiv: 1601.08169"

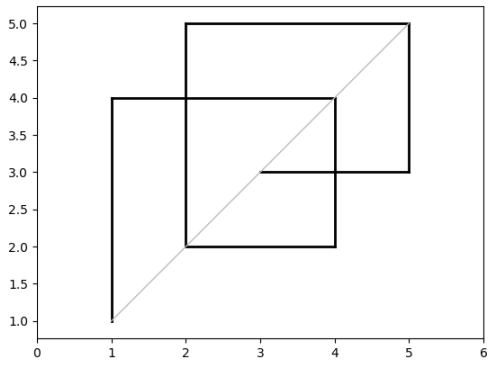
[13] Terry Lyons, Michael Caruana, and Thierry Lévy, *Differential Equations Driven by Rough Paths*, Ecole d'Eté de Porbabilités de Saint-Flour XXXIV - 2004, Lecture Notes in Mathematics, Springer, 2007.

[14] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: Uniqueness. arXiv:1406.7871, August 2014. Preprint.

A Codi que computa la transformació Lead-Lag

```
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import matplotlib.patches as patches

X=[[1,1], [4,1], [4,4], [2,4], [2,2], [5,2], [5,5], [3,5], [3,3]]
def leadlag(X):
    '''
    Torna la transformacio lead-lag de X

    Arguments:
        X: llista

    '''

    l=[]

    for j in range(2*(len(X))-1):
        i1=j//2
        i2=j//2
        if j%2!=0:
            i1+=1
        l.append((X[i1][1], X[i2][1]))

    return l

def plotLeadLag(X, diagonal=True):
    '''
    Grafica el cami lead-lag

    Arguments:
        X: llista de tuples (X^lead, X^lag)
        diagonal es una variable booleana inicialitzada com a certa.
        Si es certa, es grafica una linia que junta els punts d'inici i comenÃ§ament-
    '''
    for i in range(len(X)-1):
        plt.plot([X[i][1], X[i+1][1]], [X[i][0], X[i+1][0]],
                    color='k', linestyle='-', linewidth=2)


    # Grafica la diagonal
    if diagonal:
        plt.plot([min(min([p[0] for p in X]), min([p[1]
                for p in X])), max(max([p[0] for p in X]),
                max([p[1] for p in X]))], [min(min([p[0]
                for p in X]), min([p[1] for p in X])),
                max(max([p[0] for p in X]), max([p[1] for
                p in X]))], color='#BDBDBD', linestyle='-',
```

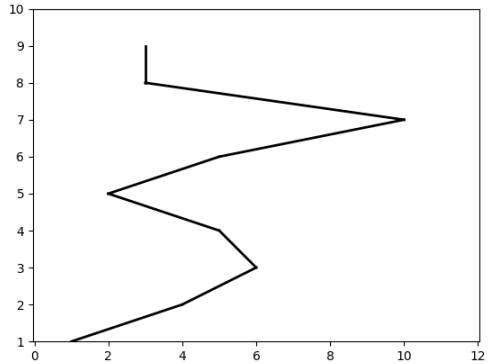
```
                linewidth=1)

    axes=plt.gca()
    axes.set_xlim([min([p[1] for p in X])-1, max([p[1] for
                  p in X])+1])
    axes.set_ylim([min([p[0] for p in X])-1, max([p[0] for
                  p in X])+1])
    axes.get_yaxis().get_major_formatter().set_useOffset(False)
    axes.get_xaxis().get_major_formatter().set_useOffset(False)
    axes.set_aspect('equal', 'datalim')
    plt.show()

print(leadlag(X))
plotLeadLag(X,)
```



B Codi que computa la transformació time-joined

```
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import matplotlib.patches as patches

X=[[1,1], [4,2], [6,3], [5,4], [2,5], [5,6], [10,7], [3,8], [3,9]]
def timejoined(X):
    '''
    Torna la transformacio time-joined de X (punts)

    '''
    X.append(X[-1])
    l=[]

    for j in range(2*(len(X))+1+2):
        if j==0:
            l.append((X[j][0], 0))
```

```
            continue
        for i in range(len(X)-1):
            if j==2*i+1:
                l.append((X[i][0], X[i][1]))
                break
            if j==2*i+2:
                l.append((X[i+1][0], X[i][1]))
                break
    return l

def plottimejoined(X):
    '''
    Grafica la transformacio time-joined de X

    Arguments:
        X: llista de tuples (t,X)
    '''

    for i in range(len(X)-1):
        plt.plot([X[i][0], X[i+1][0]], [X[i][1], X[i+1][1]],
                color='k', linestyle='-', linewidth=2)

    axes=plt.gca()
    axes.set_xlim([min([p[0] for p in X]), max([p[0] for p
                  in X])+1])
    axes.set_ylim([min([p[1] for p in X]), max([p[1] for p
                  in X])+1])
    axes.get_yaxis().get_major_formatter().set_useOffset(False)
    axes.get_xaxis().get_major_formatter().set_useOffset(False)
    axes.set_aspect('equal', 'datalim')
    plt.show()

print(timejoined(X))
plottimejoined(X,)
```



C Codi que implementa el model de la secció 5

```
import pandas as pd
import yfinance as yf
import numpy as np
from sklearn import linear_model
import numbers
import random
from iisignature import sig
from datetime import datetime

class SigLearn:

    def __init__(self, ordre, alpha=0.1):
        if not isinstance(ordre, numbers.Integral) or ordre<1:
            raise NameError('Lordre ha de ser un enter positiu')
        if not isinstance(alpha, numbers.Real) or alpha<=0.0:
            raise NameError('Alpha ha de ser un valor real positiu')

        self.ordre=int(ordre)
        self.reg=None
        self.alpha=alpha

    def train(self, x, y):
        '''
        Entrena el model utilitzant signatures.

        x es una llista de inputs, on cada element de la llista es una llista de tuples.
        y es la llista of outputs.
        '''

        # Comprovem que x i y son del tipus apropiat
        if x is None or y is None:
            return
        if not (type(x) is list or type(x) is tuple) or not (type(y) is list or type(y) i
            raise NameError('Les entrades i les sortides han de ser tuples')
        if len(x)!=len(y):
            raise NameError('El nombre de entrades i de sortides ha de ocincidir.')
        ###

        X=[list(sig(np.array(stream), self.ordre,0)) for stream in x]
        #apliquem regressio lineal amb Lasso a la llista de signatures
        self.reg = linear_model.Lasso(alpha = self.alpha,max_iter=1000000000, tol=0.5)
        self.reg.fit(X, y)

    def predict(self, x):
        '''
        Prediu els outputs dels inputs x usant el model pre-entrenat.
```

```
        Retorna llista d'outputs predits.
        '''
        if self.reg is None:
            raise NameError('El model no esta entrenat')

        X=[list(sig(np.array(stream), self.ordre,0)) for stream in x]

        return self.reg.predict(X)

# Dates de comenÃ§ament i final.
start="2022-08-01"
end="2023-01-01"

# Llegim un full de calcul amb els tickers de cada accio i el pais al que pertany.
df = pd.read_excel("Paises_TickersJP.xlsx",  header=None)

tickers = df.values
class Stock:
    '''
    Classe amb la informacio d'una accio
    '''

    def __init__(self, dades, sector):
        # Caracteristica de la classe que guarda el flux de dades
        self.dades=np.array(dades,dtype='float32')

        # Caracteristica de la classe que guarda el sector de l'accio.
        self.sector=sector

         #El output per a entrenar ha de ser un vector
         #assignem a cada sector un punt
         #amb la segÃ¼ent funcio

        self.punt=sector_a_punt(sector)



def exact(predictions, y):
    #funcio que calcula l'exactitud de les prediccions
    p=0

    for i in range(len(y)):
        if round(prediccions[i][0])==y[i][0] and round(prediccions[i][1])==y[i][1]:
            #print(prediccions[i][0],prediccions[i][1])
            p+=1


    return p/float(len(y))
def sector_a_punt(sector):
```

```
    '''
    Transforma un sector en un punt
    '''
    dictionary={"Japon": (1,0), "EEUU": (-1, 0), "RU": (0,1)}
    return dictionary[sector]
def yDades(ticker, start, end):
    '''
    Obte les dades per a un ticker i periode de temps donat.
    '''
    dat=[]
    p=yf.Ticker(ticker)
    stock=p.history(str=start,nd=end)[['Close', 'Volume']]
    valors=stock[["Close","Volume"]].reset_index().values
    for i in range(len(valors)):
        valors[i][0]=valors[i][0].strftime("%Y%m%d%H%M%S")
        valors[i][0]=i

    return valors
dades=[]

for sector in tickers:
    print("Carregant "+sector+"...")
    EmpresaDades=yDades(sector[1], start, end)

    # Si l'empresa no te dades, la ignorem.
    if len(EmpresaDades)==0:  continue

    dades.append(Stock(EmpresaDades, sector[0]))
    print("Fet.")

# Barrejem el conjunt de dades i el dividim en dos conjunts
# el training_set, que te el 70% de les dades,
# i el testing_set, amb l'altre 30%.
random.shuffle(dades)

training_set=dades[0:int(0.7*len(dades))]
testing_set=[empresa for empresa in dades
            if empresa not in training_set]
# construim les entrades i sortides per a entrenar
inputs=[empresa.dada for empresa in training_set]
outputs=[empresa.punt for empresa in training_set]

# construim les entrades i sortides per a test.
inputsTEST=[empresa.dades for empresa in testing_set]
outputsTEST=[empresa.punt for empresa in testing_set]
#Apliquem el model per a signatures d'ordre de 1 a 5.
for ordre_signatura in range(1, 6):
    # Entrenem el model
    model=SigLearn(ordre=ordre_signatura)
```

```
model.train(inputs, outputs)

#Fem les prediccions
predictions=model.predict(inputsTEST)

# imprimim l'exactitud,

print(exact(prediccions, outputsTEST))
#imprimim les prediccions i els outputs de test
j=0
e=0
r=0
j1=0
e1=0
r1=0
for i in range(len(prediccions)):
    if round(prediccions[i][0])==1:# and round(prediccions[i][1])==0:
         j+=1
    if round(prediccions[i][0])==-1: #and round(prediccions[i][1])==0:
         e+=1
    if round(prediccions[i][0])==0:# and round(prediccions[i][1])==1:
         r+=1
    #if round(outputsTest[i][0])==1:# and round(prediccions[i][1])==0:
         #j+=1
    #if round(outputsTest[i][0])==-1: #and round(prediccions[i][1])==0:
         #e+=1
    #if round(outputsTest[i][0])==0:# and round(prediccions[i][1])==1:
         #r+=1
print("predits japo:")
print(j)
#print("reals japo:" j1)
print("predits EEUU:" "+e+")
print(e)
#print("reals EEUU:" e1)
print("predits RU:" "+r+")
print(r)
#print("reals RU:" r1)
```