



UNIVERSITAT DE
BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

INTRODUCCIÓ AL MÈTODE DE LA SIGNATURA I LES SEVES APLICACIONS

Autor: Francesc Parcerisas Vela

Director: Dr. José Manuel Corcuera Valverde

Realitzat a: Departament
de Matemàtiques i Informàtica

Barcelona, 10 de juny de 2024

Resum

En aquest treball farem una breu introducció sobre els conceptes bàsics de la signatura d'un camí utilitzada en un flux de dades. En repassarem les propietats fonamentals i explorarem les aplicacions que pot tenir, en particular, com a vector de característiques d'un model d'aprenentatge automàtic.

Índex

1	Introducció teòrica	2
1.1	Signatura d'un camí	4
1.2	Signatura a partir d'iteracions de Picard	9
1.3	Propietats importants de la signatura	12
1.3.1	Invariància per reparametritzacions	12
1.3.2	Identitat de Chen	13
1.3.3	Propietat d'inversió temporal	14
2	Signatura discreta fent servir interpolació lineal	16
2.1	Signatura discreta plana	17
2.2	Signatura discreta	20
3	Implementació del càlcul de la signatura discreta	23
3.1	Implementació naïve	23
3.2	Implementació millorada	24
4	Signatura com a vector de característiques	27
4.1	Resultats experimentals	29
5	Signatura d'un camí aplicada a distribucions de probabilitat	31
5.1	Obtenció d'un camí a partir de d'una mostra d'una distribució de probabilitat	33
5.2	Ajustament de paràmetres d'una distribució de probabilitat a través de la signatura	35
5.2.1	Exemple d'aplicació	36
5.3	Signatura d'una distribució de probabilitats com a vector de característiques	38
6	Conclusions	42

Introducció

La signatura, un concepte procedent de la teoria de camins, ha guanyat popularitat els últims anys per la seva capacitat que té de sintetitzar les propietats d'un flux de dades i les múltiples aplicacions que pot tenir, en especial en el món financer. És aquest ús en el món financer el que inicialment ens ha despertat l'interés cap a la signatura, en particular la seva capacitat de ser utilitzada com a vector de característiques representatiu d'un flux de dades. L'any 2016, Ilya Chevyrev i Andrei Kormilitzin presentaven l'article *A primer on the signature method in machine learning* [2], on introduïen justament aquest concepte: la signatura com a vector de característiques d'entrada a un model d'aprenentatge automàtic. En aquest treball repassarem la publicació i n'extreurem els conceptes fonamentals que servien com a base per aplicacions futures de la signatura.

L'any 2022, Takanori Adachi i Yusuke Naritome publicaven *Discrete signature and its application to finance*[1] on introduïen un cas particular de la signatura d'un flux de dades: la signatura discreta. En aquest article mostraven una aplicació de la signatura discreta amb dades financeres i en quins casos el seu mètode és computacionalment més ràpid que d'altres que també utilitzen la signatura. En aquest treball també repassarem la seva feina i realitzarem una implementació del càlcul de la signatura discreta que plantegen, a banda d'aplicar-la a casos experimentals.

Finalment, també veurem com podem aplicar la signatura a dades obtingudes a partir d'observacions de variables amb distribucions de probabilitat.

Estructura de la memòria

Estructurarem aquest treball en tres blocs: (i) repassarem els conceptes teòrics en els que està fonamentada la signatura d'un camí de la firma com els descriu Chevyrev [2], (ii) introduïrem el concepte de la signatura discreta d'Adachi [1] i un exemple d'implementació i, (iii) veurem com es pot aplicar la signatura com a vector de característiques d'un camí i a una mostra d'observacions de variables amb distribucions de probabilitat.

1 Introducció teòrica

Començarem aquesta primera secció repassant el concepte de camí i d'integral de línia per un camí. Continuarem definint altres conceptes que ens permetran, finalment, introduir el concepte de signatura. En línies generals, al llarg d'aquest capítol, anirem seguint el capítol 1 l'article de Chevyrev, I. [2] on s'introdueixen els conceptes bàsics de la teoria de la signatura.

Definició 1.1. Un *camí* X en $\mathbb{R}^d$ és una aplicació contínua, diferenciable a trossos, $X : [a, b] \longrightarrow \mathbb{R}^d$. Habitualment, posem $X_t = X(t)$ per denotar la dependència del paràmetre $t \in [a, b]$.

Direm que un camí X és un *camí suau* si és diferenciable a qualsevol ordre.

Notació 1. Donat un camí $X : [a, b] \longrightarrow \mathbb{R}^d$, escriurem $X_t = (X_t^1, \dots, X_t^d)$ per denotar les seves components.

Exemple 1.2. Dos exemples de camins suaus a $\mathbb{R}^2$ són $X, Y : [0, 2\pi] \longrightarrow \mathbb{R}^2$:

$$X_t = (X_t^1, X_t^2) = (2 \cos(t), 3 \sin(t)) \quad (1.1)$$

$$Y_t = (Y_t^1, Y_t^2) = (t, 3 \sin(t)) \quad (1.2)$$

A la Figura 1 podem veure la representació a $\mathbb{R}^2$ de cada camí.

Exemple 1.3. A la Figura 2 podem veure un exemple de camí no suau.

Definició 1.4. Sigui $X : [a, b] \longrightarrow \mathbb{R}$ un camí unidimensional i $f : \mathbb{R} \longrightarrow \mathbb{R}$ una funció fixada. La *integral de línia* de la funció f pel camí X es defineix com

$$\int_a^b f(X_t) dX_t = \int_a^b f(X_t) \dot{X}_t dt, \quad (1.3)$$

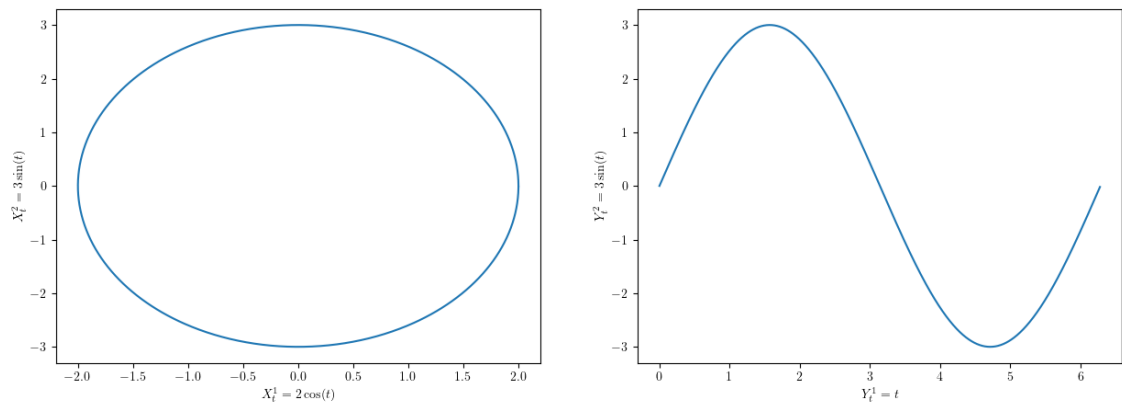


Figura 1: Exemples de camins suaus a $\mathbb{R}^2$

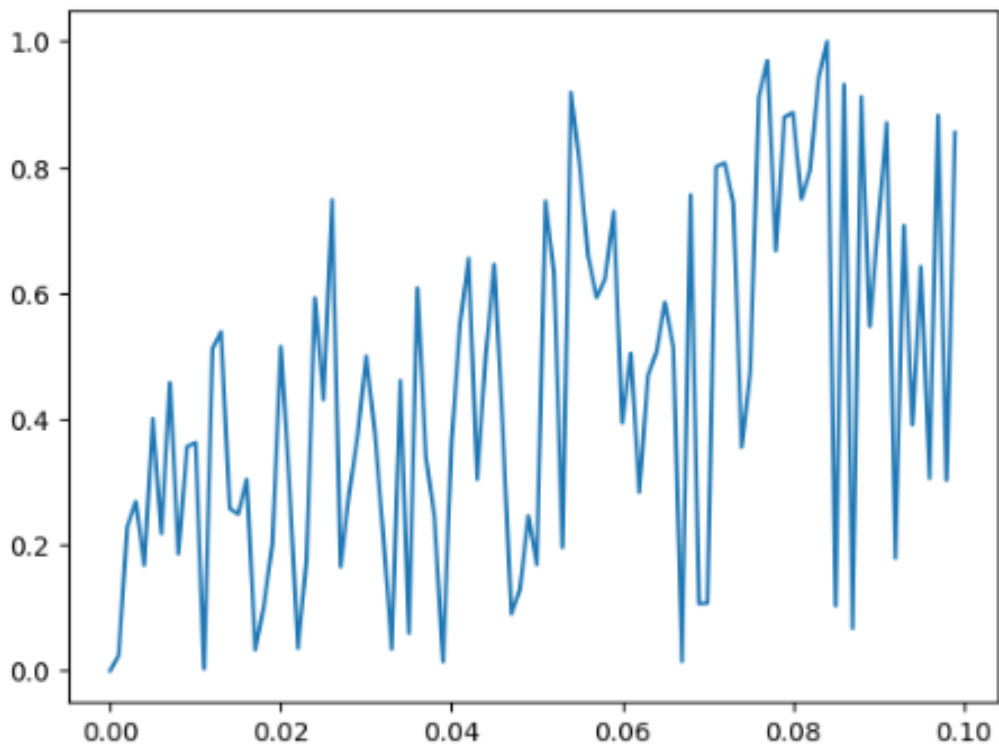


Figura 2: Camí no suau

on la segona integral és l'habitual integral de Riemann d'una funció contínua i acotada. Fem servir la notació $\dot{X}$ per designar diferenciació respecte una única variable: $\dot{X}_t = \frac{dX_t}{dt}$.

Observació 1.5. En general, podem integrar un camí $Y : [a, b] \rightarrow \mathbb{R}$ per un camí $X : [a, b] \rightarrow \mathbb{R}$. Definim la integral

$$\int_a^b Y_t dX_t = \int_a^b Y_t \dot{X}_t dt. \quad (1.4)$$

Passem a mostrar tres exemples que considerem força il·lustratius per ajudar a comprendre millor aquest concepte d'integral.

Exemple 1.6. Considerem el camí constant $Y_t = 1, \forall t \in [a, b]$. Llavors la integral de línia d' Y per qualsevol camí $X : [a, b] \rightarrow \mathbb{R}$ esdevé simplement l'increment de X entre a i b :

$$\int_a^b dX_t = \int_a^b \dot{X}_t dt = X_b - X_a. \quad (1.5)$$

Exemple 1.7. Ara considerem el camí $X_t = t, \forall t \in [a, b]$. Llavors, $\dot{X}_t = 1$ i, per tant, la integral de línia de qualsevol camí $Y : [a, b] \rightarrow \mathbb{R}$ per X_t és l'habitual integral de Riemann:

$$\int_a^b Y_t dX_t = \int_a^b Y_t dt. \quad (1.6)$$

Exemple 1.8. Finalment, introduïm un exemple que involucra càlculs numèrics. Considerem el camí bidimensional

$$X_t = (X_t^1, X_t^2) = (t^2, t^3), t \in [0, 1]. \quad (1.7)$$

Llavors, calculem la integral

$$\int_0^1 X_t^1 dX_t^2 = \int_0^1 t^2 3t^2 dt = \frac{3}{5} \quad (1.8)$$

Aquest exemple, de fet, és un cas particular d'integral iterada doble, que representem a continuació.

1.1 Signatura d'un camí

Definició 1.9. Sigui $X : [a, b] \rightarrow \mathbb{R}^d$ un camí amb $X^i : [a, b] \rightarrow \mathbb{R}$ camins unidimensionals per $i \in \{1, \dots, d\}$. Per cada i definim la quantitat

$$S(X)_{a,t}^i = \int_{a < s < t} dX_s^i = X_t^i - X_0^i, \quad (1.9)$$

que és l'increment de la coordenada i -èssima del camí a temps $t \in [a, b]$. Ara, per un parell $i, j \in \{1, \dots, d\}$, podem definir la **integral iterada doble** com

$$S(X)_{a,t}^{i,j} = \int_{a < s < t} S(X)_{a,s}^i dX_s^j = \int_{a < r < s < t} dX_r^i dX_s^j, \quad (1.10)$$

i la **integral iterada triple** com

$$S(X)_{a,t}^{i,j,k} = \int_{a < s < t} S(X)_{a,s}^{i,j} dX_s^k = \int_{a < q < r < s < t} dX_q^i dX_r^j dX_s^k. \quad (1.11)$$

Així doncs, d'aquesta manera podem continuar trobant integrals iterades.

Donats un enter $k \geq 1$ i una col·lecció d'indexos $i_1, \dots, i_k \in \{1, \dots, d\}$, de forma recursiva, definim la **integral iterada k vegades** al llarg dels indexos $i_1, \dots, i_k$ com

$$S(X)_{a,t}^{i_1, \dots, i_k} = \int_{a < s < t} S(X)_{a,s}^{i_1, \dots, i_{k-1}} dX_s^{i_k}. \quad (1.12)$$

Equivalentment, podem escriure

$$S(X)_{a,t}^{i_1, \dots, i_k} = \int_{a < t_k < t} \dots \int_{a < t_1 < t_2} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k}. \quad (1.13)$$

Definició 1.10. Anomenarem **signatura** d'un camí $X : [a, b] \rightarrow \mathbb{R}^d$ a la col·lecció d'integrals iterades d' X . Formalment,

$$S(X)_{a,b} = (1, S(X)_{a,b}^1, \dots, S(X)_{a,b}^d, S(X)_{a,b}^{1,1}, S(X)_{a,b}^{1,2}, \dots), \quad (1.14)$$

on, per convenció, el terme 0-èssim és 1 i la resta d'indexos (o **lletres**) són els elements del conjunt

$$W = \{(i_1, \dots, i_k) | k \geq 0, i_1, \dots, i_k \in \{1, \dots, d\}\}. \quad (1.15)$$

Al conjunt W l'anomenarem conjunt de **paraules** de l'**alfabet** $A = \{1, \dots, d\}$ i representarem la **paraula buida** (cas $k = 0$) amb λ .

Si $w = (i_1, \dots, i_k) \in W$ és una paraula, de forma natural, a l'enter k l'anomenarem *mida de la paraula* w .

Exemple 1.11. Si considerem un alfabet d'únicament dues lletres $\{1, 2\}$, tenim un nombre infinit de paraules que podem compondre:

- Les de mida 1: $\{1, 2\}$
- Les de mida 2: $\{11, 12, 21, 22\}$
- Les de mida 3: $\{111, 112, 121, \dots\}$
- ...

Exemple 1.12. L'exemple més senzill de signatura d'un camí és el d'un camí unidimensional. L'alfabet en aquest cas és $A = \{1\}$ i el conjunt de paraules corresponent és $W = \{(1, \dots, 1) | k \geq 1\}$, on 1 apareix k vegades a cada paraula. Si, a més, considerem el camí $X : [a, b] \longrightarrow \mathbb{R}$, $X_t = t$, la signatura ve donada per:

$$S(X)_{a,b}^1 = X_b - X_a \quad (1.16)$$

$$S(X)_{a,b}^{1,1} = \frac{(X_b - X_a)^2}{2!} \quad (1.17)$$

$$S(X)_{a,b}^{1,1,1} = \frac{(X_b - X_a)^3}{3!} \quad (1.18)$$

$\vdots$

És important tenir en compte aquest cas que, si bé és senzill, farem servir freqüentment al llarg d'aquest treball.

Exemple 1.13. Considerem el camí $X : [0, \frac{\pi}{2}] \longrightarrow \mathbb{R}$, definit explícitament com $X_t = (X_t^1, X_t^2) = (t, \sin t)$ i, representat a la Figura 3. El nostre alfabet ara és $\{1, 2\}$; i el conjunt de paraules,

$$W = \{(i_1, \dots, i_k) | k \geq 0, i_1, \dots, i_k \in \{1, 2\}\}.$$

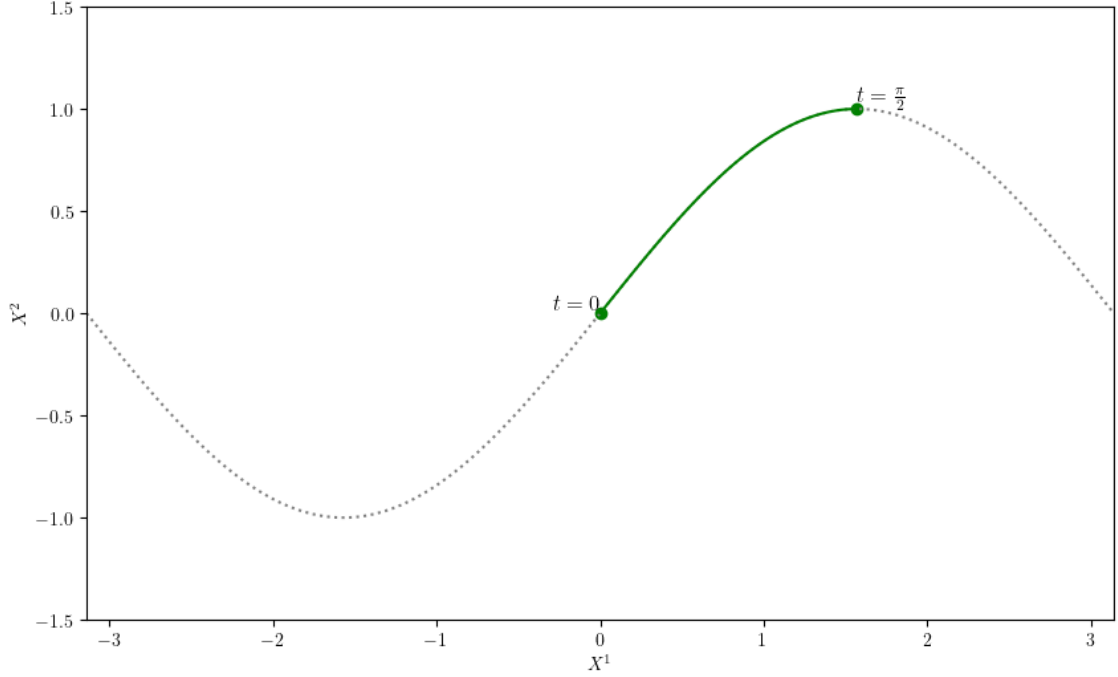


Figura 3: Parametrització d'un camí bidimensional a $\mathbb{R}^2$

Un càlcul ràpid dels elements de la signatura ens dona:

$$S(X)_{0, \frac{\pi}{2}}^1 = \int_{0 < t < \frac{\pi}{2}} dX_t^1 = \int_0^{\frac{\pi}{2}} dt = X_{\frac{\pi}{2}}^1 - X_0^1 = \frac{\pi}{2}, \quad (1.19)$$

$$S(X)_{0, \frac{\pi}{2}}^2 = \int_{0 < t < \frac{\pi}{2}} dX_t^2 = \int_0^{\frac{\pi}{2}} \cos t dt = X_{\frac{\pi}{2}}^2 - X_0^2 = \sin \frac{\pi}{2} - \sin 0 = 1, \quad (1.20)$$

$$S(X)_{0, \frac{\pi}{2}}^{1,1} = \int \int_{0 < t_1 < t_2 < \frac{\pi}{2}} dX_{t_1}^1 dX_{t_2}^1 = \int_0^{\frac{\pi}{2}} \left[\int_0^{t_2} dt_1 \right] dt_2 = \int_0^{\frac{\pi}{2}} t_2 dt_2 = \frac{\pi^2}{8}, \quad (1.21)$$

$$\begin{aligned} S(X)_{0, \frac{\pi}{2}}^{1,2} &= \int \int_{0 < t_1 < t_2 < \frac{\pi}{2}} dX_{t_1}^1 dX_{t_2}^2 = \int_0^{\frac{\pi}{2}} \left[\int_0^{t_2} dt_1 \right] \cos t_2 dt_2 = \\ &= \int_0^{\frac{\pi}{2}} t_2 \cos t_2 dt_2 = \frac{\pi}{2} - 1, \end{aligned} \quad (1.22)$$

$$\begin{aligned}
S(X)_{0, \frac{\pi}{2}}^{2,1} &= \int \int_{0 < t_2 < t_1 < \frac{\pi}{2}} dX_{t_1}^2 dX_{t_2}^1 = \int_0^{\frac{\pi}{2}} \left[\int_0^{t_2} \cos t_1 dt_1 \right] dt_2 = \\
&= \int_0^{\frac{\pi}{2}} \sin t_2 dt_2 = 1
\end{aligned} \tag{1.23}$$

$$\begin{aligned}
S(X)_{0, \frac{\pi}{2}}^{2,2} &= \int \int_{0 < t_1 < t_2 < \frac{\pi}{2}} dX_{t_1}^2 dX_{t_2}^2 = \int_0^{\frac{\pi}{2}} \left[\int_0^{t_2} \cos t_1 dt_1 \right] \cos t_2 dt_2 = \\
&= \int_0^{\frac{\pi}{2}} \sin t_2 \cos t_2 dt_2 = \frac{1}{2},
\end{aligned} \tag{1.24}$$

$$\begin{aligned}
S(X)_{0, \frac{\pi}{2}}^{1,1,1} &= \int \int \int_{0 < t_1 < t_2 < t_3 < \frac{\pi}{2}} dX_{t_1}^1 dX_{t_2}^1 dX_{t_3}^1 = \\
&= \int_0^{\frac{\pi}{2}} \left[\int_0^{t_3} \left[\int_0^{t_2} dt_1 \right] dt_2 \right] dt_3 = \frac{\pi^3}{48},
\end{aligned} \tag{1.25}$$

⋮

D'aquesta manera, podem seguir calculant tots els elements $S(X)_{0, \frac{\pi}{2}}^{i_1, \dots, i_k}$ de la signatura.

Tot i l'interés de la signatura, aquesta consta d'un nombre infinit d'elements. A la pràctica, sobretot quan volem obtenir un vector de característiques com a entrada per un model d'aprenentatge automàtic, ens caldrà fer ús d'un nombre reduït d'aquests elements. És per això que necessitem definir un concepte complementari a la signatura com és la signatura truncada.

Definició 1.14. Sigui $S(X)_{a,b}$ la signatura d'un camí $X : [a, b] \longrightarrow \mathbb{R}^d$. Llavors definim la ***signatura truncada*** de nivell k del camí X com

$$S(X)_{a,b}|_k = (1, S(X)_{a,b}^1, \dots, S(X)_{a,b}^{d, \dots, d}), \tag{1.26}$$

on l'últim element té l'índex d repetit k vegades.
 És a dir, prenem els elements de la signatura amb mida de paraula $\leq k$.

Exemple 1.15. En el cas de la signatura truncada a nivell 2, si tenim un alfabet de dos elements $A = \{1, 2\}$ com el de l'exemple 1.11, el conjunt de paraules serà

$$\{1, 2, 11, 12, 21, 22\}.$$

Com ja hem explicat a la introducció i com veurem en més detall a seccions posteriors, un dels interessos principals de la signatura és la seva capacitat de capturar propietats geomètriques essencials d'un camí $X : [a, b] \rightarrow \mathbb{R}^d$. A la secció 1.2 veurem que la signatura d'un camí és l'únic que cal per determinar la solció d'una equació diferencial ordinària controlada per X . És per això que una pregunta legítima que podem arribar a fer-nos és: un camí està completament determinat per la seva signatura? Tenint en compte les propietats de la signatura que exposarem a la secció 1.3, la resposta, generalment és no. La velocitat a la que es recorre un camí, per exemple no està determinada per la signatura. Tantmateix, raonaments més complexos ens porten a veure que, essencialment, aquesta és l'única propietat que es perd [2]. No entrarem en detall en això, però sí que exposarem alguns d'aquests aspectes més importants de la signatura.

1.2 Signatura a partir d'iteracions de Picard

Fem una breu exposició de com sorgeix la signatura, de forma natural, en el context de la teoria clàssica d'equacions diferencials ordinàries (EDOs), sense entrar massa detall en molts dels termes usats.

Considerem l'EDO

$$\frac{dx}{dy} f(x; y), \quad y(x_0) = y_0, \tag{1.27}$$

on $y(x_0) : \mathbb{R} \longrightarrow \mathbb{R}$, $x_0, y_0 \in \mathbb{R}$. Podem posar la solució d'aquesta EDO en termes de l'equació integral de Volterra

$$y(x) = y(x_0) + \int_{x_0}^x d(t; y(t))dt. \quad (1.28)$$

La versió clàssica del teorema de Picard-Lindelöf determina aquesta solució com a límit de la successió

$$y_0(x) = y(x_0), \quad y_k(x) = y(x_0) + \int_{x_0}^x f(t; y_{k-1}(t))dt, \forall k \geq 1, \quad (1.29)$$

anomenada iteracions de Picard.

Considerem, ara, un camí $X : [a, b] \longrightarrow \mathbb{R}^d$ i sigui $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^e)$ l'espai vectorial d'aplicacions lineals de $\mathbb{R}^d$ a $\mathbb{R}^e$. També podem pensar aquest espai vectorial com el de matrius reals de dimensions $d \times e$.

Sigui $Z : [a, b] \longrightarrow \mathcal{L}(\mathbb{R}^d, \mathbb{R}^e)$ un camí. De la mateixa manera com hem definit la integral de línia, podem considerar la integral

$$\int_a^b Z_t dX_t \quad (1.30)$$

com a element de $\mathbb{R}^e$. Sigui $V : \mathbb{R}^e \longrightarrow \mathcal{L}(\mathbb{R}^d, \mathbb{R}^e)$ una funció i $Y : [a, b] \longrightarrow \mathbb{R}^e$ un camí. Diem que Y resol l'equació diferencial

$$dY_t = V(Y_t)dX_t, Y_a = y \in \mathbb{R}^e, \quad (1.31)$$

justament quan és de la forma

$$Y_t = y + \int_a^t V(Y_s)dX_s, \forall t \in [a, b]. \quad (1.32)$$

Un dels procediments estàndard per trobar aquesta solució és via iteracions de Picard. Sigui $Y : [a, b] \longrightarrow \mathbb{R}^e$ una funció arbitrària. Definim el camí $F(Y) : [a, b] \longrightarrow \mathbb{R}^e$ com

$$F(Y)_t = y + \int_a^t V(Y_s)dX_s. \quad (1.33)$$

Si considerem la successió $Y_t^n = F(Y^{n-1})_t$, amb Y_t^0 arbitrari, sota les hipòtesis

adequades, podem veure que Y_t^n convergeix a Y quan $n \rightarrow \infty$. Veiem quina forma tenen aquestes iteracions de Picard, començant per $Y - t^0 = y, \forall t \in [a, b]$ (indiquem amb I_e la matriu identitat a $\mathcal{L}(\mathbb{R}^e, \mathbb{R}^e)$):

$$Y_t^0 = y, \quad (1.34)$$

$$Y_t^1 = y + \int_a^t V(Y_s^0) dX_s = \left(\int_a^t dV(X_s) + I_e \right) (y), \quad (1.35)$$

$$\begin{aligned} Y_t^2 &= y + \int_a^t V(Y_s^1) dX_s = \\ &= \left(\int_a^t \int_a^s dV(X_u) dV(X_s) + \int_a^t dV(X_s) + I_e \right) (y), \end{aligned} \quad (1.36)$$

$\vdots$

$$\begin{aligned} Y_t^n &= y + \int_a^t V(Y_s^{n-1}) dX_s = \\ &= \left(\sum_{k=1}^n \int_{a < t_1 < \dots < t_k < t} dV(X_{t_1}) \dots dV(X_{t_k} + I_e) \right) (y), \end{aligned} \quad (1.37)$$

$\ddots$

Al ser $\mathcal{L}(\mathbb{R}^e, \mathbb{R}^e)$ una àlgebra de matrius, cada

$$\int_{a < t_1 < \dots < t_k < t} dV(X_{t_1}) \dots dV(X_{t_k}) \quad (1.38)$$

pot ser definida com a element de $\mathcal{L}(\mathbb{R}^e, \mathbb{R}^e)$ que, està determinat completament pel k -èssim nivell de la signatura $S(X)_{a,t}$ de X a temps $t \in [a, b]$. D'aquí poddem extreure la conclusió que, de fet, la solució Y_t està completament determinada per la signatura $S(X)_{a,t}$.

1.3 Propietats importants de la signatura

A continuació repassem breument tres de les propietats més importants de la signatura d'un camí: la invariància de la signatura per reparametritzacions temporals, la identitat de Chen i la propietat d'inversió temporal.

1.3.1 Invariància per reparametritzacions

Definició 1.16. Anomenem *reparametrització* a una funció $\psi : [a, b] \longrightarrow [a, b]$ contínua, exhaustiva i no decreixent.

Al llarg d'aquesta exposició, considerem únicament les reparametritzacions suaus.

Teorema 1.17. (*Invariància per reparametritzacions temporals*). Sigui $X : [a, b] \longrightarrow \mathbb{R}^d$ un camí i $\psi : [a, b] \longrightarrow [a, b]$ una reparametrització. Definim el camí reparametritzat $\tilde{X} : [a, b] \longrightarrow \mathbb{R}^d$ com $\tilde{X}_t = X_{\psi(t)}$. Llavors tenim

$$S(\tilde{X})_{a,b} = S(X)_{a,b}. \quad (1.39)$$

Demostració: Siguin $X, Y : [a, b] \longrightarrow \mathbb{R}$ dos camins a $\mathbb{R}$ i $\psi : [a, b] \longrightarrow [a, b]$ una reparametrització. Definim els camins $\tilde{X}, \tilde{Y} : [a, b] \longrightarrow \mathbb{R}$ com $\tilde{X} = X_{\psi(t)}$ i $\tilde{Y} = Y_{\psi(t)}$. Tenim que la derivada de $\tilde{X}$ respecte t és

$$\dot{\tilde{X}}_t = \dot{X}_{\psi(t)} \dot{\psi}(t). \quad (1.40)$$

Per tant, la integral de línia de $\tilde{Y}$ per $\tilde{X}$ és

$$\int_a^b \tilde{Y}_t d\tilde{X}_t = \int_a^b Y_{\psi(t)} \dot{X}_{\psi(t)} \dot{\psi}(t) dt = \int_a^b Y_u dX_u, \quad (1.41)$$

on fem el canvi $u = \psi(t)$.

Ara, si tenim el camí $X[a, b] \longrightarrow \mathbb{R}^d$ multidimensional i la seva reparametrització per ψ , $\tilde{X} : [a, b] \longrightarrow \mathbb{R}^d$, $\tilde{X}_t = X_{\psi(t)}$. Llavors com que cada terme de la signatura $S(X)^{i_1, \dots, i_k}$ està definit com a una integral iterada k vegades de X , tenim

$$S(\tilde{X})_{a,b}^{i_1, \dots, i_k} = S(X)_{a,b}^{i_1, \dots, i_k}, \forall k \geq 0, i_1, \dots, i_k \in \{1, \dots, d\}. \quad (1.42)$$

□

1.3.2 Identitat de Chen

Una propietat important de la signatura és la identitat de Chen, que proporciona una relació algebraica entre camins i les seves signatures. Per arribar a formular-la, però, ens cal abans introduir el concepte d'àlgebra d'una sèrie de potències formal.

Definició 1.18. Siguin $e_1, \dots, e_d$ d indeterminacions formals. L'àlgebra de **sèries de potències formals** (no commutatives) a d indeterminacions, és l'espai vectorial de totes les sèries de potències de la forma

$$\sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} \lambda_{i_1, \dots, i_k} e_{i_1} \dots e_{i_k}, \quad (1.43)$$

on el segon sumatori recorre totes les paraules $(i_1, \dots, i_k)$, amb índexos a $i_1, \dots, i_k \in \{1, \dots, d\}$, i els $\lambda_{i_1, \dots, i_k}$ són nombres reals.

Una sèrie de potències formal amb únicament un nombre finit de coeficients $\lambda_{i_1, \dots, i_k}$ no nuls és anomenat polinomi formal. Els seus termes $e_{i_1}, \dots, e_{i_k}$ s'anomenen monomis i el terme corresponent a $k = 0$ és simplement un $\lambda_0 \in \mathbb{R}$. Habitualment, també, s'anomena àlgebra tensorial d' $\mathbb{R}^d$ a l'espai de sèries de potències formals.

Podem dotar a aquest espai de sèries de potències formals d'una estructura d'espai vectorial, de forma natural definint suma i producte escalar com

$$\left(\sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} \lambda_{i_1, \dots, i_k} e_{i_1} \dots e_{i_k} \right) + \left(\sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} \mu_{i_1, \dots, i_k} e_{i_1} \dots e_{i_k} \right) \\ \sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} ((\lambda_{i_1, \dots, i_k} + \mu_{i_1, \dots, i_k}) e_{i_1} \dots e_{i_k})$$

i,

$$c \left(\sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} \lambda_{i_1, \dots, i_k} e_{i_1} \dots e_{i_k} \right) = \sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} c \lambda_{i_1, \dots, i_k} e_{i_1} \dots e_{i_k}$$

respectivament.

També podem definir el **producte** $\otimes$ entre monomis de la forma següent:

$$e_{i_1} \dots e_{i_k} \otimes e_{j_1} \dots e_{j_m} = e_{i_1} \dots e_{i_k} \otimes e_{j_1} \dots e_{j_m}. \quad (1.44)$$

Observació 1.19. Si tenim un camí $X : [a, b] \longrightarrow \mathbb{R}^d$, podem expressar la seva signatura $S(X)_{a,b}$ fent servir un conjunt de monomis $e_{i_1}, \dots, e_{i_k}$, $i_1, \dots, i_k \in \{1, \dots, d\}$ com:

$$S(X)_{a,b} = \sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \{1, \dots, d\}} S(X)_{a,b}^{i_1, \dots, i_k} e_{i_1} \dots e_{i_k}. \quad (1.45)$$

Definició 1.20. Siguin $X : [a, b] \longrightarrow \mathbb{R}^d$ i $Y : [b, c] \longrightarrow \mathbb{R}^d$, definim la **concatenació de camins** X i Y com el camí $X * Y : [a, c]$:

$$(X * Y)_t = \begin{cases} X_t, & t \in [a, b] \\ X_b + (Y_t - Y_b), & t \in [b, c] \end{cases}. \quad (1.46)$$

Teorema 1.21. (*Identitat de Chen*). Siguin $X : [a, b] \longrightarrow \mathbb{R}^d$ i $Y : [b, c] \longrightarrow \mathbb{R}^d$. Llavors,

$$S(X * Y)_{a,c} = S(X)_{a,b} \otimes S(Y)_{b,c}. \quad (1.47)$$

Dit de forma informal, la identitat de Chen ens permet passar del producte concatenació de camins $*$ al producte $\otimes$.

1.3.3 Propietat d'inversió temporal

De forma col·loquial, podem dir que la propietat d'inversió temporal consisteix en que la signatura $S(X)_{a,b}$ d'un camí $X : [a, b] \longrightarrow \mathbb{R}^d$ és l'inversa sota el producte $\otimes$, de la signatura obtinguda en recórrer el camí X enrere en el temps.

Definició 1.22. Sigui $X : [a, b] \longrightarrow \mathbb{R}^d$ un camí a $\mathbb{R}^d$, definim la seva ***inversió temporal*** com el camí $\overleftarrow{X} : [a, b] \longrightarrow \mathbb{R}^d$, $\overleftarrow{X} = X_{a+b-t}$, $\forall t \in [a, b]$.

Teorema 1.23. (***Signatura de la inversió temporal***). Sigui $X : [a, b] \longrightarrow \mathbb{R}$ un camí multidimensional. Llavors tenim que

$$S(X)_{a,b} \otimes S(\overleftarrow{X})_{a,b} = 1. \quad (1.48)$$

Entenent que 1 és la sèrie de potències formal que té $\lambda_1 = 1$ i $\lambda_{i_1, \dots, i_k} = 0$, $\forall k \geq 1$ i $i_1, \dots, i_k \in \{1, \dots, d\}$, que és l'element neutre pel producte $\otimes$.

2 Signatura discreta fent servir interpolació lineal

Al llarg d'aquest capítol introduïrem el concepte de signatura discreta plana i de signatura discreta com a cas particular de la signatura, seguint l'article d'Adachi, T. [1]

Al capítol anterior hem definit el producte $\otimes$ entre monomis d'un polinomi formal. A continuació, anàlogament, farem servir el mateix símbol per representar la operació *concatenació* entre paraules d'un alfabet.

Definició 2.1. Siguin $u = \{i_1, \dots, i_j\}, v = \{i_{j+1}, \dots, i_k\}$ paraules de l'alfabet A , la **concatenació de paraules** u i v , que denotem $u \otimes v$, és la paraula w definida com

$$w = \{i_1, \dots, i_j, i_{j+1}, \dots, i_k\}. \quad (2.1)$$

Observació 2.2. Ràpidament podem veure com donat un camí $X : [a, b] \rightarrow \mathbb{R}^d$, una paraula $w \in W$ i $i \in A$, $S(X)_{a,t}^{w \otimes i}$ és

$$S(X)_{a,t}^{w \otimes i} = \int_{a < s < t} S(X)_{a,s}^w dX_s^i. \quad (2.2)$$

Un cop definida una paraula i la signatura d'un camí, ens centrarem en el cas particular de camins discrets i les seves signatures, cosa que ens permetrà obtenir la signatura d'un conjunt de dades d'una sèrie temporal. Fixem el domini temporal discret $T = \{t_0, t_1, t_2, \dots\}$ amb $0 = t_0 < t_1 < t_2 < \dots$. A partir d'ara doncs, considerarem camins del tipus $X : T \rightarrow \mathbb{R}^d$.

Com que tenim un camí discret, per que la signatura tingui sentit, haurem de fer-lo continu. En aquesta secció considerem l'opció de la interpolació lineal.

Si tenim un camí $X : T \rightarrow \mathbb{R}^d$, per $t \in [t_n, t_{n+1}]$, podem generalitzar-lo assignant valors

$$X_t^i = \frac{X_{t_n}^i (t_{n+1} - t) + X_{t_{n+1}}^i (t - t_n)}{t_{n+1} - t_n}, \quad (2.3)$$

llavors,

$$\dot{X}_t^i = \frac{X_{t_{n+1}}^i - X_{t_n}^i}{t_{n+1} - t_n}. \quad (2.4)$$

Definició 2.3. Sigui A un alfabet, definim l'*alfabet ampliat* $\bar{A}$ com

$$\bar{A} = A \times \{-, +\}. \quad (2.5)$$

Sigui $i \in \bar{A}$, anomenarem a $i^- = (i, -) \in \bar{A}$ i $i^+ = (i, +) \in \bar{A}$ *cap* i *cua* d' i , respectivament.

D'ara endavant, passarem a considerar el conjunt de paraules W de l'alfabet ampliat $\bar{A}$.

2.1 Signatura discreta plana

Definició 2.4. Sigui A un alfabet, $X : T \longrightarrow \mathbb{R}^d$ un camí discret, $m, n \in \mathbb{N}$ amb $m < n$, $w \in W$ i $i \in A$. Llavors definim, inductivament, $S(X)_{t_m, t_n}^w$ de la forma següent:

$$S(X)_{t_m, t_n}^\lambda = 1, \quad (2.6)$$

$$S(X)_{t_m, t_n}^{w \otimes i^-} = \sum_{l=m}^{n-1} S(X)_{t_m, t_l}^w (X_{t_{l+1}}^i - X_{t_l}^i), \quad (2.7)$$

$$S(X)_{t_m, t_n}^{w \otimes i^+} = \sum_{l=m}^{n-1} S(X)_{t_m, t_{l+1}}^w (X_{t_{l+1}}^i - X_{t_l}^i), \quad (2.8)$$

La *signatura discreta plana* de X a l'interval $[t_m, t_n]$, és el vector

$$S(X)_{t_m, t_n} = (1, S(X)_{t_m, t_n}^1, \dots, S(X)_{t_m, t_n}^d, S(X)_{t_m, t_n}^{1-}, \dots, S(X)_{t_m, t_n}^{d-}, \dots), \quad (2.9)$$

on els indexos són les paraules del conjunt de paraules W .

Tal i com hem vist amb la signatura, podem definir la **signatura discreta plana truncada** de nivell k com

$$S(X)_{t_m, t_n}|_k = (1, S(X)_{t_m, t_n}^1, \dots, S(X)_{t_m, t_n}^{d-, \dots, d-}). \quad (2.10)$$

Hom pot comprovar com aquesta definició és un cas particular de la signatura d'un camí continu feta prèviament.

A partir d'ara quan fem servir la notació $S(X)$ ens referirem específicament a la signatura discreta plana. A vegades ens referirem també a la signatura truncada simplement com a signatura.

Exemple 2.5. Suposem que tenim els indexos $i, i_1, i_2, i_3 \in A$ un alfabet i sigui $*$ $\in \{-, +\}$, un càlcul ràpid d'alguns dels elements de la signatura discreta plana d'un camí X per $m, n \in \mathbb{N}$, $m < n$ ens dona

$$S(X)_{t_m, t_n}^{i*} = \sum_{m \leq l < n} (X_{t_{l+1}}^i - X_{t_l}^i) = X_{t_n}^i - X_{t_m}^i, \quad (2.11)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^-} = \sum_{m \leq l_1 < l_2 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}), \quad (2.12)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^-} = \sum_{m \leq l_1 \leq l_2 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}), \quad (2.13)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^- \otimes i_3^-} = \sum_{m \leq l_1 < l_2 < l_3 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2})(X_{t_{l_3+1}}^{i_3} - X_{t_{l_3}}^{i_3}), \quad (2.14)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^- \otimes i_3^+} = \sum_{m \leq l_1 < l_2 \leq l_3 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2})(X_{t_{l_3+1}}^{i_3} - X_{t_{l_3}}^{i_3}), \quad (2.15)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^+ \otimes i_3^-} = \sum_{m \leq l_1 \leq l_2 < l_3 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2})(X_{t_{l_3+1}}^{i_3} - X_{t_{l_3}}^{i_3}) \quad (2.16)$$

$$S(X)_{t_m, t_n}^{i_1^* \otimes i_2^+ \otimes i_3^+} = \sum_{m \leq l_1 \leq l_2 \leq l_3 < n} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1})(X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2})(X_{t_{l_3+1}}^{i_3} - X_{t_{l_3}}^{i_3}). \quad (2.17)$$

Fixem-nos en la correspondència que hi ha entre $\{-, +\}$ i $\{<, \leq\}$ en els sumatoris.

Exemple 2.6. Proposem un exemple incorporant un flux de dades numèriques. Considerem un flux bidimensional X del que tenim les observacions de la taula 1.

t	0	1	1.5	2
X_1	0		1	1.5
X_2	0	1	1.5	

Taula 1: Exemple de mostra de flux bidimensional amb dades faltants

Deliberadament, en aquest exemple ometem algunes d'aquestes observacions. Al món real és habitual la manca de certes dades. Hi ha diverses opcions pel tractament d'aquestes situacions, al llarg d'aquest escrit considerem una de senzilla consistent a 'emplenar els buits' amb la última instància coneguda d'aquella component de la sèrie. Ho il·lustrem a la taula 2

t	0	1	1.5	2
X_1	0	0	1	1.5
X_2	0	1	1.5	1.5

Taula 2: Exemple de mostra de flux bidimensional

Trobem els valors de la signatura $S(X)_{0,2}$ de nivell ≤ 2 , considerant $* \in \{-, +\}$:

$$S(X)_{0,2}^\lambda = 1,$$

$$\begin{aligned}
S(X)_{0,2}^{1*} &= 1.5, & S(X)_{0,2}^{2*} &= 1.5, \\
S(X)_{0,2}^{1*1+} &= 1.75, & S(X)_{0,2}^{1*1-} &= 0.5, & S(X)_{0,2}^{1*2+} &= 0.5, & S(X)_{0,2}^{1*2-} &= 0, \\
S(X)_{0,2}^{2*1+} &= 2.25, & S(X)_{0,2}^{2*1-} &= 1.75, & S(X)_{0,2}^{2*2+} &= 1.75, & S(X)_{0,2}^{2*2-} &= 0.5.
\end{aligned}$$

2.2 Signatura discreta

Un cop definida la signatura discreta plana, procedim a definir el concepte de signatura discreta incorporant una **constant de decadència** μ . Aquesta constant de decadència, permetrà tenir en compte el temps recorregut a la sèrie temporal assignant pesos iguals o inferiors a 1 als termes desfassats. Aprofitarem la naturalesa recursiva de la definició de la signatura discreta per fer-ho.

Definició 2.7. Sigui A un alfabet, $X : T \longrightarrow \mathbb{R}^d$ un camí discret, $m, n \in \mathbb{N}$ amb $m < n$, $w \in W$, $i \in A$ i $\mu \geq 0$. Llavors definim, inductivament, $S^\mu(X)_{t_m, t_n}^w$ de la forma següent:

$$S^\mu(X)_{t_m, t_n}^\lambda = 1, \quad (2.18)$$

$$S^\mu(X)_{t_m, t_n}^w = \begin{cases} 1 & \text{si } w = \lambda \\ 0 & \text{en la resta de casos} \end{cases}, \quad (2.19)$$

$$S^\mu(X)_{t_m, t_n}^{w \otimes i^-} = e^{-\mu(t_n - t_{n-1})} (S^\mu(X)_{t_m, t_{n-1}}^{w \otimes i^-} + (X_{t_n}^i - X_{t_{n-1}}^i) S^\mu(X)_{t_m, t_{n-1}}^w), \quad (2.20)$$

$$S^\mu(X)_{t_m, t_n}^{w \otimes i^+} = e^{-\mu(t_n - t_{n-1})} S^\mu(X)_{t_m, t_{n-1}}^{w \otimes i^+} + (X_{t_n}^i - X_{t_{n-1}}^i) S^\mu(X)_{t_m, t_n}^w. \quad (2.21)$$

La **signatura discreta** de X a l'interval $[t_m, t_n]$, és el vector

$$S^\mu(X)_{t_m, t_n} = (1, S^\mu(X)_{t_m, t_n}^1, \dots, S^\mu(X)_{t_m, t_n}^d, S^\mu(X)_{t_m, t_n}^{1^-}, \dots, S^\mu(X)_{t_m, t_n}^{d^-}, \dots). \quad (2.22)$$

on els indexos són les paraules del conjunt de paraules W .

De la mateixa manera que amb la signatura discreta plana, la **signatura discreta truncada** de nivell k és

$$S^\mu(X)_{t_m, t_n}|_k = (1, S^\mu(X)_{t_m, t_n}^1, \dots, S^\mu(X)_{t_m, t_n}^{d^-, \dots, d^-}). \quad (2.23)$$

Observació 2.8. És trivial veure que per $\mu = 0$, la signatura discreta coincideix amb la signatura discreta plana.

Exemple 2.9. Suposem que tenim els indexos $i, i_1, i_2, \bar{A}$ un alfabet ampliat, alguns dels elements de la signatura discreta d'un camí X per $m, n \in \mathbb{N}$, $m < n$ són

$$S^\mu(X)_{t_m, t_n}^{i^-} = \sum_{m \leq l < n} e^{-\mu(t_n - t_l)} (X_{t_{l+1}}^i - X_{t_l}^i), \quad (2.24)$$

$$S^\mu(X)_{t_m, t_n}^{i^+} = \sum_{m \leq l < n} e^{-\mu(t_n - t_{l+1})} (X_{t_{l+1}}^i - X_{t_l}^i), \quad (2.25)$$

$$S^\mu(X)_{t_m, t_n}^{i_1^- \otimes i_2^-} = \sum_{m \leq l_1 < l_2 < n} e^{-\mu(t_n - t_{l_1})} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1}) (X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}), \quad (2.26)$$

$$S^\mu(X)_{t_m, t_n}^{i_1^- \otimes i_2^+} = \sum_{m \leq l_1 \leq l_2 < n} e^{-\mu(t_n - t_{l_1})} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1}) (X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}), \quad (2.27)$$

$$S^\mu(X)_{t_m, t_n}^{i_1^+ \otimes i_2^-} = \sum_{m \leq l_1 < l_2 < n} e^{-\mu(t_n - t_{l_1+1})} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1}) (X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}), \quad (2.28)$$

$$S^\mu(X)_{t_m, t_n}^{i_1^+ \otimes i_2^+} = \sum_{m \leq l_1 \leq l_2 < n} e^{-\mu(t_n - t_{l_1+1})} (X_{t_{l_1+1}}^{i_1} - X_{t_{l_1}}^{i_1}) (X_{t_{l_2+1}}^{i_2} - X_{t_{l_2}}^{i_2}). \quad (2.29)$$

Exemple 2.10. Calculem els elements de la signatura discreta de fins a nivell 2 del flux de dades X de l'exemple 2.6. Fem servir una constant de decadència $\mu = \log 2 \approx 0.693$:

$$S^\mu(X)_{0,2}^\lambda = 1,$$

$$S^\mu(X)_{0,2}^{1+} = 1.21, \quad S^\mu(X)_{0,2}^{1-} = 0.85, \quad S^\mu(X)_{0,2}^{2+} = 0.85, \quad S^\mu(X)_{0,2}^{2-} = 0.5,$$

$$S^\mu(X)_{0,2}^{1+1+} = 1.31, \quad S^\mu(X)_{0,2}^{1+1-} = 0.35, \quad S^\mu(X)_{0,2}^{1+2+} = 0.35, \quad S^\mu(X)_{0,2}^{1+2-} = 0,$$

$$S^\mu(X)_{0,2}^{1-1+} = 0.93, \quad S^\mu(X)_{0,2}^{1-1-} = 0.25, \quad S^\mu(X)_{0,2}^{1-2+} = 0.25, \quad S^\mu(X)_{0,2}^{1-2-} = 0,$$

$$S^\mu(X)_{0,2}^{2+1+} = 1.28, \quad S^\mu(X)_{0,2}^{2+1-} = 0.92, \quad S^\mu(X)_{0,2}^{2+2+} = 0.92, \quad S^\mu(X)_{0,2}^{2+2-} = 0.25,$$

$$S^\mu(X)_{0,2}^{2-1+} = 0.75, \quad S^\mu(X)_{0,2}^{2-1-} = 0.5, \quad S^\mu(X)_{0,2}^{2-2+} = 0.5, \quad S^\mu(X)_{0,2}^{2-2-} = 0.12.$$

3 Implementació del càlcul de la signatura discreta

En aquesta secció es presentaran dues alternatives per calcular els valors dels elements de la signatura discreta d'un flux de dades: (i) una versió naïve que implementa l'algoritme de càlcul de forma recursiva directament de la definició i (ii) una versió una mica més optimitzada introduïda per Adachi, T i Naritomi, Y. a l'article ja citat [1].

En seccions posteriors es farà servir aquest algoritme per $\mu = 0$ (cas signatura discreta plana) i, tot i que aquí no s'explicita, creiem que un algoritme que implementi aquest cas específicament seria força millor que l'algoritme de la signatura discreta.

El codi complet implementant els algoritmes que s'especificaran a continuació pot ser trobat a https://github.com/fparcerisasv/signature_method o a la pàgina de github dels autors <https://github.com/takanori-adachi/discrete-signature>. El codi original està implementat directament en *Python*, la nostra proposta, en canvi, està implementada en *C* i pot ser accedida des de *Python* a través d'una llibreria que fa l'*embedding*. Creiem que el canvi de llenguatge pot ajudar a reduir encara més els temps d'execució.

3.1 Implementació naïve

Tot i que no entrarem massa en detall en la implementació naïve i ens centrarem en la implementació millorada del càlcul, a continuació enumerem els passos generals que es segueixen i que, en gran mesura, es reaprofiten a la implementació millorada.

1. En primer lloc es llegeixen els valors del camí $X : T \rightarrow \mathbb{R}^d$ i el valor de μ .
Per cada $t_n \in T$:
 - (a) Es calculen els elements $e^{-\mu(t_n - t_{n-1})}$ i es guarden en un vector.
 - (b) Per cada $i \in \{1, \dots, d\}$ es calculen les diferències $X_{t_n}^i - X_{t_{n-1}}^i$ i es guarden en un vector.

Aquest pas previ ens permetrà estalviar recalculacions innecessàries.

2. Es calculen totes les paraules de l'alfabet ampliat $\{1, \dots, d\} \times +, -$, fins a un nivell donat k i es guarden en un vector.

3. Finalment, per cada paraula w es calculen els elements $S^\mu(X)_{t_m, t_n}^w$, pels m i n triats inicialment, fent crides recursives a la mateixa funció, que implementa explícitament la definició de la signatura discreta.

Aquest últim pas implica recalculacions innecessàries d'elements $S^\mu(X)_{t_m, t_n}^w$ intermitjos. D'una forma senzilla, guardant els valors ja calculats, podem estalviar-nos aquests casos accelerant així el càlcul global. A continuació passem a exposar-ho amb més detall.

3.2 Implementació millorada

Per aquesta implementació millorada, anàlogament a la implementació de *Python* d'Adachi, T., nosaltres plantegem quatre structs **Index**, **Data**, **Word** i **Signature**. A **Index**, simplement guardem un enter $i \in \{0, \dots, d-1\}$ i un booleà en funció si és un *head* o un *tail*:

```
typedef struct {
    int i;
    bool head;
}Index;
```

A **Data** guardem els valors $t \in T$ i els X_t^i . També calculem els elements $X_{t_n}^i - X_{t_{n-1}}^i$ per cada i i cada t_n i els guardem a la variable **delta_X**:

```
typedef struct {
    double *times;
    double **values;
    int d;
    int num_times;
    double **delta_X;
} Data;
```

Si el nostre objectiu és calcular la signatura $S^\mu(X)_{t_m, t_n}|_k$ per cada paraula w creem un objecte **Word**. Fixat un cert t_m , tenim en compte els vectors $c_w := (b_0, \dots, b_{n-m})$ i $v_w := (s_0, \dots, s_{n-m})$ que, per cada element $S^\mu(X)_{t_m, t_r}^w$, $r \in \{0, \dots, n-m\}$, guarden si aquest element ha estat calculat i el seu valor. Dins el nostre **Word**, aquests es guarden a **is_calculated** i **value**. Aquest struct també desa els índexs que componen la paraula per poder fer crides recursives durant el càlcul:

```

typedef struct{
    Index * indexes;
    int length,prev_word_index,num_times;
    double *value;
    bool *is_calculated;
} Word;

```

L'struct `Signature` desa una instància de `Data` i totes les paraules, `Words`, de fins a nivell k que es corresponen amb els elements de la signatura. Aquest struct també guarda els valors $e^{-\mu(t_n-t_{n-1})}$ per cada $t_n \in T$ a la variable `delta_mu`:

```

typedef struct {
    int k,d,num_combinations;
    Index *possible_indexes;
    Data *data;
    double mu;
    double *delta_mu;
    Word *words;
} Signature;

```

La funció `calculate_signature()` encapsula el càlcul de la signatura. Per cada paraula, crida la funció assistent `signature_helper()` que retorna el valor $S^\mu(X)_{t_m,t_n}^w$:

```

double* calculate_signature(Signature *signature, int m, int n) {
    for (int i = 0; i < signature->num_combinations; i++) {
        reset_is_calculated(&signature->words[i]);
    }
    double *sig = malloc(signature->num_combinations * sizeof(double));
    for(int i = 0; i<signature->num_combinations;i++){
        sig[i] = signature_helper(signature,n,m,&signature->words[i]);
    }
    return sig;
}

```

Finalment, la funció `signature_helper()` comprova si l'element $S^\mu(X)_{t_m,t_n}^w$ ha estat calculat. En cas que ho estigui, el retorna. En cas que no, es crida a ella mateixa de forma recursiva i implementa el càlcul de la definició de signatura discreta:

```

double signature_helper(Signature *signature, int n,int m,Word *word){

```

```

if(word->length == 0){ return 1; }
if(word->is_calculated[n]){ return word->value[n];}
if (n==m) {
    word->value[n] = 0;
    word->is_calculated[n] = true;
    return 0;}
if (word->indexes[word->length - 1].head) {
    word->value[n] = signature->delta_mu[n-1] *
    (signature_helper(signature,n-1,m,word) +
    signature->data->delta_X[n-1][word->indexes[word->length - 1].i]*
    signature_helper(signature,n-1,m,
        &signature->words[word->prev_word_index]));
} else {
    word->value[n] = signature->delta_mu[n-1] *
    signature_helper(signature,n-1,m,word) +
    signature->data->delta_X[n-1][word->indexes[word->length - 1].i]*
    signature_helper(signature,n,m,
        &signature->words[word->prev_word_index]);
}
word->is_calculated[n] = true;
return word->value[n];
}

```

4 Signatura com a vector de característiques

Com ja hem esbossat a la introducció, un dels principals interessos de la signatura és la seva utilitació com a vector de característiques en un model de classificació. Aquest mètode pot ser aplicat amb facilitat a qualsevol conjunt de dades seqüencial. En aquest apartat mostrarem un exemple d'ús proporcionat a la secció 2.2.1 de l'article de Chevyrev, I. i Kormilitzin, A. i el reproduïrem fent servir la nostra implementació de la signatura discreta amb la incrustació head-tail, en lloc de la incrustació lead-lag.

L'objectiu és entrenar un classificador capaç de discriminar entre dos tipus de sèries temporals univariants, simulades de forma sintètica fent ús del model $ARMA(p, q)$. Considrem una col·lecció d'observacions de sèries temporals $\{Y_{i,j}\}$ on i indica la sèrie distinta i j indica la mesura a temps t_j .

El model $ARMA(p, q)$ és un model de mitjana mòbil integrada autoregressiva (**A**uto**R**egressive **M**oving **A**verage) amb paràmetres p (l'ordre del model autoregressiu i q (l'ordre del model de mitjana mòbil). Aquest és una superposició dels models autoregressiu $AR(p)$ i de mitjana mòbil $MA(q)$:

$$AR(p) : Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t, \quad (4.1)$$

$$MA(q) : Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}, \quad (4.2)$$

on μ és la mitjana de la sèrie temporal, θ_i i ϕ_i són els paràmetres del model i ϵ_{t-i} els termes d'error de soroll blanc. En aquesta secció no pretenem indagar en la base matemàtica del model sinó centrar-nos en la seva aplicació pràctica. Per mostrar com funciona l'extracció del vector de característiques pot funcionar en aquest cas, plantegem un problema de classificació binària fent servir el model $ARMA$.

Generem una mostra de 1000 sèries temporals (500 de cada tipus) de mida 100. En aquest exemple en particular, les dues classes corresponen als models:

$$classe "0" : Y_t - 0.4Y_{t-1} = 0.5 + \epsilon_t + 0.5\epsilon_{t-1}, \quad (4.3)$$

$$classe "1" : Y_t - 0.8Y_{t-1} = 0.5 + \epsilon_t + 0.7\epsilon_{t-1}, \quad (4.4)$$

on ϵ_t està generat amb una distribució normal de mitjana 0 i variància 1. Podem

veure un exemple de cada classe a la Figura 4.

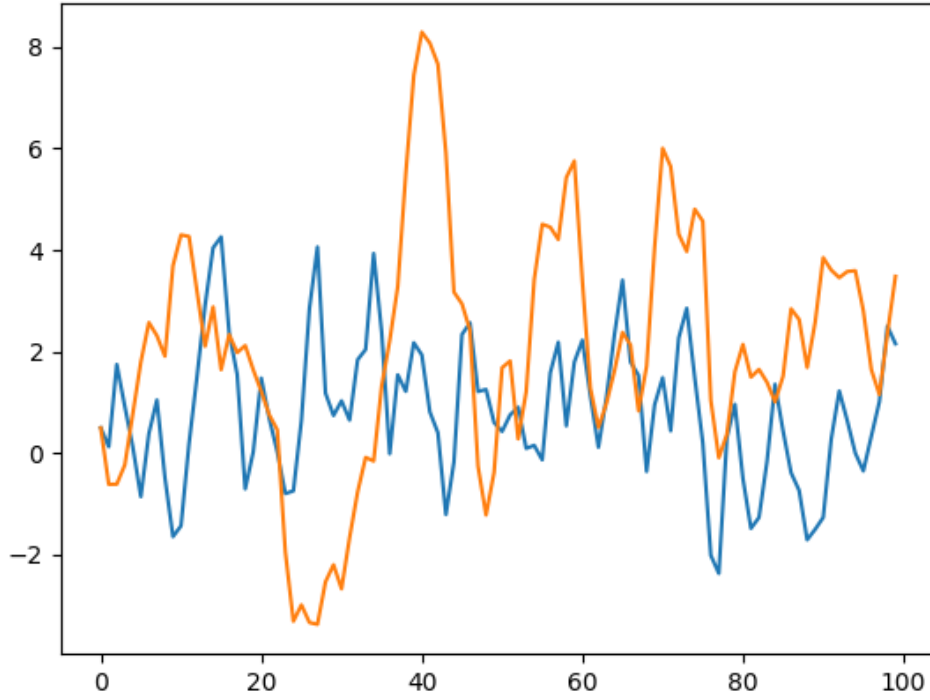


Figura 4: Exemples de sèries temporals de cada classe

De forma intuïtiva podem organitzar les sèries temporals de cada classe en una matriu d'un estil semblant al de la taula 3

Remarquem els passos fonamentals del procediment proposat:

- Crear un camí continu X_i per cada sèrie temporal $\{Y_i\}$.
- Computar la signatura truncada (nosaltres farem servir la signatura discreta plana) $S(X_i)|_k$ del camí a nivell k .
- Fer servir els elements de la signatura $\{S(X_i)^w\}$ com a característiques.

A l'article de Chevyrev [2] fan servir la transformació de suma cummulativa i lead-lag. Nosaltres, farem directament interpolació lineal i head-tail per trobar la signatura discreta plana de cada camí:

t_1	t_2	t_3	...	t_{100}	Classe
$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	...	$Y_{1,100}$	0
$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	...	$Y_{2,100}$	0
$Y_{3,1}$	$Y_{3,2}$	$Y_{3,3}$	...	$Y_{3,100}$	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$Y_{500,1}$	$Y_{500,2}$	$Y_{500,3}$	...	$Y_{500,100}$	0
$Y_{501,1}$	$Y_{501,2}$	$Y_{501,3}$	...	$Y_{501,100}$	1
$Y_{502,1}$	$Y_{502,2}$	$Y_{1,3}$	...	$Y_{502,100}$	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$Y_{1000,1}$	$Y_{1000,2}$	$Y_{1000,3}$	...	$Y_{1000,100}$	1

Taula 3: Organització de sèries en les classes definides a (4.3) i (4.4)

$$S(X_i)_{t_1, t_{100}}|_k = (1, S(X_i)_{t_1, t_{100}}^1, \dots, S(X_i)_{t_1, t_{100}}^{d^- \dots d^-}), \quad (4.5)$$

Estandaritzem els valors de la signatura (per columnes) i ometem el primer terme. Obtenim una matriu de característiques com la de la taula 4.

4.1 Resultats experimentals

Entrenem un classificador de regressió logística, en particular *LogisticRegression* de la llibreria *sklearn.linear_model* de *Python*. La divisió entre conjunts d'entrenament i avaluació és de 70% i 30% respectivament. A les taules 5 i 6 podem veure les matrius de confusió obtingudes pel model de l'article i a les taules 7 i 8 les obtingudes en el nostre cas.

Cal notar que, si bé els resultats obtinguts en els dos casos són força positius, l'objectiu aquí no és el de trobar el millor model si no el de mostrar com la signatura pot ser fàcilment utilitzada com a vector de característiques. El model utilitzat és més aviat simple, un model més complex proporcionaria precisions fins i tot millors.

Vector de característiques					Classe
S_1^{1+}	S_1^{1-}	S_1^{1+1+}	...	$S_1^{d-...d-}$	0
S_2^{1+}	S_2^{1-}	S_2^{1+1+}	...	$S_2^{d-...d-}$	0
S_3^{1+}	S_3^{1-}	S_3^{1+1+}	...	$S_3^{d-...d-}$	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
S_{500}^{1+}	S_{500}^{1-}	S_{500}^{1+1+}	...	$S_{500}^{d-...d-}$	0
S_{501}^{1+}	S_{501}^{1-}	S_{501}^{1+1+}	...	$S_{501}^{d-...d-}$	1
S_{502}^{1+}	S_{502}^{1-}	S_{502}^{1+1+}	...	$S_{502}^{d-...d-}$	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
S_{1000}^{1+}	S_{1000}^{1-}	S_{1000}^{1+1+}	...	$S_{1000}^{d-...d-}$	1

Taula 4: Organització dels vectors de característiques en classes

Real \ Predit	0	1
	0	1
0	326	20
1	35	319

Taula 5: Dades d'entrenament amb precisió 0.92 [2]

Real \ Predit	0	1
	0	1
0	139	15
1	14	132

Taula 6: Dades d'avaluació amb precisió 0.90 [2]

Real \ Predit	0	1
	0	1
0	314	17
1	41	328

Taula 7: Dades d'entrenament amb precisió 0.92

Real \ Predit	0	1
	0	1
0	162	7
1	14	117

Taula 8: Dades d'avaluació amb precisió 0.92

5 Signatura d'un camí aplicada a distribucions de probabilitat

En aquest apartat ens plantegem adaptar el concepte de signatura discreta d'un camí a la signatura discreta d'una certa distribució de probabilitat fent servir els punts amb probabilitat com a punts del camí. Després de definir aquests conceptes proposem dues aplicacions pràctiques d'aquest concepte de signatura discreta d'una família de probabilitats: (i) trobar els paràmetres òptims que ajustin una mostra a una distribució de probabilitats paramètrica via minimització de la funció d'error i (ii) utilització de la signatura com a vector de característiques d'un model de classificació d'aprenentatge automàtic per discriminar a quina família de distribucions de probabilitat paramètriques pertany una mostra.

En aquesta secció fixem una constant de decadència $\mu = 0$, és a dir, el cas signatura discreta plana. Farem servir els termes signatura discreta i signatura discreta plana de forma indistinta.

Definició 5.1. Sigui $(\Omega, \mathcal{F}, \mathcal{P}_\theta)$ un espai de probabilitats paramètric, amb $\theta = (\theta_1, \dots, \theta_\tau) \in \Theta$, on Θ és l'espai de paràmetres i $\Omega = \{x_1, x_2, \dots\}$ un conjunt discret de punts ordenats de tal manera que $x_1 < x_2 < \dots$

Siguin $f_\theta := f(\theta; x) : \Theta \times \Omega \longrightarrow [0, 1]$ la família de funcions de massa de probabilitat corresponents a $\mathcal{P}$, amb

$$\sum_{x \in \Omega} f_{\theta}(x) = 1. \quad (5.1)$$

Ara, siguin $m, n \in \mathbb{N}$ amb $m < n$. llavors, per cada θ , veiem que de fet, la funció de massa de probabilitat restringida a l'interval $[x_m, x_n]$, $f_{\theta} : [x_m, x_n] \longrightarrow [0, 1] \subset \mathbb{R}$ és un camí discret X_t a $\mathbb{R}$, definit com $X_t(\theta) := f_{\theta}$. Si fixem un Ω , podem considerar trobar la signatura d'aquest camí únicament amb dependència respecte $\theta \in \Theta$.

Per tant, definim la **signatura discreta de la família de probabilitats** $\mathcal{P}$, en funció de θ com

$$S_{\Omega, \mathcal{P}}(\theta)_{x_m, x_n} := S(X_t(\theta))_{x_m, x_n} = S(f_{\theta})_{x_m, x_n}. \quad (5.2)$$

A la pràctica, fixarem un conjunt finit de punts i farem servir la signatura truncada a un cert nivell en lloc de fer servir la signatura completa

Exemple 5.2. Suposem que tenim l'espai de probabilitats paramètric $(\Omega, \mathcal{F}, \mathcal{P}_{\theta})$, amb $\Omega = \mathbb{N} \cup \{0\}$ i $\theta \in \Theta = [0, +\infty)$. Les funcions de massa de probabilitat $f_{\theta} : \Omega \longrightarrow [0, 1]$ són de la forma:

$$f_{\theta}(n) = e^{-\theta n}(1 - e^{-\theta}). \quad (5.3)$$

Veiem que efectivament està ben definida ja que

$$\sum_{n=0}^{\infty} e^{-\theta n}(1 - e^{-\theta}) = 1. \quad (5.4)$$

Considerem el subconjunt $\{0, 1, 2\} \subset \Omega$ i tenim la taula 9 de valors del camí $X_n(\theta) = f_{\theta}(n)$.

Ara podem calcular-ne la signatura, fins a nivell 2, amb $*$ $\in \{-, +\}$ com

$$S(\theta)_{0,2}^{\lambda} = 1, \quad (5.5)$$

$$S(\theta)_{0,2}^{1*} = X_2(\theta) - X_0(\theta) = e^{-2\theta} - e^{-3\theta} + e^{-\theta} - 1, \quad (5.6)$$

t_n	$X_{t_n}(\theta)$
0	$1 - e^{-\theta}$
1	$e^{-\theta} - e^{-2\theta}$
2	$e^{-2\theta} - e^{-3\theta}$

Taula 9: Exemple valors del camí 5.2

$$S(\theta)_{0,2}^{1*1-} = S(\theta)_{0,1}^{1*}(X_2(\theta) - X_1(\theta)) = e^{-5\theta} - 4e^{-4\theta} + 6e^{-3\theta} - 4e^{-2\theta} + e^{-\theta}, \quad (5.7)$$

$$\begin{aligned} S(\theta)_{0,2}^{1*1+} &= S(\theta)_{0,1}^{1*}(X_1(\theta) - X_0(\theta)) + S(\theta)_{0,2}^{1*}(X_2(\theta) - X_1(\theta)) = \\ &= -e^{-8\theta} + 6e^{-7\theta} - 15e^{-6\theta} + 20e^{-5\theta} - 14e^{-4\theta} + 2e^{-3\theta} + 5e^{-2\theta} - 4e^{-\theta} + 1. \end{aligned} \quad (5.8)$$

5.1 Obtenció d'un camí a partir de d'una mostra d'una distribució de probabilitat

Acabem de definir la signatura d'una família de probabilitats paramètrica. Per això necessitem trobar una transformació que, a partir d'una mostra d'observacions, generi un camí continu a partir del qual calcular la signatura. Nosaltres proposem trobar les freqüències relatives.

Suposem que tenim n observacions $\{x_1, x_2, \dots, x_n\}$, $x_i \in \mathbb{R}$, $\forall i \in \{1, \dots, n\}$ d'una variable aleatòria amb una distribució concreta desconeguda. Per això escollim m intervals disjunts $\{I_j\}_{j \leq m}$ arbitraris tals que

$$\bigcup_{j \leq m} I_j = \mathbb{R}$$

i per cada interval un $t_j \in I_j$ arbitrari de tal manera que tenim $T := \{t_1, \dots, t_m\}$. Podem triar els t_j de forma arbitrària, però potser una forma interessant de triar-los es agafant el punt mig de l'interval o la mitjana de les observacions que hi pertanyen.

Ara, considerem crear un camí discret a $\mathbb{R}$, $\tilde{X} : T \longrightarrow \mathbb{R}$ i per cada t_j determinem el valors que pren, $\tilde{X}_{t_j}^1$, en base a les **freqüències relatives**

$$\tilde{X}_{t_j}^1 := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{I_j}(x_i), \quad (5.9)$$

on $\mathbb{1}_I(x)$ és la **funció característica**

$$\begin{aligned} \mathbb{1}_I : \mathbb{R} &\longrightarrow \{0, 1\} \\ \mathbb{1}_I(x) &= \begin{cases} 1 & \text{si } x \in I \\ 0 & \text{si } x \notin I \end{cases} \end{aligned} \quad (5.10)$$

D'aquesta manera hem construït un camí $\tilde{X}$ unidimensional associat a la mostra amb el que calcularem la signatura. Podem fer-lo continu via interpolació lineal.

Exemple 5.3. Suposem que tenim les 10 observacions següents:

-0.22	-0.42	0.48	-0.48	-0.5
0.11	-1.34	-1.50	0	-1.05

Triem $T = \{-1.5, -1.25, -1, -0.75, -0.49, -0.25, 0.01, 0.25\}$ i els intervals $\{(-\inf, -1.25), [-1.25, -1], (-1, -0.75], (-0.75, -0.5], (-0.5, -0.25], (-0.25, 0], (0, 0.25), [0.25, +\inf)\}$.

En calculem les freqüències relatives per crear els punts del camí:

I_j	X_{t_j}
$(-\inf, -1.25)$	0.2
$[-1.25, -1]$	0.1
$(-1, -0.75]$	0
$(-0.75, -0.5]$	0

I_j	X_{t_j}
$(-0.5, -0.25]$	0.3
$(-0.25, 0]$	0.1
$(0, 0.25)$	0.2
$[0.25, +\inf)$	0.1

Després de fer interpolació lineal, obtenim el camí de la figura 5:

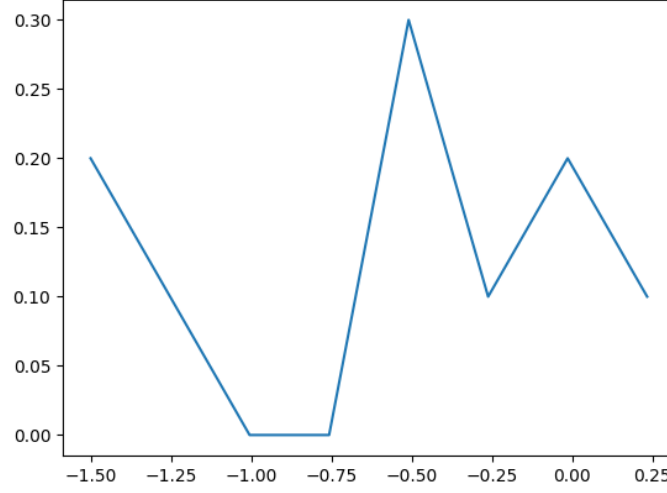


Figura 5: Camí a partir d'una mostra d'observacions

5.2 Ajustament de paràmetres d'una distribució de probabilitat a través de la signatura

Suposem que tenim un camí $\tilde{X} : T \rightarrow \mathbb{R}$ generat a partir d'una mostra d'observacions i volem veure com s'ajusta a una família de distribucions de probabilitat. La idea intuïtiva del mètode que proposem és, a partir d'aquest camí $\tilde{X}$ calcular la signatura discreta $S(\tilde{X})_{a,b}$, $a, b \in T$ i trobar els paràmetres òptims que l'ajustin a la signatura discreta $S_{\Omega, \mathcal{P}}(\theta)_{x_m, x_n}$. Per això definirem la funció **resta**

$$r_w(\theta) = S(\tilde{X})_{a,b}^w - S_{\Omega, \mathcal{P}}(\theta)_{x_m, x_n}^w, \quad (5.11)$$

per cada $w \in W$, on W és un conjunt finit de paraules de l'alfabet $\{1^+, 1^-\}$. A partir d'aquesta funció resta, només ens cal definir una funció d'error, per exemple, l'**error de mínims quadrats**

$$EQM = \frac{1}{N} \sum_{w \in W} r_w(\theta)^2, \quad (5.12)$$

on N és el nombre d'elements de W , i minimitzar-la fent servir alguna tècnica d'optimització. A la fase experimental, nosaltres farem servir l'algoritme de Levenberg-Marquardt [3], que és el que hi ha implementat a la funció `least_squares()` de la llibreria `scipy.optimize` de *Python*.

Nota: És important notar que de les tries de T i Ω que fem en dependrà la velocitat de convergència i la bondat d'ajustament de la mostra a la distribució de probabilitat escollida.

5.2.1 Exemple d'aplicació

En aquesta secció ens proposem aplicar el mètode exposat a l'apartat anterior a un cas real. Per fer-ho, generem una mostra de 10000 elements d'una variable aleatòria que segueix una distribució exponencial de paràmetre θ_0 , generat aleatòriament. Triem un espai d'unitats de temps escollides de uniformement entre 0 i 5 amb un pas de 0.1, $T = \{0, \dots, 5\}$ i trobem un camí $\tilde{X}$ calculat de la forma que hem explicat anteriorment. Prenem $a = 0$, $b = 5$ i $\mu = 0$ per trobar la signatura discreta objectiu $S(\tilde{X})_{0,5}$. La tria de paràmetres és completament arbitrària, però hipotetitzem que en aquest cas és una tria prou convenient en comparació amb d'altres que hem provat. Valors més grans de b , per exemple, implicaria fer molts càlculs amb punts $x \in \Omega$ propers a 0 (veure figura 6, cosa que augmenta el temps d'execució i no 'és útil' en la computació global del procediment).

Per implementar el mètode, triem $\Omega = T = \{x_0, \dots, x_{50}\} = \{0, \dots, 5\}$ i $x_m = 0$, $x_n = 5$ i $W = \{i_1 \dots i_l | i_1, \dots, i_l \in \{1, \dots, k\} \times \{+, -\}\}, k \in \mathbb{N}$, de tal manera que per cada $w \in W$, la funció resta a minimitzar serà $r_w(\theta) = S(\tilde{X})_{0,5}^w - S_\Omega(\theta)_{0,5}^w$. Com ja hem explicat, la nostra implementació es farà en *Python* i amb la funció *least_squares()*.

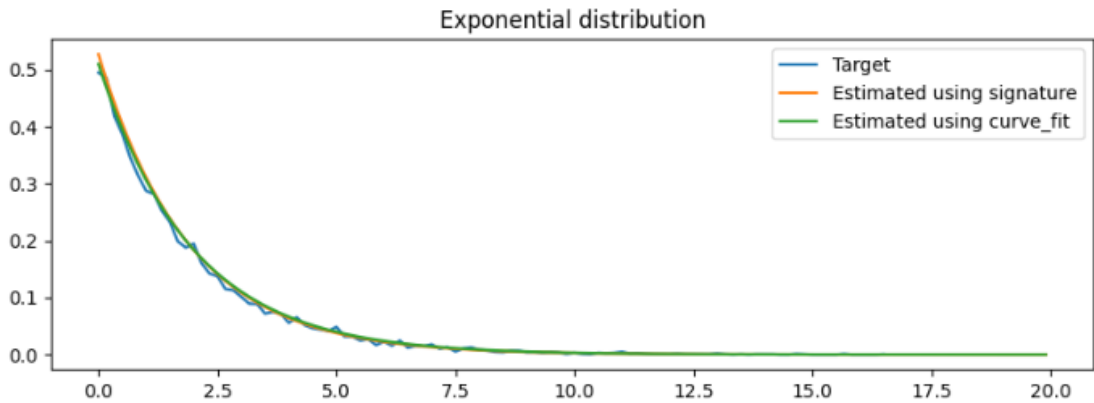


Figura 6: Il·lustració del mètode amb una distribució exponencial de paràmetre $\theta_0 = 1.8945$. En blau: el camí obtingut d'una mostra de. En taronja: la representació gràfica de $f(x) = \theta_1 e^{\theta_1 x}$, on θ_1 és el valor trobat pel procediment que hem descrit. En verd: la representació gràfica de $g(x) = \theta_2 e^{\theta_2 x}$, on θ_2 és el valor trobat per la funció *curve_fit()*.

Hem repetit l'experiment 100 vegades per cada $k \in \{2, 3, 4, 5\}$ i comparat els

resultats en termes de temps i precisió amb la funció *curve_fit()* de la mateixa llibreria *scipy.optimize* de *Python* passant-li com a entrada el camí sencer.

A les taules 10 i 11, comparem els mètodes en funció del valor k de la signatura truncada tot i que aquest valor no influeix en *curve_fit()*. S'expressa d'aquesta manera per denotar que els mètodes utilitzen la mateixa mostra inicial en cada un dels casos.

Per avaluar el resultat obtingut fem servir l'error relatiu. És a dir, si tenim un paràmetre objectiu θ_0 i obtenim un θ_1 resultant del mètode d'ajustament, l'error relatiu que utilitzarem és

$$err = \frac{|\theta_0 - \theta_1|}{|\theta_0|}. \quad (5.13)$$

Error relatiu mitjà

Mètode	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Signatura	0.0673	0.0735	0.0730	0.0721
Curve fit	0.1022	0.1026	0.0998	0.1032

Taula 10: Resultat obtingut en termes d'error relatiu. Per cada valor de k , anomenem error relatiu mitjà a la mitjana dels errors relatius obtinguts en realitzar l'experiment amb aquest k .

Temps mitjà (en segons)

Mètode	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Signatura	0.0190	0.0179	0.0210	0.0197
Curve fit	0.0084	0.0080	0.0068	0.0070

Taula 11: Resultats obtinguts termes de temps d'execució. Per cada valor de k , anomenem temps mitjà a la mitjana dels temps d'execució obtinguts en realitzar l'experiment amb aquest k .

Dels resultats obtinguts podem veure com fent servir la signatura s'obtenen valors d'error lleugerament inferiors als obtinguts mitjançant *curve_fit()*. Pel que

fa al temps d'execució dels dos mètodes, el que nosaltres proposem queda enrere respecte *curve_fit()*.

Si bé es cert que la principal motivació de l'ús de la signatura és el de reduir els temps computacionals i d'execució, en aquest experiment ens proposàvem únicament el d'explorar la viabilitat de la signatura per a determinar paràmetres que ajustin la mostra a una distribució de probabilitats paramètrica. No obstant això, creiem que tant una implementació més eficient del càlcul de la signatura i de les funcions de pre-processament com una tria més encertada de Ω i T podria reduir substancialment els temps d'execució, i fins i tot arribar a batre *curve_fit()*.

5.3 Signatura d'una distribució de probabilitats com a vector de característiques

A la secció 4 d'aquest treball hem mostrat com, de forma senzilla, podem fer servir la signatura d'un camí aplicada com a vector de característiques en un model de classificació. Allà hem generat una mostra de sèries temporals seguint el model $ARMA(1,1)$ amb diferents paràmetres. En aquesta secció, ens plantegem aplicar la signatura de distribucions de probabilitat per discriminar entre mostres de dues distribucions de probabilitat diferents.

De forma experimental, generem un conjunt de dades de 10000 camins etiquetats segons les classes Lognormal i Exponencial. Cada camí ha estat generat a partir d'una mostra de 10000 observacions distribuïdes segons $Exponential(\lambda)$, $\lambda \in [\frac{1}{2}, 2]$ i $Lognormal(0, \sigma)$, $\sigma \in [\frac{1}{4}, 1]$. El procediment per generar cadascun d'aquests camins és el que hi ha explicat a la secció 5.1. A la figura 7 hi ha exemples d'aquests camins.

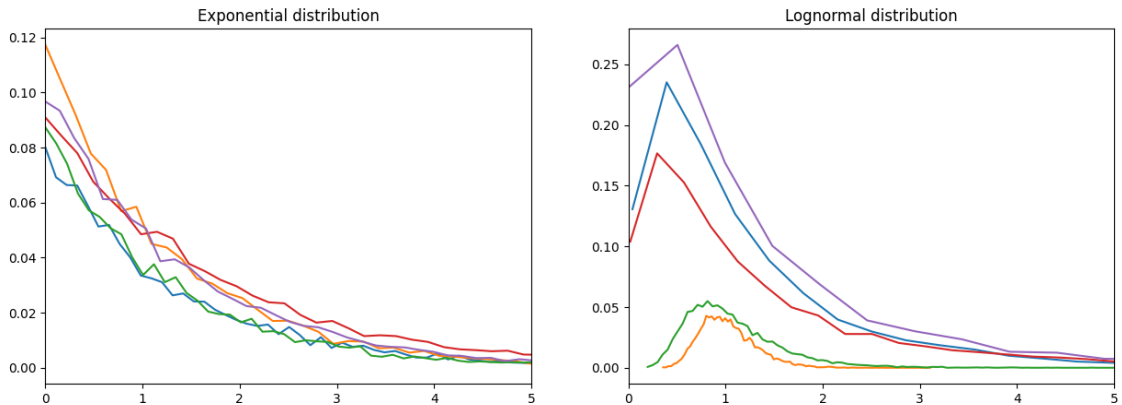


Figura 7: Mostra de distribucions exponencial i lognormal

Un cop generat el dataset, entrenem un model de classificació per a la tasca proposada. A mode de comparació, entrenem models fent servir els valors de les sèries temporals en cru i un altre mètode de reducció de la dimensionalitat de les dades inicials com és l'anàlisi de components principals, PCA (**P**rincipal **C**omponent **A**nalysis). El model és el *RandomForestClassifier* de la llibreria *sklearn.ensemble* de *Python*. Repetim l'experiment 25 vegades per cada valor de k , nivell de la signatura discreta truncada.

Per determinar la validesa dels resultats obtinguts fem servir les mètriques AUC-ROC i AUC-PRC que són comunment utilitzades per avaluar tasques de classificació binàries com aquesta. AUC (**A**rea **U**nder the **C**urve), com el seu nom indica és el valor de l'àrea sota la corba, en aquest cas les corbes ROC i PRC. Incorporem a continuació una breu explicació d'aquestes corbes:

- ROC (**R**eceiver **O**perating **C**haracteristic) és una corba que representa la VPR (raó de veritables positius) en funció de la FPR (raó de falsos negatius). VPR i FPR es defineixen com:

– $VPR = \frac{VP}{VP + FN}$, on VP és el nombre de positius predits correctament i FN el nombre de falsos negatius. És a dir, el denominador és el nombre total de positius reals.

– $FPR = \frac{VN}{VN + FP}$, on VN és el nombre de negatius predits correctament i FP el nombre de falsos positius. És a dir, el denominador és el nombre total de negatius reals.

- PRC (**P**recision-**R**ecall **C**urve) és una corba que representa la precisió (Precision) en funció de l'exhaustivitat (Recall). La precisió P i l'exhaustivitat R es defineixen com:

– $P = \frac{VP}{VP + FP}$, on VP és el nombre de positius predits correctament i FP el nombre de falsos positius.

– $R = \frac{VP}{VP + FN}$, on VP és el nombre de positius predits correctament i FN el nombre de falsos negatius.

La taula de confusió 12 ens permet ajudar a clarificar aquests conceptes.

Fem servir els termes positiu i negatiu per designar la pertinença o no a una

Real \ Predit	Positiu	Negatiu
	VP	FN
Positiu	VP	FN
Negatiu	FP	VN

Taula 12: Taula de confusió

classe arbitrària. En el nostre cas triem la classe 'Exponencial'. És a dir, un element serà un fals positiu si és 'Lognormal' però ha estat etiquetat com a 'Exponencial'.

Ara sí, un cop introduïts els conceptes que fem servir per avaluar la tasca de classificació, a la taula 13 bolquem els resultats obtinguts en termes de mètriques d'avaluació i temps d'execució. Pel que fa a PCA, hem fet servir una mida de vector igual a la signatura per cada nivell k .

Mètriques

Mètode	$k = 2$		$k = 3$		$k = 4$	
	AUCROC	AUCPRC	AUCROC	AUCPRC	AUCROC	AUCPRC
En cru	1.0	0.4421	1.0	0.4417	1.0	0.4418
Signatura	1.0	0.4869	1.0	0.4922	1.0	0.4929
PCA	1.0	0.4745	1.0	0.4458	1.0	0.4203

Temps mitjà (en segons)

Mètode	$k = 2$	$k = 3$	$k = 4$
En cru	4.9345	4.8753	5.0228
Signatura	0.9306	0.9306	2.0488
PCA	1.4334	2.4721	4.6264

Taula 13: Resultats obtinguts en la classificació de distribucions

Dels resultats obtinguts, veiem com en termes d'ajustament de la classificació,

els tres datasets han generat valors similars per tots els valors de k , obtenint un $\text{AUCROC} = 1$. En un cas real, evidentment no tindria sentit explorar mides de vector més grans que els de $k = 2$ si amb un de tant petit s'obtenen resultats tant positius. En aquest experiment però, el que preteniem visualitzar és com, fent servir la signatura, podem obtenir resultats de classificació igual de bons però amb temps d'execució molt inferiors, cosa que és de vital importància quan treballem amb datasets molt més grans i amb moltes més dimensions.

Nota: En aquesta comparació només estem tenint en compte temps d'entrenament dels models. Els temps de pre-processament de les dades han quedat fora de la taula.

6 Conclusions

Al llarg d'aquest treball hem pogut veure les diverses propietats i aplicacions que presenta la signatura d'un camí, sobretot aplicada a un flux de dades.

Hem comprovat com podem implementar un algorisme de càlcul per trobar-la de forma senzilla i com també pot ser utilitzada com a vector de característiques d'un model d'aprenentatge automàtic. A la secció 4 hem pogut veure com un classificador binari és capaç de discriminar entre dues mostres de models *ARMA* diferents. De la mateixa manera, a la secció 5.3 hem pogut veure que la signatura pot ser utilitzada per un classificador binari que discrimini entre mostres d'observacions de dues distribucions de probabilitat diferents i assolir bons resultats en termes de temps i de precisió en comparació amb altres vectors d'entrada.

També hem pogut comprovar com, a través del càlcul de la signatura podem trobar els paràmetres òptims que ajusten millor una mostra a una determinada distribució de probabilitat. Els resultats però, en termes d'eficiència i velocitat de càlcul no han estat tan bons com s'esperava. Caldria determinar com i en quins escenaris aquest càlcul es pot ser més eficient.

Inicialment aquest treball també plantejava l'utilització de dades del món real i realitzar experiments que impliquessin l'ús de la signatura. En particular, volíem utilitzar dades financeres, d'accions i adaptar l'experiment d'Adachi [1]. Aquestes dades, però, al nivell de detall necessari per a que tingui sentit un experiment d'aquestes característiques, són difícils d'obtenir i costoses de processar. Finalment, doncs no ha estat possible dur a terme part dels objectius formulats en un principi.

Referències

- [1] Adachi, T., & Naritomi, Y. (2022). *Discrete signature and its application to finance*.
- [2] Chevyrev, I., & Kormilitzin, A. (2016). *A primer on the signature method in machine learning*.
- [3] More, J. J. (1977). *The Levenberg-Marquardt Algorithm: Implementation and Theory*.
- [4] Cots, D. (2023). *El mètode de la signatura i aplicacions a dades de tipus financer*.
- [5] Lajos, G. G. (2014). *Extracting information from the signature of a financial data stream*.
- [6] Levin, D. (2016) *Learning from the past, predicting the statistics for the future, learning an evolving system*.
- [7] Pérez, I. (2020) *Sig-SDEs model for quantitative finance*.
- [8] K. W. Vugrin (2007) *Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies*