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 in space is not proportional to change in time.
Why does roughness matter?

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

$$\int_{s}^{t} Z_\nu dX_\nu?$$
Integrals and Areas

What is the meaning of

$$\int_s^t Z \, dX_u?$$
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.

(From Terry Lyons, 1998.)

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

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

(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

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

(Terry Lyons, 1998).

- Extension to systems that involve changes in time and space: theory of regularity structures (Martin Hairer, Fields Medal 2014).
The ‘iterated integrals’ provide the basic blocks for understanding any system driven by any path.

First application: Chinese Handwriting Recognition (Ben Graham, winner of 2013 competition).
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

The Signature of Sound

Introduction

In modern society, the rise of digital communications has replaced many traditional technologies e.g. Phone calls replaced with VXML 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 \( X : \mathbb{R}^d \to \mathbb{R}^n \), the N-step truncated signature of the path is given by

\[ S_N^{(n)}(X) = \left\{ X_0, X_1, X_2, \ldots, X_N \right\} \]

where

\[ X_{n+j} = \int_{t=n}^{n+j} \alpha_t \, d\mathbf{x}, \]

The final element of \( S_N^{(n)}(X) \) can be expressed element wise as a sum

\[ X_{n+j} = X_j \quad \text{for} \quad j = 1, \ldots, n, \]

and

\[ X_{n+j} = \frac{1}{n} \left( X_j - X_0 \right) \quad \text{for} \quad j = n+1, \ldots, N, \]

for \( j = 1, \ldots, d \) and \( X_j^{(n)} \) is the value of the \( j \)-th coordinate of the path at time \( t \).

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, samples at times \( t_i \), where

\[ t_i = \min \left\{ t \mid |A(t)| > K \right\}, \]

and \( t_1 = 0 \), for some threshold \( K > 0 \), where \( A(t) \) is the area of the extended path, for fixed \( s \).

Extending a Path

The signature of a path \( X : [0, T] \to \mathbb{R}^n \), is defined for \( d \geq 2 \). Define a new path \( \tilde{X} : [0, T] \to \mathbb{R}^n \), for \( s > 0 \), define \( \tilde{X} \) by

\[ t \mapsto (sX(t), X(t)), \]

Using this, a 2-D path becomes a 3-D path.

Error Quantification

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.

To evaluate data amounts, if there are \( N \) adaptive samples, \( N \) homogeneously 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 \( K \)
- Optimize choice of \( s \) and \( K \)
- Sample stereo data with and without delay
- Implement method in ADC for use when recording sound

References

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 *IEEE 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?

**Statistical Challenge:** We need to use single cell data to infer variability in RNA editing among cells.

**Conclusion:** RNA editing on specific sites varies significantly from cell to cell (Nature Communications, Harjanto et al. in collaboration with Immune Diversity group in DKFZ, Heidelberg).
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 significantly from cell to cell (*Nature Communications*, Harjanto et al - in collaboration with Immune Diversity group in DKFZ, Heidelberg).
The Signature of an RNA molecule

Statistical Context

Based on inputs from the literature and observations, we focus on the case where habit appears in the first half, in the fourth and sixth quarter of the trajectory.

Results

Classification Rule

\[ \alpha(x) = \frac{1}{2} (x + \lambda) \]

where \( \lambda \) is the minimum of the random variable \( x \) in the range of model 1.

Data Simulation


Conclusions

Nikolas Constantinou, 3rd year MORSE student.

3rd year MORSE student.