Recent Advances in Statistics: Theory of Rough Paths

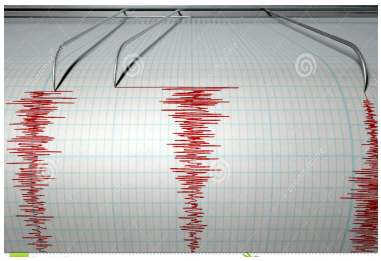
Tessy Papavasiliou

Department of Statistics, University of Warwick

Some Examples of Real Data

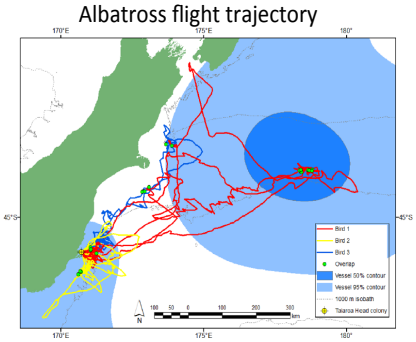
Internet traffic data

Seismic data



Financial data

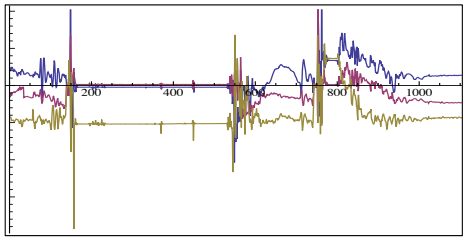
Albatross flight trajectory



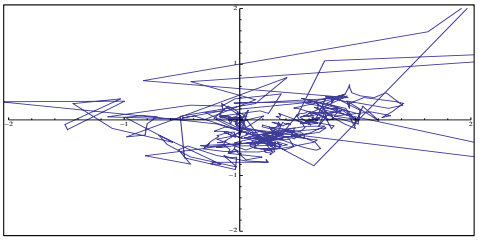
- A path is 'rough' if change is space is not proportional to change in time.

Why does roughness matter?

A home-made earthquake



Two of the signals plotted as a path in 2d



- To understand the effects of such data to systems, we need to make sense of differential equations

$$dY_t = f(Y_t) \cdot dX_t,$$

where X is rough.

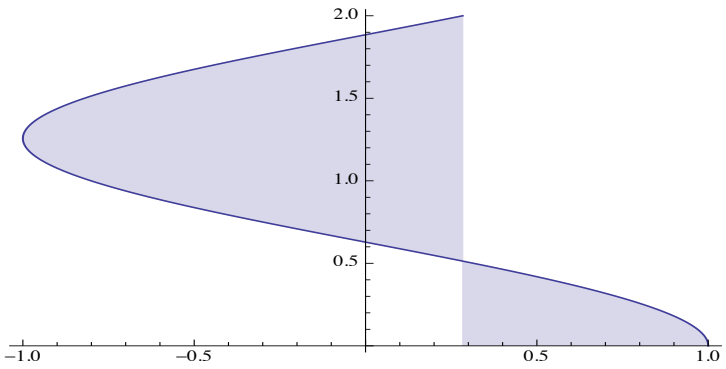
- In order to make sense of such equations, we need to make sense of integrals

$$\int_s^t f(X_u) \cdot dX_u.$$

Integrals and Areas

What is the meaning of

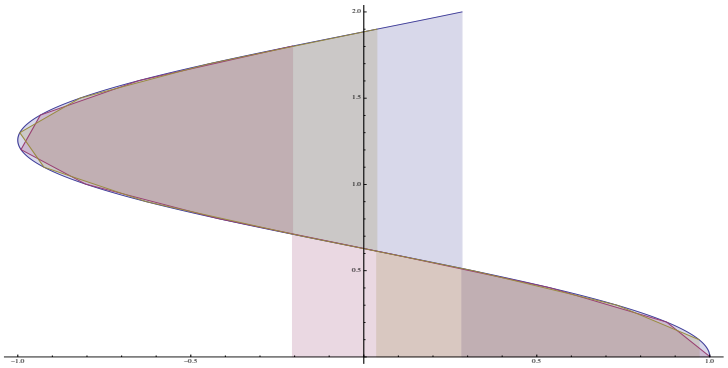
$$\int_s^t Z_u dX_u?$$



Integrals and Areas

What is the meaning of

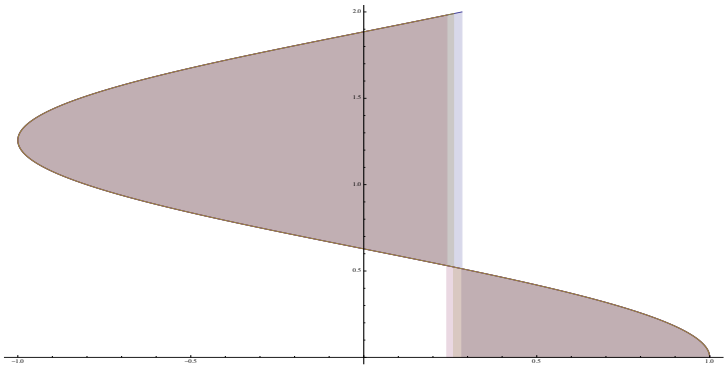
$$\int_s^t Z_u dX_u?$$



Integrals and Areas

What is the meaning of

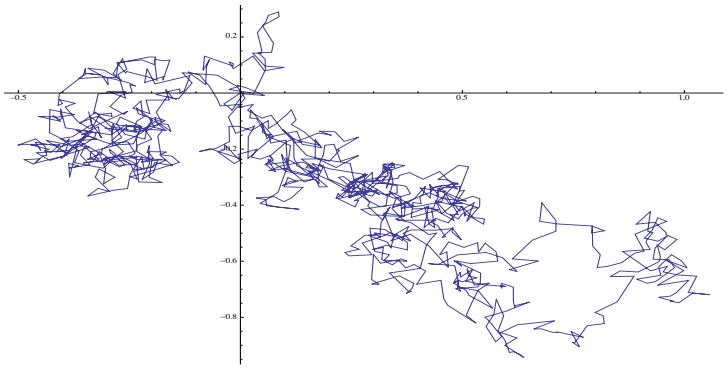
$$\int_s^t Z_u dX_u?$$



Integrals and Areas

What is the meaning of

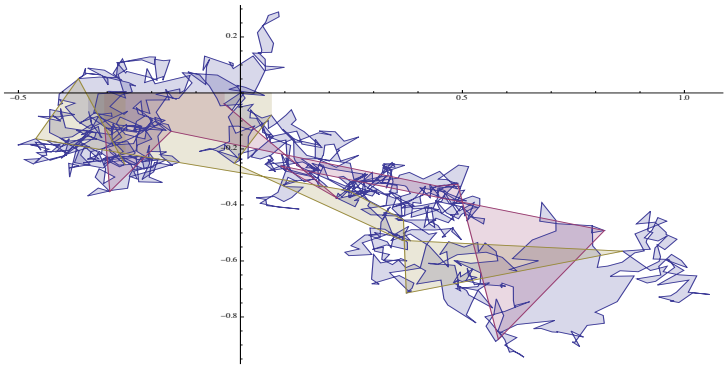
$$\int_s^t Z_u dX_u?$$



Integrals and Areas

What is the meaning of

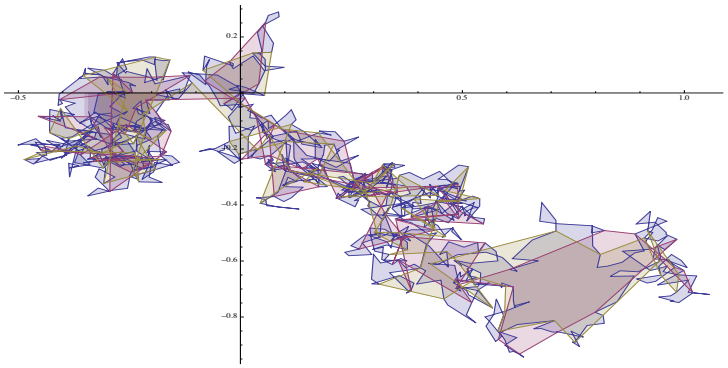
$$\int_s^t Z_u dX_u?$$



Integrals and Areas

What is the meaning of

$$\int_s^t Z_u dX_u?$$



Rough Paths

- When a path is rough, the integrals are not uniquely defined.

Rough Paths

- When a path is rough, the integrals are not uniquely defined.
- The first result of the theory of rough paths was to identify the minimum number of integrals needed to identify the rest: for a path of 'roughness p', any integral

$$\int_{0,T} f(X_u) \cdot dX_u$$

is uniquely defined, once we fix

$$\left(\int_s^t dX_u, \int_{s<u_1<u_2<t} dX_{u_1} dX_{u_2}, \ldots, \int_{s<u_1<\cdots<u_p<t} dX_{u_1} \ldots dX_{u_p} \right).$$

(Terry Lyons, 1998).

Rough Paths

- When a path is rough, the integrals are not uniquely defined.
- The first result of the theory of rough paths was to identify the minimum number of integrals needed to identify the rest: for a path of 'roughness p', any integral

$$\int_{0,T} f(X_u) \cdot dX_u$$

is uniquely defined, once we fix

$$\left(\int_s^t dX_u, \int_{s<u_1<u_2<t} dX_{u_1} dX_{u_2}, \ldots, \int_{s<u_1<\cdots<u_p<t} dX_{u_1} \ldots dX_{u_p} \right).$$

(Terry Lyons, 1998).

- Extension to systems that involve changes in time and space: theory of regularity structures (Martin Hairer, Fields Medal 2014).

Signature of a Path

- The ‘iterated integrals’ provide the basic blocks for understanding any system driven by any path.

Signature of a Path

- The 'iterated integrals' provide the basic blocks for understanding any system driven by any path.
- All these integrals together (the 'signature' of the path) provide an alternative way of describing the path. (Boedihardjo et al, 2016)

Signature of a Path

- The 'iterated integrals' provide the basic blocks for understanding any system driven by any path.
- All these integrals together (the 'signature' of the path) provide an alternative way of describing the path. (Boedihardjo et al, 2016)
- In many cases, by replacing a data stream (path) by the corresponding signature, it is possible to capture information more efficiently.

Signature of a Path

- The 'iterated integrals' provide the basic blocks for understanding any system driven by any path.
- All these integrals together (the 'signature' of the path) provide an alternative way of describing the path. (Boedihardjo et al, 2016)
- In many cases, by replacing a data stream (path) by the corresponding signature, it is possible to capture information more efficiently.
- First application: Chinese Handwriting Recognition (Ben Graham, winner of 2013 competition).

Capturing Sound

Warwick Statistics

The Signature of Sound

WARWICK THE UNIVERSITY OF WARWICK

Introduction

In modern society, the rise of digital communications has replaced many traditional technologies e.g.

Phone calls: replaced with VOIP calls

Books: replaced with recorded audiobooks

Aim to find a more data efficient way to store and transmit spoken audio.

Signature and Area

For a path $X : [s,t] \to \mathbb{R}^d$, $d \geq 2$, the N-step truncated signature of the path is given by

$$S^{(N)}_{[s,t]}(X) = \left(1, X^1_{[s,t]}, X^2_{[s,t]}, \dots, X^N_{[s,t]}\right)$$

where

$$X^n_{[s,t]} = \int\limits_{\substack{u_1 \leq \cdots \leq u_n \\ u_1,\dots,u_n \in [s,t]}} \mathrm{d}X_{u_1} \otimes \cdots \otimes \mathrm{d}X_{u_n}.$$

The final element of $S^{(2)}_{[s,t]}(X)$ can be expressed element wise as a sum

$$X^{(i,j)}_{[s,t]} = A^{(i,j)}_{[s,t]} + B^{(i,j)}_{[s,t]},$$

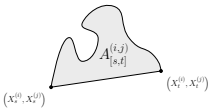
where

$$A^{(i,j)}_{[s,t]} = \frac{1}{2} \int\limits_{\substack{u_1 \leq u_2 \\ u_1,u_2 \in [s_1,t_1]}} \mathrm{d}X^{(i)}_{u_1} \mathrm{d}X^{(j)}_{u_2} - \mathrm{d}X^{(j)}_{u_1} \mathrm{d}X^{(i)}_{u_2},$$

and

$$B^{(i,j)}_{[s,t]} = \frac{1}{2}\left(X^{(i)}_t - X^{(i)}_s\right)\left(X^{(j)}_t - X^{(j)}_s\right),$$

for $i,j = 1,\dots,d$, and $X^{(i)}_t$ is the value of the i-th coordinate of the path at time t.
$A^{(i,j)}_{[s,t]}$ is the area enclosed by the the planar curve $\left(X^{(i)}_u, X^{(j)}_u\right)$ for $u \in [s,t]$, and the chord from $\left(X^{(i)}_s, X^{(j)}_s\right)$ to $\left(X^{(i)}_t, X^{(j)}_t\right)$, as shown below.

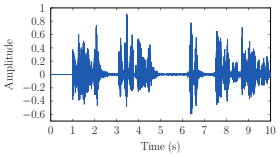


Extending a Path

The signature of a path $X : [0,T] \to \mathbb{R}$, is defined for $d \geq 2$. Define a new path $\widetilde{X} : [0,T] \to \mathbb{R}^{2d}$, fix $\varepsilon > 0$, define $\widetilde{X}$ by $t \mapsto (X_{t-\varepsilon}, X_t)$. Using this, a 1-$D$ path becomes a 2-D path.

Adaptive Sampling

In spoken audio, there are many periods of no sound - as seen in the clip of an audiobook below.

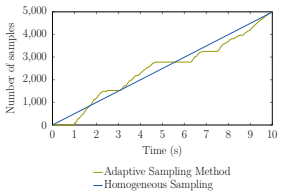


When a sound is homogeneously sampled, data is recorded when no sound is produced. By not recording data when no sound is produced, data is saved. Instead of sampling homogeneously, sample at times $\{\tau_i\}$, where

$$\tau_{i+1} = \min\left\{\tau : \left|A^{(1,2)}_{[\tau_i,\tau]}\right| \geq K\right\},$$

and $\tau_0 = 0$, for some threshold $K > 0$, where $A^{(1,2)}_{[\tau_i,\tau]}$ is the area of the extended path, for fixed ε.

A comparison of the methods when applied to the audiobook clip is shown below. When no sound is produced, no samples are taken; when sound is produced, samples are taken more frequently.



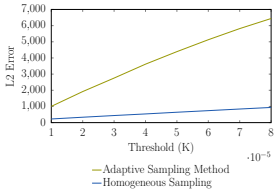
Error Quantification

Compare difference between path from adaptive sampling and homogeneous sampling. Carried out in two ways on a 56 second test clip.

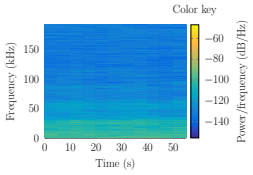
L^2 norm: norm of difference between paths

Spectrogram: compare the spectrograms - based on wavelets

For 8 different thresholds, $K \times 10^{-5} = 1,\dots,8$, and fixed ε, the error in the L^2 sense is shown below for both adaptive and homogeneous samples.



To equalise data amounts, if there are N adaptive samples, $2N$ homogeneous samples are used.



The above spectrogram shows the difference between the original path and the linear interpolation of the adaptively sampled path. A lower valued colour indicates less loss of data in that frequency range at that time.

Future Work

- Find error based on spectrogram
- Reconstruct path using knowledge of K
- Optimise choice of ε and K
- Sample stereo data with and without delay
- Implement method in ADC for use when recording sound

References

[1] T. J. Lyons, M. J. Caruana, and T. Lévy.
Differential Equations Driven by Rough Paths: Ecole d'Eté de Probabilités de Saint-Flour XXXIV-2004.
Springer, April 2007.

[2] P. K. Friz and N. B. Victoir.
Multidimensional Stochastic Processes as Rough Paths: Theory and Applications.
Cambridge University Press, February 2010.

Arabic Handwriting Recognition

- Different challenge than Chinese: sequence of strokes matters.

Arabic Handwriting Recognition

- Different challenge than Chinese: sequence of strokes matters.
- By including the signature as a feature of the data, Daniel achieved 92.5% recognition, which is an improvement to the state of the art (D Wilson-Nunn et al, in I*EEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 2018.)

RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.

RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.
- **Hypothesis:** Random changes in the RNA (editing) create variability, making cells behave differently, thus allowing an individual to adapt to change in the same way that DNA mutations help species adapt to change.

RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.
- **Hypothesis:** Random changes in the RNA (editing) create variability, making cells behave differently, thus allowing an individual to adapt to change in the same way that DNA mutations help species adapt to change.
- **Experiment:** We are given RNA molecules from a large number of genetically identical cells (same DNA) of an individual organism, subject to change. We do not know which cell each RNA molecule comes from except for a small number of cells (single cell experiments).

RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.
- **Hypothesis:** Random changes in the RNA (editing) create variability, making cells behave differently, thus allowing an individual to adapt to change in the same way that DNA mutations help species adapt to change.
- **Experiment:** We are given RNA molecules from a large number of genetically identical cells (same DNA) of an individual organism, subject to change. We do not know which cell each RNA molecule comes from except for a small number of cells (single cell experiments).
- **Question:** Are cells different based on the percentage of editing of their RNA molecules?

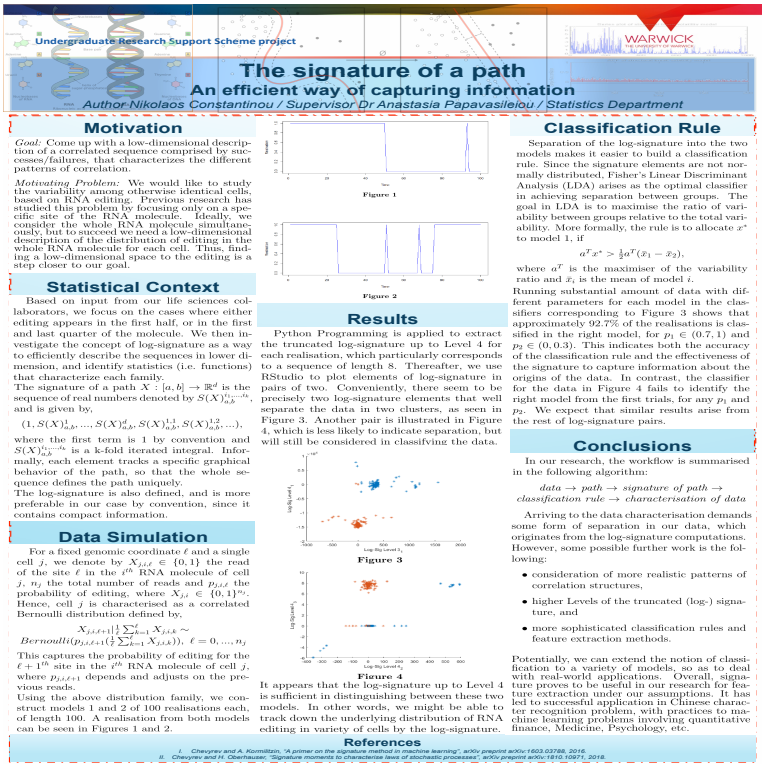
RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.
- **Hypothesis:** Random changes in the RNA (editing) create variability, making cells behave differently, thus allowing an individual to adapt to change in the same way that DNA mutations help species adapt to change.
- **Experiment:** We are given RNA molecules from a large number of genetically identical cells (same DNA) of an individual organism, subject to change. We do not know which cell each RNA molecule comes from except for a small number of cells (single cell experiments).
- **Question:** Are cells different based on the percentage of editing of their RNA molecules?
- **Statistical Challenge:** We need to use single cell data to infer variability in RNA editing among cells.

RNA editing - an adapting mechanism

- **Main Dogma:** DNA to RNA to protein expression.
- **Hypothesis:** Random changes in the RNA (editing) create variability, making cells behave differently, thus allowing an individual to adapt to change in the same way that DNA mutations help species adapt to change.
- **Experiment:** We are given RNA molecules from a large number of genetically identical cells (same DNA) of an individual organism, subject to change. We do not know which cell each RNA molecule comes from except for a small number of cells (single cell experiments).
- **Question:** Are cells different based on the percentage of editing of their RNA molecules?
- **Statistical Challenge:** We need to use single cell data to infer variability in RNA editing among cells.
- **Conclusion:** RNA editing on specific sites varies sagnificantly from cell to cell (*Nature Communications*, Harjanto et al - in collaboration with Immune Diversity group in DKFZ, Heidelberg).

The Signature of an RNA molecule



Undergraduate Research Support Scheme project

WARWICK

The signature of a path

An efficient way of capturing information

Author Nikolaos Constantinou / Supervisor Dr Anastasia Papavasileiou / Statistics Department

Motivation

Goal: Come up with a low-dimensional description of a correlated sequence comprised by successes/failures, that characterizes the different patterns of correlation.

Motivating Problem: We would like to study the variability among otherwise identical cells, based on RNA editing. Previous research has studied this problem by focusing only on a specific site of the RNA molecule. Ideally, we consider the whole RNA molecule simultaneously, but to succeed we need a low-dimensional description of the distribution of editing in the whole RNA molecule for each cell. Thus, finding a low-dimensional space to the editing is a step closer to our goal.

Statistical Context

Based on input from our life sciences collaborators, we focus on the cases where either editing appears in the first half, or in the first and last quarter of the molecule. We then investigate the concept of log-signature as a way to efficiently describe the sequences in lower dimension, and identify statistics (i.e. functions) that characterize each family.

The signature of a path $X : [a, b] \to \mathbb{R}^d$ is the sequence of real numbers denoted by $S(X)^{i_1,\dots,i_k}_{a,b}$, and is given by,

$$(1, S(X)^1_{a,b}, \dots, S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \dots),$$

where the first term is 1 by convention and $S(X)^{i_1,\dots,i_k}_{a,b}$ is a k-fold iterated integral. Informally, each element tracks a specific graphical behavior of the path, so that the whole sequence defines the path uniquely.

The log-signature is also defined, and is more preferable in our case by convention, since it contains compact information.

Data Simulation

For a fixed genomic coordinate ℓ and a single cell j, we denote by $X_{j,i,\ell} \in \{0, 1\}$ the read of the site ℓ in the i^{th} RNA molecule of cell j, n_j the total number of reads and $p_{j,i,\ell}$ the probability of editing, where $X_{j,i} \in \{0, 1\}^{n_j}$. Hence, cell j is characterised as a correlated Bernoulli distribution defined by,

$$X_{j,i,\ell+1}|\tfrac{1}{\ell}\textstyle\sum_{k=1}^{\ell} X_{j,i,k} \sim Bernoulli(p_{j,i,\ell+1}(\tfrac{1}{\ell}\textstyle\sum_{k=1}^{\ell} X_{j,i,k})),\ \ell = 0, \dots, n_j$$

This captures the probability of editing for the $\ell+1^{th}$ site in the i^{th} RNA molecule of cell j, where $p_{j,i,\ell+1}$ depends and adjusts on the previous reads.

Using the above distribution family, we construct models 1 and 2 of 100 realisations each, of length 100. A realisation from both models can be seen in Figures 1 and 2.

Figure 1

Figure 2

Results

Python Programming is applied to extract the truncated log-signature up to Level 4 for each realisation, which particularly corresponds to a sequence of length 8. Thereafter, we use RStudio to plot elements of log-signature in pairs of two. Conveniently, there seem to be precisely two log-signature elements that well separate the data in two clusters, as seen in Figure 3. Another pair is illustrated in Figure 4, which is less likely to indicate separation, but will still be considered in classifying the data.

Figure 3

Figure 4

It appears that the log-signature up to Level 4 is sufficient in distinguishing between these two models. In other words, we might be able to track down the underlying distribution of RNA editing in variety of cells by the log-signature.

Classification Rule

Separation of the log-signature into the two models makes it easier to build a classification rule. Since the signature elements are not normally distributed, Fisher's Linear Discriminant Analysis (LDA) arises as the optimal classifier in achieving separation between groups. The goal in LDA is to maximise the ratio of variability between groups relative to the total variability. More formally, the rule is to allocate x^* to model 1, if

$$a^T x^* > \tfrac{1}{2} a^T (\bar{x}_1 - \bar{x}_2),$$

where a^T is the maximiser of the variability ratio and $\bar{x}_i$ is the mean of model i.

Running substantial amount of data with different parameters for each model in the classifiers corresponding to Figure 3 shows that approximately 92.7% of the realisations is classified in the right model, for $p_1 \in (0.7, 1)$ and $p_2 \in (0, 0.3)$. This indicates both the accuracy of the classification rule and the effectiveness of the signature to capture information about the origins of the data. In contrast, the classifier for the data in Figure 4 fails to identify the right model from the first trials, for any p_1 and p_2. We expect that similar results arise from the rest of log-signature pairs.

Conclusions

In our research, the workflow is summarised in the following algorithm:

data $\to$ *path* $\to$ *signature of path* $\to$ *classification rule* $\to$ *characterisation of data*

Arriving to the data characterisation demands some form of separation in our data, which originates from the log-signature computations. However, some possible further work is the following:

- consideration of more realistic patterns of correlation structures,
- higher Levels of the truncated (log-) signature, and
- more sophisticated classification rules and feature extraction methods.

Potentially, we can extend the notion of classification to a variety of models, so as to deal with real-world applications. Overall, signature proves to be useful in our research for feature extraction under our assumptions. It has led to successful application in Chinese character recognition problem, with practices to machine learning problems involving quantitative finance, Medicine, Psychology, etc.

References

I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning", arXiv preprint arXiv:1603.03788, 2016.
II. Chevyrev and H. Oberhauser, "Signature moments to characterise laws of stochastic processes", arXiv preprint arXiv:1810.10971, 2018.

3rd year MORSE student.