

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM
KHOA TOÁN - TIN HỌC

00



BÁO CÁO SEMINAR

MEAN REVERSION - APPLICATION
FOR MARKET MAKING ALGORITHMS

Môn học: Seminar cho Khoa học Dữ Liệu

Giảng viên hướng dẫn: TS. Tô Đức Khánh

Sinh viên thực hiện: 21280059 - Trần Thị Bích Tuyền
21280119 - Phạm Ngọc Phương Uyên
21280123 - Nguyễn Thị Lan Diệp

Lớp: 21KDL1

TPHCM, ngày 16 tháng 01 năm 2025

Lời cảm ơn

Trước tiên, chúng em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến Tiến sĩ Tô Đức Khanh – người đã tận tình hướng dẫn, góp ý và hỗ trợ chúng em trong suốt quá trình thực hiện seminar. Những kiến thức và kinh nghiệm mà thầy chia sẻ đã giúp chúng em hoàn thiện đồ án này một cách tốt nhất.

Bên cạnh đó, chúng em cũng xin cảm ơn đến các anh chị tại Công ty Vamient đã giúp chúng em có cơ hội tiếp xúc thực tiễn, học hỏi thêm kiến thức và kinh nghiệm thực tế để có thông tin hoàn thành bài báo cáo cũng như luôn sẵn sàng hỗ trợ chúng em trong suốt thời gian thực hiện seminar.

Những sự giúp đỡ và hỗ trợ quý báu này đã góp phần quan trọng để chúng em hoàn thành tốt đồ án seminar của mình.

Mặc dù bản thân đã rất cố gắng nhưng do thời gian, kiến thức và kinh nghiệm có hạn, và cũng là lần đầu tiên được tiếp xúc, làm việc thực tế với công ty nên bài nghiên cứu của chúng em còn có nhiều thiếu sót trong việc trình bày, đánh giá và đề xuất ý kiến. Chúng em rất mong nhận được sự thông cảm và đóng góp ý kiến của quý thầy cô và anh chị.

Chúng em xin chân thành cảm ơn!

Các thành viên nhóm

Mục lục

1 Giới thiệu về tài	1
1.1 Bối cảnh nghiên cứu	1
1.2 Mục tiêu nghiên cứu	1
2 Các khái niệm và định nghĩa liên quan	3
2.1 Khái niệm Mean Reversion	3
2.2 Các chỉ số	4
2.3 Lợi ích và hạn chế của Mean Reversion	7
2.4 Market Making	8
2.4.1 Định nghĩa	8
2.4.2 Thành phần trong một thuật toán market making	10
2.4.3 Ví dụ minh họa	11
2.5 Ornstein Uhlenbeck Processes	11
2.5.1 Giới thiệu	11
2.5.2 Mô phỏng theo phương pháp số	12
2.5.3 Ước lượng tham số sử dụng Maximum log-likelihood	13
2.5.4 Kalman Filter trong Ornstein-Uhlenbeck Process	14
3 Các nghiên cứu trước đây	17
4 Phương pháp nghiên cứu	18
4.1 Dữ liệu sử dụng	18
4.2 Kiểm định giả thuyết	19
5 Tạo tín hiệu giao dịch	22
5.1 Orderbook Signal	22
5.2 Mean reversion Signal	25
6 Chiến lược Market Making	29
6.1 Tính Fair Value	29
6.2 Tính giá trị Spread	29
6.3 Tìm giá Bid & Ask tối ưu	30
7 Backtesting	31
7.1 Mục tiêu	31
7.2 Quy trình Backtest	31

8 Kết quả	33
8.1 Kết quả đạt được	33
8.2 Hiệu suất các tín hiệu Order Book và Mean Reversion.	34
8.3 So sánh với các benchmark khác	36
8.4 Nhận xét chung	39
9 Kết luận	41
10 Ngoài lề	42
10.1 Ước lượng tham số MLE - cải tiến	42
10.2 Chuyển đổi tham số OU sang ma trận của Bộ lọc Kalman	44
11 Tài liệu tham khảo	46

1 Giới thiệu đề tài

1.1 Bối cảnh nghiên cứu

Mean Reversion là một lý thuyết trong tài chính, giả định rằng giá của tài sản sẽ di chuyển theo một chu kỳ và có xu hướng quay trở lại mức trung bình dài hạn của nó sau một khoảng thời gian. Trong các thị trường tài chính, các tài sản như cổ phiếu, ngoại hối, và hàng hóa thường xuyên trải qua các giai đoạn biến động, nhưng với thời gian, những biến động này có xu hướng giảm bớt và giá sẽ quay lại giá trị trung bình ban đầu. Đây là một yếu tố quan trọng trong phân tích kỹ thuật và trong việc xây dựng các chiến lược giao dịch.

Tầm quan trọng của Mean Reversion trong tài chính thể hiện ở khả năng dự đoán và tận dụng các sai lệch ngắn hạn từ mức giá trung bình. Khi giá tài sản vượt ra ngoài một mức độ nào đó so với giá trị trung bình, các nhà giao dịch có thể tận dụng sự đảo chiều của giá để tối đa hóa lợi nhuận. Vì vậy, hiểu rõ và áp dụng lý thuyết này giúp các nhà đầu tư có được những chiến lược giao dịch hiệu quả hơn.

Market Making là hoạt động cung cấp thanh khoản cho thị trường bằng cách luôn luôn sẵn sàng mua và bán tài sản với giá cả xác định. Trong chiến lược Market Making, việc dự đoán sự quay lại của giá tài sản về mức trung bình (Mean Reversion) có thể giúp xác định các cơ hội giao dịch. Nếu thị trường bị định giá quá cao hoặc quá thấp so với mức trung bình, các nhà market maker có thể tận dụng điều này để thu lợi nhuận từ sự quay lại của giá.

Bằng cách sử dụng các tín hiệu Mean Reversion, các chiến lược Market Making có thể được tối ưu hóa để mua tài sản khi giá thấp hơn mức trung bình và bán khi giá vượt quá mức trung bình, từ đó giảm thiểu rủi ro và cải thiện hiệu quả giao dịch. Do đó, Mean Reversion đóng một vai trò quan trọng trong việc xây dựng và triển khai chiến lược Market Making hiệu quả.

1.2 Mục tiêu nghiên cứu

Mục tiêu đầu tiên của nghiên cứu là kiểm định giả thuyết Mean Reversion trên các loại dữ liệu tài chính khác nhau như cổ phiếu, ngoại hối và các sản phẩm tài chính khác. Việc kiểm định này sẽ giúp xác định liệu có sự quay lại của giá tài sản về mức trung bình sau các biến động ngắn hạn và trong trường hợp nào, điều này có thể xảy ra.

Mục tiêu thứ hai là tích hợp tín hiệu từ cả Mean Reversion và Order Book (sổ lệnh) để xây dựng một chiến lược Market Making tối ưu. Dữ liệu từ Order Book cung cấp

thông tin về các lệnh mua và bán trong thị trường, giúp dự đoán cung cầu và xu hướng giá. Việc kết hợp các tín hiệu này sẽ cung cấp một cái nhìn toàn diện hơn về thị trường và giúp cải thiện chiến lược giao dịch.

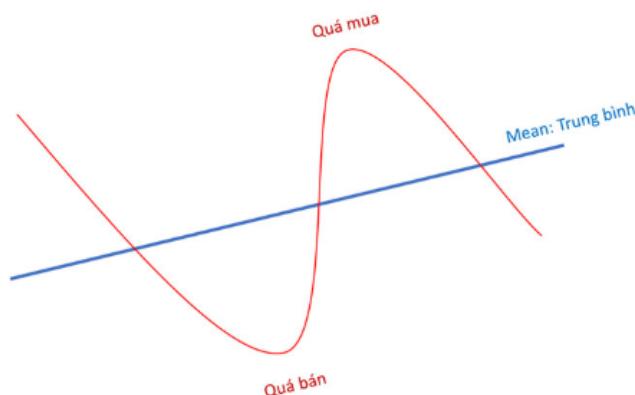
Mục tiêu cuối cùng của nghiên cứu là thực hiện đánh giá hiệu quả của chiến lược Market Making áp dụng Mean Reversion được tối ưu hóa qua Backtesting, so với các chiến lược benchmark hiện có. Việc sử dụng phương pháp Backtesting sẽ cho phép kiểm tra chiến lược trong các điều kiện thị trường khác nhau, đánh giá mức độ thành công và so sánh với các chiến lược giao dịch khác để xác định hiệu quả thực tế của chiến lược này.

2 Các khái niệm và định nghĩa liên quan

2.1 Khái niệm Mean Reversion

Mean reversion là một lý thuyết tài chính cho rằng giá của tài sản sẽ cuối cùng trở về mức trung bình hoặc giá trị trung bình dài hạn của nó. Khái niệm này dựa trên niềm tin rằng giá tài sản và lợi suất lịch sử sẽ có xu hướng hướng về một mức trung bình dài hạn theo thời gian. Sự lệch lạc càng lớn so với mức trung bình này, xác suất giá của tài sản sẽ di chuyển gần hơn đến mức trung bình trong tương lai càng cao.

Mean reversion trong giao dịch là một khái niệm đề cập đến giả định rằng giá tài sản và lợi suất lịch sử cuối cùng sẽ trở về mức trung bình dài hạn. Khái niệm này dựa trên hiện tượng thống kê được gọi là hồi quy về mức trung bình. Các nhà giao dịch sử dụng chiến lược này thường tìm kiếm cơ hội mua hoặc bán tài sản khi giá lệch đáng kể so với mức trung bình lịch sử của nó với kỳ vọng rằng giá sẽ quay trở lại mức trung bình đó.



Hình 1: Hình ảnh minh họa Mean Reversion

Lý do đứng sau lý thuyết mean reversion là theo thời gian, những giá cả bất thường cao hoặc thấp sẽ có xu hướng trở về các giá trị trung bình của chúng. Sự chuẩn hóa này có thể do nhiều yếu tố khác nhau, bao gồm sự thay đổi trong tâm lý thị trường, các yếu tố kinh tế, hoặc đơn giản là những biến động ngẫu nhiên xảy ra trong các thị trường tài chính.

Ví dụ: Nếu giá cổ phiếu đã dao động quanh mức 50 đô la nhưng đột ngột tăng vọt lên 80 đô la mà không có sự thay đổi đáng kể nào trong các yếu tố cơ bản của công ty, một nhà giao dịch theo chiến lược mean reversion có thể xem đây là một cơ hội để bán cổ phiếu, kỳ vọng rằng giá cổ phiếu sẽ giảm trở lại mức trung bình lịch sử. Ngược lại, nếu giá cổ phiếu giảm xuống 20 đô la, nhà giao dịch có thể mua vào, dự đoán rằng giá

sẽ tăng trở lại mức 50 đô la.

Và đó là ý tưởng cơ bản của việc sử dụng các chiến lược mean reversion - tìm kiếm những thay đổi cực đoan trong giá của một tài sản với giả định rằng giá sẽ quay trở lại mức trung bình của chúng.

Tương tự như những gì đã đề cập trước đó, chiến lược giao dịch mean reversion dựa trên hiện tượng thống kê rằng những biến động cực đoan trong giá tài sản có khả năng sẽ được theo sau bởi một chuyển động ngược lại, đưa giá trở lại mức trung bình theo thời gian. Để áp dụng hiệu quả chiến lược này, các nhà giao dịch trước tiên phải xác định "mức trung bình" cho thị trường và khung thời gian mà họ đã chọn.

Việc tính toán mức trung bình liên quan đến việc lấy trung bình của các giá lịch sử của tài sản trong một khoảng thời gian cụ thể. Điều này có thể được thực hiện bằng cách sử dụng các công cụ khác nhau, chẳng hạn như một bảng tính đơn giản để tính toán trung bình một cách thủ công, một chỉ báo mean reversion tự động hóa quá trình này, hoặc bằng cách kiểm tra trực quan một biểu đồ để xác định mức giá trung bình mà giá đã có xu hướng dao động xung quanh.

Khi mức giá trung bình đã được thiết lập, các nhà giao dịch có thể theo dõi những sự lệch đáng kể từ mức trung bình này. Các nhà giao dịch trong ngày thường tìm kiếm những biến động giá trong ngày lệch khỏi mức trung bình để tận dụng các sự đảo chiều nhanh chóng. Ngược lại, các nhà giao dịch swing và nhà đầu tư dài hạn có thể tìm kiếm những cực trị giá rõ rệt hơn xảy ra trong vài ngày, vài tuần hoặc vài tháng, dự đoán một sự di chuyển đáng kể trở lại mức trung bình.

2.2 Các chỉ số

Khi nói đến hệ thống giao dịch mean reversion, việc sử dụng các chỉ báo kỹ thuật là rất quan trọng để xác định khi nào giá đang lệch khỏi mức trung bình của chúng và có khả năng quay trở lại. Dưới đây là một số chỉ báo nhóm sử dụng trong chiến lược giao dịch mean reversion:

1. Moving Averages

Các đường trung bình động rõ ràng là một trong những chỉ báo tốt nhất để xác định các tín hiệu cho mean reversion. Chúng phục vụ như một đại diện mượt mà cho giá điển hình của một tài sản trong một khoảng thời gian đã chọn và là một công cụ cơ bản để xác định các xu hướng mean reversion.

Các nhà giao dịch theo dõi những sự lệch khỏi mức trung bình này: khi giá tăng lên đáng kể trên đường trung bình động, đạt vào vùng quá mua, điều này gợi ý

rằng giá sẽ giảm trở lại mức trung bình. Tương tự, giá giảm xuống dưới đường trung bình động đến một mức được coi là quá bán cho thấy một khả năng phục hồi về mức giá trung bình.

Có hai dạng đường trung bình động được sử dụng phổ biến là:

a. Đường trung bình đơn giản (SMA – Simple Moving Average)

Là loại đơn giản nhất với trọng số được chia đều cho những mức giá gần đây. Nó được tính bằng cách lấy giá đóng cửa (close) của N phiên giao dịch rồi chia cho N.

Công thức tính:

$$SMA = \frac{A_1 + A_2 + \dots + A_n}{n}$$

b. Đường trung bình hàm mũ (EMA – Exponential Moving Average)

Đôi khi đường trung bình đơn giản (SMA) quá đơn giản và chưa giúp lọc hết những tín hiệu nhiễu. Cho nên bạn cần phải dùng đến đường trung bình động hàm mũ (EMA).

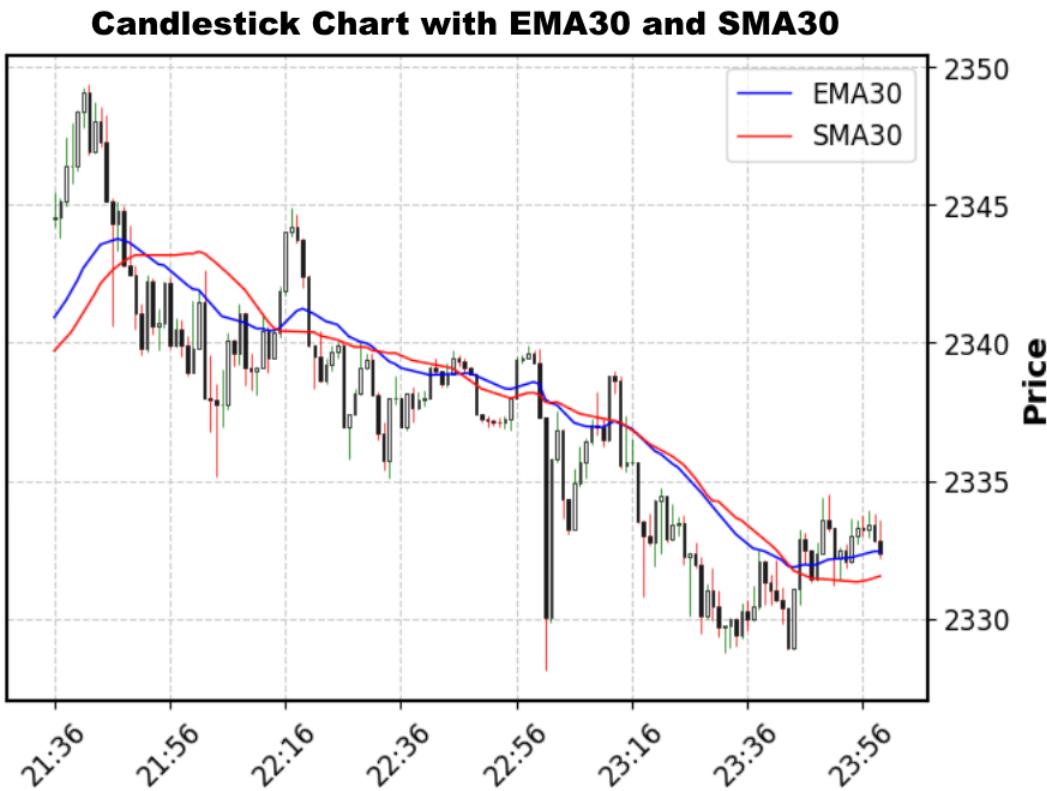
Công thức tính:

$$EMA_t = \alpha \times P_t + (1 - \alpha) \times EMA_{t-1}$$

Với n là số ngày và α được tính như sau:

$$\alpha = \frac{2}{n + 1}$$

(EMA) chú tâm nhiều hơn đến những hành động giá gần hơn là những dữ liệu quá xa trong quá khứ.



Hình 2: Biểu đồ giá đóng của ETH/USDT với SMA30 và EMA30

Nhìn vào biểu đồ trên, ta nhận thấy đường màu xanh (*EMA₃₀*) gần giá hơn so với đường màu đỏ (*SMA₃₀*). Điều này có nghĩa nó đại diện chính xác hơn về những biến động giá gần đây nhất.

2. Bollinger Bands

Bollinger Bands là một trong những chỉ báo phổ biến nhất cho các nhà giao dịch mean reversion. Được tạo ra bởi John Bollinger vào những năm 1980, các dải này bao gồm một dải giữa là trung bình động đơn giản (SMA) và hai dải ngoài cách xa SMA một số độ lệch chuẩn. Cài đặt tiêu chuẩn là SMA 20 ngày với các dải ngoài được đặt ở hai độ lệch chuẩn trên và dưới.

- **Dải giữa (Middle Band)** là đường trung bình động đơn giản (SMA) chu kỳ 20 ngày, tính bằng giá trị trung bình của giá đóng cửa trong 20 ngày qua.

$$\text{Middle_Band} = \text{SMA}_{20} = \frac{1}{20} \sum_{i=1}^{20} \text{Close}_i$$

- **Dải trên (Upper Band)** được tính bằng cách lấy SMA20 cộng với 2 lần độ

lệch chuẩn của giá đóng cửa trong chu kỳ 20 ngày.

$$\text{Upper_Band} = \text{SMA}_{20} + 2 * \sigma_{20}$$

- **Dải dưới (Lower Band)** được tính bằng cách lấy SMA20 trừ đi 2 lần độ lệch chuẩn của giá đóng cửa trong chu kỳ 20 ngày.

$$\text{Lower_Band} = \text{SMA}_{20} - 2 * \sigma_{20}$$

Chiến lược mean reversion của chỉ báo Bollinger Bands xoay quanh ba dải của chỉ báo này. Về mặt kỹ thuật, khi giá chạm hoặc vượt qua một trong các dải Bollinger, các nhà giao dịch có thể coi đây là một sự lệch cực đoan so với mức trung bình, báo hiệu một điểm đảo chiều tiềm năng. Càng gần giá di chuyển đến dải ngoài, giá càng được coi là quá mua hoặc quá bán, do đó, khả năng quay trở lại càng cao.

Như hình ảnh ví dụ bên dưới:



Hình 3: Cách sử dụng Bollinger Band

2.3 Lợi ích và hạn chế của Mean Reversion

Giao dịch mean reversion là một chiến lược có thể vừa mang lại lợi ích vừa mang lại thách thức, cung cấp những lợi thế độc đáo và những nhược điểm tiềm ẩn. Dưới đây là một cái nhìn nhanh về một số ưu điểm và nhược điểm của chiến lược này.

Bảng 1: Lợi ích và Hạn chế của Mean Reversion

Ưu điểm	Nhược điểm
Tỷ lệ thắng cao vì tận dụng sự dao động giá quanh mức trung bình.	Khó khăn tâm lý khi mua vào thị trường giảm và bán ra thị trường tăng.
Tiêu chí rõ ràng cho các điểm vào và ra, đơn giản hóa giao dịch.	Hiệu suất kém trong thị trường có xu hướng.
Hiệu quả trong thị trường không có xu hướng mạnh.	Dễ bị ảnh hưởng bởi các sự kiện bất ngờ.
Giao dịch ngắn hạn, mang lại lợi nhuận nhanh.	Yêu cầu giám sát liên tục, tốn thời gian.
Tự động hóa, phù hợp với kiểm tra lại và giao dịch có hệ thống.	

2.4 Market Making

2.4.1 Định nghĩa

Market Making

Market Making là hoạt động cung cấp thanh khoản cho thị trường bằng cách tạo ra các lệnh mua và bán liên tục. Các nhà tạo lập thị trường (market makers) giúp duy trì sự ổn định của giá cả và giảm thiểu sự biến động.

Khái niệm market making đã xuất hiện từ những ngày đầu của thị trường chứng khoán, khi các nhà giao dịch (jobbers) có vai trò tương tự như market maker hiện nay. Họ là những người duy trì tính thanh khoản cho thị trường bằng cách mua và bán chứng khoán. Sự kiện Big Bang tại London vào năm 1986 đã thay đổi cách thức hoạt động của thị trường chứng khoán, cho phép các công ty chứng khoán hoạt động như market makers chính thức, thay thế cho các jobbers trước đó. Điều này đã dẫn đến sự gia tăng tính cạnh tranh và cải thiện tính thanh khoản trên thị trường. Với sự phát triển của công nghệ và giao dịch điện tử, market making đã trở nên tự động hóa hơn, với sự xuất hiện của các market maker tự động (automated market makers) trong các thị trường phi tập trung (decentralized markets). Các quy định và luật lệ cũng đã được cập nhật để điều chỉnh hoạt động của market makers, nhằm bảo vệ nhà đầu tư và duy trì sự ổn định của thị trường.

Bảng 2: Lợi Ích và Rủi Ro của thuật toán Market Making

Danh mục	Mô tả
Lợi ích	Tăng thanh khoản thị trường Thu nhập ổn định từ Spread
Rủi ro	Rủi ro khi thị trường biến động mạnh Chi phí giao dịch và trượt giá

Trong thị trường tài chính, thanh khoản chính là chìa khóa quan trọng giúp đảm bảo tính bền vững của các loại tài sản. Do đó, trong thị trường, ngoài 2 bên là người mua và người bán thì còn tồn tại một bên thứ 3 giúp duy trì tính thanh khoản của thị trường, đó chính là Market Maker.

Market Marker

Market Maker (MM) là cá nhân hoặc tổ chức với nguồn vốn lớn, chuyên cung cấp thanh khoản cho thị trường bằng cách mua và bán tài sản như cổ phiếu, trái phiếu, ngoại hối, và tiền mã hóa. Họ duy trì mức chênh lệch nhỏ giữa giá mua (bid) và giá bán (ask), giúp tài sản được giao dịch dễ dàng hơn. Market Maker thường được trả tiền bằng cách tính phí hoa hồng cho các giao dịch mà họ thực hiện. Họ cũng có thể kiếm tiền bằng cách chênh lệch giữa giá mua và giá bán của họ.

Ví dụ: Một nhà đầu tư tìm mua cổ phiếu để giao dịch tại một sàn môi giới chứng khoán. Lúc này nhà môi giới đã niêm yết giá mua cổ phiếu đó là 100 USD/cổ phiếu và giá chào bán ở mức 100,05 USD/ cổ phiếu. Điều này có nghĩa là Market Maker trong vai trò là một nhà môi giới sẽ mua cổ phiếu với giá là 100 USD, sau đó họ bán cho người mua với giá 100,05 USD/ cổ phiếu. Thông qua các giao dịch có khối lượng lớn, những khoản chênh lệch nhỏ kết hợp với nhau sẽ tạo lợi nhuận rất lớn.

Market Maker đóng vai trò quan trọng đối với các loại tài sản, đặc biệt là tiền mã hóa, bởi vì thanh khoản là yếu tố cần thiết để đảm bảo tính liên tục và ổn định của thị trường tài chính. Khi một loại tài sản có thanh khoản cao, nhà đầu tư có thể dễ dàng mua bán tài sản đó với giá cả hợp lý và ít rủi ro hơn. Trong giai đoạn đầu khi một loại tài sản mới được niêm yết, thanh khoản thường rất thấp. MM sẽ là bên thứ 3 cung cấp thanh khoản cho thị trường bằng cách mua và bán tài sản đó. Nói cách khác, MM là người đóng vai trò trung gian giữa người mua và người bán, giúp kết nối hai bên với nhau và tạo điều kiện cho giao dịch diễn ra thuận lợi.

2.4.2 Thành phần trong một thuật toán market making

Thuật toán Market Making là một chiến lược giao dịch trong đó nhà giao dịch cung cấp thanh khoản cho thị trường bằng cách đặt các lệnh mua (Bid) và bán (Ask) quanh một mức giá trung tâm (“Fair Value”). Việc cung cấp thanh khoản giúp giảm chênh lệch giá (Spread) và duy trì sự ổn định của thị trường.

Các thành phần cấu tạo nên thuật toán:

a. Fair value

Được tính dựa trên giá Bid và Ask hiện tại hoặc các mô hình định giá. Bên cạnh đó Fair value cũng có thể được điều chỉnh theo mô hình dự báo xu hướng, như trung bình trượt (Moving Average) hoặc trung bình trọng số (Exponential Moving Average). Đây là tham số quan trọng để quyết định vị trí đặt lệnh.

Công thức tính:

$$FairValue = \frac{BestBid + BestAsk}{2}$$

b. Spread

Là khoảng chênh lệch giữa giá Bid và Ask. Giá trị Spread thường được điều chỉnh để phù hợp với điều kiện thị trường.

Công thức tính:

$$Spread = FairValue * SpreadPercentage$$

c. Skew (Độ lệch)

Là sự chênh lệch giữa giá Bid và giá Ask so với Fair Value. Skew được sử dụng để điều chỉnh giá Bid và Ask tựa theo xu hướng thị trường hoặc vị thế hiện tại. Nó phản ánh tâm lý thị trường và có thể cho thấy liệu thị trường đang nghiêng về phía mua hay bán.

Ví dụ: Thị trường đang có xu hướng tăng \Rightarrow giá Bid/Ask sẽ được điều chỉnh lên cao hơn Fair Value.

d. Risk Management (Quản lý rủi ro)

Các quy tắc để kiểm soát vị thế và đảm bảo thanh khoản:

- Max Inventory: Số lượng tài sản tối đa nắm giữ.
- Stop Loss/Take Profit: Giới hạn lỗ hoặc chốt lời.

e. Order Execution (Thực hiện lệnh):

Quyết định khi nào đặt lệnh mua và bán trong sổ lệnh

- Limit Order: Đặt lệnh tại mức giá cố định.
- Market Order: Thực hiện giao dịch tại giá thị trường.

2.4.3 Ví dụ minh họa

Thuật toán được sử dụng cho cặp BTC/USDT với giả định:

$$FairValue = 25,005; \ Spread = 25.005 \ Skew = 0$$

Bước Thực Hiện:

- Tính Fair Value: $Fair Value = (25,000 + 25,010) / 2 = 25,005$
- Tính Spread: $Spread = 25,005 \times 0.001 = 25.005$
- Tính giá Bid/Ask:

$$\begin{aligned} Bid &= FairValue - \frac{Spread}{2} = 25,005 - 12.5025 = 24,992.5 \\ Ask &= FairValue + \frac{Spread}{2} = 25,005 + 12.5025 = 25,017.5. \end{aligned}$$

Đặt Lệnh:

- Đặt lệnh mua (Bid) tại giá 24,992.5.
- Đặt lệnh bán (Ask) tại giá 25,017.5

Lệnh điều chỉnh: Nếu thị trường có xu hướng tăng, market maker có thể điều chỉnh giá bid và ask để duy trì spread và phản ánh fair value.

2.5 Ornstein Uhlenbeck Processes

2.5.1 Giới thiệu

Mô hình Ornstein-Uhlenbeck (OU) là một quá trình ngẫu nhiên với tính chất *mean-reverting* (hồi quy về trung bình), trong đó giá trị của biến có xu hướng quay trở lại mức trung bình dài hạn μ theo thời gian. OU thường được sử dụng trong tài chính để mô phỏng các hiện tượng mà giá trị tài sản dao động quanh một mức trung bình cụ thể. Trong *market making*, việc nhận biết xu hướng hồi quy này rất quan trọng vì thuật toán market making sẽ đặt giá mua (Bid) và bán (Ask) dựa trên dự đoán rằng giá sẽ dao động quanh mức trung bình dài hạn đó.

Phương trình Ornstein-Uhlenbeck:

$$dQ_t = \gamma(\mu - Q_t)dt + \sigma dW_t$$

Trong đó: dQ_t : là sự thay đổi trong giá trị tại thời điểm t ; γ là hệ số hồi quy, cho biết tốc độ quay trở lại mức giá trung bình ; μ là mức giá trung bình dài hạn mà quá trình OU có xu hướng quay lại; Q_t là giá trị của quá trình tại thời điểm t . σ là độ biến động (volatility) của quá trình. dW_t là biến động ngẫu nhiên (diffusion term), mô tả bởi quá trình Wiener.

Sự thay đổi của quá trình bằng lực kéo hướng về giá trị trung bình (drift) cộng với phần nhiễu ngẫu nhiên. Phần drift $-\gamma(Q_t - \mu)$ trong phương trình **Ornstein-Uhlenbeck** đóng vai trò như một lực kéo luôn hướng quá trình về phía giá trị trung bình μ . Cường độ của lực kéo này tỉ lệ thuận với khoảng cách giữa giá trị hiện tại và giá trị trung bình, đồng thời phụ thuộc vào hệ số hồi quy γ . σdW_t (diffusion term) là biến động ngẫu nhiên, không thể dự đoán trước, tác động lên quá trình. Nhờ có phần nhiễu ngẫu nhiên, **quá trình O-U** không đơn thuần chỉ tiến về giá trị trung bình một cách đều đặn mà còn có những biến động ngẫu nhiên xung quanh giá trị trung bình này. Điều này làm cho quá trình trở nên phù hợp hơn để mô hình hóa nhiều hiện tượng thực tế.

Quá trình OU không có bộ nhớ, nghĩa là giá trị trong tương lai chỉ phụ thuộc vào giá trị hiện tại mà không phụ thuộc vào lịch sử

$$P(Q_t|Q_0) = P(Q_{x+t}|Q_x)$$

Dưới giả định rằng quá trình bắt đầu tại một giá trị cụ thể Q_0 , giá trị tại thời điểm t tuân theo phân phối chuẩn với kỳ vọng và phương sai xác định

$$\begin{aligned} E[Q_t] &= \mu + (Q_0 - \mu)e^{-\gamma t} \xrightarrow{t \rightarrow \infty} \mu \\ Var[Q_t] &= \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t}) \xrightarrow{t \rightarrow \infty} \frac{\sigma^2}{2\gamma} \end{aligned}$$

Công thức chứng minh được tham khảo tại [1]

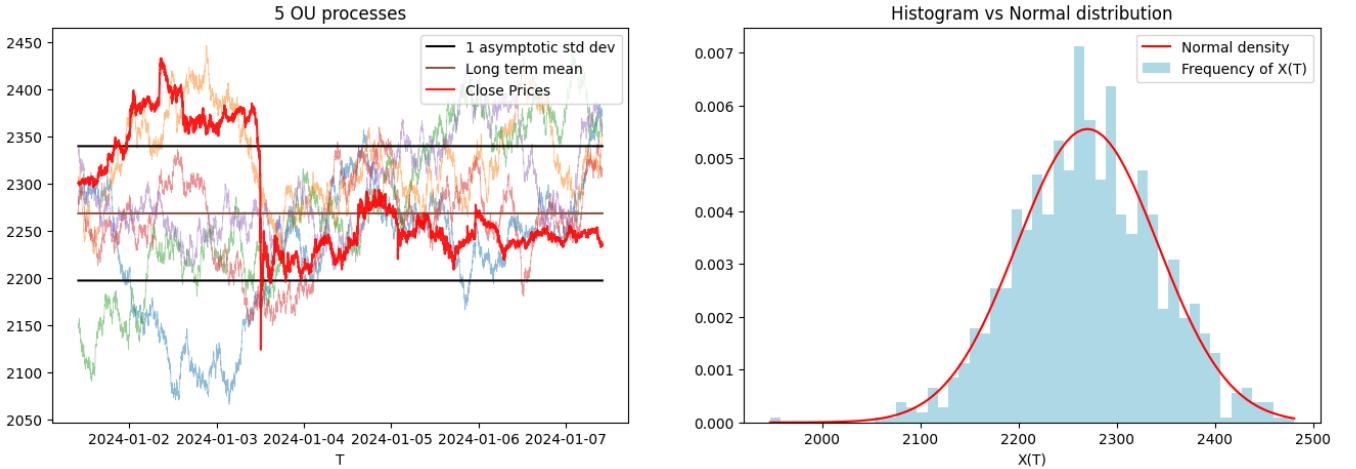
Quá trình OU có xu hướng quay trở lại mức giá trung bình μ với tốc độ được xác định bởi hệ số hồi quy γ . Giá trị γ càng lớn, tốc độ quay về trung bình càng nhanh.

2.5.2 Mô phỏng theo phương pháp số

Có rất nhiều cách để mô phỏng quá trình OU, ở đây nhóm đề cập đến cách tiếp cận theo Phương pháp phân tích (analytical solution). Mô phỏng trực tiếp từ lời giải chứng minh phân phối chuẩn của quá trình OU

$$Q_{t+\Delta t} = \mu + (Q_t - \mu)e^{-\gamma\Delta t} + \sqrt{\frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma\Delta t})}\epsilon_t$$

với $\epsilon \sim N(0, 1)$



Hình 4: Mô phỏng quá trình OU

2.5.3 Ước lượng tham số sử dụng Maximum log-likelihood

Để ước lượng các tham số (μ, γ, σ) của mô hình Ornstein-Uhlenbeck (OU), ta sử dụng một phương pháp khép kín dựa trên tính toán các tổng và tích của dữ liệu chuỗi thời gian. Phương pháp này đơn giản, dễ triển khai và đảm bảo tính toán nhanh chóng.

Gọi $X = X_1, X_2, \dots, X_{n+1}$ là chuỗi dữ liệu giá đóng cửa (close price) của một tài sản, và ta có các giá trị sau:

- \mathbf{XX} là giá của chuỗi X tại thời gian t , với $\mathbf{XX} = X[: -1]$
- \mathbf{YY} : là giá trị của chuỗi X tại thời gian $t + 1$, với $\mathbf{YY} = X[1 :]$
- N : là số lượng phần tử trong chuỗi X trừ đi 1, tức là $N = \text{size}(X) - 1$

Các giá trị cần thiết (tổng và tích):

$$S_x = \sum \mathbf{XX}, \quad S_y = \sum \mathbf{YY}, \quad S_{xx} = \mathbf{XX} \cdot \mathbf{XX}, \quad S_{xy} = \mathbf{XX} \cdot \mathbf{YY}, \quad S_{yy} = \mathbf{YY} \cdot \mathbf{YY}$$

Dựa vào các giá trị thống kê trên, các tham số của mô hình OU được ước lượng như sau:

1. Ước lượng μ :

$$\mu = \frac{S_y S_{xx} - S_x S_{xy}}{N(S_{xx} - S_{xy}) - (S_x^2 - S_x S_y)}$$

Giá trị μ đại diện cho mức trung bình dài hạn mà quá trình OU có xu hướng quay lại.

2. Ước lượng γ :

$$\gamma = -\frac{1}{\Delta t} \ln \left(\frac{S_{xy} - \mu S_x - \mu S_y + N\mu^2}{S_{xx} - 2\mu S_x + N\mu^2} \right)$$

Trong đó Δt là khoảng thời gian giữa hai quan sát liên tiếp. Tham số γ thể hiện tốc độ điều chỉnh của chuỗi dữ liệu về giá trị trung bình μ .

3. Ước lượng $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{S_{yy} - 2e^{-\gamma\Delta t}S_{xy} + e^{-2\gamma\Delta t}S_{xx} - 2\gamma(1 - e^{-\gamma\Delta t})(S_y - e^{-\gamma\Delta t}S_x) + N\gamma^2(1 - e^{-\gamma\Delta t})^2}{N}$$

4. Ước lượng σ :

$$\sigma = \sqrt{\hat{\sigma}^2 \cdot \frac{2\gamma}{1 - e^{-2\gamma\Delta t}}}$$

Giá trị σ phản ánh mức độ dao động của chuỗi dữ liệu quanh giá trị trung bình.

Phương pháp ước lượng này, dựa trên các công thức khép kín, là một cách tiếp cận đơn giản và hiệu quả để xác định các tham số của mô hình Ornstein-Uhlenbeck. Nó nổi bật với ưu điểm về tốc độ tính toán và tính khả dụng, đặc biệt trong các ứng dụng thực tế như phân tích chuỗi thời gian và tài chính. Tuy nhiên, nhược điểm của phương pháp này là có thể nhạy cảm với sai số số học khi áp dụng trên dữ liệu lớn hoặc trong các trường hợp mà dữ liệu không đáp ứng các giả định cơ bản của mô hình OU, chẳng hạn như tính phân phối Gaussian của nhiễu và tính tuyến tính trong dữ liệu. Điều này có thể dẫn đến ước lượng sai lệch và ảnh hưởng đến tính chính xác của mô hình.

Phương pháp sẽ được áp dụng trong việc triển khai Kalman Filter để ước lượng và cập nhật tham số của mô hình OU một cách đệ quy. Kalman Filter cho phép dự đoán giá trị tài sản bằng cách kết hợp ước lượng từ mô hình OU và quan sát thực tế theo thời gian. Chúng tôi sẽ trình bày chi tiết về Kalman Filter và ứng dụng của nó trên chuỗi dữ liệu mô phỏng từ mô hình OU trong phần tiếp theo.

2.5.4 Kalman Filter trong Ornstein-Uhlenbeck Process

Kalman Filter là một thuật toán tái hồi tuyến tính giúp ước lượng trạng thái của một hệ thống động với các phép đo bị nhiễu. Nó được sử dụng rộng rãi trong các bài toán xử lý tín hiệu, dự đoán, điều khiển tự động và tài chính. Kết hợp giá trị trung bình trước đó (dựa trên mô hình) và dữ liệu đo lường hiện tại (có nhiễu) để tính toán trạng thái hiện tại của quá trình. Khi phép đo có nhiễu lớn, bộ lọc ưu tiên dựa vào mô hình. Ngược lại, khi phép đo có độ bất định thấp, bộ lọc dựa nhiều hơn vào dữ liệu đo lường. Kalman Filter là công cụ hiệu quả trong việc giảm nhiễu, ngay cả khi dữ liệu bị

tác động bởi các yếu tố bất định. Thuật toán này đặc biệt phù hợp với các hệ thống có tính động học cao và hoạt động tốt trong môi trường có nhiễu Gaussian. Ngoài ra, Kalman Filter dễ dàng triển khai trong nhiều lĩnh vực thực tế như định vị GPS, xử lý ảnh, dự đoán tài chính, và điều hướng robot tự hành.

(1) Mô hình hóa quá trình OU

Để áp dụng bộ lọc Kalman vào quá trình OU, trước tiên cần xác định ma trận chuyển trạng thái F_k . Quá trình này bắt đầu bằng cách sắp xếp lại phương trình rời rạc hóa. Gọi k là bước thời gian tại t_k và $k+1$ là bước thời gian tại $t_k + \Delta t$ (với Δt là hằng số cho tất cả các k). Khi đó phương trình cho X_{k+1} sẽ được viết lại như sau:

$$X_{k+1} = \mu + (X_k - \mu)e^{-\alpha\Delta t} + \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha\Delta t})\epsilon_k$$

Trong đó: μ : Giá trị trung bình dài hạn của quá trình OU, α : tốc độ quay về μ , σ : Độ dao động của quá trình, ϵ_k : Nhiễu ngẫu nhiên, phân phối chuẩn $\mathcal{N}(0, 1)$

Phương trình có thể được viết lại thành:

$$X_{k+1} = \mu(1 - e^{-\alpha\Delta t}) + X_k e^{-\alpha\Delta t} + \sigma_p \epsilon_k$$

Với: $\sigma_p = \sqrt{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha\Delta t})}$, $A = \mu(1 - e^{-\alpha\Delta t})$, $B = e^{-\alpha\Delta t}$

Phương trình trên được rút gọn lại thành:

$$X_{k+1} = A + BX_k + \sigma_p \epsilon_k$$

(2) Đưa vào bộ lọc Kalman

Để biểu diễn trong phương trình Kalman, đầu tiên ta sẽ định nghĩa 1 số thành phần:

$$Z_k = \begin{bmatrix} 1 \\ X_k \end{bmatrix}, \quad Z_{k+1} = \begin{bmatrix} 1 \\ X_{k+1} \end{bmatrix}$$

Khi đó phương trình được viết lại thành:

$$Z_{k+1} = \begin{bmatrix} 1 & 0 \\ A & B \end{bmatrix} Z_k + \begin{bmatrix} 0 \\ \sigma_p \epsilon_k \end{bmatrix}$$

Khi đó: Ma trận trạng thái $F = \begin{bmatrix} 1 & 0 \\ A & B \end{bmatrix}$, là hằng số cho mọi k

(3) Xác định ma trận và các tham số trong Kalman

- Ma trận hiệp phương sai nhiễu quá trình $Q = \sigma_p^2 I_2$ (ma trận đường chéo).
- Ma trận hiệp phương sai nhiễu quan sát: $R = \sigma_o^2 I_2$ với σ_o là độ lệch chuẩn của nhiễu quan sát.
- Ma trận quan sát : $H = I_2$ (ma trận đơn vị).

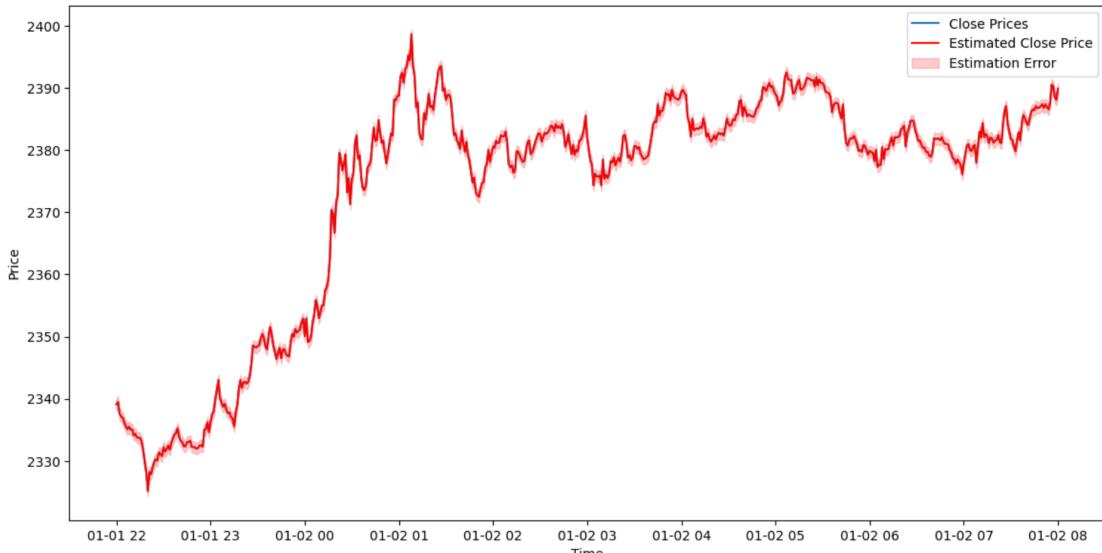
Xem thêm, Chuyển đổi tham số OU sang ma trận của bộ lọc Kalman [10.2](#).

(4) Quan sát nhiễu và mô phỏng

Để mô phỏng các quan sát nhiễu, ta thêm một thành phần nhiễu mới vào quá trình OU được mô phỏng, với độ lệch chuẩn $\sigma_o = 0.1$. Gọi quan sát tại bước k là:

$$\tilde{Z}_k = Z_k + \sigma_o \epsilon_k$$

Với $\epsilon_k \sim \mathcal{N}(0, 1)$.



Hình 5: Hình ảnh minh họa Kalman Filter

3 Các nghiên cứu trước đây

Avellaneda & Stoikov Model

Mô hình **Avellaneda-Stoikov** được giới thiệu vào năm 2008 trong bài báo nổi tiếng "*High-Frequency Trading in a Limit Order Book*" của Marco Avellaneda và Sasha Stoikov. Đây là một trong những mô hình đầu tiên được phát triển để tối ưu hóa chiến lược Market Making dựa trên lý thuyết tài chính và toán học ứng dụng.

Mục tiêu của mô hình là giúp các nhà tạo lập thị trường (Market Makers) đưa ra mức giá chào mua (Bid) và chào bán (Ask) tối ưu trong sổ lệnh giới hạn (Limit Order Book), tối ưu hóa lợi nhuận trong khi kiểm soát rủi ro tồn kho (Inventory Risk).

Mô hình này là công cụ mạnh mẽ cho các chiến lược Market Making, đặc biệt trong môi trường giao dịch tốc độ cao (High-Frequency Trading). Nó cho phép các Market Makers đưa ra mức giá Bid và Ask hiệu quả, hỗ trợ giao dịch nhanh chóng, cải thiện tính thanh khoản của thị trường và tối ưu hóa tốc độ khớp lệnh. Đây là yếu tố quan trọng trong các thị trường có tính thanh khoản cao, nơi mà độ trễ trong giao dịch có thể ảnh hưởng đáng kể đến lợi nhuận.

Mô hình dựa trên một phương trình cơ bản cho vị trí giá chào mua và chào bán như sau:

$$\begin{aligned} P_{bid}(t) &= S(t) + \Delta(t) \\ P_{ask}(t) &= S(t) + \Delta(t) \end{aligned}$$

Trong đó: $P_{bid}(t)$: Giá chào mua tối ưu tại thời điểm t ; $P_{ask}(t)$: Giá chào bán tối ưu tại thời điểm t . $S(t)$: Giá trung bình (mid-price) tại thời điểm t , thường được tính bằng $\frac{P_{bid}+P_{ask}}{2}$. $\Delta(t)$: Điều chỉnh spread động để tối ưu lợi nhuận.

Công thức cho Spread tối ưu $\Delta(t)$:

$$\Delta(t) = \frac{\gamma\sigma^2}{k} + \frac{2}{\gamma} \ln\left(1 + \frac{\gamma}{k}\right)$$

Trong đó: γ : Hệ số rủi ro tồn kho (risk aversion coefficient). σ : Độ biến động của giá (volatility). k : Tốc độ khớp lệnh trung bình (order execution rate).

Công thức điều chỉnh giá Mid-price $S(t)$:

$$S(t) = S_0 - \frac{q\gamma\sigma^2}{2k}$$

Trong đó: S_0 : Giá mid-price ban đầu. q : Hàng tồn kho hiện tại (inventory).

4 Phương pháp nghiên cứu

4.1 Dữ liệu sử dụng

Dữ liệu sử dụng trong nghiên cứu được lấy từ cùng một nguồn [Bybit](#), trong khoảng thời gian từ 1/1/2024 đến 7/1/2024 (7 ngày), nhằm đảm bảo tính nhất quán và độ chính xác trong phân tích. Ba loại dữ liệu chính bao gồm:

a. OHLC(Open/High/Low/Close) data

OHLC - Open/High/Low/Close (Bybit/Index Price Kline) là một loại dữ liệu cung cấp thông tin toàn diện về biến động giá và khối lượng giao dịch trong thị trường tài chính.

Các biến số chính trong dữ liệu bao gồm: *timestamp* (thời gian ghi nhận từng mốc dữ liệu), *Open* (giá mở cửa tại thời điểm đầu kỳ), *High* (giá cao nhất đạt được trong kỳ), *Low* (giá thấp nhất đạt được trong kỳ), *Close* (giá đóng cửa tại thời điểm cuối kỳ), *Volume* (khối lượng giao dịch trong kỳ), và *Turnover* (tổng giá trị giao dịch, được tính bằng *Volume* \times Giá). Dữ liệu này giúp cung cấp cái nhìn chi tiết về sự biến động của giá và khối lượng giao dịch trong một khoảng thời gian cụ thể.

# start_at	Δ symbol	# period	# open	# high	# low	# close
1.70b	1 unique value	1	2.12k	2.14k	2.1k	2.12k
1.70b	1.70b	1	2.43k	2.43k	2.43k	2.43k
1704067200	ETHUSD	1	2281.49	2282.48	2280.88	2281.45
1704067260	ETHUSD	1	2281.45	2283.51	2281.43	2283.37
1704067320	ETHUSD	1	2283.37	2284.21	2283.08	2284.02
1704067380	ETHUSD	1	2284.02	2285.88	2284.01	2285.85
1704067440	ETHUSD	1	2285.85	2287.39	2285.65	2287.29
1704067500	ETHUSD	1	2287.29	2289.38	2286.36	2288.81
1704067560	ETHUSD	1	2288.81	2290.3	2288.8	2289.81
1704067620	ETHUSD	1	2289.81	2291.11	2289.25	2289.73

Hình 6: OHLC data

b. Order Book data

Dữ liệu Order Book (Bybit/OrderBook) phản ánh mức độ thanh khoản và trạng thái của thị trường tại một thời điểm cụ thể.

Các biến số chính bao gồm: *Timestamp* (thời gian chụp nhanh trạng thái số lệnh), *Price* (Bid/Ask) (mức giá của các lệnh mua và bán), và *Size* (Bid/Ask) (số lượng tài sản tương ứng với từng mức giá).

Hình 7: Orderbook data

c. Trade data

Trade Data (Bybit/Public Trading History) được sử dụng để thực hiện backtesting, cung cấp thông tin chi tiết về các giao dịch đã diễn ra.

Dữ liệu bao gồm các trường chính như sau: *id* (Mã định danh của thông điệp giao dịch); *price* (Giá thực hiện giao dịch); *side* (Loại giao dịch của người thực hiện lệnh, với các giá trị Buy-mua hoặc Sell-bán); *timestamp* (Thời gian giao dịch được thực hiện, tính bằng mili-giây (ms)); *volume* (Khối lượng giao dịch).

# id	# timestamp	# price	# volume	▲ side
1 333k	1705b 1705b	2.17k 2.36k	0 176	buy sell 53% 47%
1	1704672002285	2221.58	0.37811	sell
2	1704672002285	2221.58	0.02189	sell
3	1704672002567	2221.58	0.4	sell
4	1704672002804	2221.58	0.4	sell
5	1704672003064	2221.58	0.4	sell
6	1704672003551	2221.59	0.03994	buy
7	1704672003551	2221.59	0.001	buy
8	1704672003872	2221.58	0.398	sell

Hình 8: Trade data

4.2 Kiểm định giả thuyết

Mục tiêu của phân kiểm định giả thuyết trong nghiên cứu này là để xác định liệu các yếu tố như Mean Reversion có tồn tại trong dữ liệu tài chính hay không. Cụ thể, mục tiêu là kiểm tra giả thuyết rằng các chuỗi thời gian tài chính có xu hướng quay lại mức giá trung bình theo một quy trình cụ thể (tính dừng), thông qua các kiểm định thống kê.

Tính dừng mô tả chuỗi thời gian mà các đặc tính thống kê (như trung bình và phương sai) không thay đổi theo thời gian. Trong ngữ cảnh tài chính, giá có tính dừng khi chúng khuếch tán chậm hơn so với chuyển động ngẫu nhiên hình học (geometric random walk).

Phương pháp kiểm định

a. Augmented Dickey-Fuller (ADF)

Mục tiêu chính của ADF là kiểm tra liệu một chuỗi thời gian có tính hồi quy trung bình hay không. Tập trung vào việc kiểm tra mối quan hệ giữa sự thay đổi giá tại thời điểm hiện tại Δ_t và giá tại thời điểm trước đó y_{t-1} . Quá trình phân tích dựa trên mô hình tuyến tính được biểu diễn như sau:

$$\Delta y(t) = \lambda y(t-1) + \mu + \beta t + \alpha_1 \Delta y(t-1) + \dots + \alpha_k \Delta y(t-k) + \epsilon_t$$

Trong đó: y_t (Giá tại thời điểm t); λ (Hệ số hồi quy xác định xu hướng hồi quy trung bình); μ, β_t (Các tham số mô tả xu hướng tuyến tính trong chuỗi thời gian); $\alpha_1, \dots, \alpha_k$ (Các tham số điều chỉnh ảnh hưởng của độ trễ - lag); ϵ_t (Nhiều ngẫu nhiên tại thời điểm t)

Phương pháp hồi quy được áp dụng để ước tính hệ số λ và sai số chuẩn SE của nó. Bước tiếp theo là kiểm tra ý nghĩa thống kê bằng cách tính toán tỷ số $\frac{\lambda}{SE}$. Tỷ số này được so sánh với các giá trị ngưỡng tin cậy, chẳng hạn ở mức 95%, để đưa ra kết luận:

- Nếu $\lambda < 0$: Chuỗi thời gian có tính hồi quy trung bình, nghĩa là giá có xu hướng quay về giá trị trung bình theo thời gian.
- Nếu $\lambda \geq 0$: Chuỗi thời gian không hồi quy trung bình, cho thấy xu hướng tăng/giảm dài hạn hoặc ngẫu nhiên mà không quay lại giá trị trung bình.

b. Hurst Exponent

Hệ số Hurst (H) đo lường tốc độ khuếch tán của log giá $z(t)$ trong một khoảng trễ bất kỳ τ :

$$\text{Var}(\tau) = \langle |z(t+\tau) - z(t)|^2 \rangle \sim \tau^{2H}$$

Ý nghĩa của hệ số Hurst - H :

- Nếu $H = 0.5$: Chuỗi giá tuân theo chuyển động ngẫu nhiên hình học (random walk), không có xu hướng rõ rệt

- Nếu $H < 0.5$: Chuỗi giá có xu hướng hồi quy về trung bình (mean reversion), tức là giá có khả năng quay lại giá trị trung bình sau những biến động.
- Nếu $H > 0.5$: Chuỗi giá có xu hướng (trend) rõ rệt, tức là có sự thay đổi liên tục theo một hướng nhất định trong tương lai.

So sánh hệ số **Hurst** với chuyển động ngẫu nhiên hình học giúp xác định tính dừng và xu hướng của chuỗi thời gian

Kết quả kiểm định

Kiểm định ADF trên dữ liệu OHLC của đồng ETH/USDT từ 01/01/2024 - 07/01/2024.

- ADF Statistic: -15.528754953255493
- p_value: 2.2345676828495575e-28

Nhận xét: Với p-value rất nhỏ ($2.23e-28$), thấp hơn mức ý nghĩa thông thường (0.05) cho thấy chuỗi dữ liệu có tính dừng.

5 Tạo tín hiệu giao dịch

Mục tiêu của việc tạo tín hiệu giao dịch là xác định các thời điểm mua (Bid) và bán (Ask) tài sản tài chính một cách tối ưu. Quá trình này dựa trên các tín hiệu từ dữ liệu sổ lệnh (Order Book) và tín hiệu dựa vào đặc điểm Mean Reversion của giá tài sản, nhằm cải thiện hiệu suất và giảm rủi ro trong chiến lược Market Making. Bên cạnh đó kết hợp các mô hình học máy để có thể khai thác dữ liệu từ nhiều nguồn khác nhau và nâng cao độ chính xác.

Trước khi thực hiện việc xây dựng các mô hình học phân loại để tạo tín hiệu (signal), chúng ta cần tự gán nhãn tín hiệu bằng cách sử dụng dữ liệu OHLC. Việc gán nhãn dữ liệu này giúp mô hình học được xu hướng giá tại thời điểm tiếp theo. Cụ thể, dữ liệu được gán nhãn như sau:

- Lớp (1) (Giá tăng): nếu sự chêch lệch giữa 2 thời điểm giá vượt qua ngưỡng (ϵ) xác định.
- Lớp (-1) (Giá giảm): nếu sự chêch lệch giữa 2 thời điểm giá thấp dưới ngưỡng âm (ϵ) xác định.
- Lớp (0) (Giá đi ngang): nếu sự chênh lệch giá nằm trong ngưỡng (ϵ) xác định.

5.1 Orderbook Signal

Trong giao dịch tài chính, việc dự báo xu hướng giá tại thời điểm tiếp theo là một bài toán quan trọng giúp các nhà đầu tư đưa ra quyết định giao dịch tối ưu. Một trong những nguồn dữ liệu hữu ích là order book, cung cấp thông tin chi tiết về các lệnh mua (bid) và bán (ask) trên thị trường.

1. Chuẩn bị dữ liệu

Sử dụng dữ liệu từ orderbook, bao gồm các mức giá mua (bid) và bán (ask) cao nhất trong mỗi phút cùng với khối lượng tương ứng, chúng ta tiến hành resample dữ liệu orderbook để đồng bộ với thời gian của dữ liệu OHLC.

Ngoài ra, nhóm còn tạo ra các biến phụ thuộc vào thời gian và không phụ thuộc vào thời gian, bao gồm

Bảng 3: Tập các thuộc tính mới trong signal orderbook

Tập thuộc tính	Mô tả
v1	Danh sách các thuộc tính cơ bản của sổ lệnh (giá và khối lượng cho 5 cấp độ đầu tiên)
v2	Spread và Midprice tại mỗi mức trong sổ lệnh
v3	Khoảng cách giá giữa các mức trong sổ lệnh và mức giá đầu tiên (level 1)
v4	Các biến liên quan đến giá trị trung bình
v5	Chênh lệch tích lũy
v6	Giá và khối lượng phái sinh
v7	Orderbook pressure

Sử dụng timestamp để đồng bộ nhãn từ dữ liệu OHLC với đặc trưng từ order book. Đảm bảo rằng mỗi mẫu trong order book có nhãn tương ứng.

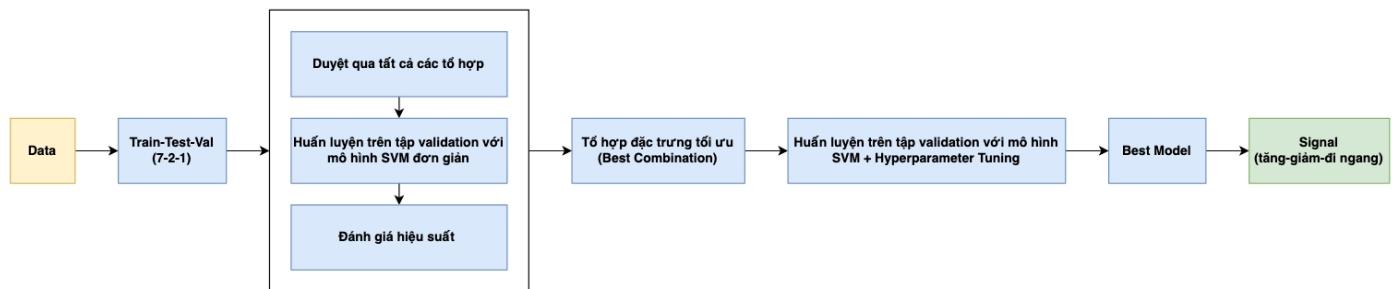
2. Xây dựng mô hình Support Vector Machine (SVM)

Input:

- X, y : Dữ liệu đặc trưng và nhãn tín hiệu ($1, -1, 0$).
- Feature = V_1, V_2, \dots, V_6 : tập các đặc trưng

Output: BestCombination, BestScore, y_{pred}

Thuật toán: Tìm tổ hợp đặc trưng tối ưu bằng cách duyệt tất cả tổ hợp, huấn luyện SVM, đánh giá hiệu suất, và chọn tổ hợp có điểm cao nhất.



Hình 9: Quy trình xây dựng mô hình SVM cho Orderbook Signal

Dánh giá hiệu suất trên các tổ hợp ở tập huấn luyện, nhận thấy rằng tổ hợp v5, v6 cho điểm accuracy cao nhất. Kết hợp hai tổ hợp trên với thuộc tính order-book_pressure để xây dựng mô hình.

	V1	V2	V3	V4	V5	V6	Accuracy	Precision	Recall	Time Taken
0	Y	N	N	N	N	N	0.3877	0.3845	0.3657	5.438285
1	N	Y	N	N	N	N	0.3847	0.3724	0.3727	6.800811
2	N	N	Y	N	N	N	0.3745	0.3678	0.3629	5.095467
3	N	N	N	Y	N	N	0.3884	0.3816	0.3687	3.875007
4	N	N	N	N	Y	N	0.3715	0.3711	0.3514	5.358988
5	N	N	N	N	N	Y	0.3824	0.3760	0.3602	6.078897
6	Y	Y	N	N	N	N	0.3817	0.3732	0.3575	7.362250
7	Y	N	Y	N	N	N	0.3652	0.2643	0.3370	7.758727
8	Y	N	N	Y	N	N	0.3880	0.3848	0.3666	5.690516
9	Y	N	N	N	Y	N	0.3897	0.4074	0.3684	7.607327

Hình 10: Kết quả trên tập huấn luyện với các tổ hợp thuộc tính khác nhau

Sử dụng GridSearchCV để thực hiện tuning siêu tham số cho một mô hình SVM trên tập dữ liệu thử nghiệm (valid set), nhằm tìm ra tổ hợp các giá trị tham số tốt nhất để tối ưu hóa hiệu suất của mô hình trên tập dữ liệu.

```

1 # Hyperparameter tuning using GridSearchCV
2 param_grid = {
3     'C': [0.01, 0.1, 1, 10], # Regularization parameter
4     'gamma': [0.001, 0.01, 0.1, 1], # Kernel coefficient
5     'kernel': ['sigmoid', 'rbf', 'linear'] # Kernel type
6 }
7
8 cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
9 svm = SVC(probability=True)
10 grid = GridSearchCV(estimator=svm, param_grid=param_grid,
11                      cv=cv, scoring='accuracy', verbose=2, n_jobs
12                      =-1)
13 grid.fit(X_val, y_val)

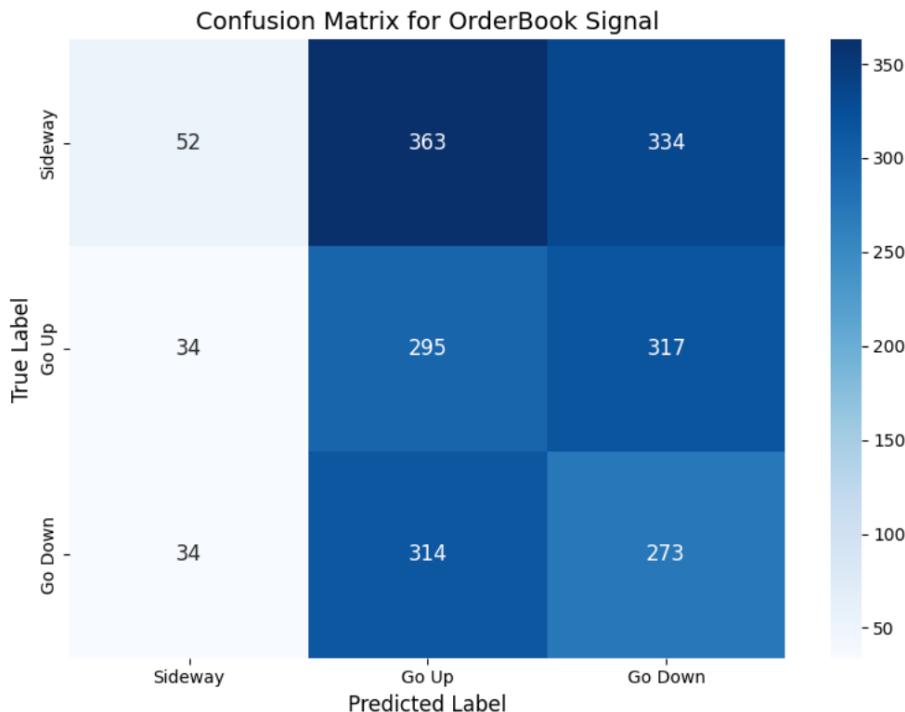
```

Listing 1: Hyperparameter tuning using GridSearchCV

Để xây dựng mô hình dự đoán Orderbook Signal, bộ siêu tham số tối ưu đã được lựa chọn thông qua quá trình tìm kiếm là $C : 0.01, gamma : 0.01, kernel : sigmoid$.

Việc sử dụng các tham số này giúp tối ưu hóa hiệu suất của mô hình, đảm bảo khả năng dự đoán chính xác hơn đối với các tín hiệu từ dữ liệu orderbook.

Kết quả khi chạy mô hình trên tập dữ liệu kiểm tra (test set) để dự đoán OrderBook Signal:



Hình 11: Confusion Matrix trên tập kiểm tra của OrderBook Signal Model

Từ biểu đồ confusion matrix ta có thể thấy rằng mô hình được xây dựng để tạo Signal Orderbook có hiệu suất còn thấp. Mô hình còn gặp khó khăn trong việc phân loại chính xác các lớp 'Sideway', 'Go Up', và 'Go Down'. Số lượng nhầm lẫn giữa 'Go Up' và 'Go Down' còn lớn, cho thấy sự tương đồng cao giữa hai lớp này. Nguyên nhân có thể là do sự biến động của thị trường dẫn đến các tín hiệu không nhất quán trong dữ liệu được sử dụng, làm cho mô hình khó dự đoán chính xác.

5.2 Mean reversion Signal

Tín hiệu Mean Reversion (hồi quy về trung bình) được sử dụng để dự đoán khi nào giá sẽ quay lại mức giá trung bình sau khi có sự biến động mạnh. Để thực hiện chiến lược này, nhóm đã sử dụng LSTM (Long Short-Term Memory), một mô hình học sâu hiệu quả trong việc xử lý dữ liệu chuỗi thời gian dài hạn và học các phụ thuộc thời gian. LSTM giúp dự đoán sự quay lại mức giá trung bình trong tương lai, đặc biệt khi giá có

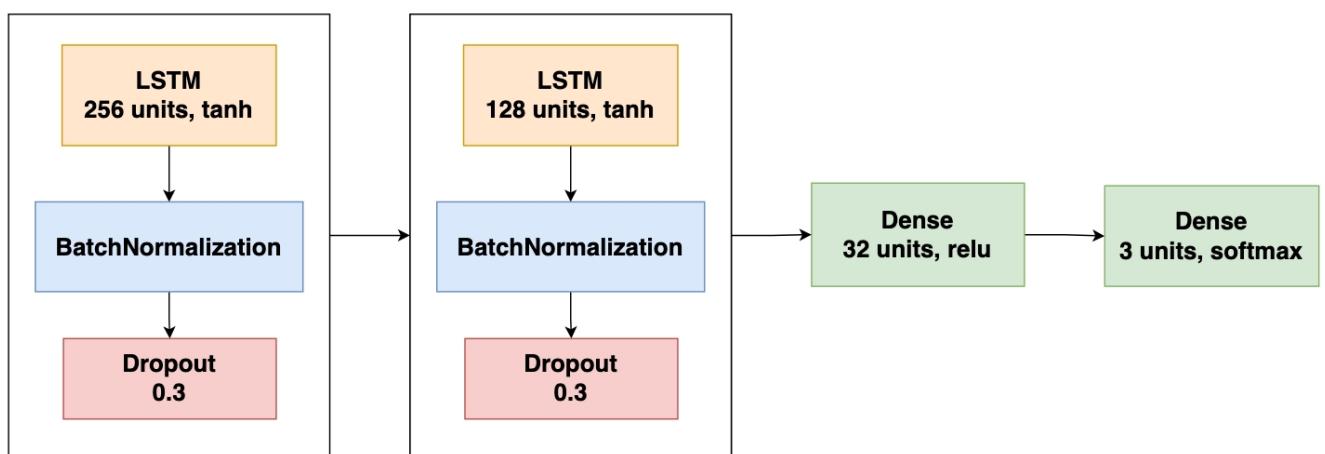
xu hướng dao động xung quanh mức trung bình, từ đó hỗ trợ việc xác định các điểm đảo chiều tiềm năng trong giao dịch tài chính.

1. Chuẩn bị dữ liệu

Sử dụng dữ liệu OHLC (Open, High, Low, Close) từ thị trường, chúng ta tính toán các chỉ số EMA (Exponential Moving Average) ngắn hạn và dài hạn, chẳng hạn như EMA_{12} và EMA_{26} . Đồng thời, tính toán sự chênh lệch giữa hai thời điểm liên tiếp của giá đóng cửa (Close) để đánh giá mức độ biến động của giá. Cuối cùng, chúng ta sử dụng nhãn dữ liệu vừa được gán để đồng bộ với các đặc trưng đã tính toán, đảm bảo rằng mỗi mẫu trong dữ liệu có nhãn tương ứng với sự biến động giá tại thời điểm tiếp theo.

2. Xây dựng mô hình LSTM

Mô hình bao gồm hai lớp LSTM, với 256 và 128 đơn vị, và sử dụng hàm kích hoạt 'tanh'. Cả hai lớp LSTM đều có `return_sequences=True` trong lớp đầu tiên để truyền tiếp dữ liệu qua lớp thứ hai. Sau mỗi lớp LSTM, một lớp BatchNormalization và một lớp Dropout với tỷ lệ 0.3 được thêm vào để giảm overfitting và tăng độ chính xác của mô hình. Lớp Dense đầu tiên với 32 đơn vị sử dụng hàm kích hoạt 'relu', trong khi lớp Dense cuối cùng có 3 đơn vị và sử dụng hàm kích hoạt 'softmax' để phân loại thành 3 lớp: Go Up, Go Down và Sideway. Mô hình được biên dịch với bộ tối ưu hóa Adam và hàm mất mát 'sparse_categorical_crossentropy', và được đánh giá bằng chỉ số độ chính xác.



Hình 12: Mô hình LSTM 2 layers - 1 unit

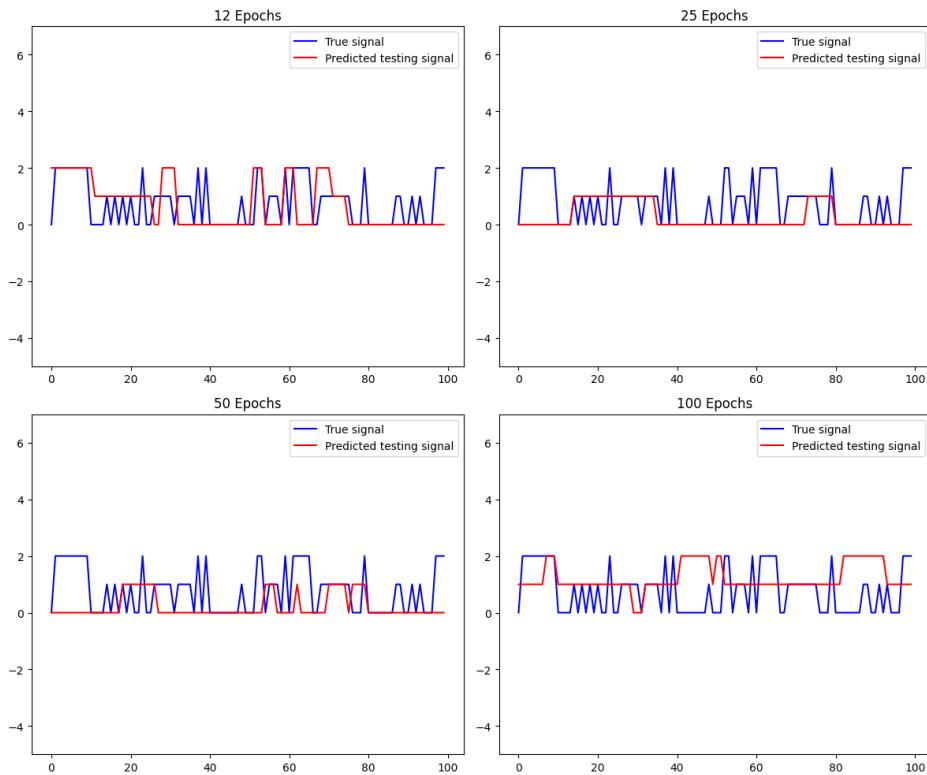
Dữ liệu được chia thành 80% để huấn luyện và 20% để kiểm tra. Mô hình được huấn luyện với số epochs lần lượt là 12, 25, 50, 100, và sẽ dừng sớm nếu callback

EarlyStopping phát hiện rằng mô hình không cải thiện trên tập kiểm tra. Lưu checkpoint tốt nhất (best model weights) và ghi lại thời gian xử lý.

Bảng 4: Kết quả huấn luyện của các mô hình

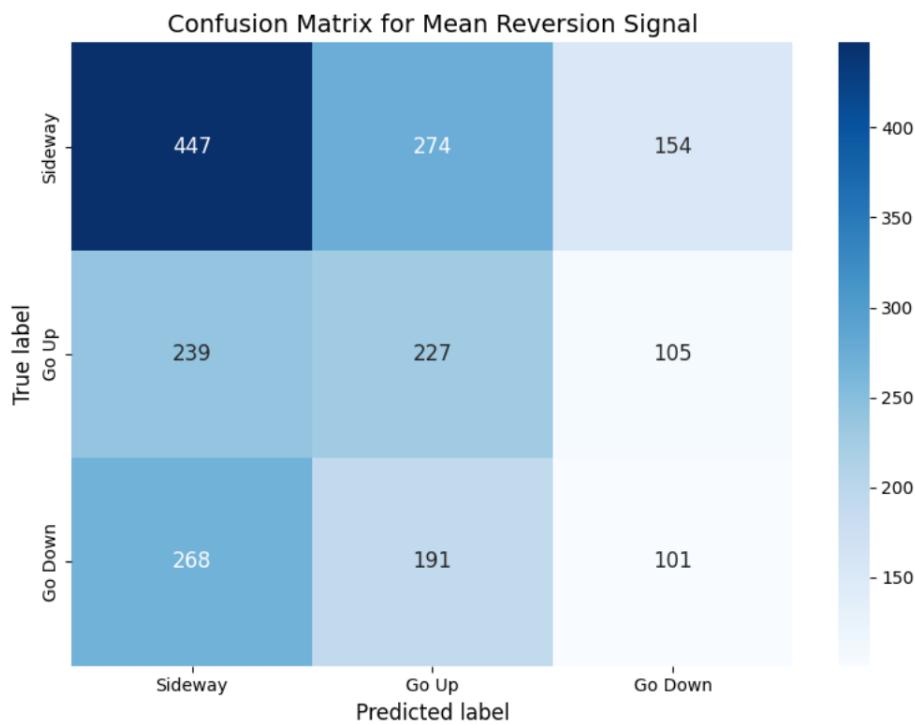
Num Epochs	Processing Time (sec)	Loss
12	52.394194	1.080756
25	43.562836	1.102369
50	46.914188	1.076651
100	52.624454	1.099664

Dựa trên bảng kết quả huấn luyện các mô hình, có thể nhận thấy rằng mô hình với 12 và 50 epochs đạt được giá trị val_loss thấp nhất, lần lượt là 1.081 và 1.077. Tuy nhiên, khi so sánh tín hiệu dự đoán (y_{pred}) với tín hiệu thực tế (y_{test}) đối với 100 giá trị đầu tiên ở mỗi checkpoint, mô hình huấn luyện trong 50 epochs thường như không thể dự đoán chính xác xu hướng giảm giá (go down), mặc dù có giá trị val_loss thấp nhất.



Hình 13: Kết quả trên tập kiểm tra với các epochs khác nhau

Do đó, nhóm đã chọn mô hình với 12 epochs để xây dựng Mean Reversion Signal



Hình 14: Confusion Matrix trên tập kiểm tra của Mean Reversion Signal Model

Mô hình nhận diện lớp 'Sideway' với độ chính xác cao nhất, cho thấy khả năng phân biệt khi thị trường không có xu hướng rõ ràng. Số lượng dự đoán đúng cho hai lớp ('Go Up' và 'Go Down') còn khá thấp. Điều này cho thấy có sự nhầm lẫn đáng kể giữa chúng, cho thấy rằng mô hình còn gặp khó khăn trong việc phân biệt xu hướng tăng và giảm. Do đó mô hình cần được tối ưu hóa cho các xu hướng tăng và giảm nhằm nâng cao hiệu quả phân loại tổng thể.

6 Chiến lược Market Making

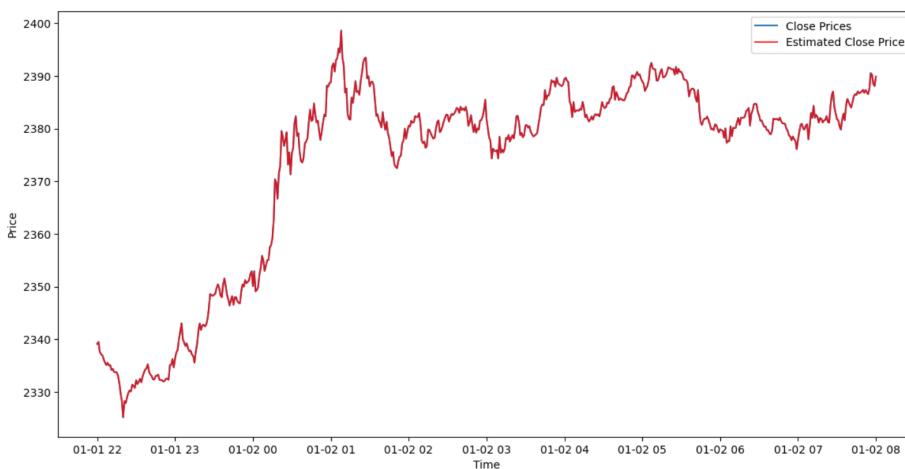
6.1 Tính Fair Value

Fair Value (Q_t) là giá trị hợp lý của tài sản tại thời điểm t , đại diện cho mức giá "thực" mà tài sản đó sẽ đạt được nếu không có các yếu tố ảnh hưởng ngẫu nhiên từ thị trường. Để ước lượng *Fair Value* (Q_t), ta sử dụng mô hình Kalman Filter, kết hợp với các tham số được ước lượng từ phương pháp Maximum Likelihood Estimation (MLE).

Các bước thực hiện:

- (1) Ước lượng tham số OU (θ, μ, σ) bằng phương pháp MLE từ chuỗi thời gian của giá.
- (2) Chuyển đổi tham số của quy trình OU thành các ma trận cần thiết cho Kalman Filter.
- (3) Ứng dụng Kalman Filter để lọc và ước tính giá trị hợp lý (Fair Value) qua thời gian.
- (4) Lấy giá trị ước tính từ Kalman Filter làm giá trị hợp lý (Q_t) của tài sản.

Xem thêm, phần Ornstein Uhlenbeck Processes [2.5.1](#).



Hình 15: Ước lượng giá trị hợp lý của tài sản tại thời điểm t

Giá trị Fair Value ước tính được xấp xỉ bằng giá đóng cửa (Close price).

6.2 Tính giá trị Spread

Spread (C_t) trong lĩnh vực tài chính, đặc biệt trong giao dịch chứng khoán, ngoại hối, hay các thị trường tài chính khác, là sự chênh lệch giữa hai giá quan trọng: Giá mua

(Bid price) và Giá bán (Ask price). Spread là một chỉ báo quan trọng trong việc xác định chi phí giao dịch và độ thanh khoản của thị trường. Spread được tính theo công thức:

$$C_t = k\sigma$$

Trong đó: k là hệ số thường được chọn phụ thuộc vào mức độ chấp nhận rủi ro hoặc kỳ vọng của nhà giao dịch (ví dụ: $k = 1$ trong Bollinger Band); σ là độ lệch chuẩn của chuỗi giá hoặc lợi nhuận, phản ánh mức độ biến động của tài sản (được ước lượng từ phương pháp MLE).

6.3 Tìm giá Bid & Ask tối ưu

Giá Bid và Ask tối ưu trong giao dịch tài chính là hai mức giá quan trọng giúp xác định điểm mua và bán của một tài sản. Được tính theo công thức:

$$Bid_{t+1} = Q_t - \frac{C_t}{2} + (sign_{OB} * weight_{OB} + sign_{MR} * weight_{MR}) * ticksize$$

$$Ask_{t+1} = Q_t + \frac{C_t}{2} + (sign_{OB} * weight_{OB} + sign_{MR} * weight_{MR}) * ticksize$$

Trong đó:

- Q_t : Fair value tại thời điểm t , được ước lượng thông qua mô hình Kalman Filter. Đây là giá trị tham chiếu cho giá trị hợp lý của tài sản.
- C_t : Spread tại thời điểm t , được tính từ sự chênh lệch giữa giá Bid và Ask.
- $sign_{OB}$: Tín hiệu từ Orderbook Signal.
- $weight_{OB}$: Trọng số cho tín hiệu từ Orderbook Signal.
- $sign_{MR}$: Tín hiệu từ Mean Reversion Signal.
- $weight_{MR}$: Trọng số cho tín hiệu từ Mean Reversion Signal.

Qua việc thử nghiệm với các giá trị khác nhau của các trọng số, nhóm đã tiến hành tính toán lợi nhuận (PnL) cho mỗi cặp trọng số. Kết quả cho thấy cặp trọng số tối ưu, mang lại lợi nhuận cao nhất, được xác định thông qua việc tính toán giá bid và ask.

$$weight_{OB} = 0.5, weight_{MR} = 0.5$$

Từ đó, nhóm chọn cặp trọng số này để tối ưu hóa chiến lược giao dịch.

7 Backtesting

7.1 Mục tiêu

Mục tiêu của quy trình backtesting là đánh giá hiệu quả của chiến lược giao dịch, tối ưu hóa chiến lược, và giảm thiểu rủi ro. Đầu tiên, backtesting giúp xác minh tính khả thi của chiến lược giao dịch bằng cách kiểm tra xem chiến lược phát triển có hoạt động hiệu quả trên dữ liệu lịch sử hay không, từ đó đánh giá tính khả thi của chiến lược trong môi trường thực tế. Đồng thời, qua backtest, ta có thể đo lường lợi nhuận kỳ vọng và các yếu tố rủi ro đi kèm như Sharpe Ratio và Sortino Ratio, các chỉ số quan trọng trong việc đánh giá sự thành công của chiến lược.

Tiếp theo, backtesting hỗ trợ tối ưu hóa chiến lược giao dịch bằng cách điều chỉnh các tham số mô hình như đường SMA, độ dài EMA, ngưỡng Z-score, v.v., nhằm cải thiện kết quả dự báo và ra quyết định. Quá trình này cũng giúp phát hiện các lỗi tiềm ẩn trong chiến lược, chẳng hạn như việc tạo ra quá nhiều hoặc quá ít tín hiệu giao dịch, từ đó không tận dụng hết các cơ hội.

Cuối cùng, backtesting giúp giảm thiểu rủi ro bằng cách đánh giá các yếu tố rủi ro của chiến lược, bao gồm tỷ lệ thắng/thua, sự biến động của giá trị danh mục, khả năng thua lỗ và sự phục hồi sau các giai đoạn thua lỗ lớn. Kết quả từ quá trình backtest cho phép tối ưu hóa quản lý rủi ro, xác định các mức độ rủi ro chấp nhận được và điều chỉnh tham số chiến lược để bảo vệ vốn và tối ưu hóa lợi nhuận khi triển khai chiến lược trong thực tế.

7.2 Quy trình Backtest

Quy trình Backtest bao gồm các bước cụ thể nhằm đánh giá hiệu quả của chiến lược giao dịch. Đầu tiên, giá trị Fair Value được tìm ra bằng cách sử dụng mô hình Kalman Filter, đã được xây dựng và hiệu chỉnh trước đó, để ước lượng mức giá Fair Value từ dữ liệu giá lịch sử và các tham số ước lượng từ mô hình Ornstein-Uhlenbeck. Sau đó, một khung thời gian cụ thể trong bộ dữ liệu sẽ được lựa chọn để thực hiện backtest. Tiếp theo, tín hiệu Mean Reversion được tạo ra bằng cách sử dụng mô hình LSTM đã được huấn luyện và lưu lại trước đó, để dự đoán tín hiệu mean reversion signal trên bộ dữ liệu backtest. Tín hiệu Order Book cũng được tạo ra từ mô hình SVM đã huấn luyện và lưu lại trước đó, giúp dự đoán tín hiệu order book signal.

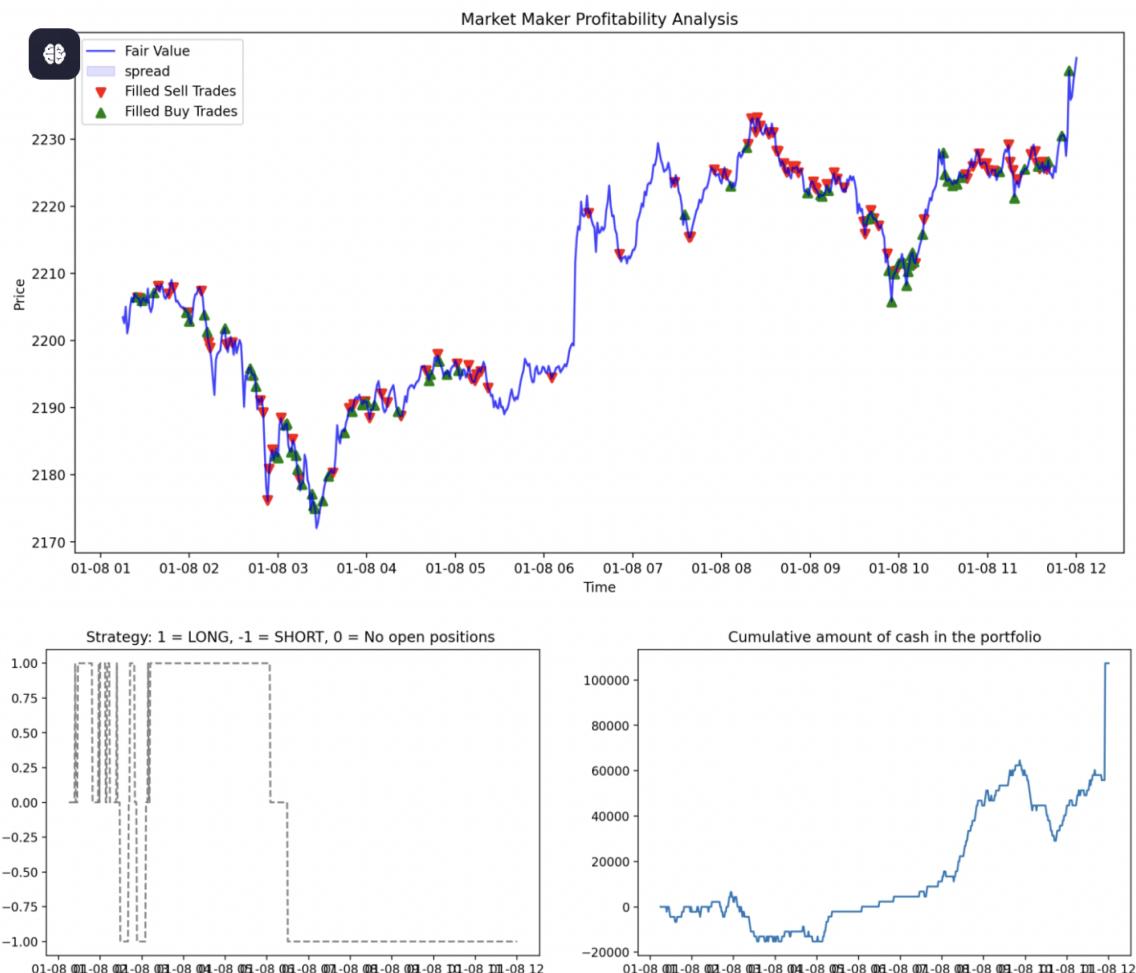
Sau khi có các tín hiệu cần thiết, lợi nhuận và thua lỗ (PnL) từ các giao dịch được thực hiện sẽ được tính toán để đánh giá hiệu quả chiến lược, sử dụng các chỉ số như tổng lợi nhuận, tỷ lệ Sharpe, Drawdown, và tỷ lệ thắng/thua. Để làm benchmark, chiến

lược Mean Reversion Market Making sẽ được so sánh với các chiến lược khác như EMA Crossover, Bollinger Bands. Việc so sánh này giúp đánh giá hiệu suất của chiến lược trong các tiêu chí như lợi nhuận, tỷ lệ thắng/thua, và khả năng quản lý rủi ro.

Kết quả từ quy trình backtest sẽ cung cấp thông tin chi tiết về hiệu quả và tiềm năng của chiến lược, đồng thời giúp phát hiện các điểm cần cải thiện để tối ưu hóa hiệu suất trong tương lai.

8 Kết quả

8.1 Kết quả đạt được



Hình 16: Profit and Loss backtest trên dữ liệu trade

Hình trên bao gồm ba biểu đồ đưa ra cái nhìn tổng quan về hiệu suất, số lượng các lệnh trade được fill thành công và tổng lợi nhuận/lỗ (PnL) của chiến lược đã được đề xuất.

Qua đó, có thể thấy tổng số lệnh trade được fill là 213 và trung bình giá trị khớp lệnh là 2213.97643. Và các lệnh trade có xu hướng được fill nhiều hơn ở các thời điểm giá không có quá nhiều sự biến động bất thường.

Biểu đồ bên trái cho thấy vị thế của chiến lược thay đổi liên tục (từ vị thế mua sang vị thế bán và ngược lại) tại các thời điểm khác nhau. Điều này thể hiện đúng vai trò của một chiến lược Market Making, đó là cung cấp và duy trì thanh khoản cho thị trường.

Tuy nhiên, chiến lược chưa cho thấy hiệu suất ổn định khi mà lượng tiền mặt nắm giữ có những thời điểm giảm và tăng bất thường. Điều này có thể do nhiều yếu tố chưa được xét tới, ví dụ như lượng tồn kho (inventory).

8.2 Hiệu suất các tín hiệu Order Book và Mean Reversion.

Tín hiệu Mean Reversion

	Performance Metrics for Best Model (50 Epochs):			
	precision	recall	f1-score	support
Sideway	0.47	0.51	0.49	875
Go Up	0.33	0.40	0.36	571
Go Down	0.28	0.18	0.22	560
accuracy			0.39	2006
macro avg	0.36	0.36	0.36	2006
weighted avg	0.38	0.39	0.38	2006

Hình 17: Classification Report của LSTM model cho Mean Reversion signal

Về hiệu suất của từng lớp:

- Lớp 0 (giá giảm - go down): Độ thu hồi (recall) rất thấp (0.18) và độ chính xác (precision) ở mức thấp (0.28), dẫn đến F1-score thấp 0.22.
- Lớp 2 (giá tăng - go up): Độ chính xác trung bình (0.33) nhưng độ thu hồi trung bình (0.22), dẫn đến F1-score chưa được tối ưu là 0.36.
- Lớp 1 (giá đi ngang - sideway): Lớp đạt độ thu hồi tốt nhất (0.47) nhưng độ chính xác cao nhất (0.51), cho F1-score cao nhất là 0.49.

Về các chỉ số tổng thể: Độ chính xác (accuracy) chỉ đạt khoảng 39%, khá thấp cho một bài toán phân loại. Cả macro và weighted average đều chưa thực sự tốt, với F1-score lần lượt là 0.36 và 0.38.

Qua đó, nhìn chung có sự đánh đổi rõ rệt giữa độ chính xác và độ thu hồi, đặc biệt ở lớp 2. Độ thu hồi (recall) dao động lớn (từ 0.18 đến 0.51) cho thấy khả năng dự đoán các lớp không nhất quán. Điều này có thể là do chất lượng dữ liệu thu thập chưa được đảm bảo, và có thể được cải thiện bằng cách mở rộng quy mô tập dữ liệu, đặc biệt chú trọng việc thu thập thêm mẫu cho các lớp còn hạn chế.

Tín hiệu Order Book

	precision	recall	f1-score	support
0	0.45	0.06	0.11	749
1	0.30	0.47	0.37	646
2	0.30	0.44	0.35	621
accuracy			0.31	2016
macro avg	0.35	0.32	0.28	2016
weighted avg	0.35	0.31	0.27	2016

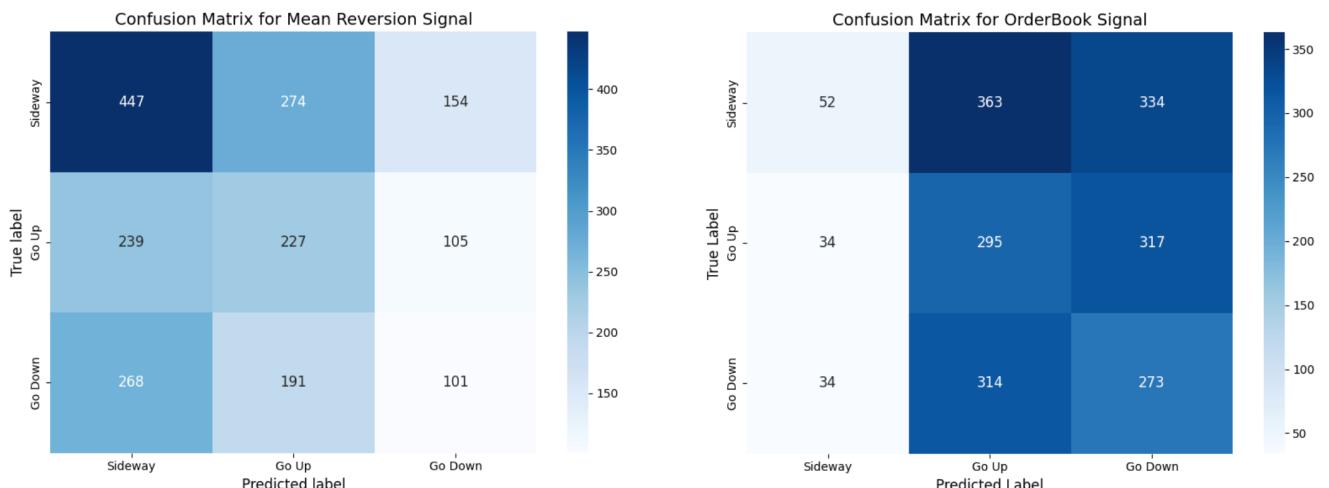
Hình 18: Classification Report của SVM model cho Order Book signal

Về mặt độ chính xác (precision), mô hình đạt hiệu suất tốt nhất ở lớp 0 với độ chính xác là 0.45. Xét về độ thu hồi (recall), mô hình thể hiện khả năng nhận diện tương đối tốt đối với lớp 1 và lớp 2, với các chỉ số lần lượt là 0.47 và 0.44.

Độ chính xác tổng thể (accuracy) của mô hình đạt 0.31, phản ánh khả năng dự đoán tương đối trên toàn bộ tập dữ liệu. Các chỉ số macro average và weighted average duy trì được sự ổn định với precision đạt 0.35 ở cả hai phương pháp đánh giá.

Tương tự như mô hình LSTM của phần Mean Reversion signal, hiệu suất mô hình chưa thực sự tốt có thể là do chất lượng dữ liệu thu thập chưa được đảm bảo, và có thể được cải thiện bằng cách mở rộng quy mô tập dữ liệu.

Kết hợp cả hai tín hiệu



Hình 19: Confusion Matrix cho mô hình Mean Reversion Signal và OrderBook Signal

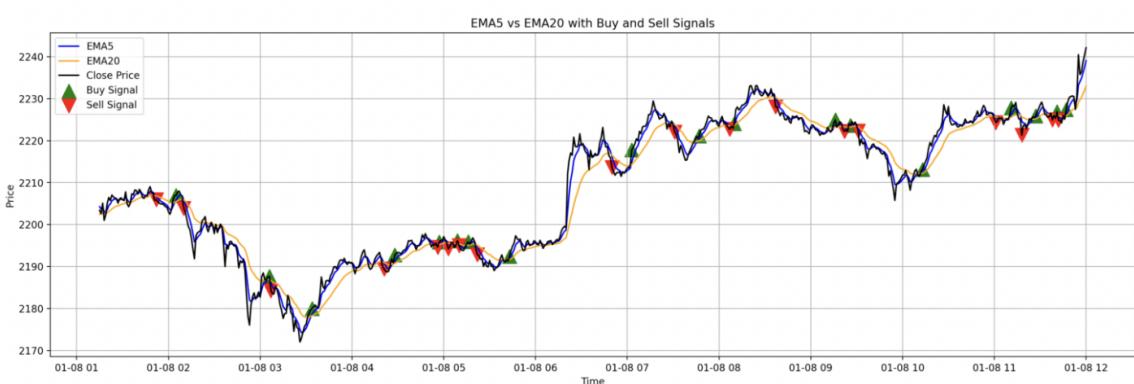
Việc kết hợp cả hai tín hiệu Order Book và Mean Reversion mang lại những lợi ích đáng kể trong việc cải thiện hiệu suất dự đoán xu hướng thị trường. Tín hiệu Mean Reversion cho thấy khả năng phân biệt tốt hơn khi thị trường đi ngang, với F1-score cao nhất đạt 0.49, trong khi tín hiệu Order Book thể hiện sự ổn định và khả năng nhận diện xu hướng tăng và giảm tốt hơn. Sự kết hợp này giúp tận dụng điểm mạnh của từng mô hình riêng lẻ, tăng độ bao phủ và giảm nhầm lẫn, đặc biệt giữa hai lớp giá tăng và giá giảm. Ngoài ra, việc đồng thời sử dụng hai tín hiệu cũng giúp cải thiện tính nhất quán của mô hình và giảm sự dao động trong hiệu suất dự đoán trên các lớp.

Tuy nhiên, việc kết hợp hai tín hiệu cũng đặt ra những thách thức đáng kể. Hiệu quả của mô hình kết hợp phụ thuộc nhiều vào chất lượng dữ liệu đầu vào. Do đó, cần xử lý kỹ lưỡng, cân bằng dữ liệu giữa các lớp và đảm bảo tính đồng nhất trong tập dữ liệu. Hơn nữa, việc kết hợp tín hiệu sẽ làm tăng độ phức tạp của mô hình và đòi hỏi các phương pháp tối ưu hóa phù hợp, chẳng hạn như học kết hợp (ensemble learning) để cân bằng trọng số giữa hai tín hiệu.

Dù vậy, kết quả thử nghiệm cho thấy sự kết hợp này có tiềm năng cải thiện đáng kể các chỉ số hiệu suất như độ chính xác, F1-score, và độ ổn định trên toàn bộ tập dữ liệu. Để phát huy tối đa hiệu quả, cần tiếp tục mở rộng tập dữ liệu, đặc biệt ở các lớp bị hạn chế, và thử nghiệm nhiều chiến lược kết hợp khác nhau để tối ưu hóa mô hình trong các điều kiện thị trường thực tế.

8.3 So sánh với các benchmark khác

1. Exponential Moving Average



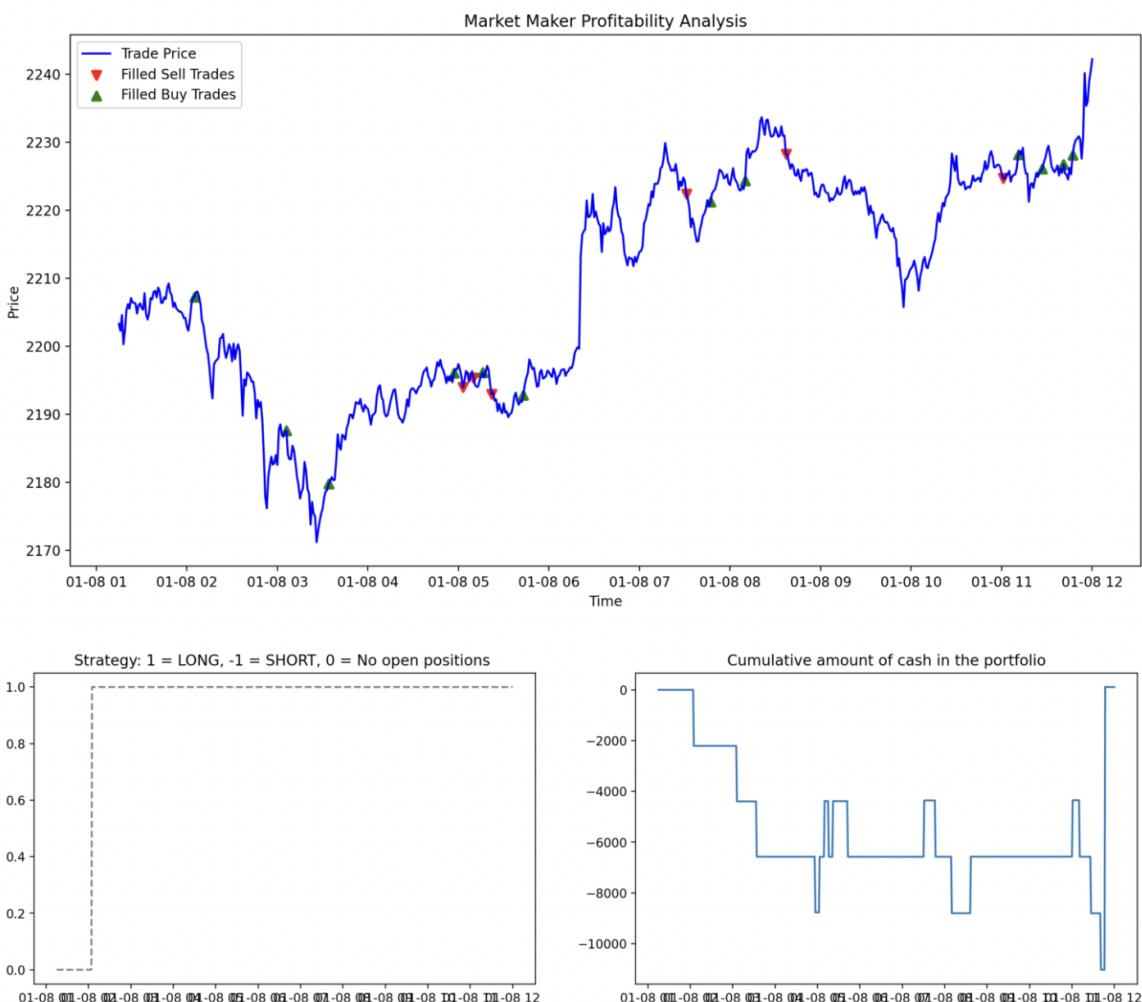
Hình 20: Tín hiệu Buy/Sell dựa vào tín hiệu EMA

EMA (Exponential Moving Average) là benchmark đầu tiên được sử dụng để so

sánh dùng để phát hiện xu hướng. Nếu EMA5 cắt lên EMA20, đây là tín hiệu mua (uptrend), nếu EMA5 cắt xuống EMA20, đây là tín hiệu bán (downtrend).

Biểu đồ giúp ta theo dõi trạng thái các vị thế giao dịch trong suốt khoảng thời gian đang phân tích, giúp hiểu rõ khi nào chiến lược của đang mở một vị thế mua (LONG) hoặc bán (SHORT), và khi nào không có giao dịch (No positions).

Biểu đồ 'Profit & Loss' khi sử dụng EMA' cho thấy tổng quan về lợi nhuận/lỗ (PnL) trong danh mục giao dịch theo thời gian khi sử dụng chiến lược EMA.



Hình 21: Profit & Loss khi sử dụng EMA

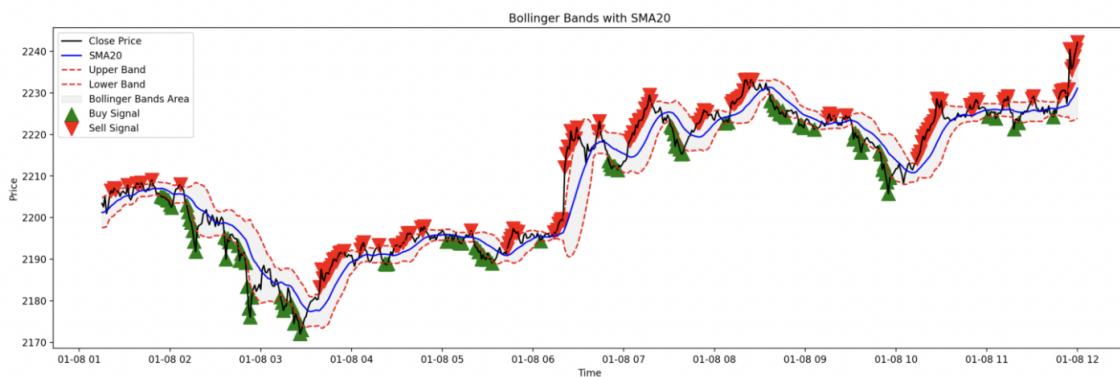
Biểu đồ bên trái cho thấy tình trạng giao dịch với các giá trị 1 (long), -1 (short) và 0 (không có vị thế mở).

Biểu đồ bên phải thể hiện tổng số tiền trong danh mục theo thời gian. Có thể thấy rằng trạng thái tiền mặt có những thời điểm giảm và tăng, cho thấy sự biến động

trong lợi nhuận/lỗ.

2. Bollinger Bands

Tín hiệu mua được phát sinh khi thị trường bán quá mức, giá dưới Lower Band. Tương tự, tín hiệu bán được phát sinh khi thị trường mua quá mức, giá trên Upper Band.

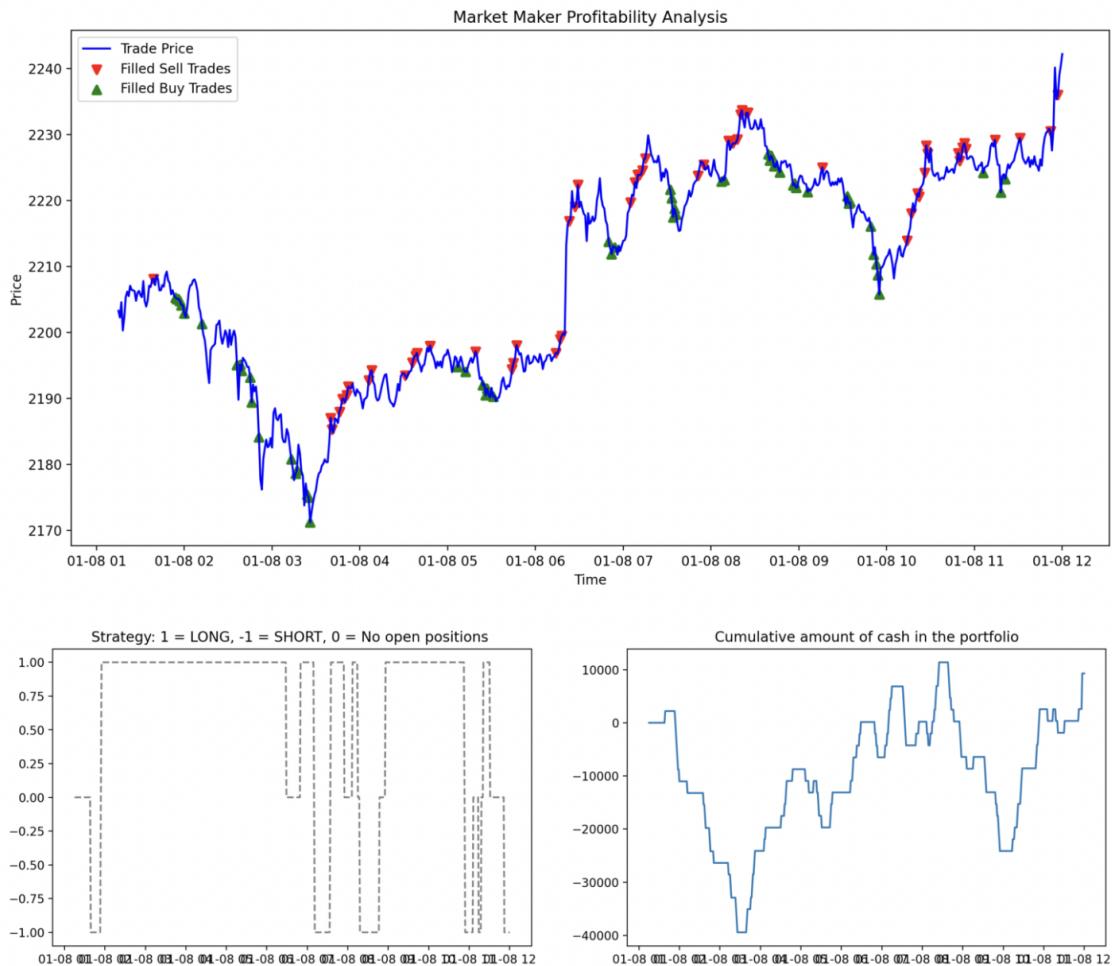


Hình 22: Tín hiệu Buy/Sell dựa vào tín hiệu Bollinger Bands

Biểu đồ 'Profit & Loss khi sử dụng tín hiệu Bollinger Bands' cho thấy tổng quan về lợi nhuận/lỗ (PnL) trong danh mục giao dịch theo thời gian khi sử dụng chiến lược Bollinger Bands.

Biểu đồ bên trái cho thấy các trạng thái giao dịch với giá trị 1 (long), -1 (short) và 0 (không có vị thế mở). Sự dao động cho thấy có nhiều cơ hội giao dịch được thực hiện dựa trên chiến lược Bollinger Bands.

Biểu đồ bên phải thể hiện số tiền tích lũy trong danh mục theo thời gian. Xu hướng đi lên cho thấy một giai đoạn lợi nhuận tích cực.



Hình 23: Profit & Loss khi sử dụng tín hiệu Bollinger Bands

8.4 Nhận xét chung

Mỗi chiến lược Market Making đều có những ưu điểm và nhược điểm riêng, phù hợp với các điều kiện thị trường khác nhau. Đối với tín hiệu từ EMA (EMA5 cắt lên/xuống EMA20), chiến lược này phản ứng nhanh với sự thay đổi của thị trường và dễ dàng phát hiện xu hướng mới, đồng thời đơn giản và nhanh chóng trong việc áp dụng. Tuy nhiên, trong trường hợp thị trường đi ngang hoặc không có xu hướng rõ ràng, tín hiệu cắt nhau có thể không chính xác, dẫn đến các tín hiệu sai (false signals), và việc nhận ra sự đảo chiều của thị trường có thể mất thời gian.

Tín hiệu Bollinger Bands lại phù hợp với các thị trường dao động quanh một mức giá trung bình, giúp nhận diện mức giá thấp và cao dễ dàng. Tuy nhiên, chiến lược này không hiệu quả trong thị trường có xu hướng mạnh, khi giá có thể không quay lại mức trung bình mà tiếp tục di chuyển theo xu hướng hiện tại.

Chiến lược tính toán (Fair Value & Optimized Bid/Ask) kết hợp nhiều tín hiệu từ cả Order Book và Mean Reversion, đồng thời sử dụng mô hình Kalman Filter để đưa ra các quyết định giao dịch thông minh và tối ưu hóa giá trị Bid/Ask. Chiến lược này tạo ra tín hiệu mạnh mẽ dựa trên cả phân tích thị trường và các yếu tố ngoại vi (Order Book), tuy nhiên, đòi hỏi tính toán phức tạp, dữ liệu lớn và quá trình backtesting dài. Mô hình cũng có thể gặp khó khăn khi thị trường không ổn định hoặc không tuân theo các giả thuyết của mô hình, ví dụ như không dừng hoặc quá nhiều biến động, điều này có thể làm giảm hiệu quả của chiến lược.

9 Kết luận

Đồ án đã thực hiện được việc áp dụng Mean Reversion vào thuật toán Market Making, mang lại một cách tiếp cận hiệu quả để dự đoán giá và tối ưu hóa giao dịch. Mặc dù vẫn tồn tại một số hạn chế, nhưng chiến lược này cho thấy tiềm năng lớn trong việc khai thác cơ hội từ thị trường tài chính, đồng thời tạo nền tảng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực giao dịch thuật toán.

Chiến lược này có thể được ứng dụng rộng rãi trong giao dịch tự động, giúp các quỹ đầu tư và nhà giao dịch tối ưu hóa các quyết định giao dịch thông qua việc kết hợp tín hiệu từ mô hình Kalman Filter, SVM và LSTM. Hơn nữa, với khả năng cập nhật liên tục các tham số của mô hình, chiến lược này duy trì tính hiệu quả trong các điều kiện thị trường thay đổi nhanh chóng. Ngoài ra, chiến lược còn hỗ trợ trong quản lý rủi ro, như dự báo và điều chỉnh tỷ lệ đòn bẩy, giúp các tổ chức tài chính giảm thiểu rủi ro và tối ưu hóa lợi nhuận. Bên cạnh đó, việc áp dụng chiến lược Mean Reversion còn giúp nhà đầu tư phát hiện các cơ hội giao dịch trái ngược với xu hướng ngắn hạn, từ đó tối ưu hóa danh mục đầu tư.

Chiến lược này không chỉ có thể áp dụng cho các tài sản chứng khoán mà còn mở rộng ra các thị trường ngoại hối, hàng hóa và sản phẩm phái sinh, giúp nâng cao hiệu quả đầu tư trong nhiều lĩnh vực khác nhau. Đặc biệt, việc tối ưu hóa giá Bid/Ask rất hữu ích trong các thị trường thiếu thanh khoản, giúp giảm thiểu chi phí giao dịch và cải thiện khả năng tham gia vào các giao dịch lớn.

Định hướng nghiên cứu trong tương lai có thể tập trung vào việc nâng cao mô hình dự báo giá, đặc biệt là cải thiện mô hình Kalman Filter để ước lượng giá trị Fair Value chính xác hơn, cũng như ứng dụng các mô hình học sâu như LSTM, GRU hay Transformer để nâng cao độ chính xác trong dự báo. Ngoài ra, việc cải thiện tín hiệu từ Order Book cũng sẽ mở rộng khả năng phân tích và dự báo giao dịch phức tạp hơn, thông qua việc áp dụng các phương pháp học máy như Autoencoders hoặc Reinforcement Learning.Thêm vào đó, nghiên cứu về khả năng backtest và mô phỏng với dữ liệu lớn và thực tế sẽ giúp kiểm tra hiệu quả của chiến lược trong nhiều điều kiện thị trường khác nhau, đồng thời tăng cường việc đánh giá hiệu suất chiến lược qua các chỉ số như Sharpe Ratio và Maximum Drawdown để có cái nhìn chính xác hơn về mức độ rủi ro và lợi nhuận của chiến lược.

10 Ngoài lề

10.1 Ước lượng tham số MLE - cải tiến

Các tham số tối ưu để mô hình hoá một quá trình hoàn nguyên trung bình là không xác định và cần được ước tính. Trên thực tế, các tham số có thể sẽ phát triển theo thời gian, đòi hỏi chúng phải được ước tính lại trong một số khoảng thời gian.

Phương pháp MLE được sử dụng để ước lượng các tham số chưa biết của mô hình $\Theta = [\gamma, \mu, \sigma]$. Dữ liệu được thu thập là một tập hợp chuỗi giá $q = q_0, q_1, \dots, q_n$ tương ứng với các thời điểm t_0, t_1, \dots, t_n (các thời điểm thu thập không nhất thiết phải cách đều nhau)

Dựa trên giả thiết rằng điều kiện phân phối của $Q_t|q_{t-1}$ tuân theo phân phối chuẩn, hàm log-likelihood được biểu diễn như sau:

$$L(\Theta, q) = -\frac{n}{2} \log \frac{\sigma^2}{2\gamma} - \frac{1}{2} \sum_{i=1}^n \log(1 - e^{-2\gamma\Delta t}) - \frac{\gamma}{\sigma^2} \sum_{i=1}^n \frac{(q_i - \mu - (q_{i-1} - \mu)e^{-\gamma\Delta t})^2}{1 - e^{-2\gamma\Delta t}} \quad (10.1)$$

Trong đó:

- $\Delta t_i = t_i - t_{i-1}$ với $i = 1, \dots, n$.
- q_i : Giá trị quan sát tại thời điểm t_i

Ước tính tham số OU tối ưu bao gồm giải bài toán

$$\max_{\Theta} L(\Theta, q) \quad (10.2)$$

Từ lý thuyết tối ưu, nghiệm của bài toán này phải thỏa mãn điều kiện cần bậc nhất. Điều này nghĩa là đạo hàm riêng của hàm log-likelihood $\frac{\delta L}{\delta \mu}(\Theta, q) = 0$. Khi đó, giá trị tối ưu của μ là:

$$\mu = \left(\sum_{i=1}^n \frac{x_{t_i} - x_{t_{i-1}} e^{-\gamma\Delta t_i}}{1 + e^{-\gamma\Delta t_i}} \right) \left(\sum_{i=1}^n \frac{1 - e^{-\gamma\Delta t_i}}{1 + e^{-\gamma\Delta t_i}} \right) \quad (10.3)$$

Tương tự, điều kiện cần bậc nhất đối với $\sigma^2 \frac{\delta L}{\delta \sigma^2}(\Theta, q) = 0$ cho ta giá trị tối ưu của σ^2 là:

$$\sigma^2 = \frac{2\gamma}{n} \sum_{i=1}^n \frac{(x_{t_i} - \mu - (x_{t_{i-1}} - \mu)e^{-\gamma\Delta t_i})^2}{1 - e^{-2\gamma\Delta t_i}} \quad (10.4)$$

Để thỏa mãn điều kiện tối ưu cuối cùng, đạo hàm riêng $\frac{\delta L}{\delta \gamma}(\Theta, q) = 0$ cũng phải được thỏa mãn tại nghiệm tối ưu. Tuy nhiên, biểu thức của $\frac{\delta L}{\delta \gamma}(\Theta, q)$ khá phức tạp và không dẫn đến nghiệm tường minh như trường hợp của các tham số μ, σ^2 .

Dẫu vậy, các mối quan hệ trên đóng vai trò quan trọng trong việc thiết kế thuật toán để giải bài toán tối ưu.

[2] trình bày ba cách giải khác nhau để ước lượng tham số của mô hình Ornstein-Uhlenbeck (OU).

- (1) Sử dụng một bộ giải số đa chiều để tìm tập hợp các tham số (μ, σ^2, γ) sao cho giảm thiểu giá trị nghịch đảo (số âm) của phương trình (2.1), trong đó Q_t và $Var[Q_t]$ được xác định dựa trên mô hình OU.
- (2) Thay các điều kiện bậc nhất (2.3) và (2.4) của μ và σ^2 vào hàm log-likelihood (2.1). Giải bài toán tối ưu một chiều để tìm γ_{MLE} . Sau đó thay γ_{MLE} vào (2.3), (2.4) để tính μ_{MLE}, σ_{MLE}
- (3) Tương tự phương pháp 2, Thay các điều kiện bậc nhất (2.3) và (2.4) của μ và σ^2 vào hàm phương trình $\frac{\Delta L}{\Delta \gamma}(\Theta, q) = 0$. Sử dụng một thuật toán tìm nghiệm (root finder) để tìm γ_{MLE} . Sau đó thay γ_{MLE} vào (2.3), (2.4) để tính μ_{MLE}, σ_{MLE}

Bảng 5: Thư viện python cho ba cách ước lượng tham số

	Solver	Options
Multi	scipy.optimize.minimize	method = ‘L-BFGS-B’, bounds = ((None, None), (0.05, None), (0.05,None)), jac = supplied
Scalar	scipy.optimize.minimize	method = ‘bounded’, bounds = (0, 10)
Scalar root	scipy.root scalar	

Bài báo cho thấy cả ba phương pháp đều tương đương về mặt thống kê, với các tham số ước lượng tương đồng. Tuy nhiên, có sự khác biệt nhỏ về tốc độ (từ nhanh nhất đến chậm nhất: 2, 3, 1). Do đó, chúng tôi chọn phương pháp thứ 1 để đơn giản hóa quá trình thực hiện.

Phương pháp sẽ được áp dụng trong việc triển khai Kalman Filter để ước lượng và cập nhật tham số của mô hình OU một cách đệ quy. Kalman Filter cho phép dự đoán giá trị tài sản bằng cách kết hợp ước lượng từ mô hình OU và quan sát thực tế theo thời gian.

10.2 Chuyển đổi tham số OU sang ma trận của Bộ lọc Kalman

Giả sử ta đã có các tham số của quá trình Ornstein-Uhlenbeck (OU): γ , σ , và Δt , chúng ta sẽ chuyển đổi các tham số này thành các ma trận Kalman sau:

- **Ma trận chuyển tiếp F** : Đây là ma trận mô tả sự thay đổi trạng thái theo thời gian, thể hiện tốc độ quay lại trung bình trong quá trình OU:

$$F = \exp(-\gamma \cdot \Delta t)$$

- **Ma trận quan sát H** : Đây là ma trận mô tả sự quan sát trực tiếp của quá trình, thường là 1 nếu quan sát là giá trị thực tế của quá trình OU:

$$H = 1$$

- **Độ lệch chuẩn ban đầu của trạng thái P_0** : Được tính từ phương trình sau:

$$P_0 = \frac{\sigma^2}{2\gamma}$$

- **Độ biến động chuyển tiếp Q** : Đây là sự không chắc chắn trong mô hình quá trình OU:

$$Q = \frac{\sigma^2 \cdot (1 - \exp(-2\gamma \cdot \Delta t))}{2\gamma}$$

- **Độ nhiễu quan sát R** : Đây là tham số điều chỉnh độ nhiễu trong phép đo. Giá trị này thường được gán là một số cố định, ví dụ:

$$R = 0.1$$

Sau khi tính toán các ma trận trên, ta có thể sử dụng chúng trong bộ lọc Kalman. Bộ lọc Kalman có các tham số sau:

- **Ma trận chuyển tiếp F**
- **Ma trận quan sát H**
- **Giá trị ban đầu của trạng thái $\mu_0 = \text{resampled_close_price}[0]$**
- **Độ lệch chuẩn ban đầu P_0**
- **Độ nhiễu quan sát R**
- **Độ biến động chuyển tiếp Q**

Bộ lọc Kalman sẽ được áp dụng lên chuỗi dữ liệu giá đóng cửa đã được làm mẫu lại (resampled_close_price) qua các vòng lặp tối ưu hóa với phương pháp EM (Expectation-Maximization). Sau đó, các giá trị trạng thái được ước tính bằng cách sử dụng bộ lọc Kalman:

$$\hat{x}_t = \text{KalmanFilter}(F, H, \mu_0, P_0, R, Q)$$

11 Tài liệu tham khảo

Tài liệu

- [1] Michael Sekatchev and Zhengxiang Zhou. (2024). *Stochastic approaches to asset price analysis*. [arXiv:2407.06745](https://arxiv.org/abs/2407.06745),
- [2] Franco, J. C. G., & Onward. (n.d.). *Maximum Likelihood Estimation of a Mean Reverting Process*. Semantic Scholar, Corpus ID: 15749742. http://www.investmentscience.com/MLE_for_OR_mean_reverting.pdf
- [3] Mitchell, C. (2024, June 13). *How to Use a Moving Average to Buy Stocks?*. Investopedia. <https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp>
- [4] Tanmoy Chakraborty and Michael J. Kearns. (2011). *Market Making and Mean Reversion*. Conference on Electronic Commerce (EC-2011). [10.1145/1993574.1993622](https://doi.org/10.1145/1993574.1993622).
- [5] Krause, F., & Calliess, J. (2021). *Neural Policy Learning of Interpretable Trading Strategies using Inductive Prior Knowledge*. SSRN Electronic Journal. [abstract_id=3953228](https://doi.org/10.2139/ssrn.3953228)
- [6] Xiaodong, Li, Xiaotie, Deng, Shanfeng, Zhu, Feng, Wang, Haoran, & Xie. (2014). *An intelligent market making strategy in algorithmic trading*. Department of Computer Science, City University of Hong Kong. <http://www.cqvip.com/QK/71018X/201404/662063727.html>