ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА»

МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА ТЕОРИИ ВЕРОЯТНОСТЕЙ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

специалиста

СИМУЛИРОВАНИЕ СЛУЧАЙНЫХ ПРОЦЕССОВ С ИСПОЛЬЗОВАНИЕМ СИГНАТУРНЫХ МЕТОДОВ

Выполнил студент 609 группы

Мащенко Кирилл Алексеевич

_____

подпись студента

Научный руководитель:

к.ф.-м.н.

Житлухин Михаил Валентинович

_____

подпись научного руководителя

Москва

2022 год

Содержание

1 Введение

Настоящая дипломная работа посвящена исследованию и обобщению одного генеративного метода для случайных процессов, основанного на теории сигнатур и предложенного в работе [1]. Рассматриваемую задачу в общих словах можно описать следующим образом. Пусть имеется результат наблюдения за одной или несколькими траекториями некоторого случайного процесса, который описывает изменение какой-либо величины - как численно произвести достаточно большое число новых траекторий, которые были бы "похожи" на наблюдаемый процесс? Можно сказать, что в работе предлагается непараметрический метод Монте-Карло, где под непараметричностью имеется в виду то, что для симуляции траекторий не производится оценка параметров некоторой конкретной модели случайного процесса, а случайным образом выбираются траектории в некотором смысле близкие к наблюдаемой. Ввиду важности метода Монте-Карло в различных прикладных задачах, рассматриваемая задача представляет существенный интерес для приложений.

Центральную роль в предложенном методе играет преобразование траектории случайного процесса, состоящее в вычислении ее сигнатуры. Сигнатурой пути в n-мерном пространстве (то есть некоторой функции, аргументом которой является время – например, траектории случайного процесса), является семейство повторных интегралов от компонент этого пути по другим компонентам пути (точное определение приводится далее). Известно, что при выполнении некоторых достаточно общих условий, наложенных на рассматриваемый класс путей, существует взаимно однозначное соответствие между сигнатурами и путями. При этом в различных приложениях оказывается, что для достаточно хорошей степени восстановления траектории по сигнатуре достаточно использовать лишь несколько первых повторных интегралов из всего набора, что позволяет описывать траектории весьма экономным образом. Иными словами, сигнатуру можно рассматривать как способ сжатия информации, содержащейся в траектории.

Предложенный метод симуляции состоит в том, чтобы по имеющейся траектории, разбив ее сначала на меньшие фрагменты, вычислить n первых элементов (то есть повторных интегралов) сигнатуры каждого фрагмента. Такие наборы можно рассматривать как случайные векторы. В силу их небольшой размерности, стано-

вится удобным оценить их распределение и производить выборку новых случайных векторов из данного распределения с помощью генеративной модели, основанной на вариационном автокодировщике, производя, таким образом, случайные сигнатуры. Затем нужно восстановить случайный процесс по сигнатуре, что дает искомые случайные траектории.

В работе все случайные процессы, по сути, представляются своими значениями в дискретном множестве моментов времени. Таким образом все траектории являются кусочно-линейными, что весьма упрощает определение сигнатуры и использование ее свойств. Например, для определения сигнатуры нет необходимости привлекать повторные стохастические интегралы или интегралы по грубым траекториям - достаточно обычного интеграла Римана.

Исследование сигнатур траекторий восходит к работам [2] [3], в которых рассматривались достаточно регулярные траектории. Для грубых траекторий (например, с показателем Гельдера меньше $1/2$) сигнатуры исследовались в работах Т. Лайонсом, М. Хайрера и др. [4] [5] [6] в связи с теорией интегрирования по грубым траекториям. Оказывается, что ряд результатов, например, взаимно-однозначное соответствие между траекториями и сигнатурами, остается верным и для грубых траекторий [7] [8].

В последнее десятилетие сигнатурные методы находят применение в машинном обучении. Например, в сочетании со сверточными нейронными сетями завоевали первый приз в онлайн конкурсе ICDAR 2013 по распознаванию изолированных китайских символов [9]. В сочетании с моделью регрессии на градиентном бустинге выиграли первый приз в конкурсе PhysioNet 2019 по вычислительной технике в кардиологии [10]. Также в последние годы сигнатурные методы нашли применение в задачах финансовой математики, связанных с хеджированием производных инструментов [11].

Настоящая дипломная работа устроена следующим образом. В разделе 2 сформулирована теория сигнатурных методов. В разделе 3 сделан обзор генеративных моделей и области их применения, а также представлены основные шаги по применению генеративной модели к финансовым данным. В разделе 4 описаны результаты по расширению области применения генеративной модели на многомерные пути, и

продемонстрированы результаты её работы на примере цен акций компании ПАО “Газпром” и отраслевого индекса нефти и газа Московской биржи.

2 Сигнатурные методы

Сигнатурный подход представляет собой непараметрический робастный способ извлечения характерных признаков из данных, которые впоследствии можно использовать для моделей машинного обучения. С помощью сигнатурного подхода по последовательно поступающим данным, представленным в виде параметризованного пути, можно сгенерировать признаки, хорошо обобщающие и полноценно описывающие эти данные, поскольку сигнатуры содержат в себе полную информацию об аналитических и геометрических свойствах пути.

Множество типов данных можно считать последовательно поступающими, или упорядоченными по времени, например финансовые временные ряды, тексты на естественных языках, и др.

2.1 Определение и свойства пути

Определение 2.1. *Путь в* $\mathbb{R}^d$ – это непрерывное отображение X из некоторого интервала $[a, b]$ в $\mathbb{R}^d$. Чтобы подчеркнуть зависимость от времени будем использовать обозначение $X_t = X(t) : [a, b] \mapsto \mathbb{R}^d$.

В дальнейшем будем предполагать, что рассматриваемые пути являются “достаточно хорошими” отображениями, а именно они являются кусочно-дифференцируемыми (вообще говоря, справедливость большинства результатов сохранится, если считать, что пути имеют ограниченную вариацию). Будем называть путь гладким, если он бесконечно дифференцируем.

Определение 2.2. Интегралом от функции $f : \mathbb{R} \mapsto \mathbb{R}$ по одномерному пути $X : [a, b] \mapsto \mathbb{R}$ называется величина

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

где последний интеграл является обычным (римановым) интегралом непрерывной ограниченной функции. Обозначение “верхняя точка” здесь и далее используется для дифференцирования по одной переменной: $\dot{X}_t = dX_t/dt$.

Заметим, что $f(X_t)$ тоже является путем на $[a,b]$, поэтому можно естественным образом определить интеграл от пути по пути.

Определение 2.3. Интегралом от пути $Y : [a,b] \mapsto \mathbb{R}$ по пути $X : [a,b] \mapsto \mathbb{R}$ называется величина

$$\int_a^b Y_t dX_t = \int_a^b Y_t \dot{X}_t dt.$$

2.2 Определение сигнатуры пути

Обозначим координаты пути $X : [a,b] \mapsto \mathbb{R}^d$ за $\left(X_t^1, \ldots, X_t^d\right)$, где каждая координата $X^i : [a,b] \mapsto \mathbb{R}$ является путем. Для каждого $i \in \{1, \ldots, d\}$ и $t \in [a,b]$ определим величину

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

которая является приращением i-ой координаты пути до момента времени t. Отметим, что $S(X)_{a,\cdot}^i$ – это тоже путь.

Теперь определим для любой пары $i,j \in \{1,...,d\}$ двойной повторный интеграл

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

Можно продолжить по индукции: для любого $k \geq 1$ и набора индексов $i_1, ..., i_k \in \{1,...,d\}$ определим

$$S(X)_{a,t}^{i_1,\ldots,i_k} = \int_{a<s<t} S(X)_{a,s}^{i_1,\ldots,i_{k-1}} dX_s^{i_k} = \int_{a<t_k<t} \ldots \int_{a<t_1<t_2} dX_{t_1}^{i_1} \ldots dX_{t_k}^{i_k}.$$

Величина $S(X)_{a,t}^{i_1,\ldots,i_k}$ называется k-кратным повторным интегралом от пути X по индексам $i_1, ..., i_k$.

Определение 2.4. *Сигнатурой* пути $X : [a,b] \mapsto \mathbb{R}^d$ называется бесконечный набор $S(X)_{a,b}$ всех повторных интегралов от X:

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

где первый элемент сигнатуры (соответствующий пустому индексу) по определению считается равным 1, а верхние индексы остальных элементов пробегают набор всевозможных мульти-индексов

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

Множество W называется множеством слов в алфавите $A = \{1, ..., d\}$. Конечный набор чисел $S(X)_{a,b}^{i_1,...,i_k}$ для всевозможных мульти-индексов длины k будем называть k-ым уровнем сигнатуры.

Пример 2.5. Рассмотрим произвольный одномерный путь $X : [a, b] \mapsto \mathbb{R}$. Тогда сигнатура этого пути вычисляется следующим образом:

$$\begin{aligned}
&S(X)_{a,b}^{1} = X_b - X_a, \\
&S(X)_{a,b}^{1,1} = \frac{(X_b - X_a)^2}{2!}, \\
&S(X)_{a,b}^{1,1,1} = \frac{(X_b - X_a)^3}{3!}, \\
&... \\
&S(X)_{a,b}^{1,1,...,1} = \frac{(X_b - X_a)^k}{k!}.
\end{aligned}$$

Отсюда можно видеть, что сигнатура по повторным индексам выражается через приращение по соответствующей координате, что верно и для многомерного случая.

Пример 2.6. Рассмотрим следующий конкретный двумерный путь $X : [a, b] \mapsto \mathbb{R}^2$:

$$\begin{aligned}
X_t &= \{X_t^1, X_t^2\} = \{t, k \cdot t + d\}, \\
dX_t &= \{dX_t^1, dX_t^2\} = \{dt, k \cdot dt\},
\end{aligned}$$

где $k \neq 0$ – параметр. Тогда элементы сигнатуры $S(X)_{a,b}^{1,2}$ и $S(X)_{a,b}^{2,1}$ этого пути вычисляются следующим образом:

$$\begin{aligned}
S(X)_{a,b}^{1,2} &= \int_a^b (\int_a^{t_2} dt_1) k dt_2 = \frac{k \cdot (b-a)^2}{2}, \\
S(X)_{a,b}^{2,1} &= \int_a^b (\int_a^{t_2} k dt_1) dt_2 = \frac{k \cdot (b-a)^2}{2}.
\end{aligned}$$

2.3 Шафл-произведение

Важное свойство сигнатуры состоит в том, что произведение двух элементов $S(X)_{a,b}^{i_1,...,i_k}$ и $S(X)_{a,b}^{j_1,...,j_m}$ может быть всегда представлено в виде суммы других элементов сигнатуры $S(X)_{a,b}$, которая зависит только от мульти-индексов $(i_1, ..., i_k)$ и $(j_1, ..., j_m)$. Это свойство показывает, что члены сигнатуры не являются алгебраически независимыми, а также дает возможность работать с линейными объектами вместо произведений.

Далее требуется определить понятие шафл-произведения двух мульти-индексов.

Определение 2.7. Перестановка σ множества $\{1, \dots, k+m\}$ называется (k, m)-шафлом, если $\sigma^{-1}(1) < ... < \sigma^{-1}(k)$ и $\sigma^{-1}(k+1) < ... < \sigma^{-1}(k+m)$. Для множества всех (k, m)-шафлов будем использовать обозначение $\text{Shuffles}(k, m)$.

Определение 2.8. Рассмотрим два мульти-индекса $I = (i_1, ..., i_k)$ и $J = (j_1, ..., j_m)$, $i_1, ..., i_k, j_1, ..., j_m \in \{1, ..., d\}$. Определим мульти-индекс

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

Шафл-произведением I и J (обозначение: $I \text{ш} J$) называется конечный набор мульти-индексов длины $k+m$ вида

$$I \text{ш} J = \{(r_{\sigma(1)}, ..., r_{\sigma(k+m)} \mid \sigma \in \text{Shuffles}(k, m)\}.$$

Теорема 2.9. *Для любого пути $X : [a, b] \mapsto \mathbb{R}^d$ и мульти-индексов $I = (i_1, ..., i_k)$ и $J = (j_1, ..., j_m)$, $i_1, ..., i_k, j_1, ..., j_m \in \{1, ..., d\}$ верно равенство*

$$S(X)_{a,b}^I S(X)_{a,b}^J = \sum_{K \in I \text{ш} J} S(X)_{a,b}^K.$$

Доказательство. По определению сигнатуры имеем

$$\begin{aligned} S(X)_{a,b}^I S(X)_{a,b}^J &= \int_a^b S(X)_{a,s}^{i_1,...,i_{k-1}} dX_s^{i_k} \int_a^b S(X)_{a,t}^{j_1,...,j_{m-1}} dX_t^{j_m} \\ &= \int_{s,t \in [a,b]^2} S(X)_{a,s}^{i_1,...,i_{k-1}} S(X)_{a,t}^{j_1,...,j_{m-1}} dX_s^{i_k} dX_t^{j_m}, \end{aligned}$$

где последнее равенство верно по теореме Фубини. Разделив промежуток интегрирования на два промежутка $s < t$ и $s > t$, получаем

$$S(X)_{a,b}^I S(X)_{a,b}^J = \int_a^b S(X)_{a,t}^I S(X)_{a,t}^{j_1,...,j_{m-1}} dX_t^{j_m} + \int_a^b S(X)_{a,s}^J S(X)_{a,s}^{i_1,...,i_{k-1}} dX_s^{i_k}.$$

По индукции можно разложить произведение элементов сигнатуры внутри интегралов следующим образом:

$$S(X)^I_{a,t}S(X)^{j_1,\dots,j_{m-1}}_{a,t} = \sum_{K\in I\text{ш}j_1,\dots,j_{m-1}} S(X)^K_{a,t},$$

$$S(X)^J_{a,s}S(X)^{i_1,\dots,i_{k-1}}_{a,s} = \sum_{K\in I\text{ш}i_1,\dots,i_{k-1}} S(X)^K_{a,s}.$$

Подставив полученные выражения в интегралы, получаем

$$S(X)^I_{a,b}S(X)^J_{a,b} = \int_a^b \sum_{K\in I\text{ш}j_1,\dots,j_{m-1}} S(X)^K_{a,t}dX_t^{j_m} + \int_a^b \sum_{K\in I\text{ш}i_1,\dots,i_{k-1}} S(X)^K_{a,s}dX_s^{i_k}.$$

Заметим, что

$$I\text{ш}J = \{Kj_m : K \in I\text{ш}j_1, ..., j_{m-1}\} \sqcup \{Ki_k : K \in I\text{ш}i_1, ..., i_{k-1}\},$$

откуда следует утверждение теоремы. □

Пример 2.10. Произведение элементов сигнатуры первого порядка выражается через элементы второго порядка следующим образом:

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

2.4 Геометрический смысл сигнатур

Предложение 2.11. *(Независимость от начальной точки). Рассмотрим путь* $X : [a, b] \mapsto \mathbb{R}^d$ *и* $h \in \mathbb{R}^d$. *Пусть путь* $Y : [x, y] \mapsto \mathbb{R}^d$ *имеет вид* $Y_t = X_t + h$. *Тогда*

$$S(X)_{a,b} = S(Y)_{a,b}.$$

Доказательство. $S(X)_{a,b}$ зависит только от $\dot{X}_t$, поэтому прибавление константы не влияет на результат. □

Предложение 2.12. *(Независимость от репараметризации времени). Рассмотрим путь* $X : [a, b] \mapsto \mathbb{R}^d$ *и биективную непрерывную неубывающую функцию* $\psi : [x, y] \mapsto [a, b]$. *Пусть путь* $Y : [x, y] \mapsto \mathbb{R}^d$ *имеет вид* $Y_t = X_{\psi_t}$. *Тогда*

$$S(X)_{a,b} = S(Y)_{x,y}.$$

Доказательство. По определению сигнатуры

$$S(Y)^{i_1,..,i_k}_{x,y} = \int_x^y S(Y)^{i_1,..,i_{k-1}}_{x,s} dY^{i_k}_s = \int_x^y S(Y)^{i_1,..,i_{k-1}}_{x,s} \dot{\psi}(s) dX^{i_k}_{\psi(s)}.$$

По индукции верно равенство

$$S(Y)^{i_1,..,i_{k-1}}_{x,s} = S(X)^{i_1,..,i_{k-1}}_{a,\psi(s)}.$$

Подставив полученное выражение в интеграл, получаем

$$S(Y)^{i_1,..,i_k}_{x,y} = \int_x^y S(X)^{i_1,..,i_{k-1}}_{a,\psi(s)} \dot{\psi}(s) dX^{i_k}_{\psi(s)} \overset{[u=\psi(s)]}{=} \int_a^b S(X)^{i_1,..,i_{k-1}}_{a,u} dX^{i_k}_u = S(X)^{i_1,...,i_k}_{a,b}.$$

□

Оба эти свойства показывают, что сигнатура зависит от геометрических свойств пути. Можно также заметить, что первый уровень сигнатуры $(S(X)^1_{a,b}, ..., S(X)^d_{a,b})$ по определению является приращением пути по каждому аргументу, а второй уровень сигнатуры связан с так называемой площадью Леви.

Определение 2.13. Пусть $X : [a, b] \mapsto \mathbb{R}^2$ - двумерный путь, где $X_t = (X^1_t, X^2_t)$. Проведём прямую от начальной до конечной точки пути, после чего все получившиеся площади при пересечении этой прямой и пути рассмотрим со знаком минус, если они выше прямой, и со знаком плюс, если они ниже прямой. *Площадью Леви* называется сумма данных площадей с учетом знаков.

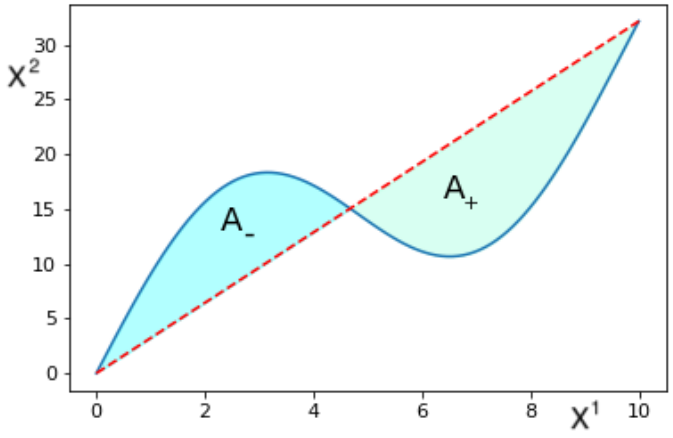

Рис. 1: Площадь Леви $A = A_+ - A_-$

Предложение 2.14. *Площадь Леви выражается через элементы сигнатуры следующим образом:*

$$A = \frac{1}{2}\left(S(X)^{1,2}_{a,b} - S(X)^{2,1}_{a,b}\right).$$

Доказательство. Элементы сигнатуры $S(X)^{1,2}_{a,b}$ и $S(X)^{2,1}_{a,b}$ пути по определению вычисляются следующим образом:

$$S(X)^{1,2}_{a,b} = \int_a^b (\int_a^{t_2} dX^1_{t_1}) dX^2_{t_2} = \int_a^b X^1_{t_2} dX^2_{t_2} - X^1_a (X^2_b - X^2_a),$$
$$S(X)^{2,1}_{a,b} = \int_a^b (\int_a^{t_2} dX^2_{t_1}) dX^1_{t_2} = \int_a^b X^2_{t_2} dX^1_{t_2} - X^2_a (X^1_b - X^1_a).$$

Заметим, что определенные интегралы в данных выражениях являются площадями под графиком пути: в первом выражении если рассматривать в качестве оси $x - X^2$, а оси $y - X^1$, во втором выражении - если рассматривать в качестве оси $x - X^1$, а оси $y - X^2$. Вычитая из них соответствующие площади прямоугольников получаем, что $S(X)^{1,2}_{a,b}$ и $S(X)^{2,1}_{a,b}$ равны усеченным площадям над и под графиком (как показано на Рис. 2)

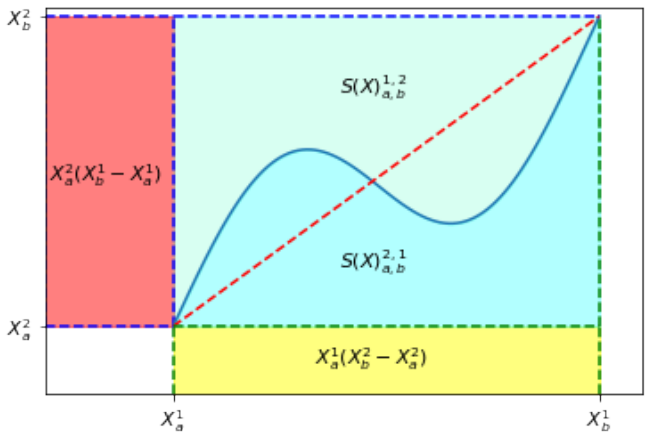

Рис. 2: Геометрический смысл $S(X)^{1,2}_{a,b}$ и $S(X)^{2,1}_{a,b}$

Используя это, выразим $S(X)^{1,2}_{a,b}$ и $S(X)^{2,1}_{a,b}$ следующим образом:

$$S(X)^{1,2}_{a,b} = A_+ + \frac{(X^2_b - X^2_a)(X^1_b - X^1_a)}{2} - A_-,$$
$$S(X)^{2,1}_{a,b} = A_- + \frac{(X^2_b - X^2_a)(X^1_b - X^1_a)}{2} - A_+.$$

Вычитая из первого равенства второе, получаем

$$S(X)^{1,2}_{a,b} - S(X)^{2,1}_{a,b} = 2(A_+ - A_-) = 2A.$$

□

2.5 Тождество Чена

Для эффективного вычисления сигнатур можно использовать несколько подходов. Если известна аналитическая параметризация пути, и интегралы по нему считаются явно, то можно непосредственно аналитически их вычислить. Однако для множества реальных данных, которые являются результатом последовательных наблюдений за некоторыми явлениями, пути являются кусочно-линейными функциями и обладают неудобной для вычисления параметризацией. Для таких случаев эффективнее использовать следующее тождество Чена.

Теорема 2.15. *(Тождество Чена). Пусть $a < b < c$ и $X : [a, c] \mapsto \mathbb{R}^d$. Тогда для любых $i_1, ..., i_k \in W$ выполнено равенство*

$$S(X)^{i_1,...,i_k}_{a,c} = \sum_{m=0}^{k} S(X)^{i_1,...,i_m}_{a,b} S(X)^{i_{m+1},...,i_k}_{b,c}.$$

Доказательство. Пользуясь определением элемента сигнатуры и разбивая интеграл на два промежутка, получаем

$$S(X)^{i_1,...,i_k}_{a,c} = \int_a^b S(X)^{i_1,...,i_{k-1}}_{a,t} dX^{i_k}_t + \int_b^c S(X)^{i_1,...,i_{k-1}}_{a,t} dX^{i_k}_t.$$

По индукции верно равенство

$$S(X)^{i_1,...,i_{k-1}}_{a,t} = \sum_{m=0}^{k-1} S(X)^{i_1,...,i_m}_{a,b} S(X)^{i_{m+1},...,i_{k-1}}_{b,t}.$$

Подставив полученное выражение во второй интеграл, а также применив определение элемента сигнатуры, получаем

$$\begin{aligned} S(X)_{a,c}^{i_1,\dots,i_k} &= S(X)_{a,b}^{i_1,\dots,i_k} + \int_b^c \sum_{m=0}^{k-1} S(X)_{a,b}^{i_1,\dots,i_m} S(X)_{b,t}^{i_{m+1},\dots,i_{k-1}} dX_t^{i_k} \\ &= S(X)_{a,b}^{i_1,\dots,i_k} + \sum_{m=0}^{k-1} S(X)_{a,b}^{i_1,\dots i_m} S(X)_{b,c}^{i_{m+1},\dots,i_{k-1},i_k} \\ &= \sum_{m=0}^{k} S(X)_{a,b}^{i_1,\dots,i_m} S(X)_{b,c}^{i_{m+1},\dots,i_k}. \end{aligned}$$

□

Можно легко вычислить все сигнатуры на каждом одномерном отрезке кусочно-линейной функции, а потом по очереди вычислить сигнатуру всего пути с помощью тождества Чена, по очереди присоединяя каждый следующий отрезок.

Пример 2.16. Рассмотрим путь $X : [0, 2] \mapsto \mathbb{R}^2$ следующего вида:

$$X_t = \{t, 2 \cdot t\}, t \in [0, 1],$$
$$X_t = \{t, 3 - t\}, t \in [1, 2].$$

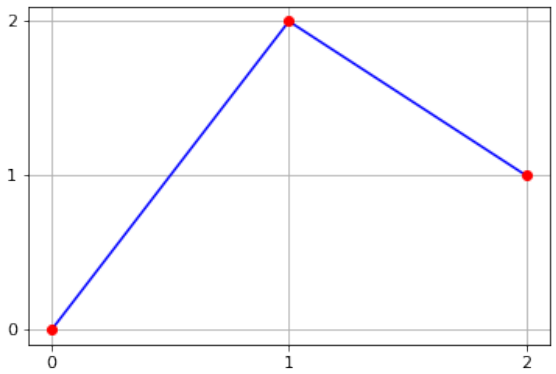

Рис. 3: Кусочно-линейный путь

Посчитаем первые два уровня сигнатуры пути отдельно на каждом из отрезков $[0, 1], [1, 2]$. На этих отрезках путь является одномерным и линейным, поэтому, согласно примерам 1.5 и 1.6, первые два уровня сигнатуры этих участков выражаются следующим образом:

на отрезке $[0,1]$:

$$
\begin{aligned}
&S(X)^{\emptyset}_{0,1} = 1 \,(\text{по определению})\,,\\
&S(X)^{1}_{0,1} = \Delta X^{1}_{0,1} = 1,\\
&S(X)^{1,1}_{0,1} = \frac{\left(\Delta X^{1}_{0,1}\right)^2}{2} = 0.5,\\
&S(X)^{2}_{0,1} = \Delta X^{2}_{0,1} = 2,\\
&S(X)^{2,2}_{0,1} = \frac{\left(\Delta X^{2}_{0,1}\right)^2}{2} = 2,\\
&S(X)^{1,2}_{0,1} = \frac{2\cdot(1-0)^2}{2} = 1,\\
&S(X)^{2,1}_{0,1} = \frac{2\cdot(1-0)^2}{2} = 1.
\end{aligned}
$$

на отрезке $[1,2]$:

$$
\begin{aligned}
&S(X)^{\emptyset}_{1,2} = 1 \,(\text{по определению})\,,\\
&S(X)^{1}_{1,2} = \Delta X^{1}_{1,2} = 1,\\
&S(X)^{1,1}_{1,2} = \frac{\left(\Delta X^{1}_{1,2}\right)^2}{2} = 0.5,\\
&S(X)^{2}_{1,2} = \Delta X^{2}_{1,2} = -1,\\
&S(X)^{2,2}_{1,2} = \frac{\left(\Delta X^{2}_{1,2}\right)^2}{2} = 0.5,\\
&S(X)^{1,2}_{1,2} = \frac{-1\cdot(2-1)^2}{2} = -0.5,\\
&S(X)^{2,1}_{1,2} = \frac{-1\cdot(2-1)^2}{2} = -0.5.
\end{aligned}
$$

Применим тождество Чена:

$$
\begin{aligned}
&S(X)^{1}_{0,2} = S(X)^{1}_{0,1}\cdot S(X)^{\emptyset}_{1,2} + S(X)^{\emptyset}_{0,1}\cdot S(X)^{1}_{1,2} = 1\cdot 1 + 1\cdot 1 = 2 = \Delta X^{1}_{0,2},\\
&S(X)^{2}_{0,2} = S(X)^{2}_{0,1}\cdot S(X)^{\emptyset}_{1,2} + S(X)^{\emptyset}_{0,1}\cdot S(X)^{2}_{1,2} = 2\cdot 1 + 1\cdot(-1) = 1 = \Delta X^{2}_{0,2},\\
&S(X)^{1,2}_{0,2} = S(X)^{1,2}_{0,1}\cdot S(X)^{\emptyset}_{1,2} + S(X)^{1}_{0,1}\cdot S(X)^{2}_{1,2} + S(X)^{\emptyset}_{0,1}\cdot S(X)^{1,2}_{1,2}\\
&\qquad = 1\cdot 1 + 1\cdot(-1) + 1\cdot(-0.5) = -0.5,\\
&S(X)^{2,1}_{0,2} = S(X)^{2,1}_{0,1} + S(X)^{2}_{0,1}\cdot S(X)^{1}_{1,2} + S(X)^{2,1}_{1,2} = 1 + 2\cdot 1 - 0.5 = 2.5,\\
&S(X)^{1,1}_{0,2} = S(X)^{1,1}_{0,1} + S(X)^{1}_{0,1}\cdot S(X)^{1}_{1,2} + S(X)^{1,1}_{1,2} = 0.5 + 1\cdot 1 + 0.5 = 2 = \frac{\left(\Delta X^{1}_{0,2}\right)^2}{2},\\
&S(X)^{2,2}_{0,2} = S(X)^{2,2}_{0,1} + S(X)^{2}_{0,1}\cdot S(X)^{2}_{1,2} + S(X)^{2,2}_{1,2} = 2 + 2\cdot(-1) + 0.5 = 0.5 = \frac{\left(\Delta X^{2}_{0,2}\right)^2}{2}.
\end{aligned}
$$

2.6 Логарифмические сигнатуры

Из теоремы о шафл-произведении видно, что члены сигнатуры не являются алгебраически независимыми, поэтому не любой конечный набор чисел может представлять из себя набор элементов сигнатуры соответствующей размерности. В частности, если случайным образом производить выборку из распределения элементов сигнатуры, то нельзя гарантировать, что полученные случайные векторы будут являться сигнатурами некоторых путей. Этой проблемы можно избежать, если вместо сигнатур рассматривать логарифмические сигнатуры, которые представляют из себя независимый набор признаков.

Для определения понятия логарифмической сигнатуры введем сначала алгебру формальных степенных рядов.

Определение 2.17. Рассмотрим d формальных неопределенных величин $e_1, \ldots, e_d$. Алгеброй некоммутативных формальных степенных рядов из d неопределенных называется векторное пространство всех рядов вида

$$\sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k=1}^{d} \lambda_{i_1,\ldots,i_k} e_{i_1} \ldots e_{i_k},$$

где параметр второй суммы пробегает по всем мульти-индексам $(i_1, \ldots, i_k), i_1, \ldots, i_k \in \{1, \ldots, d\}$, и коэффициенты $\lambda_{i_1,\ldots,i_k}$ являются вещественными числами.

На этом пространстве определены стандартные операции сложения рядов и умножения ряда на коэффициент, а также операция умножения рядов $\otimes$. При этом элементы пространства некоммутативны, т.е., например, $e_1 e_2$ и $e_2 e_1$ являются разными элементами.

Сигнатура может быть "закодирована" как элемент этого пространства (т.е. существует взаимно-однозначное соответствие между сигнатурами и формальными степенными рядами):

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

Иными словами, элементы сигнатуры рассматриваются как соответствующие коэффициенты ряда, что позволяет "перейти" в алгебру рядов. В ней можно выполнять

операции сложения и умножения рядов, а коэффициенты полученного ряда интерпретировать как элементы сигнатуры.

Далее для формального ряда, в котором первый коэффициент равен 1, а остальные равны 0 (что соответствует сигнатуре $\{1, 0, 0, \ldots\}$) будем использовать обозначение $\mathbf{1}$.

Определение 2.18. *Логарифмической сигнатурой* пути $X : [a, b] \mapsto \mathbb{R}^d$ называется формальный степенной ряд

$$\log S(X)_{a,b} = \sum_{n=1} \frac{(-1)^n}{n} \left(\mathbf{1} - S(X)_{a,b}\right)^{\otimes n}.$$

Замечание 2.19. В общем случае путь не определяется однозначно его сигнатурой. Например, по сигнатуре нельзя восстановить точную скорость, с которой этот путь проходится (из-за инвариантности сигнатуры при репараметризации времени). Однако для непересекающихся путей можно полностью определить по сигнатуре все точки, через которые пройдёт путь, и порядок их обхода.

3 Генеративные модели

Генерация новых данных, похожих по распределению на исходные, может быть необходима во многих сферах. Например, для анонимизации данных, когда данные конфиденциальны и нужно сгенерировать другие примеры того же распределения. Примером этого могут являются финансовые и медицинские данные. Помимо этого, бывают ситуации, когда количество данных заведомо мало из-за ограничения на число экспериментов или доступ к данным. Также генеративные модели могут использоваться для тестирования стратегий, которое нельзя проводить на исторических данных во избежание переобучения.

Целью работы [1] является изучение методов моделирования финансовых временных рядов без предположений о лежащей в их основе стохастической динамики. Авторами была разработана гибкая генеративная модель для финансовых временных рядов, основанная на сигнатурах, которая работает на маленьком количестве данных и создаёт синтетические данные, основанные на рыночных показателях. Когда большой объем данных недоступен, генеративная модель, использующая малое

количество данных, может сгенерировать большое количество новых данных, статистически неотличимых от изначальных, которые впоследствии можно передать другим моделям, которым в свою очередь необходимо много данных.

В статье показывается различие между классическими подходами и подходом машинного обучения. В классических моделях известна стохастическая основа и математические свойства модели. Однако такие модели недостаточно гибки к рынку. К тому же на практике неизвестно распределение финансовых данных. Данные модели можно значительно усложнять, добавляя в них новые параметры, однако в реальности они не могут полностью описать поведение рынка.

В случае же генеративных моделей, основанных на машинном обучении, мы генерируем новые данные, похожие по распределению на исходные. Тогда не нужно знать распределение исходных данных, мы генерируем новые и сравниваем их по распределению с исходными, что предлагается также делать с помощью сигнатурных методов.

Зафиксируем обозначения, которые в дальнейшем будут использоваться. Пусть $S(t)$ обозначает цену финансового актива, а $X(t) = \ln S(t)$. Тогда логарифмическая доходность обозначается как $r(t, \Delta t) = X(t + \Delta t) - X(t)$.

В случае финансовых данных, например индекса фондового рынка $S\&P500$, имеется только один доступный поток данных. При этом, даже если поделить этот поток данных по различным промежуткам времени (по дням, по неделям, по месяцам), то данных зачастую все равно оказывается недостаточно для применения многих методов машинного обучения.

Авторы статьи предлагают свою модель вариационного автокодировщика VAE и его обусловленную версию CVAE, которой на вход еще подается сигнатура предыдущего участка пути. Авторы показывают, что данные модели хорошо работают на маленьком количестве данных, а их недостатки по сравнению с генеративно-состязательными сетями GAN на временных рядах не имеют значения.

Предлагаются следующие основные шаги при обучении данной генеративной модели на финансовых данных:

1) **Извлечение данных из временного потока.** Полный временной ряд необходимо разделить на интервалы равной длины: 1 день, 5 дней (соответствует рабочей

неделе), 20 дней (соответствует рабочему месяцу) и посчитать логарифмическую доходность на данных промежутках времени.

2) **Предобработка данных.** На этом шаге к данным применяются необходимые преобразования, например, lead-lag (см. [11], стр. 20), и считаются логарифмические сигнатуры.

3) **Обучение модели.** Использовав в качестве входных данных логарифмические сигнатуры исходного набора путей, на данном этапе обучается генеративная модель, выходными данными которой являются сгенерированные логарифмические сигнатуры из того же распределения. Дополнительно можно обучить обусловленную модель, добавив к базовой модели различные рыночные показатели в данный момент времени, такие как уровень мгновенной волатильности и уровень индекса в начале пути, а также логарифмическую сигнатуру на предыдущем участке. Благодаря этой обусловленной модели возможно генерировать данные на длинном промежутке времени, по очереди его расширяя новым интервалом. Помимо этого, можно использовать тождество Чена для получения сигнатуры на длинном промежутке времени из маленьких промежутков.

4) **Использование полученных данных.** Полученные сгенерированные данные можно передать другой модели в качестве входных данных, оставив их в виде логарифмических сигнатур. Однако, если необходимо получить конкретные пути, в статье предлагается применить эволюционный алгоритм восстановления пути по его логарифмической сигнатуре, подробное описание которого приводится далее.

4 Основные результаты

Целью дипломной работы является обобщение алгоритма, предложенного в работе [1] (см. также репозиторий с программным кодом [12]) для одномерных данных, на многомерный случай, а также расширение класса допустимых предварительных преобразований путей и улучшение работы модели обращения логарифмической сигнатуры в случае известного распределения исходного набора путей. В качестве результата приводится готовый программный код для генерации многомерных путей, протестированный на реальных ценах активов.

4.1 Обзор алгоритма Buhler, Horvath, Lyons, Arribas, Wood. Постановка задачи

Алгоритм, предложенный в работе [1], выполнен на языке Python, и его можно условно разделить на следующие четыры части, реализованные в виде классов:

1) Класс, отвечающий за загрузку данных, их обработку, а также содержащий генеративную модель и предоставляющий интерфейс для её использования.

2) Класс для получения путей по логарифмическим сигнатурам (т.е. решающий задачу обращения логарифмических сигнатур).

3) Класс, предоставляющий генеративную модель CVAE (conditional variational autoencoder – обусловленный вариационный автокодировщик).

4) Класс для проверки гипотезы о близости вероятностного распределения исходного множества путей и путей, соответствующих сгенерированным логарифмическим сигнатурам.

Отметим, что область применимости алгоритма работы [1] ограничивается одномерными путями, что делает невозможным его применение, например, для изучения статистической зависимости между зависимыми временными рядами. Помимо этого, алгоритм обладал рядом ограничений в подходах к предобработке данных, а также в настройке параметров моделей, таких как используемая генеративная модель, распределение, которым приближаются пути при обращении логарифмической сигнатуры, и других.

Классы в пунктах 3 и 4 не нуждались в изменениях, поскольку они работают непосредственно с сигнатурами. При увеличении размерности путей или при изменениях в процессе обработки данных меняется только длина сигнатуры, что происходит, например, и при увеличении уровня, до которого мы берем элементы сигнатуры, поэтому работа данных классов от этого независит.

Однако, классы в пунктах 1 и 2 требовали значительных обобщений и доработок в их программном коде.

4.2 Класс для обработки рыночных данных

Соответствующий класс в программном коде, сопровождающем работу [1], имел следующие недостаки.

1) Мог быть применен только к одномерным временным рядам.

2) Использовалось жестко заданное преобразование временного ряда – алгоритм lead-lag (см. [11], стр. 20). Преобразование lead-lag в случае одномерных данных помогает увеличить размерность пути, что необходимо для корректной работы сигнатурных методов (в случае одномерного пути сигнатуры кодируют только приращение пути). Помимо этого, данное преобразование связано с выборочной дисперсией (см. [11], стр. 25) и имеет большое значение в анализе временных рядов [14]. Увеличения размерности пути можно также добиться с помощью присоединения времени дополнительной размерностью, что в случае финансовых данных может иметь и самостоятельное значение. Также эти и другие преобразования данных, например, обработку пропусков в данных (см. [11], стр. 32), можно объединять, последовательно применяя их друг за другом, при этом порядок применения преобразований также важен и влияет на итоговый результат.

3) Использовался определенный встроенный алгоритм генерации логарифмических сигнатур, в котором нельзя поменять параметры, а также нельзя заменить его на другой.

В улучшенной мной модели проведена значительная работа по обобщению кода на несколько размерностей, а функция преобразования данных и генеративная модель передаются в класс параметрами.

В качестве иллюстрации применения улучшенного алгоритма, а также проверки его работоспособности, приведём пример симуляции совместных траекторий процессов цен акций компании ПАО “Газпром” (код GAZP) и отраслевого индекса нефти и газа Московской биржи (код MOEXOG).

Данные выгружаются с 01.01.2005 по 01.01.2019 с Московской биржи (moex) с помощью библиотеки pandas_datareader, после чего все элементы пути делятся на значение цены в первый момент времени.

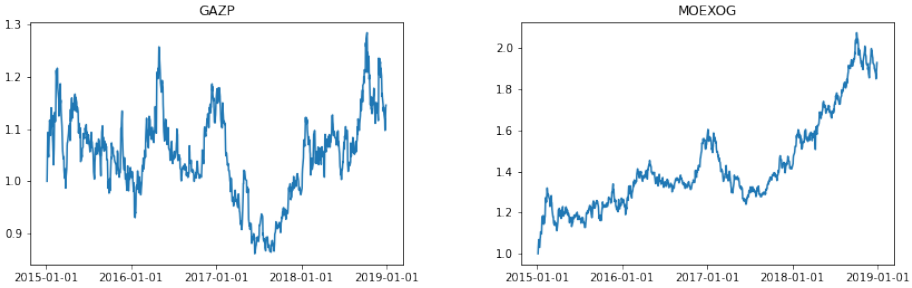

Рис. 4: Цены активов GAZP и MOEXOG с Московской биржи.

Далее делим временной ряд на интервалы равной длины, соответсвующие рабочим месяцам. При этом необходимо учесть, что активы могут торговаться в разные дни, таким образом, в какие-то дни данных о цене некоторых из набора активов может не быть. Для того, чтобы получить многомерный путь, необходимо взять пересечение по дате цен активов, исключив, таким образом, из итогового пути дни, в которые хотя бы один из активов не торговался. Помимо этого, пропуски в цене активов можно заполнить ценой в предыдущий момент времени, в таком случае исключать цены других активов не придётся. В итоге получаем набор многомерных совместных траекторий процессов цен этих активов (см. рис. 5).

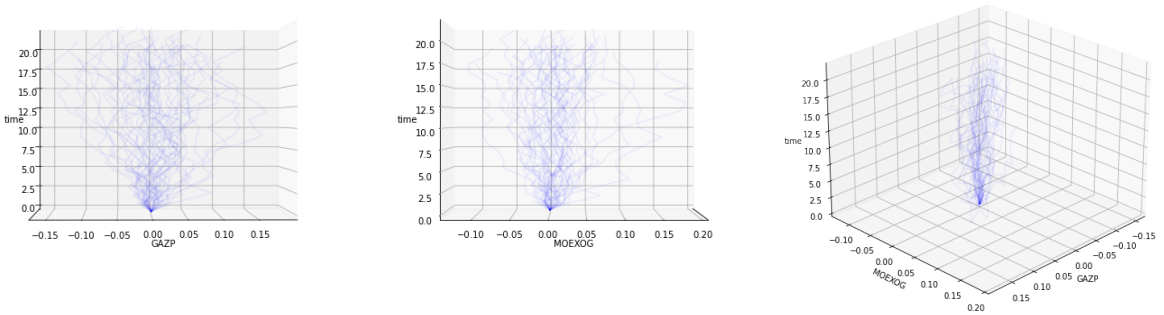

Рис. 5: Семейство совместных траекторий процессов цен активов GAZP и MOEXOG

На графиках на рис. 5 все пути приведены к точке (0,0). Поскольку сигнатуры независимы от начальной точки - то и сгенерированные пути могут начинаться с любой начальной точки. Однако для реальной работы с полученными данными это не имеет значения, поскольку путь можно перенести к цене в текущий момент времени.

После этого применяем к полученным данным переданное в модель преобразование данных и считаем элементы логарифмической сигнатуры до необходимого уровня, после чего масштабируем их на отрезок $[0, 1]$ для корректной работы предложенной авторами генеративной модели.

Составляем наборы "данных" и "условий" для обусловленной модели генерации: набором данных будем считать логарифмические сигнатуры участков от первого до последнего, набором условий - логарифмические сигнатуры участков от нулевого до предпоследнего. Таким образом, получаем, что для логарифмической сигнатуры участка пути условием будет являться логарифмическая сигнатура предыдущего участка.

После этого по полученным данным и условиям запускается тренировка предложенного авторами обусловленного вариационного автокодировщика CVAE. Для использования модели ей необходимо передать масштабированную на отрезок $[0, 1]$ логарифмическую сигнатуру в качестве условия, после чего модель вернет логарифмическую сигнатуру следующего участка, которую необходимо масштабировать с отрезка $[0, 1]$ обратным преобразованием.

Полученный таким образом набор логарифмических сигнатур можно использовать в чистом виде, например, передать другой модели для обучения или её тестирования. Однако в случае, если необходимы непосредственно совместные траектории, нужно применить к полученным логарифмическим сигнатурам алгоритм их обращения.

4.3 Алгоритм обращения логарифмической сигнатуры

Соответствующий алгоритм в программном коде, сопровождающем работу [1], имел следующие недостаки.

1) Работал только с одномерными путями.

2) Аналогично первому классу, использовалось жестко заданное преобразование временного ряда – алгоритм lead-lag.

3) Путь можно было приближать только одним распределением, зафиксированным в коде алгоритма, что неэффективно, если известно распределение исходных дан-

ных. Помимо этого, для разных размерностей пути чаще всего необходимы разные распределения для их приближения.

Для обращения логарифмической сигнатуры авторы предлагают использовать эволюционный алгоритм подбора нужного пути, чтобы его логарифмическая сигнатура была наиболее похожа на изначальную.

Если распределения приращений каждой координаты пути известны, их можно передать в модель в качестве параметров. Однако такое решение подойдет только для теоретической модели, ведь на практике распределения приращений траекторий чаще всего неизвестны.

Пусть мы ищем путь $X : [a, b] \mapsto \mathbb{R}^d$, $\left(X_t^1, \ldots, X_t^d\right)$, где каждая координата $X^i : [a, b] \mapsto \mathbb{R}$. В общем случае путь не является кусочно-линейным и может принимать любые вещественные значения.

Требуется найти кусочно-линейный путь $\widehat{X} : [a, b] \mapsto \mathbb{R}^d$, который наилучшим образом приближает путь X в смысле минимизации метрики

$$\sup_{t \in \mathbb{R}_+} |X(t) - \widehat{X}(t)| = \sup_{t \in \mathbb{R}_+} (\sum_{i=1}^{d} |X_i(t) - \widehat{X}_i(t)|).$$

Известно, что любой непрерывный путь можно аппроксимировать сколь угодно близко кусочно-линейным путем, однако в рамках данной модели у нас есть ряд ограничений. Координаты "изломов" $\widehat{X}$ можно рассматривать только с фиксированным шагом Δt, который отвечает за то, с каким промежутком времени делаются измерения цены активов. Помимо этого, на практике возможно подбирать значения приращения каждой из координат пути $\widehat{X_i}$ только из ограниченной сетки $\{-n_i \cdot h_i, \ldots, n_i \cdot h_i\}$ с шагом h_i. Если взять слишком маленькие h_i и слишком большие n_i - то алгоритм будет сходиться слишком долго. Поэтому в зависимости от реальных целей необходимо ограничить значения h_i и n_i для каждой координаты пути. При этом при ограничении n_i накладывается ограничение на модуль производной приближаемого пути в зависимости от значения Δt: чем меньше Δt, тем пути с большей производной возможно приблизить при фиксированных h_i и n_i. Итоговое утверждение можно сформулировать следующим образом.

Теорема 4.1. *Пусть даны фиксированные числа* $h_i \in \mathbb{R}_+, n_i \in \mathbb{N}, \Delta t \in \mathbb{R}_+$. *Пусть* $X : [a,b] \mapsto \mathbb{R}^d$ *- непрерывный и кусочно-дифференцируемый путь, у которого* $\forall i : |X_i'(t)| \leq \dfrac{h_i \cdot n_i}{\Delta t}$. *Тогда существует кусочно-линейный путь* $\widehat{X} : [a,b] \mapsto \mathbb{R}^d$ *с приращениями по каждой координате* i *со значениями из множества* $\{-n_i \cdot h_i, \ldots, 0, \ldots, n_i \cdot h_i\}$ *и "изломами" в точках* $k \cdot \Delta t,\ k \in \mathbb{Z}$ *такой, что* $\forall i : \sup\limits_{t \in \mathbb{R}_+} |X_i(t) - \widehat{X}_i(t)| \leq h_i \cdot (2n_i + \frac{1}{2})$.

Доказательство. Построим у искомого пути $\widehat{X}_i$ значения в узловых точках с координатами $k \cdot \Delta t$ следующим образом: возьмем значения пути X_i в этих точках и выберем к ним ближайшие точки со значениями из сетки $\{h_i \cdot l,\ l \in \mathbb{Z}\}$.

Таким образом, из построения $\widehat{X}_i$ получаем, что $|X_i(k \cdot \Delta t) - \widehat{X_i}(k \cdot \Delta t)| \leq \dfrac{h_i}{2}$, так как в узловых точках значение пути $X_i(k \cdot \Delta t)$ может лежать только в окрестности $[-\dfrac{h_i}{2}, \dfrac{h_i}{2}]$ значения пути $\widehat{X_i}(k \cdot \Delta t)$ (иначе в этой точке можно было бы увеличить или уменьшить значение подбираемого пути $\widehat{X_i}$ так, чтобы значение исходного пути попало бы в эту окрестность его значения).

На каждом из отрезков $[k \cdot \Delta t, (k+1) \cdot \Delta t]$ по теореме Лагранжа для каждой координаты пути X_i имеем:

$$X_i(t) = X_i(k \cdot \Delta t) + \xi \cdot (t - k \cdot \Delta t),$$
$$\widehat{X}_i(t) = \widehat{X_i}(k \cdot \Delta t) + \widehat{\xi} \cdot (t - k \cdot \Delta t),$$

где ξ и $\widehat{\xi}$ - значения производных $X_i(t)$ и $\widehat{X}_i(t)$ соответственно в некоторых точках на отрезке $[k \cdot \Delta t, t]$.

Имеем $|\xi| \leq \dfrac{h_i \cdot n_i}{\Delta t}$ по формулировке предложения.

Также $|\widehat{\xi}| \leq \dfrac{h_i \cdot n_i}{\Delta t}$, потому что приращения $\widehat{X}_i(t)$ на отрезках длиной Δt берутся из множества $\{-n_i \cdot h_i, \ldots, n_i \cdot h_i\}$, поэтому неравенство следует из определения производной как тангенса угла наклона касательной.

Из построения $\widehat{X}_i$ имеем, что $|X_i(k \cdot \Delta t) - \widehat{X_i}(k \cdot \Delta t)| \leq \dfrac{h_i}{2}$.

Итого, получаем:

$$\begin{aligned} |X_i(t) - \widehat{X}_i(t)| &\leq |X_i(k \cdot \Delta t) - \widehat{X_i}(k \cdot \Delta t)| + |t - k \cdot \Delta t||\xi - \widehat{\xi}| \\ &\leq \frac{h_i}{2} + \Delta t \cdot 2\frac{h_i n_i}{\Delta t} = h_i \cdot (2n_i + \frac{1}{2}). \end{aligned}$$

□

Замечание 4.2. Если провести достаточное количество итераций генерирования приращений каждой координаты пути $\widehat{X}_i$ из арифметического распределения со значениями из соответствующего множества $\{-n_i \cdot h_i, \ldots, n_i \cdot h_i\}$, то с вероятностью, равной единице, в один момент будет получен оптимальный путь $\widehat{X}$.

Замечание 4.3. Итоговую метрику разности исходного пути и его приближаения можно оценить следующим образом:

$$\sup_{t\in\mathbb{R}_+} |X(t) - \widehat{X}(t)| = \sup_{t\in\mathbb{R}_+} (\sum_{i=1}^{d} |X_i(t) - \widehat{X}_i(t)|) \leq d \cdot \max_i \left(h_i \cdot (2n_i + \frac{1}{2}) \right)$$

Перейдём к практическому применению данной теории. Необходимо оптимальным образом подобрать параметры h_i и n_i для каждой координаты пути. Параметр h_i, отвечающий за окрестность, в пределах которой мы отождествляем значения приращений, подбирается, исходя из реально поставленной задачи -в пределах какой погрешности изменения цены не имеют значения для этого актива.

Параметр n_i, отвечающий за то, какой максимальный скачок может быть сделан за один шаг времени (ведь $|h_i \cdot n_i|$ - это модуль максимально возможного приращения, при этом h_i уже фиксировано), подбирается тоже исходя из реально поставленной задачи - насколько сильно мы можем ограничить класс генерируемых путей. Если выбрать значение параметра слишком большим, то подбор будет идти путями со слишком большим размахом приращений, что затрудняет сходимость алгоритма, а если сделать его слишком маленьким, то теряется возможность приблизить пути с большими приращениями. Для помощи в подборе этого параметра предлагается посчитать среднее и максимальное приращение по каждой из координат.

После того, как подобраны нужные распределения и их параметры, подбираем оптимальным образом количество путей в популяции и количество итераций алгоритма эволюции. Эти параметры необходимо подбирать в зависимости от допустимых ограничений по времени работы алгоритма: чем они больше, тем дольше будет выполняться обращение одной логарифмической сигнатуры.

Далее у каждого пути из текущей популяции считается логарифмическая сигнатура и смотрится метрика *loss* разницы исходной логарифмической сигнатуры и посчитанной, по этой метрике пути сортируются и выбирается настраиваемый процент лучших.

Теперь лучшие пути нужно размножить до исходного размера популяции. Для этого, пока не получено изначальное количество путей, из лучших случайно выбираются два различных пути, после чего вызывается оператор их сложения, который заключается в том, что с вероятностью 0.5 берутся приращения одного или другого пути. После этого дополнительно вызывается функция “мутации” полученного пути. В процессе этой функции настраиваемый процент приращений заново определяется случайным образом из заданных в модели распределений. После этого считается путь как последовательная сумма его приращений, и к нему применяется заданное в модели преобразование временного ряда.

Этот процесс повторяется с новой популяцией заданное количество итераций. В итоге модель возвращает путь с логарифмической сигнатурой, наиболее близкой к исходной.

Для проверки работы модели обращения логарифмической сигнатуры возьмём любой путь, посчитаем его логарифмическую сигнатуру, передадим в модель обращения и сравним исходный путь и полученный (см. рис. 6).

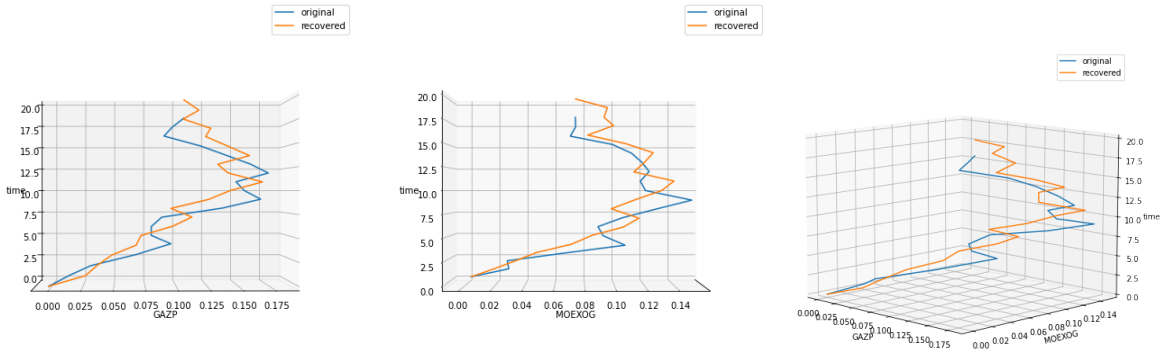

Рис. 6: Сравнение оригинального пути и пути, восстановленного по его сигнатуре.

4.4 Генерация рынка

Запустив по полученным в первом пункте логарифмическим сигнатурам модель обращения логарифмических сигнатур из второго пункта, получаем следующий набор сгенерированных путей: красным цветом нарисованы сгенерированные пути, синим - исходные (см. рис. 7).

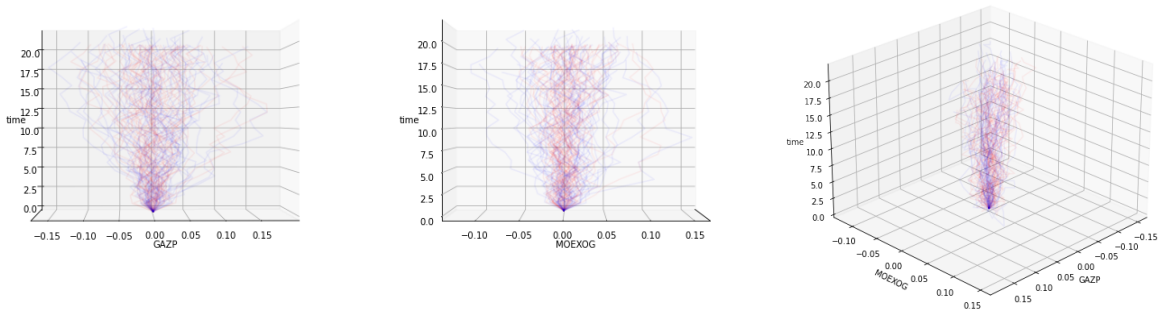

Рис. 7: Оригинальные и сгенерированные совместные траектории процессов цен активов GAZP и MOEXOG.

После того, как мы получили сгенерированные пути, необходимо непроверить, что мы получили пути из того же распределения, что исходный набор. Для этого предлагается использовать тест с метрикой максимального среднего расхождения (MMD), использующий сигнатуры путей [13]. Тест основан на сигнатурных методах, поэтому он не требует доработок под многомерный случай. Данный тест показал, что исходные и полученные данные действительно из одного вероятностного распределения.

5 Заключение

В работе приведен обзор теории сигнатурных методов и представлены примеры их применения для практических задач. Также представлен обзор различных генеративных моделей и областей их применения, в частности, генеративной модели, предложенной в работе [1], основанной на сигнатурных методах и работающей на маленьком количестве данных.

Результатом работы является обобщение предложенного алгоритма, которое состоит в обобщении модели на многомерный случай и расширении её области применения. Помимо генеративной модели, значительно улучшен был и алгоритм обращения логарифмической сигнатуры. Данные модели были успешно протестированы на реальных котировках активов Московской биржи - получены сгенерированные пути, статистически неотличимые от исходного набора совместных траекторий процессов

цен активов.

Можно выделить несколько дальнейших направлений исследования, относящихся к результату данной работы. Во-первых, было бы интересно исследовать различные алгоритмы обращения сигнатур - в общем случае это является нерешённой задачей, однако интересны практические подходы к её решению в частных, практических случаях. Во-вторых, представляют интерес возможные улучшения модели генерации сигнатур - добавление в качестве условий новых параметров, хорошо описывающих рынок, а также улучшение её архитектуры. Наконец, можно рассмотреть влияние различных преобразований данных на итоговое качество модели и сформулировать, в каких ситуациях какие преобразования являются оптимальными.

Список литературы

[1] H. Buhler, B. Horvath, T. Lyons, I. P. Arribas, B. Wood (2020). A data-driven market simulator for small data environments. arXiv:2006.14498.

[2] K.-T. Chen (1957). Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Ann. of Math. (2)*, 65:163–178.

[3] K.-T. Chen (1958). Integration of paths – a faithful representation of paths by non-commutative formal power series. *Trans. Amer. Math. Soc.*, 89:395–407.

[4] T. J. Lyons (1998). Differential equations driven by rough signals. *Rev. Mat. Iberoamericana* 14(2):215–310.

[5] T.J. Lyons, M. Caruana, T. Levy (2007). Differential equations driven by rough paths. *Lecture Notes in Mathematics*, vol. 1908. Springer.

[6] P.K. Friz, M. Hairer (2020). A course on rough paths. With an introduction to regularity structures (2nd edition). Springer.

[7] H. Boedihardjo, X. Geng, T. Lyons, D. Yang (2014). The signature of a rough path: Uniqueness. arXiv:1406.7871.

[8] B. Hambly, T. Lyons (2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. of Math. (2)*, 171(1):109–167.

[9] B. Graham (2013). Sparse arrays of signatures for online character recognition. arXiv:1308.0371.

[10] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, T.J. Lyons (2019). The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. IEEE Conference *2019 Computing in Cardiology*.

[11] I. Chevyrev, A. Kormilitzin (2016). A Primer on the Signature Method in Machine Learning. arXiv:1603.03788.

[12] Программный репозиторий market_simulator. https://github.com/imanolperez/market_simulator

[13] I. Chevyrev, H. Oberhauser (2018). Signature moments to characterize laws of stochastic processes. arXiv.org:1810.10971.

[14] George E. P. Box et al. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons