

# ỨNG DỤNG QUÁ TRÌNH NGẪU NHIÊN VÀ HỌC MÁY PHÁT TRIỂN MÔ HÌNH DỰ BÁO THỊ TRƯỜNG TÀI CHÍNH

Nguyễn Thành Trung  
Học viện Công nghệ Bưu chính Viễn thông

**Tóm tắt:** Nghiên cứu này khám phá ứng dụng các phương pháp ngẫu nhiên và học máy trong việc xây dựng mô hình dự báo thị trường tài chính, với trọng tâm là phân tích số lệnh giao dịch trên thị trường chứng khoán Việt Nam. Bằng cách kết hợp các kỹ thuật tiên tiến của giải tích ngẫu nhiên hiện đại như lý thuyết dấu ấn, lý thuyết đường nhám, các phương pháp thống kê toán học và học máy, nghiên cứu nhằm khám phá mối quan hệ thống kê giữa quá trình giá và quá trình dấu ấn từ số lệnh. Mục tiêu chính là xây dựng các phương trình định lượng mới giúp mô tả chính xác hơn cơ chế ảnh hưởng của thông tin số lệnh đến biến động giá chứng khoán. Kết quả chính của bài báo là xây dựng một mô hình phụ thuộc giữa giá chứng khoán và quá trình dấu ấn sinh bởi quá trình số lệnh.

**Từ khóa:** quá trình ngẫu nhiên, lý thuyết dấu ấn, lý thuyết đường nhám, giải tích ngẫu nhiên, học máy.

## I. MỞ ĐẦU

Trong bối cảnh toàn cầu hóa và sự phát triển nhanh chóng của khoa học và công nghệ, việc dự báo và phân tích thị trường chứng khoán đã trở thành một lĩnh vực nghiên cứu quan trọng và cấp thiết. Thị trường chứng khoán Việt Nam, với những bước phát triển năng động và tiềm năng to lớn, đang ngày càng thu hút sự quan tâm của các nhà đầu tư, nhà nghiên cứu và các tổ chức tài chính. Tuy nhiên, sự phức tạp và tính không ổn định của thị trường đòi hỏi các phương pháp phân tích tiên tiến, vượt ra ngoài các mô hình truyền thống vốn hướng đưa các thông tin về dữ liệu tài chính trong đó có dữ liệu số lệnh vào các sigma trường thông tin để nghiên cứu, thay vì tích hợp nó như một quá trình có tác động trực tiếp trong phương trình giá cổ phiếu.

Trong những năm gần đây, sự phát triển của các ngành khoa học như toán học và khoa học máy tính đã mở ra những phương pháp nghiên cứu mới trong lĩnh vực tài chính. Các kỹ thuật hiện đại trong lý thuyết quá trình ngẫu nhiên, giải tích ngẫu nhiên như lý thuyết dấu ấn và lý thuyết đường nhám đang dần thay đổi cách chúng ta tiếp cận và hiểu các quy luật vận động của thị trường tài chính. Đặc biệt, việc phân tích số lệnh giao dịch theo thời gian thực - nguồn thông tin

phong phú và chi tiết về trạng thái thị trường - đã trở thành một hướng nghiên cứu đầy triển vọng.

Bài báo này tập trung vào việc ứng dụng các lý thuyết dấu ấn, lý thuyết đường nhám và học máy để phát triển mô hình dự báo cho thị trường chứng khoán, với trọng tâm là nghiên cứu sâu về thị trường chứng khoán Việt Nam. Mục tiêu của chúng tôi là xây dựng một mô hình toán học phân tích chuyên sâu, có khả năng nắm bắt các mối quan hệ phức tạp và đa chiều trong quá trình hình thành giá chứng khoán.

Frédéric Abergel, Aymen Jedidi năm 2013 [9] đưa ra các cách tiếp cận toán học cho dữ liệu số lệnh. Với mong muốn thu hẹp khoảng cách giữa mô tả vi mô về sự hình thành giá và phương pháp Phương trình vi phân ngẫu nhiên.

Năm 2019, Oomen Roel [10] phát triển ‘Dấu ấn giá’ (price signatures) để hình thành các phép đo thống kê nhằm phát hiện các mô hình có hệ thống trong động thái giá được tập trung xung quanh điểm thực hiện giao dịch. Roel cố gắng khám phá động lực hệ thống trong thị trường tài chính thường bị che lấp sâu trong lượng dữ liệu khổng lồ.

Futter Owen and Horvath Blanka and Wiese Magnus năm 2023 [2], xây dựng khung tối ưu hóa danh mục đầu tư, trong đó việc sử dụng ‘dấu ấn đường nhám’ (rough path signatures) cung cấp một phương pháp mới để kết hợp các phụ thuộc đường dẫn dấu ấn trong động lực tài sản tín hiệu chung, mở rộng một cách tự nhiên các mô hình truyền thống.

Tháng 9 năm 2023, T. Sidogi, W. T. Mongwe, R. Mbuvha, P. Olukanmi and T. Marwala [11] khẳng định dữ liệu số lệnh là rất quan trọng trong tài chính hiện đại thông qua một loạt nghiên cứu trước đó. và họ khẳng định là những người tiên phong đưa ra một cách tiếp cận mới để dự đoán giá cổ phiếu trung bình bằng cách sử dụng dữ liệu số lệnh tần số cao và phức tạp làm đầu vào cho thuật toán học máy. Cụ thể, phương pháp được đề xuất sử dụng lý thuyết đường nhám để trích xuất các đặc điểm đường dẫn dấu ấn từ dữ liệu số lệnh và nén chúng để huấn luyện các mô hình học máy.

Tổng hợp lại các nghiên cứu trước đây:

- Các nghiên cứu về thị trường chứng khoán tại Việt Nam thường sử dụng dữ liệu về giá đóng cửa với tần suất là hằng ngày. Chưa sử dụng các dữ liệu với thời gian thực và các thông tin phản ánh chi tiết về giao dịch thời gian thực như dữ liệu số lệnh.

Tác giả liên hệ: Nguyễn Thành Trung,

Email: trungnt@ptit.edu.vn

Đến tòa soạn: 02/12/2024, chỉnh sửa: 20/01/2025, chấp nhận đăng: 08/02/2025.

• Các mô hình giá cổ phiếu truyền thống trên thế giới thông thường đưa các thông tin về số lệnh vào các sigma trường thông tin để nghiên cứu, thay vì tích hợp nó như một quá trình có tác động trực tiếp trong phương trình giá cổ phiếu.

Trong 10 năm gần đây, nhờ các tiến bộ trong lý thuyết giải tích ngẫu nhiên, đặc biệt là lý thuyết dấu ấn và lý thuyết đường nhám, đã giúp xem xét quá trình giá chứng khoán như một hàm phụ thuộc trực tiếp vào quá trình số lệnh. Trong công trình tiên phong, Cuchiero và Gazzani [1] đã phát triển mô hình giá chứng khoán dựa trên dấu ấn của quá trình một chiều - quá trình giá. Mở rộng kết quả này, nghiên cứu của chúng tôi xây dựng mô hình tổng quát hơn, trong đó giá chứng khoán được xác định bởi dấu ấn của quá trình nhiều chiều - quá trình số lệnh. Cách tiếp cận này không chỉ nắm bắt được động thái phức tạp của thị trường mà còn phản ánh chân thực hơn mối tương tác giữa các yếu tố của số lệnh và diễn biến giá.

Tên cơ sở đó, phần còn lại của bài báo được tổ chức như sau. Phần II của bài trình bày về cơ sở lý thuyết và xây dựng mô hình. Phần III trình bày về dữ liệu thu thập và các kết quả thực nghiệm trên dữ liệu cổ phiếu tại thị trường chứng khoán Việt Nam. Phần IV là phần kết luận của bài báo.

## II. CƠ SỞ LÝ THUYẾT VÀ XÂY DỰNG MÔ HÌNH

### A. CƠ SỞ LÝ THUYẾT

#### a) Lý thuyết dấu ấn

##### 1. Đại số tensor:

Với mỗi  $n \in \mathbb{N}_0$  xét tích tensor  $n$  lần của  $\mathbb{R}^d$  được cho bởi:

$$(\mathbb{R}^d)^{\otimes 0} := \mathbb{R}, (\mathbb{R}^d)^{\otimes n} := \underbrace{\mathbb{R}^d \otimes \dots \otimes \mathbb{R}^d}_n \quad (1)$$

Với  $d \in \mathbb{N}$ , ta định nghĩa đại số tensor mở rộng trên  $\mathbb{R}^d$  là:

$$T((\mathbb{R}^d)) := \{\mathbf{a} := (a_0, \dots, a_n, \dots) : a_n \in (\mathbb{R}^d)^{\otimes n}\} \quad (2)$$

Chúng tôi giới thiệu đại số tensor cắt ngắn bậc  $N \in \mathbb{N}$ :

$$T^{(N)}(\mathbb{R}^d) := \{\mathbf{a} \in T((\mathbb{R}^d)) : a_n = 0 \forall n > N\} \quad (3)$$

Chúng tôi định nghĩa đại số tensor cắt cụt có bậc  $N \in \mathbb{N}$  và đại số tensor bằng  $T^{(N)}(\mathbb{R}^d)$  và  $T(\mathbb{R}^d)$  tương ứng, bởi:

$$\begin{aligned} T^{(N)}(\mathbb{R}^d) &:= \{a = (a_n)_{n=0}^\infty \mid a_n \in (\mathbb{R}^d)^{\otimes n}\} \subset T((\mathbb{R}^d)) \\ &= \bigoplus_{n=0}^N (\mathbb{R}^d)^{\otimes n}, a_n \neq 0 \forall n \geq N \\ T(\mathbb{R}^d) &:= \bigoplus_{n=0}^\infty (\mathbb{R}^d)^{\otimes n} \subset T((\mathbb{R}^d)) \end{aligned} \quad (4)$$

##### 2. Khái niệm dấu ấn:

Cho một không gian xác suất với lọc  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ . Cho  $(X_t)_{t \in [0, T]}$  là một semimartingale

liên tục có giá trị  $\mathbb{R}^d$ . Dấu ấn của  $X$  là quá trình có giá trị  $T((\mathbb{R}^d))(s, t) \mapsto \mathbb{X}_{s,t}$  có các thành phần được định nghĩa đệ quy như sau:

$$\begin{aligned} \langle e_\emptyset, \mathbb{X}_{s,t} \rangle &:= 1 \\ \langle e_I, \mathbb{X}_{s,t} \rangle &:= \int_s^t \langle e_{I'}, \mathbb{X}_{s,r} \rangle \circ dX_{r,t}^I \end{aligned} \quad (5)$$

Với mỗi  $I = (i_1, \dots, i_n)$ ,  $I' = (i_1, \dots, i_{n-1})$ , và  $0 \leq s \leq t \leq T$ , trong đó ký hiệu tích phân Stratonovich.

#### b) Lý thuyết đường nhám

1. Phần này chúng tôi trình bày một giới thiệu ngắn gọn về lý thuyết đường đi nhám cho đường dẫn liên tục Hölder [3], [4]. Đối với khoảng thời gian nhỏ  $[a, b] \subset \mathbb{R}$ , định nghĩa:

$$\Delta^2([a, b]) := \{(s, t) : a \leq s \leq t \leq b\} \quad (6)$$

Cho  $C([a, b], \mathbb{R})$  biểu thị không gian của tất cả các đường đi liên tục  $y : [a, b] \rightarrow \mathbb{R}$  được trang bị chuẩn sup:

$$\|y\|_{\infty, [a, b]} := \sup_{t \in [a, b]} \|y_t\| \quad (7)$$

với  $\|\cdot\|$  là giá trị tuyệt đối. Chúng ta viết  $y_{s,t} := y_t - y_s$ .

Cho  $0 < \alpha < 1$ , chúng ta biểu thị bằng  $C^\alpha([a, b], \mathbb{R})$  không gian của các hàm liên tục Hölder với số mũ  $\alpha$  trên  $[a, b]$  được trang bị chuẩn:

$$\begin{aligned} \|y\|_{\alpha, [a, b]} &:= \|y_a\| + \|y\|_{\alpha, [a, b]} \\ &= \|y_a\| + \sup_{s < t \in [a, b]} \frac{\|y_{s,t}\|}{(t-s)^\alpha} \end{aligned} \quad (8)$$

Theo [5], [6] và [7], một đường dẫn  $x \in C^\alpha([a, b], \mathbb{R})$  có thể được nâng lên một bộ  $\mathbf{x} = (x, [\omega, dx], [x, dx]) \in C^\alpha([a, b], \mathbb{R}) \oplus C^{\alpha+\beta}([a, b]^2, \mathbb{R} \otimes \mathbb{R}) \oplus C^{2\alpha}([a, b]^2, \mathbb{R} \otimes \mathbb{R})$  đối với một đường dẫn bổ sung  $\omega \in C^\beta([a, b], \mathbb{R})$  với  $0 < \alpha, \beta \leq \frac{1}{2}$ , trong đó:

$$\begin{aligned} [\omega, dx] &\in C^{\alpha+\beta}([a, b]^2, \mathbb{R} \otimes \mathbb{R}) := \\ &\left\{ \Xi_{s,t} \in C([a, b]^2, \mathbb{R} \otimes \mathbb{R}) : \sup_{s,t \in [a, b], s \neq t} \frac{\|\Xi_{s,t}\|}{|t-s|^{\alpha+\beta}} < \infty \right\} \\ [x, dx] &\in C^{2\alpha}([a, b]^2, \mathbb{R} \otimes \mathbb{R}) := \\ &\left\{ \Xi'_{s,t} \in C([a, b]^2, \mathbb{R} \otimes \mathbb{R}) : \sup_{s,t \in [a, b], s \neq t} \frac{\|\Xi'_{s,t}\|}{|t-s|^{2\alpha}} < \infty \right\} \end{aligned}$$

Chúng được gọi là diện tích Levy nếu chúng thỏa mãn đẳng thức Chen, tức là đối với tất cả  $a \leq s \leq u \leq t \leq b$ :

$$\begin{aligned} \delta[\omega, dx]_{s,u,t} &= [\omega, dx]_{s,t} - [\omega, dx]_{s,u} - [\omega, dx]_{u,t} = \omega_{s,u} x_{u,t} \\ \delta[x, dx]_{s,u,t} &= [x, dx]_{s,t} - [x, dx]_{s,u} - [x, dx]_{u,t} = x_{s,u} x_{u,t} \end{aligned}$$

#### 2. Tích phân nhám

Cho  $y \in C^\beta([a, b], \mathbb{R})$  và  $x \in C^\nu([a, b], \mathbb{R})$  với  $\beta + \nu > 1$ , tích phân Young  $\int_a^b y_t dx_t$  [3] có thể được định nghĩa như sau:

$$\int_a^b y_s dx_s := \lim_{|\Pi[a,b]| \rightarrow 0} \sum_{[u,v] \in \Pi[a,b]} y_u x_{u,v} \quad (10)$$

trong đó giới hạn được lấy qua tất cả các phân hoạch hữu hạn  $\Pi[a, b]$  của  $[a, b]$ . Tích phân này thỏa mãn tính cộng tính do cấu trúc, cũng như bất đẳng thức Young Loeve [3]:

$$\left\| \int_s^t y_u dx_u - y_s x_{s,t} \right\| \leq K_{\beta, \nu} |t - s|^{\beta + \nu} \|y\|_{\beta, [s,t]} \|x\|_{\nu, [s,t]} \quad (11)$$

với tất cả  $[s, t] \subset [a, b]$ , trong đó  $K_{\beta, \nu} = (1 - 2^{1-\beta-\nu})^{-1}$ . Tiếp theo, chúng ta giới thiệu cách xây dựng tích phân nhám trong trường hợp  $0 < \beta, \alpha < \frac{1}{2}$  với điều kiện:

$$3\alpha > \beta + 2\alpha > 1 \quad (12)$$

Theo Gubinelli [6], một đường  $y \in C^\beta([a, b], \mathbb{R})$  được gọi là bị điều khiển bởi:  $(\omega, x)^T \in C^\beta([a, b], \mathbb{R}) \otimes C^\alpha([a, b], \mathbb{R})$  nếu tồn tại một bộ  $(\partial_\omega y, \partial_x y, R_y)$  với  $\partial_\omega y \in C^\alpha([a, b], \mathcal{L}(\mathbb{R}, \mathbb{R}))$ ,  $\partial_x y \in C^\beta([a, b], \mathcal{L}(\mathbb{R}, \mathbb{R}))$  và  $R_y \in C^{\alpha+\beta}([a, b]^2, \mathbb{R})$  sao cho:

$$y_{s,t} = (\partial_\omega y)_s \omega_{s,t} + (\partial_x y)_s x_{s,t} + R_{y,s,t} \quad \forall a \leq s \leq t \leq b.$$

$\partial_\omega y$  và  $\partial_x y$  được gọi là các đạo hàm Gubinelli riêng phần của  $y$  đối với  $\omega$  và  $x$  tương ứng, và chúng được xác định duy nhất khi  $\omega$  và  $x$  là các đường nhám thực sự, nghĩa là  $\omega \in C^\beta([a, b], \mathbb{R}) \setminus C^{2\beta}([a, b], \mathbb{R})$  và  $x \in C^\alpha([a, b], \mathbb{R}) \setminus C^{2\alpha}([a, b], \mathbb{R})$

Áp dụng bổ đề khâu (sewing lemma), tích phân  $\int_s^t y_u dx_u$  có thể được định nghĩa nhờ định lý tham khảo [6], và [8], với bất kỳ  $a \leq s \leq t \leq b$ , như sau:

$$\int_s^t y_u dx_u := \lim_{|\mathcal{P}[s,t]| \rightarrow 0} \sum_{[u,v] \in \mathcal{P}[s,t]} (y_u x_{u,v} + (\partial_\omega y)_u [\omega, dx]_{u,v} + (\partial_x y)_u [x, dx]_{u,v}) \quad (13)$$

trong đó giới hạn được lấy qua tất cả các phân hoạch hữu hạn  $\mathcal{P}[s, t]$  của  $[s, t]$  và do đó không phụ thuộc vào  $\mathcal{P}$ . Tích phân nhám này là cộng tính, tức là:

$$\int_s^t y dx = \int_s^\tau y dx + \int_\tau^t y dx \quad (14)$$

với mọi  $s \leq \tau \leq t$ .

#### c) Mô hình học máy

Hồi quy Lasso (Least Absolute Shrinkage and Selection Operator): là một phương pháp hồi quy tuyến tính với ràng buộc, nhằm tối thiểu hóa hàm mục tiêu:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \quad (15)$$

Hồi quy Ridge là một phương pháp hồi quy tuyến tính với ràng buộc, nhằm tối thiểu hóa hàm mục tiêu:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} \quad (16)$$

Trong đó:

- $y$  là vector kết quả
- $X$  là ma trận đặc trưng hay ma trận dữ liệu.
- $\beta$  là vector hệ số cần ước lượng.
- $\lambda > 0$  là tham số điều chỉnh.
- $\|\cdot\|_1$  và  $\|\cdot\|_2$  lần lượt là chuẩn L1 và L2.

#### d) Các mô hình giá cổ phiếu

Mô hình giá cổ phiếu Samuelson [3]:

$$dS_t = \mu S_t dt + \sigma S_t dB_t \quad (17)$$

Đối với giá cổ phiếu  $S_t$  tại thời điểm  $t$  với hệ số tăng trưởng  $\mu$ , độ biến động  $\sigma$ , và tích phân ngẫu nhiên theo nghĩa của Itô liên quan đến chuyển động Brown chuẩn  $B_t$ , chưa phù hợp với một số đặc điểm khá phổ biến của dữ liệu giá cổ phiếu theo thực nghiệm; do đó, nhiều sửa đổi đã được đề xuất kể từ đó. Đầu tiên, mô hình Hull và White [12] cho rằng hệ số tăng trưởng và độ biến động phải phụ thuộc vào thời gian và ngẫu nhiên, dẫn đến mô hình có dạng:

$$dS_t = \mu(t) S_t dt + \sigma(t) S_t dB_t \quad (18)$$

trong đó  $\mu: [0, T] \rightarrow \mathbb{R}$  được coi là tích phân Lebesgue và xác định dễ đơn giản, và  $\sigma(t)$  thỏa mãn một phương trình vi phân ngẫu nhiên:

$$d \log \sigma(t) = k(\theta - \log \sigma(t)) dt + \gamma dW_t \quad (19)$$

với các tham số  $k, \theta, \gamma$ , và một chuyển động Brown khác  $W_t$ , với một tương quan tức thời  $dB_t dW_t = \rho dt$  với một tham số  $\rho \in [0, 1]$  giữa hai chuyển động Brown khác nhau. Mô hình Heston [13] đề xuất rằng  $\sigma(t) = \sqrt{v/t}$  trong đó  $v_t$  tuân theo một phương trình Cox - Ingersoll Ross:

$$dv_t = k[\theta - v_t] dt + \lambda \sqrt{v_t} dW_t \quad (20)$$

ở đây  $k, \theta, \lambda > 0$  là các tham số và hai chuyển động Brown một lần nữa sở hữu mối tương quan tức thời  $dB_t dW_t = \rho dt$  với  $\rho \in [0, 1]$ .

#### B. XÂY DỰNG MÔ HÌNH

Chúng tôi xem xét mô hình dựa trên dấu ấn với mục tiêu cung cấp một mô hình dựa trên dữ liệu, phổ quát cho một tập hợp các tài sản giao dịch  $S$ . Để minh họa, chúng

tôi sẽ giả định rằng  $S$  là một chiều. Thành phần chính trong mô hình của chúng tôi là một quá trình cơ bản  $X$  có số chiều là  $d$  và  $X = (X_t^1, \dots, X_t^d)_{t \geq 0}$ , được mở rộng với thời gian  $t$ , trong đó  $X$  được cho là một semimartingale liên tục. Chúng tôi sẽ ký hiệu quá trình mở rộng với thời gian này là  $(\tilde{X}_t)_{t \geq 0} = (t, X_t^1, \dots, X_t^d)_{t \geq 0}$  và giả định rằng dấu ấn của  $\tilde{X}$  là cơ sở hồi quy cho  $S$ .  $S$  được mô hình hóa/xấp xỉ thông qua một quá trình  $S_n(\ell)$  [1] được định nghĩa như sau:

$$S_n(\ell)_t = \ell(\tilde{X}_t) \quad (21)$$

trong đó  $\ell$  là một ánh xạ tuyến tính của dấu ấn  $\tilde{X}$  tới bậc  $n \in \mathbb{N}$  và được suy ra từ dữ liệu.

Việc xây dựng mô hình tùy thuộc vào việc lựa chọn quá trình  $X$ . Nếu  $X$  là một vector của các chuyển động Brownian, trong đó các quỹ đạo được trích xuất từ các quan sát của  $S$ . Chúng tôi xây dựng động học của quá trình giá tài sản được mô tả bởi mô hình biến động ngẫu nhiên dưới độ đo  $\mathbb{P}$ :

$$\begin{aligned} dS_t &= \mu(S_t, V_t)dt + g(S_t, V_t)dB_t^P \\ dV_t &= h(S_t, V_t)dt + \sigma(S_t, V_t)dW_t^P \end{aligned} \quad (22)$$

với  $d[B_t^P, W_t^P] = \rho dt$ , trong đó  $\rho \in [-1, 1]$  và các hàm  $\mu, h, g, \sigma$  là từ  $\mathbb{R}^2$  đến  $\mathbb{R}$ , sao cho  $g$  và  $\sigma$  luôn dương.  $S_t$  và  $V_t$  là các dữ liệu thuộc quá trình số lệnh, ví dụ như cặp (Giá và Cung-cầu thị trường) hay (Giá và Khối lượng), hoặc (Thanh khoản và Cung-cầu thị trường).

Ta có thể ước lượng các quỹ đạo của quá trình chính  $X_t := (B_t^Q, W_t^Q)$  bằng:

$$\begin{aligned} B_t^Q &= \int_0^t \frac{\mu(S_s, V_s)}{g(S_s, V_s)} ds + B_t^P \\ W_t^Q &= \int_0^t \frac{h(S_s, V_s)}{\sigma(S_s, V_s)} ds + W_t^P \end{aligned} \quad (23)$$

Với điều kiện tích phân đủ lớn trên giá thị trường của rủi ro,  $\mathbb{Q}$  là một độ đo martingale địa phương tương đương, trong đó  $(B_t^Q, W_t^Q)$  là các chuyển động Brownian theo  $\mathbb{Q}$ , và động học của  $(S, V)$  dưới độ đo  $\mathbb{Q}$  là:

$$\begin{aligned} dS_t &= g(S_t, V_t)dB_t^Q \\ dV_t &= \sigma(S_t, V_t)dW_t^Q \end{aligned} \quad (24)$$

Để ngoại suy các quỹ đạo của  $(B^Q, W^Q)$ , chúng ta ước tính biến động bình phương điểm của quá trình giá và quá trình biến động, để có được một ước lượng của  $\Sigma_{11} := g(S_t, V_t)^2$  và  $\Sigma_{22} := \sigma(S_t, V_t)^2$ , từ đó cho phép tính toán:

$$dB_t^Q = \frac{dS_t}{\sqrt{\Sigma_{11,t}}}, \quad dW_t^Q = \frac{dV_t}{\sqrt{\Sigma_{22,t}}} \quad (25)$$

Sau khi trích xuất quỹ đạo của  $B^Q$  và  $W^Q$ , chúng ta có thể tính toán quỹ đạo của dấu ấn mở rộng với thời gian của chúng.

Ta có theo [1]:

$$[S]_t \approx \int_0^t \left( \ell_0^* + \sum_{0 < |I| \leq n} \ell_I^* \langle e_I, \tilde{X}_s \rangle \right)^2 ds \quad (26)$$

Vì theo định nghĩa  $S_t = \int_0^t g(S_s, V_s)dB_s^Q$ ,  $g(S_t, V_t) > 0$ , và  $[B^Q]_t = t$ , theo [1] điều này dẫn đến:

$$S_t \approx S_0 + \int_0^t \left( \ell_0^* + \sum_{0 < |I| \leq n} \ell_I^* \langle e_I, \tilde{X}_s \rangle \right) dB_s^Q \quad (27)$$

Theo [1] có thể viết lại phương trình trên như sau:

$$S_t \approx S_0 + \ell_0^* B_t^Q + \sum_{0 < |I| \leq n} \ell_I^* \langle \bar{e}_I, \tilde{X}_t \rangle \quad (28)$$

Với tham số  $\ell^*$  thu được qua quá trình học từ dữ liệu.

### III. KẾT QUẢ MÔ HÌNH

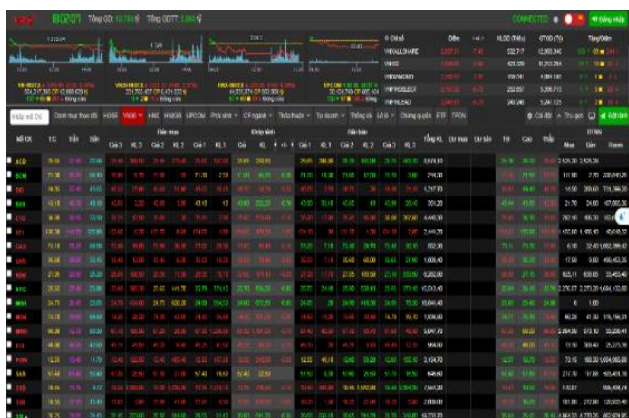
#### A. Dữ liệu

Dữ liệu trong nghiên cứu này được thu thập thông qua các giao thức kết nối để lấy dữ liệu theo thời gian thực từ trang web chính thức của công ty chứng khoán VPS. Cụ thể, các công cụ và thư viện như requests, BeautifulSoup, và Selenium được sử dụng để truy xuất dữ liệu số lệnh và dữ liệu giao dịch theo thời gian thực.

Dữ liệu nghiên cứu bao gồm 30 mã cổ phiếu thuộc rõ VN-30, đại diện cho các công ty lớn nhất trên sàn HOSE. Dữ liệu được thu thập từ ngày 25/02/2024 đến 15/08/2024, bao gồm giá cổ phiếu, khối lượng giao dịch và các mức giá, khối lượng đặt mua/bán theo thời gian thực trong các phiên giao dịch. Cụ thể, dữ liệu được lưu trữ dưới dạng bảng với 20 cột, mô tả chi tiết về từng mã cổ phiếu tại các thời điểm khác nhau.

- Mã CP: Mã cổ phiếu của công ty phát hành.
- Giá tham chiếu: Giá mở cửa.
- Giá Trần/Sàn: Giá cao nhất/thấp nhất trong phiên.
- Giá/KL mua 1, 2, 3: Giá và khối lượng đặt mua cao nhất, thứ hai, thứ ba.
- Giá/KL bán 1, 2, 3: Giá và khối lượng chào bán thấp nhất, thứ hai, thứ ba.
- Giá/KL khớp lệnh: Giá và khối lượng giao dịch thực tế.
- Tổng khối lượng: Tổng khối lượng cổ phiếu giao dịch trong phiên.
- Thời gian: Thời điểm ghi nhận dữ liệu.

Mỗi bản ghi đại diện cho một mã cổ phiếu tại một thời điểm cụ thể trong phiên giao dịch, cho phép phân tích sự biến động giá và khối lượng, cũng như hành vi của nhà đầu tư qua từng thời điểm.



Hình 1: Dữ liệu được lấy qua API công ty VPS.



Hình 2: Dữ liệu giá và khối lượng cổ phiếu FPT theo thời gian thực.

**B. Kết quả tính toán dấu ấn của quá trình số lênh**

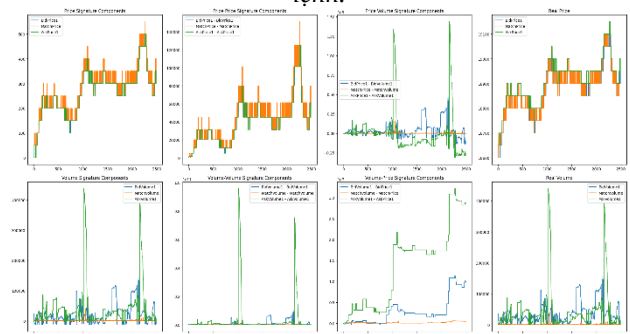
Lý thuyết dấu ấn cho phép mở rộng từ một tập dữ liệu nhỏ thành nhiều thành phần có ý nghĩa toán học riêng, giúp mô tả mối quan hệ giữa các biến. Với 6 đặc trưng đầu vào (Giá mua 1, Khối lượng mua 1, Giá khớp lệnh, Khối lượng khớp lệnh, Giá bán 1, Khối lượng bán 1), chúng tôi đã tạo ra 42 thành phần dấu ấn. Mỗi thành phần này mô tả sự tương tác và quan hệ giữa các đặc trưng theo thời gian, đặc biệt hữu ích trong phân tích tài chính.

42 thành phần dấu ẩn: Được tính toán từ 6 đặc trưng đầu vào, thể hiện khả năng mô tả nhiều mối quan hệ phức tạp giữa các biến số.

- **Bản chất dấu ấn:** Signature giúp mô tả sự tương tác và biến động của các biến theo thời gian, nắm bắt những phức tạp trong dữ liệu tài chính.
- **Khám phá mối quan hệ:** Ví dụ, mối quan hệ giữa Giá mua 1 và Giá bán 1 có thể phản ánh sự cân bằng giữa cung và cầu, trong khi Khối lượng và Giá khớp lệnh thể hiện ảnh hưởng của các giao dịch lên giá cả thị trường.
- **Lợi ích của thành phần dấu ấn:** Phân tích sự tương tác giữa các đặc trưng cho phép khám phá các mẫu phức tạp, cung cấp một cái nhìn chi tiết và toàn diện về diễn biến thị trường.



Hình 3: Kết quả tính toán 42 dấu ẩn từ quá trình số lệnh.



Hình 4: Kết quả phân tích sâu các dấu ấn có ý nghĩa.

### C. Ước lượng mô hình dấu ẩn

Bài toán hiện tại là tìm  $\ell^* \in \mathbb{R}^{d^*}$  sao cho:

$$\ell^* \in \arg \min_{\ell} L_{\text{price},a}(\ell),$$

trong đó hàm lỗi  $L_{\text{price},\alpha}(\ell)$  được cho bởi công thức (29) như sau:

$$\begin{aligned} L_{\text{price},\alpha}(\ell) &:= \sum_{i=1}^N (S_n(\ell)t_i - S_{t_i})^2 + \alpha(\ell) \\ &= \sum_{i=1}^N \left( S_0 + \ell_0 \langle e_1, \mathbb{X}_{t_i} \rangle + \sum_{0 \leq l \leq n-1} \ell_l \langle \tilde{e}_l, \mathbb{X}_{t_i} \rangle - S_{t_i} \right)^2 + \alpha(\ell) \end{aligned}$$

với  $\alpha$  là một hàm phạt cố định. Một ví dụ cho  $\alpha$  là tổ hợp lồi của các chuẩn  $L_1$  và  $L_2$ , tức là  $\alpha(\ell) = \lambda \|\ell\|_1 + (1 - \lambda) \|\ell\|_2$ , với  $\lambda \in [0, 1]$ . Số lượng tham số tương ứng với  $d^*$ , tức là chiều của dấu ẩn được cắt tại mức  $n - 1$ .

Độ đo sai số trung bình bình phương (MSE): Sai số trung bình bình phương là độ đo phổ biến để đánh giá sự khác biệt giữa các giá trị thực tế và giá trị dự đoán. Nó được tính bằng cách lấy trung bình bình phương của sự chênh lệch giữa các giá trị thực tế ( $y_i$ ) và giá trị dự đoán ( $\hat{y}_i$ ). Công thức toán học của MSE được cho bởi:

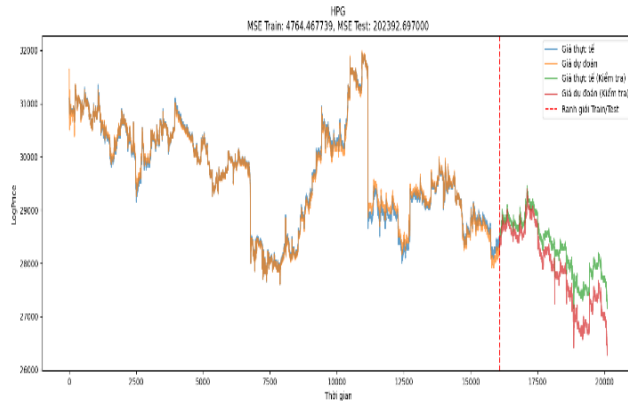
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (30)$$

Trong đó:

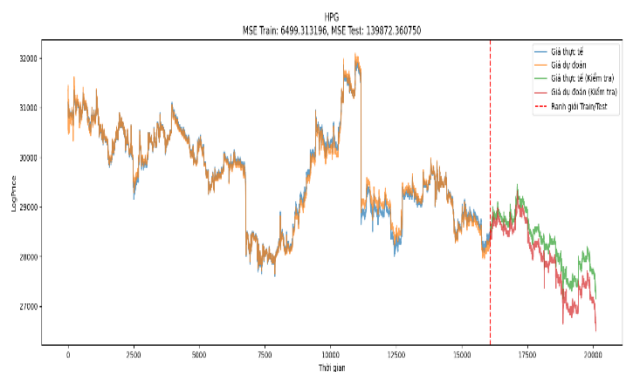
- $n$  là số lượng điểm dữ liệu.
- $y_i$  là giá trị thực tế của điểm dữ liệu thứ  $i$ .
- $\hat{y}_i$  là giá trị dự đoán của điểm dữ liệu thứ  $i$ .

#### D. Kết quả mô hình dấu ấn với quá trình số lệnh

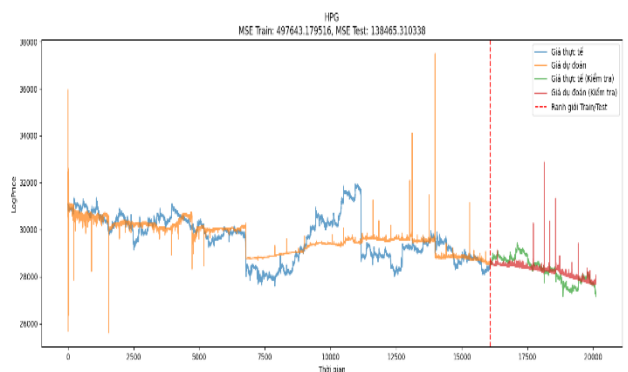
Hai mô hình Lasso và mô hình Ridge có kết quả tương đương nhau trong việc ước lượng giá cổ phiếu dựa trên mô hình dấu ấn với quá trình số lệnh.



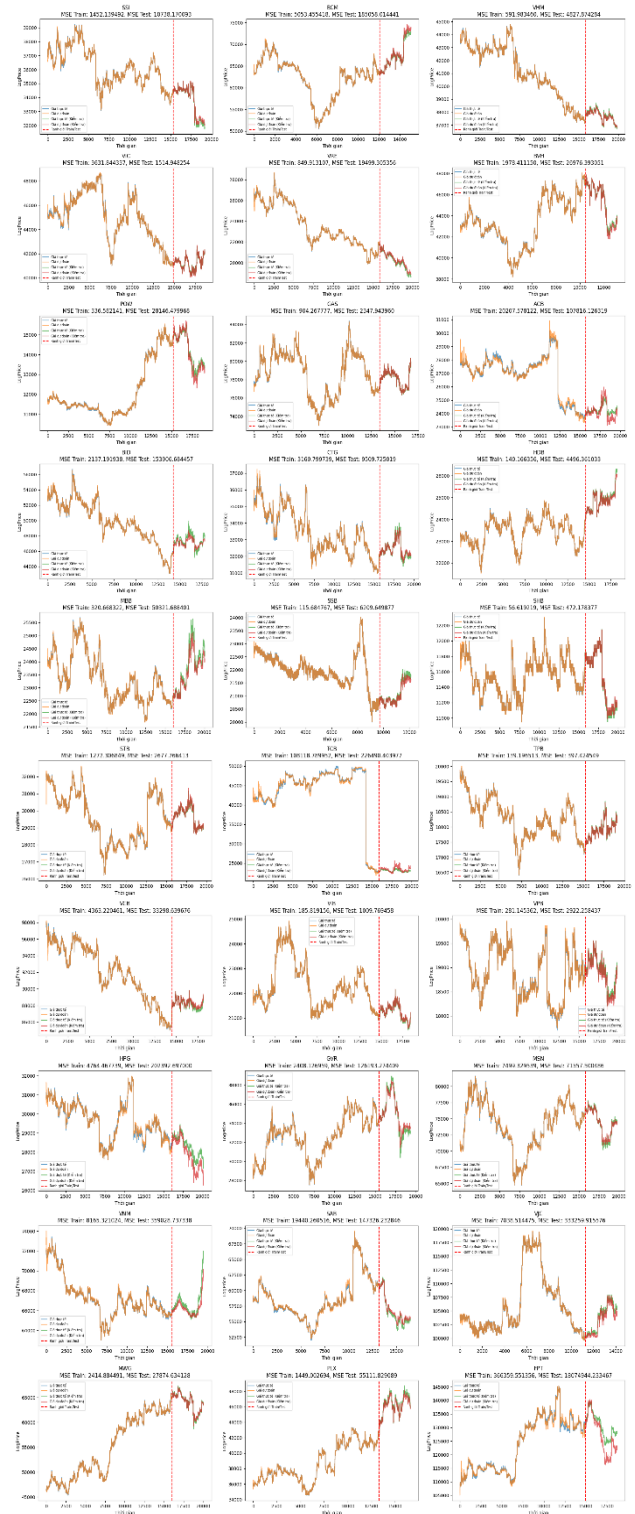
Hình 5: Kết quả với dữ liệu giá cổ phiếu và khối lượng khớp của cổ phiếu HPG



Hình 6: Kết quả với dữ liệu giá và cung-cầu thị trường cổ phiếu HPG



Hình 7: Kết quả với dữ liệu thanh khoản và cung-cầu thị trường cổ phiếu HPG



Hình 8: Kết quả với quá trình giá và khối lượng cho 30 cổ phiếu trong VN-30

#### IV. KẾT LUẬN

Nghiên cứu đã thành công trong việc ứng dụng lý thuyết dấu ấn và đường nhám vào phân tích dữ liệu số lệnh trên thị trường chứng khoán Việt Nam. Lý thuyết đường nhám và quá trình dấu ấn cho phép xử lý các dòng dữ liệu ngẫu nhiên, phù hợp với thực tế trên thị trường tài chính. Dấu ấn của đường dẫn, một khái niệm then chốt trong lý thuyết đường nhám, đã cho thấy khả năng mô tả cấu trúc của thị trường. Thông qua chuỗi các tích phân

lập của các đường đi, phương pháp này cung cấp cái nhìn sâu về cách thức biến động của thị trường theo thời gian.

Kết quả nghiên cứu đã phát triển mô hình để cải thiện các mô hình tài chính cổ điển như Black-Scholes giả định đường đi giá cổ phiếu tuân theo chuyển động Brown. Lý thuyết dấu ấn và lý thuyết đường nhám đã giúp xây dựng được các mô hình tài chính có khả năng xử lý hiệu quả tính bất thường và không chắc chắn của thị trường.

Các phương pháp và kết quả phân tích gợi ý về tiềm năng của lý thuyết dấu ấn khi được áp dụng cho dữ liệu sổ lệnh, đặc biệt trong bối cảnh giao dịch tần suất cao. Thông qua việc phân tích dữ liệu sổ lệnh và xây dựng mô hình, nghiên cứu ghi nhận khả năng dự báo khá sát với dữ liệu kiểm định, góp phần tìm hiểu mối liên hệ giữa các đặc trưng trong dữ liệu và biến động giá cổ phiếu.

Nghiên cứu này là một nỗ lực tìm hiểu thêm về cơ chế hoạt động của thị trường chứng khoán Việt Nam thông qua việc kết hợp lý thuyết dấu ấn và dữ liệu sổ lệnh. Kết quả nghiên cứu có thể đóng góp một phần vào việc hiểu rõ hơn về hành vi thị trường và hỗ trợ các quyết định đầu tư. Chúng tôi hy vọng những phát hiện từ nghiên cứu này sẽ hữu ích cho các nghiên cứu tiếp theo về thị trường chứng khoán Việt Nam.

Các hình ảnh chi tiết tính toán của nghiên cứu này được lưu trữ và cập nhật tại:

<https://github.com/trungmit13/vietnam-stock-market-analysis/blob/main/Ph%E1%BB%A5L%E1%BB%A5C%E1%BB%A5C3%ADnhTo%E1%BB%A5C3%A1n.pdf>

## TÀI LIỆU THAM KHẢO

- [1] Christa Cuchiero, Guido Gazzani, & Sara Svaluto-Ferro (2023). Signature-Based Models: Theory and Calibration. *SIAM Journal on Financial Mathematics*, 14(3), 910–957.
- [2] Owen Futer, Blanka Horvath, & Magnus Wiese (2023). Signature Trading: A Path-Dependent Extension of the Mean-Variance Framework with Exogenous Signals. *Social Science Research*.
- [3] Luu Hoang Duc & Jürgen Jost (2023). How Rough Path Lifts Affect Expected Return and Volatility: A Rough Model under Transaction Cost. *SIAM Journal on Financial Mathematics*, 14(3), 879–909.
- [4] Luu Hoang Duc & Peter Kloeden (2023). Numerical attractors for rough differential equations. *SIAM Journal on Numerical Analysis*, 61(5), 2381–2407.
- [5] P. Friz & M. Hairer (2014). A Course on Rough Path with an Introduction to Regularity Structure. *Universitext* (Vol. XIV). Springer, Berlin.
- [6] M. Gubinelli (2004). Controlling rough paths. *Journal of Functional Analysis*, 216, 86–140.
- [7] T. Lyons (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14, 215–310.
- [8] L. Coutin (2012). Rough paths via sewing lemma. *ESAIM: Probability and Statistics*, 16, 479–526.
- [9] Frédéric Abergel & Aymen Jedidi (2013). A Mathematical Approach to Order Book Modeling. *International Journal of Theoretical and Applied Finance*, 16(05).
- [10] Roel Oomen (2019). Price signatures. *Quantitative Finance*, 19(5), 733–761.
- [11] T. Sidogi, W. T. Mongwe, R. Mbuvha, P. Olukanmi, & T. Marwala (2023). A Signature Transform of Limit Order Book Data for Stock Price Prediction. *IEEE Access*, 11, 70598–70609.
- [12] J. Hull & A. White (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 3, 281–300.
- [13] S. L. Heston (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6, 327–343.

## APPLICATION OF STOCHASTIC PROCESSES AND MACHINE LEARNING IN DEVELOPING FINANCIAL MARKET PREDICTION MODELS

**Abstract:** This research explores the application of stochastic methods and machine learning in developing financial market prediction models, focusing on order book analysis in the Vietnamese stock market. By integrating advanced techniques from modern stochastic calculus such as signature theory, rough path theory, mathematical statistics, and machine learning methods, the study aims to investigate the statistical relationships between price processes and signature processes derived from the order book. The main objective is to construct new quantitative equations that more accurately describe the mechanism of how order book information influences stock price movements. The research not only maintains deep academic rigor but also provides practical analytical tools to support investors and financial institutions in making more accurate and efficient trading decisions.

**Keywords:** stochastic process, signature theory, rough path theory, stochastic calculus, machine learning



**Nguyễn Thành Trung**, Nhận học vị Thạc sĩ toán học lý thuyết – ngành lý thuyết xác suất năm 2024 tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Thạc sĩ Khoa học máy tính năm 2023 tại Đại học Công nghệ, Đại học Quốc Gia Hà Nội. Thạc sĩ Khoa học dữ liệu năm 2021 tại Đại học Khoa học Tự nhiên, Đại học Quốc Gia Hà Nội. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông.

**Lĩnh vực nghiên cứu:** Toán tài chính, học máy, xử lý ngôn ngữ tự nhiên.

**Email:** [trungnt@ptit.edu.vn](mailto:trungnt@ptit.edu.vn)