

Học máy tài chính

Đại học

Bryan Kelly Yale và AQR Capital Management

bryan.kelly@yale.edu

Đại học

DATHÀNH TÚ Chicago

dathanh.xiu@chicagobooth.edu

Nội dung

1 Giới thiệu: Trường hợp về Học máy Tài chính	1.1 Giá cả là Dự đoán .	2
1.2 Tập hợp thông tin lớn	2
1.3 Các hình thức chức năng không rõ ràng	3
1.4 Học máy so với Kinh tế lượng	5
1.5 Những thách thức của việc áp dụng Machine Learning trong Tài chính (và lợi ích của cơ cấu kinh tế)	6
1.6 Nội dung kinh tế (Hai nền văn hóa kinh tế tài chính)	10
2 Ưu điểm của các mô hình phức tạp	2.1 Công cụ phân tích các mô hình học máy .	15
2.2 Lớn hơn thường tốt hơn	16
2.3 Cái nêm phức tạp	20
2.4 Cải thiện	26
3 Dự đoán lợi nhuận	3.1 Dữ liệu .	28
3.2 Thiết kế thí nghiệm	31
3.3 Điểm chuẩn: Mô hình tuyến tính đơn giản	32
3.4 Mô hình tuyến tính bị phạt	37
3.5 Giảm kích thước	40
3.6 Cây quyết định	44
3.7 Mạng lưới thần kinh Vanilla	53
		59

3.8 Phân tích so sánh	64
3.9 Mạng lưới thần kinh phức tạp hơn	68
3.10 Mô hình dự đoán trả về cho dữ liệu "Thay thế"	70
4 Sự đánh đổi rủi ro-lợi nhuận	79
4.1 Nền tảng APT	79
4.2 Mô hình nhân tố vô điều kiện	80
4.3 Mô hình nhân tố có điều kiện	88
4.4 Mô hình nhân tố phức hợp	94
4.5 Mô hình tần số cao	96
4.6 Chữ cái	98
5 Danh mục đầu tư tối ưu 5.1	105
Danh mục đầu tư "plug-in"	107
5.2 Ước tính và tối ưu hóa tích hợp	111
5.3 Hồi quy tỷ lệ Sharpe tối đa	112
5.4 MSRR có độ phức tạp cao	116
5.5 Ước tính SDF và lựa chọn danh mục đầu tư	118
5.6 Chi phí giao dịch và Học tập cùng có	126
6. Kết luận	132
Người giới thiệu	134

Học máy tài chính Bryan Kelly¹ và DaThành Xiu² 1 Trưởng Quản

lý Yale, Quản lý vốn AQR và NBER;

bryan.kelly@yale.edu 2Trưởng Kinh doanh Booth của Đại học Chicago;

dachen.xiu@chicagobooth.edu

TRÙU TƯỢNG

Chúng tôi khảo sát các tài liệu mới về học máy trong nghiên cứu thị trường tài chính. Chúng tôi nêu bật những ví dụ điển hình nhất về những gì dòng nghiên cứu này mang lại và đề xuất những hướng đi dài hứa hẹn cho nghiên cứu trong tương lai. Cuộc khảo sát này được thiết kế cho cả các nhà kinh tế tài chính quan tâm đến việc nắm bắt các công cụ học máy cũng như cho các nhà thống kê và người học máy đang tìm kiếm bối cảnh tài chính thú vị nơi các phương pháp tiên tiến có thể được triển khai.

Sự nhìn nhận. Chúng tôi xin cảm ơn Doron Avramov, Max Helman , Antti Ilmanen, Sydney Ludvigson, Yueran Ma, Semyon Malamud, Tobias Moskowitz, Lasse Pedersen, Markus Pelger, Seth Pruitt và Guofu Zhou vì những nhận xét hữu ích. AQR Capital Management là một công ty quản lý đầu tư toàn cầu, có thể áp dụng hoặc không áp dụng các kỹ thuật đầu tư hoặc phương pháp phân tích tương tự như được mô tả ở đây. Quan điểm thể hiện ở đây là của các tác giả và không nhất thiết là của AQR.

1

Giới thiệu: Trường hợp của cỗ máy tài chính

Học hỏi

1.1 Giá là dự đoán

Phân tích hiện đại về thị trường tài chính tập trung vào định nghĩa sau về giá, xuất phát từ điều kiện tối ưu chung của nhà đầu tư:

$$P_{i,t} = E[M_{t+1}X_{i,t+1}|I_t]. \quad (1.1)$$

Nói cách khác, giá hiện tại của một tài sản, $P_{i,t}$, phản ánh đánh giá của nhà đầu tư về lợi nhuận thu được trong tương lai, $X_{i,t+1}$. Những định giá này được chiết khấu dựa trên sở thích của nhà đầu tư, được tóm tắt một cách tổng quát là tỷ lệ thay thế biên thực hiện trong tương lai, M_{t+1} . Giá sau đó được xác định bởi kỳ vọng của nhà đầu tư đối với những đối tượng này dựa trên thông tin điều hòa của chúng. Nói cách khác, giá cả là những dự đoán - chúng phản ánh những dự đoán tốt nhất của nhà đầu tư về lợi nhuận (đã chiết khấu) trong tương lai mà một tài sản mang lại.

Người ta thường phân tích giá theo tỷ suất lợi nhuận kỳ vọng tương đương, hay "tỷ lệ chiết khấu", biểu diễn chuẩn hóa (1.1) tại thời điểm giá t :

$$E[R_{i,t+1}|I_t] = \beta_{i,t}\lambda_t, \quad (1.2)$$

trong đó $R_{i,t+1} = X_{i,t+1}/P_{i,t}$ là lợi nhuận vượt mức của tài sản, $R_{f,t} = E[M_{t+1}|N_0]$ là lãi suất phi rủi ro một kỳ, $\beta_{i,t} = \frac{Cov[M_{t+1}, R_{i,t+1}|N_0]}{Var[M_{t+1}|N_0]}$

1.2. Bộ thông tin rất lớn

3

hiệp phương sai của tài sản với $Mt+1$, và $\lambda t = \frac{\text{Var}[Mt+1|It]}{\text{đặt các trung trung}}.$ Chúng ta có thể hỏi kinh tế dưới dạng giá cả hoặc tỷ lệ chiết khấu, nhưng tài liệu thường chọn cách trình bày tỷ lệ chiết khấu vì một số lý do. Giá thường không cố định trong khi tỷ lệ chiết khấu thường cố định, vì vậy khi các đặc tính thống kê của công cụ ước tính dựa trên các giả định về tính cố định thì sẽ có lợi hơn khi sử dụng tỷ lệ chiết khấu. Ngoài ra, những khác biệt không thú vị về quy mô thu nhập của tài sản sẽ dẫn đến những khác biệt không thú vị về quy mô giá cả. Nhưng tỷ lệ chiết khấu thường không bị ảnh hưởng bởi sự khác biệt trong quy mô chi trả nên nhà nghiên cứu không cần phải điều chỉnh chúng.

Tổng quát hơn, việc nghiên cứu các hiện tượng thị trường về mặt lợi nhuận làm giảm bớt một số gánh nặng lập mô hình của nhà nghiên cứu bằng cách đồng nhất một phần dữ liệu để có các đặc tính động học và quy mô có thể điều chỉnh được.

Ngoài ra, tỷ lệ chiết khấu cũng là những dự đoán và việc giải thích chúng đặc biệt đơn giản và quan trọng về mặt thực tế. $E[Ri,t+1|It]$ mô tả kỳ vọng của nhà đầu tư về sự tăng giá trị tài sản trong giai đoạn tiếp theo. Như vậy, lợi nhuận kỳ vọng là đầu vào quan trọng cho các quyết định phân bổ. Nếu chúng ta cố gắng tách biệt một mô hình thực nghiệm phù hợp chặt chẽ với dữ liệu cho kỳ vọng này, thì chúng ta đã hiểu rõ hơn về chức năng thị trường và đồng thời tìm ra một công cụ để cải thiện việc phân bổ nguồn lực trong tương lai. Đây là một ví dụ điển hình về tính hai mặt trong nghiên cứu khoa học xã hội ứng dụng: Một mô hình tốt vừa nâng cao hiểu biết khoa học vừa cải thiện việc ra quyết định trong thế giới thực.

1.2 Bộ thông tin lớn

Có hai điều kiện nghiên cứu tài chính khiến nó trở thành mảnh đất màu mỡ cho các phương pháp học máy: tập hợp thông tin có điều kiện lớn và các dạng hàm mơ hồ. Rõ ràng ngay từ (1.1) là việc nghiên cứu giá tài sản gắn bó chặt chẽ với thông tin. Các câu hỏi định hướng trong nghiên cứu kinh tế tài chính bao gồm “những người tham gia thị trường có thông tin gì và họ sử dụng thông tin đó như thế nào?” Các dự đoán thể hiện trong giá được định hình bởi thông tin có sẵn phù hợp với khoản hoàn trả tài sản trong tương lai ($Xi,t+1$) và cảm xúc của nhà đầu tư về những khoản hoàn trả đó ($Mt+1$). Nếu giá cả diễn biến giống nhau ở tất cả các quốc gia trên thế giới-ví dụ: nếu mức chi trả và sở thích gần với iid-thì thông tin

bộ mation sẽ bị loại bỏ. Nhưng ngay cả nhà đầu tư ghê bành đang tìm hiểu tài khoản trực tuyến của họ hoặc đọc ấn bản mới nhất của The Wall Street Journal cũng nhanh chóng trực giác được phạm vi rộng lớn của thông tin điều hòa ẩn giấu đằng sau giá cả thị trường. Trong khi đó, chức năng sản xuất của ngành quản lý tài sản hiện đại là minh chứng cho lượng thông tin khổng lồ chảy vào giá tài sản: Các nhà quản lý chuyên nghiệp (theo nhiều cách thủ công và tự động khác nhau) thường xuyên nghiên cứu hàng loạt nguồn cấp tin tức, công bố dữ liệu và dự đoán của chuyên gia trong để thông báo cho các quyết định đầu tư của họ.

Sự mở rộng của thông tin liên quan đến giá cả được kết hợp bởi tính chất bảng của thị trường tài chính. Giá của bất kỳ nội dung nhất định nào cũng có xu hướng thay đổi theo thời gian theo những cách có thể thú vị-điều này tương ứng với thứ nguyên chuỗi thời gian của bảng điều khiển. Trong khi đó, tại một thời điểm nhất định, giá của các tài sản khác nhau theo những cách thú vị-kích thước mặt cắt ngang của bảng dữ liệu. Sự thay đổi chuỗi thời gian trong môi trường thị trường sẽ ảnh hưởng đến nhiều tài sản theo những cách liên kết với nhau. Ví dụ, hầu hết giá tài sản đều biến động khác nhau trong điều kiện rủi ro cao và rủi ro thấp hoặc trong các chế độ chính sách khác nhau. Khi các điều kiện kinh tế vĩ mô thay đổi, giá tài sản sẽ điều chỉnh đồng loạt thông qua những tác động chung này. Ngoài ra, có những hành vi cheo khác biệt với từng tài sản hoặc nhóm tài sản. Vì vậy, thông tin điều hòa về bản chất không chỉ là chuỗi thời gian mà còn bao gồm các thuộc tính ở cấp độ nội dung. Một mô hình tài sản thành công

hành vi phải đồng thời tính đến các tác động động được chia sẻ cũng như các tác động cụ thể của tài sản (bản thân chúng có thể là tĩnh hoặc động).

Như được nhấn mạnh bởi Gu et al. (2020b),

Nghề này đã tích lũy được một danh sách đáng kính ngạc các yếu tố dự đoán mà nhiều nhà nghiên cứu cho rằng có khả năng dự đoán lợi nhuận. Số lượng các đặc điểm dự đoán mức tồn kho được báo cáo trong tài liệu là hàng trăm và các yếu tố dự đoán kinh tế vĩ mô về tổng số thị trường là hàng chục.

Hơn nữa, do xu hướng nghiên cứu kinh tế tài chính là điều tra một hoặc một vài biến số cùng một lúc, có lẽ chúng ta đã bỏ sót nhiều cơ sở. Ví dụ, chỉ gần đây nội dung thông tin của văn bản tin tức mới nổi lên như một đầu vào cho các mô hình thực nghiệm của (1.1),

1.3. Các hình thức chức năng là mơ hồ

5

và còn rất nhiều cơ hội để mở rộng biên giới này và những biên giới khác.

1.3 Các hình thức chức năng là mơ hồ

Nếu giá tài sản là kỳ vọng về kết quả trong tương lai thì các công cụ thống kê để nghiên cứu giá là các mô hình dự báo. Cách tiếp cận kinh tế lượng truyền thống đối với nghiên cứu thị trường tài chính (ví dụ: Hansen và Singleton, 1982) trước tiên chỉ định một dạng chức năng cho mô hình dự báo lợi nhuận được thúc đẩy bởi mô hình kinh tế lý thuyết, sau đó ước tính các tham số để hiểu cách các nguồn thông tin ứng viên liên kết với các dữ liệu được quan sát.

giá thị trường trong giới hạn của mô hình đã chọn. Nhưng chúng ta nên áp dụng mô hình kinh tế nào trong số rất nhiều mô hình kinh tế có sẵn trong tài liệu?

Việc xây dựng điều kiện bậc nhất, hay “phương trình Euler” trong (1.1) đủ rộng để bao gồm nhiều giả định kinh tế mang tính cấu trúc. Tính tổng quát này được đảm bảo vì không có sự đồng thuận về công thức cấu trúc cụ thể nào là khả thi. Các mô hình dựa trên tiêu dùng ban đầu không khớp với dữ liệu giá thị trường ở hầu hết các thước đo (ví dụ Mehra và Prescott, 1985). Các mô hình cấu trúc hiện đại phù hợp hơn với dữ liệu về giá nếu thước đo thành công đủ dễ chấp nhận (ví dụ Chen và cộng sự, 2022a), nhưng phạm vi của hiện tượng mà chúng mô tả có xu hướng bị giới hạn ở một số tài sản và thường chỉ được đánh giá trên cơ sở cơ sở mẫu.

Do sự thành công thực nghiệm còn hạn chế của các mô hình cấu trúc, hầu hết các nghiên cứu thực nghiệm trong hai thập kỷ qua đã chọn từ bỏ các giả định về cấu trúc để chuyển sang các khuôn khổ “dạng rút gọn” hoặc “không chênh lệch giá” ít cứng nhắc hơn. Trong khi nghiên cứu thực nghiệm về thị trường thường tránh xa việc áp đặt cấu trúc kinh tế chi tiết, nó thường áp đặt cấu trúc thống kê (ví dụ, dưới dạng mô hình nhân tố có chiều hướng thấp hoặc các giả định tham số khác). Nhưng có nhiều lựa chọn tiềm năng về cấu trúc thống kê trong các mô hình dạng rút gọn và đáng để khám phá lợi ích của các mô hình linh hoạt có thể chứa nhiều dạng hàm khác nhau và mức độ phi tuyến tính cũng như tương tác biến đổi khác nhau.

Nhập các công cụ học máy như phương pháp hạt nhân, công cụ ước tính khả năng bị phạt, cây quyết định và mạng lưới thần kinh. Bao gồm các công cụ ước tính phi tham số đa dạng và các mô hình tham số lớn, các phương pháp học máy được thiết kế rõ ràng để ước tính gần đúng dữ liệu chưa biết

tạo ra các hàm. Ngoài ra, học máy có thể giúp tích hợp nhiều nguồn dữ liệu vào một mô hình duy nhất. Theo thảo luận ở Phần 1.2, việc lập mô hình hiệu quả về giá và lợi nhuận kỳ vọng đòi hỏi

thông tin điều hòa phong phú trong nó. Về điểm này, Cochrane (2009) lưu ý rằng "Rõ ràng là chúng ta thậm chí không quan sát được tất cả thông tin điều kiện được sử dụng bởi các tác nhân kinh tế và chúng ta không thể đưa ngay cả một phần thông tin điều kiện quan sát được vào mô hình của mình". Hansen và Richard (1987) (và gần đây hơn là Martin và Nagel, 2021) nêu bật sự khác biệt về thông tin mà các nhà đầu tư trong mô hình kinh tế có thể tiếp cận so với thông tin có sẵn cho nhà kinh tế lượng ở bên ngoài mô hình¹

nhin vào. Học máy là một bộ công cụ có thể giúp thu hẹp khoảng cách giữa các tập hợp thông tin của nhà nghiên cứu và người tham gia thị trường bằng cách cung cấp các phương pháp cho phép nhà nghiên cứu tiếp thu các tập thông tin lớn hơn.

Chúng ta càng mở rộng phạm vi xem xét các bộ điều hòa lớn thì mô hình của chúng ta sẽ càng thực tế hơn. Logic tương tự này cũng áp dụng cho câu hỏi về dạng hàm. Những người tham gia thị trường không chỉ đưa thông tin phong phú vào dự báo của mình mà còn thực hiện điều đó theo những cách có thể phức tạp nhằm tận dụng sức mạnh sắc thái của lý luận và trực giác của con người. Chúng ta phải thừa nhận rằng các nhà đầu tư sử dụng thông tin theo những cách mà chúng ta, với tư cách là nhà nghiên cứu, không thể biết rõ ràng và do đó không thể chỉ định một cách thấu đáo (và chắc chắn là không chính xác) trong mô hình thống kê tham số. Giống như Cochrane (2009) nhắc nhở chúng ta phải thận trọng khi xem xét thông tin điều hòa, chúng ta cũng phải thận trọng như vậy khi xem xét các dạng chức năng.

1.4 Học máy so với Kinh tế lượng

Học máy là gì và nó khác với các thước đo kinh tế truyền thống như thế nào? Gu và cộng sự. (2020b) nhấn mạnh rằng định nghĩa về học máy còn rất mơ hồ và thuật ngữ này đôi khi bị sai lệch do mục đích tiếp thị của người dùng. Chúng tôi theo dõi Gu et al. (2020b) và sử dụng thuật ngữ này để mô tả (i) một tập hợp đa dạng các mô hình nhiều chiều cho

¹Người đọc của cuộc khảo sát này được khuyến khích xem lại chương 8 của Cochrane (2009) và nhận ra nhiều cách mà các khái niệm máy học kết hợp với phác thảo của ông về vai trò của thông tin điều hòa trong giá tài sản.

1.4. Học máy so với Kinh tế lượng

7

dự đoán thống kê, kết hợp với (ii) các phương pháp "chính quy hóa" để lựa chọn mô hình và giảm thiểu tình trạng quá phù hợp, và (iii) các thuật toán hiệu quả để tìm kiếm trong số lượng lớn các đặc tả mô hình tiềm năng.

Với định nghĩa này, cần phải rõ ràng rằng, ở bất kỳ phiên bản nào của nó, máy học tài chính đều bao gồm một tập hợp các quy trình để ước tính một mô hình thống kê và sử dụng mô hình đó để đưa ra quyết định. Vì vậy, về cốt lõi, học máy không cần phải phân biệt với kinh tế lượng hoặc thống kê nói chung. Nhiều ý tưởng làm nền tảng cho học máy đã tồn tại thoải mái dưới sự bảo trợ của thống kê trong nhiều thập kỷ (Israel và cộng sự, 2020).

Để học thông qua trải nghiệm dữ liệu, máy cần thể hiện chức năng của những gì nó đang cố gắng học. Nhà nghiên cứu phải đưa ra lựa chọn về cách trình bày đây là bức vẽ mà trên đó dữ liệu sẽ vẽ nên câu chuyện của nó.

Phần (i) trong định nghĩa của chúng tôi chỉ ra rằng học máy mang lại tự duy cởi mở cho các biểu diễn hàm có tính tham số hóa cao và thường phi tuyến tính. Các mô hình nhỏ thì cứng nhắc và đơn giản hóa quá mức, nhưng tính phân tích của chúng có những lợi ích như ước tính tham số tương đối chính xác và dễ diễn giải.

Các mô hình lớn và phức tạp linh hoạt hơn nhiều, nhưng cũng có thể nhạy hơn và có hiệu suất ngoài mẫu kém khi chúng điều chỉnh quá mức nhiều trong hệ thống. Các nhà nghiên cứu chuyển sang các mô hình lớn khi họ tin rằng lợi ích từ việc mô tả chính xác hơn mức độ phức tạp của các hiện tượng trong thế giới thực sẽ lớn hơn chi phí của việc phù hợp quá mức tiềm năng. Ở cấp độ trực quan, học máy là một cách để theo đuổi phân tích thống kê khi nhà phân tích không chắc chắn về cấu trúc cụ thể mà mô hình thống kê của họ nên sử dụng. Theo định nghĩa này, phần lớn học máy có thể được xem là mô hình hóa không tham số (hoặc bán tham số).

Phương thức hoạt động của nó xem xét nhiều thông số kỹ thuật mô hình tiềm năng và yêu cầu hướng dẫn của dữ liệu trong việc chọn mô hình nào hiệu quả nhất cho vấn đề hiện tại. Người ta có thể hỏi: khi nào thì nhà phân tích biết được cấu trúc nào phù hợp cho phân tích thống kê của họ? Tất nhiên, câu trả lời là "không bao giờ", đó là lý do tại sao học máy nói chung có giá trị trong nghiên cứu tài chính. Như Breiman (2001) đã nhấn mạnh, việc tập trung vào việc tối đa hóa độ chính xác của dự đoán khi đối mặt với mô hình dữ liệu chưa biết là đặc điểm khác biệt trung tâm của học máy với mục tiêu thống kê truyền thống là ước tính dữ liệu đã biết.

Giới thiệu: Trường hợp học máy tài chính

tạo mô hình và tiến hành kiểm tra giả thuyết.

Phần (ii) trong định nghĩa của chúng tôi nhấn mạnh rằng học máy chọn một mô hình ưa thích (hoặc kết hợp các mô hình) từ "bộ sưu tập đa dạng" các mô hình ứng cử viên. Một lần nữa, ý tưởng này có một lịch sử phong phú trong kinh tế lượng dưới tiêu đề lựa chọn mô hình (và có liên quan là lấy trung bình mô hình). Sự khác biệt là học máy đặt việc lựa chọn mô hình làm trọng tâm của thiết kế thực nghiệm. Quá trình tìm kiếm qua nhiều mô hình để tìm ra những mô hình có hiệu suất cao nhất (thường được gọi là "điều chỉnh" mô hình) là đặc điểm của tất cả các phương pháp học máy. Tuy nhiên, việc chọn từ nhiều mô hình một cách máy móc sẽ dẫn đến việc khớp quá mức trong mẫu và có thể tạo ra hiệu suất ngoài mẫu kém. Do đó, các quy trình nghiên cứu học máy đi kèm với "chính quy hóa", đây là một thuật ngữ chung để hạn chế kích thước mô hình nhằm khuyến khích hiệu suất ổn định ngoài mẫu. Như Gu và cộng sự. (2020b) đã nói, "Mô hình tối ưu là mô hình 'Goldilocks'. Nó đủ lớn để có thể phát hiện một cách đáng tin cậy các mối quan hệ dự đoán phức tạp tiềm ẩn trong dữ liệu, nhưng không linh hoạt đến mức bị chi phối bởi tình trạng quá khớp và bị thiếu mẫu." Các phương pháp chính quy hóa khuyến khích các mô hình nhỏ hơn; các mô hình phong phú hơn chỉ được chọn nếu chúng có khả năng tăng cường cường thực sự độ chính xác của dự đoán ngoài mẫu.

Yếu tố (iii) trong định nghĩa học máy có lẽ là điểm khác biệt rõ ràng nhất của nó với thống kê truyền thống, nhưng cũng có lẽ là yếu tố ít thú vị nhất về mặt kinh tế. Khi các tập dữ liệu lớn và/hoặc các mô hình được tham số hóa rất nhiều, việc tính toán có thể trở thành một nút thắt cổ chai. Học máy đã phát triển nhiều quy trình tối ưu hóa gần đúng để giảm tải tính toán. Ví dụ, các công cụ ước tính kinh tế lượng truyền thống thường sử dụng tất cả các điểm dữ liệu trong mỗi bước của quy trình tối ưu hóa lặp lại và chỉ dừng tìm kiếm tham số khi quy trình hội tụ. Các phím tắt như sử dụng tập hợp con dữ liệu và tạm dừng tìm kiếm trước khi hội tụ thường làm giảm tính toán và làm như vậy với ít mất độ chính xác (ví dụ: xem, giảm độ dốc ngẫu nhiên và dừng sớm là hai yếu tố chính trong đào tạo mạng thần kinh).

1.5. Những thách thức của việc áp dụng Machine Learning trong tài chính
(và lợi ích của cơ cấu kinh tế)

9

1.5 Những thách thức của việc áp dụng Machine Learning trong Tài chính (và Lợi ích của cơ cấu kinh tế)

Mặc dù nghiên cứu tài chính về nhiều mặt là lý tưởng cho các phương pháp học máy, nhưng một số khía cạnh tài chính cũng đặt ra những thách thức đối với học máy.

Hiểu được những trở ngại này là điều quan trọng để phát triển những kỳ vọng thực tế về lợi ích của việc học máy tài chính.

Đầu tiên, trong khi học máy thường được xem như một công cụ “dữ liệu lớn”, nhiều câu hỏi cơ bản về tài chính lại bị nắn lòng trước thực tế “dữ liệu nhỏ” rõ ràng của chuỗi thời gian kinh tế. Ví dụ, các bộ dữ liệu tiêu chuẩn về tài chính vĩ mô bị giới hạn ở vài trăm quan sát hàng tháng.

Loại khan hiếm dữ liệu này không bình thường trong các lĩnh vực học máy khác, nơi các nhà nghiên cứu thường có dữ liệu không giới hạn (hoặc khả năng tạo dữ liệu mới nếu cần). Trong nghiên cứu chuỗi thời gian, dữ liệu mới chỉ được tích lũy theo thời gian.

Thứ hai, nghiên cứu tài chính thường phải đổi mới với tỷ lệ tín hiệu liên tục âm yếu. Không nơi nào điều này rõ ràng hơn trong dự đoán lợi nhuận, nơi các lực lượng thị trường hiệu quả (tối đa hóa lợi nhuận và cạnh tranh) luôn cố gắng loại bỏ khả năng dự đoán biến động giá (Samuelson, 1965; Fama, 1970). Kết quả là, sự biến động về giá dự kiến chủ yếu xuất phát từ sự xuất hiện của những tin tức không lường trước được (là tiếng ồn không thể dự đoán được từ góc nhìn của mô hình). Thị trường cũng có thể hiện sự kém hiệu quả và sở thích của nhà đầu tư có thể làm tăng phần bù rủi ro thay đổi theo thời gian, dẫn đến khả năng dự đoán được lợi nhuận.

Tuy nhiên, chúng ta nên kỳ vọng khả năng dự đoán lợi nhuận sẽ nhỏ và bị cạnh tranh gay gắt.

Thứ ba, các nhà đầu tư học hỏi và thị trường phát triển. Điều này tạo ra một mục tiêu di động cho các mô hình dự đoán học máy. Các mô hình dự đoán đáng tin cậy trước đây có thể bị loại bỏ. Những thay đổi về quy định và công nghệ làm thay đổi cấu trúc của nền kinh tế. Sự không ổn định về cấu trúc khiến tài chính trở thành một lĩnh vực học tập đặc biệt phức tạp và tạo ra những thách thức về dữ liệu nhỏ và tỷ lệ tín hiệu trên nhiều thấp.

Những thách thức này tạo cơ hội để hưởng lợi từ kiến thức thu được từ lý thuyết kinh tế. Theo ghi nhận của Israel et al. (2020),

“Nguyên tắc cơ bản của phân tích thống kê là các tham số lý thuyết và mô hình có thể thay thế cho nhau. Bạn càng có nhiều cấu trúc

có thể áp đặt vào mô hình của bạn, bạn càng cần ước tính ít tham số hơn và mô hình của bạn có thể sử dụng các điểm dữ liệu có sẵn để loại bỏ nhiều càng hiệu quả hơn. Nghĩa là, các mô hình rất hữu ích vì chúng lọc nhiều. Tuy nhiên, một mô hình được đơn giản hóa quá mức cũng có thể lọc ra một số tín hiệu, vì vậy trong mô trường giàu dữ liệu và có tỷ lệ tín hiệu trên tạp âm cao, bạn sẽ không muốn sử dụng một mô hình nhỏ không cần thiết. Người ta có thể bắt đầu giải quyết các vấn đề về dữ liệu nhỏ và tín hiệu trên nhiều thấp bằng cách đưa lý thuyết kinh tế để mô tả một số khía cạnh của dữ liệu, được bổ sung bởi các công cụ học máy để nắm bắt các khía cạnh của dữ liệu mà lý thuyết không có tiếng nói."

1.6 Nội dung kinh tế (Hai nền văn hóa kinh tế tài chính)

Chúng tôi nhớ lại bài tiểu luận của Breiman (2001) về "hai nền văn hóa" thống kê, có điểm tương đồng trong kinh tế tài chính (với những sửa đổi phù hợp). Một là văn hóa "mô hình cấu trúc/kiem tra giả thuyết", ủng hộ việc áp đặt các giả định cấu trúc được xác định đầy đủ hoặc một phần và nghiên cứu các cơ chế kinh tế thông qua kiểm tra giả thuyết. Chương trình truyền thống về phân tích định giá tài sản theo kinh nghiệm (trước sự xuất hiện của các mô hình hệ số dạng rút gọn và mô hình dự đoán học máy) nghiên cứu giá thông qua lăng kính của các mô hình dự đoán bị ràng buộc năng nề. Các ràng buộc có dạng i) các dạng/phân bố hàm cự thể và ii) các biến giới hạn được thừa nhận vào tập thông tin điều kiện. Những mô hình này thường "khái quát hóa" kém theo nghĩa là chúng có khả năng giải thích yếu đối với các hành vi giá tài sản nằm ngoài phạm vi hẹp của thiết kế mô hình hoặc ngoài tập dữ liệu huấn luyện. Đây là một tuyên bố rõ ràng đến mức người ta hiếm khi xem xét hiệu suất ngoài mẫu của các mô hình định giá tài sản cấu trúc được chỉ định đầy đủ.

Loại còn lại là văn hóa "mô hình dự đoán", coi trọng khả năng giải thích thống kê hơn tất cả, và phần lớn được sinh ra từ những hạn chế của văn hóa cấu trúc đã được thiết lập trước đó. Văn hóa mô hình dự đoán săn sàng tân thành các thông số kỹ thuật của mô hình có thể thiếu mối liên hệ rõ ràng với lý thuyết kinh tế, miễn là chúng tạo ra những cải tiến mạnh mẽ, có ý nghĩa về mức độ phù hợp của dữ liệu so với hiện trạng. TRONG

1.6. Nội dung kinh tế (Hai nền văn hóa kinh tế tài chính)

11

Ngoài mô hình dạng rút gọn vốn đã thống trị hầu hết nền tài chính thực nghiệm kể từ những năm 1990, nghiên cứu máy học tài chính cho đến nay hoàn toàn thuộc về nền văn hóa thứ hai này.

"Không có kinh tế học" là lời buộc tội đôi khi được đưa ra đối với nghiên cứu dự đoán thống kê bởi khán giả hội thảo kinh tế, người thảo luận và người giới thiệu. Lời chỉ trích này thường sai lầm và chúng ta nên đề phòng việc nó làm mất giá trị quá mức những tiến bộ trong lĩnh vực học máy tài chính. Chúng ta đừng bỏ lỡ vai trò kinh tế quan trọng của ngay cả những ứng dụng mô hình thống kê thuần túy nhất trong tài chính. Các mô hình dự đoán tương đối phi cấu trúc khiến chúng không kém phần quan trọng về mặt kinh tế so với kinh tế lượng truyền thống trong việc kiểm tra giả thuyết cấu trúc, chúng chỉ đóng một vai trò khoa học khác. Kiểm tra giả thuyết học kinh tế bằng cách thăm dò các cơ chế kinh tế cụ thể. Nhưng kinh tế học không chỉ là kiểm tra các cơ chế lý thuyết. Các mô hình dự đoán phi lý thuyết (vì thiếu thuật ngữ tốt hơn) khảo sát bối cảnh thực nghiệm theo nghĩa rộng hơn, vạch ra các dữ kiện thực nghiệm mới mà dựa vào đó các lý thuyết có thể được phát triển và các thử nghiệm giả thuyết trong tương lai có thể điều tra các cơ chế. Hai hình thức điều tra thực nghiệm này - kiểm tra độ chính xác và bản đồ tổng quát - đóng vai trò bổ sung cho nhau trong quá trình tiên bộ khoa học của Kuhnian.

Hãy xem xét câu hỏi cơ bản của nghiên cứu định giá tài sản: Điều gì quyết định phần bù rủi ro tài sản? Ngay cả khi chúng ta có thể quan sát hoàn hảo lợi tức kỳ vọng, chúng ta vẫn cần các lý thuyết để giải thích hành vi của chúng và phân tích thực nghiệm để kiểm tra các lý thuyết đó. Nhưng chúng ta không thể quan sát phần bù rủi ro và chúng rất khó ước tính. Học máy đạt được tiến bộ trong việc đo lường phần bù rủi ro, tạo điều kiện phát triển các lý thuyết tốt hơn về cơ chế kinh tế xác định hành vi của chúng.

Lợi ích quan trọng của việc mở rộng tập hợp các đường nét đã biết trong bối cảnh thực nghiệm là ngay cả khi các chi tiết về cơ chế kinh tế vẫn bị che giấu, các tác nhân kinh tế - đặc biệt là những người tham gia thị trường tài chính - luôn có thể được hưởng lợi từ các bản đồ thực nghiệm được cải thiện. Văn hóa mô hình dự đoán có truyền thống lâu đời trong việc tạo ra các nghiên cứu để giúp các nhà đầu tư, người tiêu dùng và các nhà hoạch định chính sách đưa ra quyết định tốt hơn. Những dự đoán được cải thiện sẽ cung cấp những mô tả chính xác hơn về sự phân bổ phụ thuộc vào nhà nước mà các chủ thể kinh tế này phải đối mặt.

Kinh tế nói chung là một lĩnh vực ứng dụng. Tính kinh tế của văn hóa mô hình dự đoán nằm chính xác ở khả năng cải thiện các dự đoán.

Được trang bị những dự đoán tốt hơn-tức là, đánh giá chính xác hơn về tập hợp cơ hội kinh tế-các tác nhân có thể đánh đổi tốt hơn chi phí và lợi ích khi phân bổ các nguồn lực khan hiếm. Điều này nâng cao phúc lợi. Không ở đâu điều này được thể hiện rõ ràng hơn trong vấn đề lựa chọn danh mục đầu tư. Không phải lúc nào chúng ta cũng có thể hiểu được cơ chế kinh tế mà nhờ đó một mô hình mang lại dự báo rủi ro hoặc lợi nhuận tốt hơn; nhưng nếu có, nó sẽ nâng cao lợi ích của các nhà đầu tư và do đó có tầm quan trọng về mặt kinh tế.

Lời chỉ trích trọng tâm của Breiman về văn hóa kiểm tra giả thuyết cấu trúc là

"khi một mô hình phù hợp với dữ liệu để rút ra kết luận định lượng: kết luận là về cơ chế của mô hình chứ không phải về cơ chế của tự nhiên. Nếu mô hình mô phỏng kém về bản chất thì kết luận có thể sai."

Chúng tôi coi đây không phải là sự chỉ trích về mô hình cấu trúc, vốn vẫn phải là nền tảng của tài chính thực nghiệm, mà là động lực và sự bảo vệ cho các mô hình dự đoán. Tất nhiên, sự phân đôi hai nền văn hóa là một bức tranh biếm họa. Nghiên cứu trải rộng trên nhiều phạm vi và dựa trên nhiều công cụ, và các nhà nghiên cứu không tách biệt thành các phe hệ tư tưởng đồng nhất. Cả hai nền văn hóa đều quan trọng về mặt kinh tế. Breiman khuyến khích chúng ta xem xét các mô hình linh hoạt, thậm chí phi tham số, để tìm hiểu về các cơ chế kinh tế:

"Mục đích của một mô hình là thu được thông tin hữu ích về mối quan hệ giữa phản ứng và các biến dự đoán.

Khả năng giải thích là một cách để có được thông tin. Nhưng một mô hình không nhất thiết phải đơn giản để cung cấp thông tin đáng tin cậy về mối quan hệ giữa biến dự đoán và biến phản hồi; nó cũng không nhất thiết phải là một mô hình dữ liệu [cấu trúc]."

Các mô hình dự đoán là bước đầu tiên để hiểu các cơ chế. Hơn nữa, mô hình hóa cấu trúc có thể hướng lợi trực tiếp từ học máy mà không phải hy sinh các thử nghiệm giả thuyết nhọn hoặc tính đặc thù của các cơ chế kinh tế.² Cho đến nay, học máy chủ yếu phục vụ văn hóa mô hình dự đoán của kinh tế tài chính. Điều quan trọng là phải thừa nhận nó như một công cụ hữu hiệu tương tự để kiểm tra giả thuyết cấu trúc. Ví dụ, hãy xem thảo luận của

chúng tôi về Chen và Ludvigson (2009), trong Phần 5.5.

1.6. Nội dung kinh tế (Hai nền văn hóa kinh tế tài chính)

13

văn hóa (đây là một hướng quan trọng cho nghiên cứu máy học trong lĩnh vực tài chính trong tương lai). Chắc chắn, một chương trình nghiên cứu được thành lập chỉ dựa trên "sự do lưỡng mà không có lý thuyết" (Koopmans, 1947) sẽ được phục vụ tốt hơn bằng cách xem dữ liệu qua lăng kính lý thuyết kinh tế và với sự hiểu biết sâu sắc về bài phê bình của Lucas Jr, 1976. Tương tự như vậy, một chương trình chỉ diễn giải dữ liệu thông qua các mô hình kinh tế hiện có có thể bỏ qua các mẫu thống kê bất ngờ nhưng quan trọng về mặt kinh tế.

Hayek (1945) đối mặt với những tác động kinh tế của việc phân tán thông tin trong việc phân bổ nguồn lực. Về câu hỏi trọng tâm của ông là làm thế nào để đạt được một trật tự kinh tế hiệu quả, ông lưu ý

Nếu chúng ta có tất cả thông tin liên quan, nếu chúng ta có thể bắt đầu từ một hệ thống ưu tiên nhất định, và nếu chúng ta có được kiến thức đầy đủ về các phương tiện sẵn có, thì vấn đề còn lại thuần túy là logic... Tuy nhiên, điều này rõ ràng là không phải là vấn đề kinh tế mà xã hội phải đối mặt. Và phép tính kinh tế mà chúng ta đã phát triển để giải quyết vấn đề logic này, mặc dù là một bước quan trọng hướng tới giải pháp cho vấn đề kinh tế của xã hội, nhưng vẫn chưa đưa ra câu trả lời cho nó. Lý do cho điều này là vì 'dữ liệu' mà từ đó các phép tính kinh tế bắt đầu không bao giờ được 'trao' cho toàn bộ xã hội cho một bộ óc duy nhất có thể tìm ra các hàm ý và không bao giờ có thể được đưa ra như vậy.

Trong khi mối quan tâm chính của Hayek là giá trị của việc lập kế hoạch phi tập trung, các tuyên bố của ông cũng có ý nghĩa a đối với công nghệ thông tin nói chung và công nghệ dự đoán nói riêng. Chúng ta hãy tự phụ đến mức diễn giải lại tuyên bố của Hayek như một vấn đề thống kê: Có một sự chênh lệch giữa hiệu quả phân bổ mà các tác nhân kinh tế có thể đạt được khi biết quá trình tạo ra dữ liệu (DGP) so với khi nó phải được ước tính. Đầu tiên, có vấn đề về đặc tả mô hình - đơn giản là không thể mong đợi các tác nhân kinh tế xác định chính xác các mô hình thống kê của họ. Họ phải sử dụng một số dạng mô hình tham số được xác định sai hoặc mô hình xấp xỉ không tham số. Trong cả hai trường hợp, việc xác định sai sẽ tạo ra sự khác biệt giữa phân bổ tối ưu có thể đạt được khi biết DGP (gọi đây là "tốt nhất thứ nhất") và phân bổ bắt nguồn từ các mô hình được xác định sai của chúng (gọi đây là "tốt nhất thứ hai") .

Nhưng ngay cả điều tốt thứ hai cũng không thể tin được, bởi vì chúng ta phải ước tính những mô hình này với dữ liệu hữu hạn. Điều này dẫn đến một cái nêm khác, do sự thay đổi trong việc lấy mẫu. Ngay cả khi chúng tôi biết dạng hàm của DGP, chúng tôi vẫn phải ước tính nó và nhiều trong ước tính của chúng tôi tạo ra sai lệch so với giá trị tốt nhất đầu tiên. Kết hợp điều đó với chủ nghĩa hiện thực về việc xác định sai thông số kỹ thuật và chúng tôi nhận ra rằng trên thực tế, chúng tôi luôn phải tuân theo sự phân bổ "tốt nhất thứ ba"; tức là, các mô hình được xác định sai và được ước tính một cách ồn ào.

Những dự đoán được cải thiện bắt nguồn từ các phương pháp có thể xử lý các tập dữ liệu và thông tin không lò mang đến cơ hội giảm thiểu sự khác biệt giữa vấn đề "logic" thuận túy về phân bổ nguồn lực tốt nhất đầu tiên được Hayek lưu ý và phân bổ thực tế tốt nhất thứ ba mà các tác nhân kinh tế có thể đạt được. Các nêm không bao giờ co lại về 0 do các giới hạn thống kê về khả năng học hỏi (Da và cộng sự, 2022; Didisheim và cộng sự, 2023). Nhưng các mô hình gần đúng mạnh mẽ và các thiết bị chính quy hóa thông minh có nghĩa là việc học máy chính xác là quan trọng về mặt kinh tế vì chúng có thể đưa đến những quyết định tốt hơn. Vấn đề lựa chọn danh mục đầu tư là một ví dụ minh họa. Một nhà đầu tư có phương sai trung bình biệt ma trận lợi nhuận kỳ vọng thực và hiệp phương sai của tài sản chỉ cần thực hiện "logic" của danh mục Markowitz và đạt được phân bổ tốt nhất đầu tiên. Tuy nhiên, tương tự như Hayek, đây rõ ràng không phải là vấn đề mà các nhà đầu tư trong thế giới thực phải vật lộn. Thay vào đó, vấn đề của họ chủ yếu là do lưỡng - dự đoán. Nhà đầu tư tìm kiếm lợi nhuận kỳ vọng hợp lý và ước tính hiệp phương sai, khi kết hợp với mục tiêu Markowitz, sẽ thực hiện khá tốt ngoài mẫu. Thiếu các phép đo chất lượng cao, giải pháp Markowitz có thể hoạt động rất tai hại, như nhiều nghiên cứu đã chứng minh.

2

Ưu điểm của các mô hình phức tạp

Nhiều người trong chúng tôi được đào tạo về kinh tế lượng theo truyền thống gắn liền với "nguyên tắc tiết kiệm". Điều này được minh họa bằng phương pháp xây dựng mô hình của Box và Jenkins (1970) , ảnh hưởng của nó đến kinh tế lượng tài chính là không thể phủ nhận. Trong phần giới thiệu về ấn bản gần đây nhất của cuốn sách giáo khoa dự báo của Box và Jenkins,¹ "sự phân tích" được trình bày như là phần đầu tiên trong số "Những ý tưởng cơ bản trong xây dựng mô hình": "Điều quan trọng là, trong thực tế, chúng ta sử dụng số lượng nhỏ nhất có thể". của các tham số để thể hiện đầy đủ" [sự nhấn mạnh của họ].

Nguyên tắc phân tích dường như xung đột với các tham số hóa lớn được áp dụng bởi các thuật toán học máy hiện đại. Mô hình ngôn ngữ GPT-3 hàng đầu (Brown và cộng sự, 2020) sử dụng 175 tỷ tham số. Ngay cả các mạng lưới thần kinh dự đoán lợi nhuận tương đối sơ sài trong Gu et al. (2020b) có khoảng 30.000 tham số. Đối với một nhà kinh tế lượng thuộc dòng Box-Jenkins, các tham số hóa phong phú như vậy có vẻ hoang phí, có xu hướng phù hợp quá mức và có thể gây tai hại cho hiệu suất ngoài mẫu .

Các kết quả gần đây trong nhiều lĩnh vực phi tài chính khác nhau mâu thuẫn với quan điểm này. Trong các ứng dụng như thị giác máy tính và xử lý ngôn ngữ tự nhiên,

¹Box và cộng sự. (2015), ấn bản thứ năm của Box và Jenkins gốc (1970).

các mô hình mang số hóa thiên văn-và khớp chính xác với dữ liệu huấn luyện-thường là các mô hình ngoài mẫu hoạt động tốt nhất.

Khi đánh giá trạng thái của tài liệu về mạng thần kinh, Belkin (2021) kết luận “có vẻ như các mạng khả thi về mặt công nghệ lớn nhất luôn được ưu tiên hơn để có hiệu suất tốt nhất”. Rõ ràng, học máy hiện đại đã thay đổi nguyên tắc phân tích.

Việc tìm kiếm cơ sở lý thuyết đang được tiến hành để giải thích sự thành công của việc tham số hóa không lò và trả lời câu hỏi do Breiman (1995) đặt ra một cách ngắn gọn: “Tại sao mạng lưới thần kinh được tham số hóa nhiều không phù hợp với dữ liệu?” Trong phần này chúng tôi cung cấp một cái nhìn thoáng qua về câu trả lời. Chúng tôi dựa trên những tiền bộ gần đây trong tài liệu thống kê đặc trưng cho hành vi của các mô hình “tham số hóa quá mức” (những mô hình có tham số vượt xa các quan sát huấn luyện hiện có).²

Các tài liệu gần đây đã thực hiện bước đầu tiên trong việc tìm hiểu ý nghĩa a lý thuyết thống kê của các mô hình học máy và tập trung vào độ chính xác dự đoán ngoài mẫu của các mô hình được tham số hóa quá mức. Trong phần này, chúng tôi quan tâm đến ý nghĩa kinh tế của việc tham số hóa quá mức và sự phù hợp quá mức trong học máy tài chính. Một số bài báo tài chính đã ghi nhận những lợi ích thực nghiệm đáng kể trong dự đoán lợi nhuận của máy học. Trường hợp sử dụng kinh tế chính của những dự đoán này là xây dựng danh mục đầu tư tối ưu hóa tiện ích. Chúng tôi định hướng sự phát triển của mình theo hướng tìm hiểu sự đánh đổi rủi ro-lợi nhuận ngoài mẫu của “danh mục đầu tư máy học” bắt nguồn từ các mô hình dự đoán lợi nhuận được tham số hóa cao. Sự phát triển của chúng tôi theo sát Kelly et al. (2022a) và Didisheim và cộng sự. (2023).

2.1 Công cụ phân tích mô hình học máy

Kelly và cộng sự. (2022a) đề xuất một thí nghiệm tưởng tượng. Hãy tưởng tượng một nhà phân tích đang tìm kiếm một mô hình dự đoán lợi nhuận thành công. Lợi nhuận tài sản R được tạo ra bởi một mô hình thực sự có dạng

$$R_{t+1} = f(X_t) + \epsilon_{t+1} \quad (2.1)$$

²Nhóm công việc này thuộc các tiêu đề như “tham số hóa quá mức”, “phù hợp quá mức lành tính” và “giảm xuống kép” và bao gồm (trong số những nội dung khác) Spigler và cộng sự (2019), Belkin và cộng sự (2018), Belkin và cộng sự (2019), Belkin và cộng sự (2020), Bartlett và cộng sự (2020), Jacot và cộng sự (2019), và Allen-Zhu và cộng sự (2019).

trong đó nhà phân tích có thể biết tập hợp các biến dự đoán X nhưng hàm dự đoán thực sự f vẫn chưa được biết. Không có kiến thức về f và được thúc đẩy bởi lý thuyết gần đúng phổ quát (ví dụ, Hornik và cộng sự, 1990), nhà phân tích quyết định ước chừng f bằng mạng nơ-ron cơ bản:

$$P \\ f(Xt) \approx \sum_{t=1}^{t+1} S_i, t\beta_i .$$

Mỗi đặc điểm trong phép hồi quy này là một phép biến đổi phi tuyến tính được xác định trước của các yếu tố dự đoán thô,³

$$S_i, t = \sim f(w_i xt) . \quad (2.2)$$

Cuối cùng, nhà phân tích ước tính hồi quy gần đúng

$$P \\ R_{t+1} = \sum_{t=1}^{t+1} S_i, t\beta_i + \sim t+1 . \quad (2.3)$$

Nhà phân tích có T huấn luyện các quan sát để học hỏi và phải chọn mức độ phong phú của mô hình gần đúng để sử dụng—cô ấy phải chọn P . Các mô hình đơn giản với P nhỏ có phương sai thấp, nhưng các mô hình phức tạp với P lớn (thậm chí có thể $P > T$) có thể tốt hơn xấp xỉ sự thật. Nhà phân tích nên lựa chọn mức độ phức tạp của mô hình (P nào)?

Có lẽ đáng ngạc nhiên là Kelly và cộng sự. (2022a) chỉ ra rằng nhà phân tích nên sử dụng mô hình gần đúng lớn nhất mà cô ấy có thể tính toán! Dự báo ngoài mẫu dự kiến và hiệu suất danh mục đầu tư đang ngày càng phức tạp về mô hình.⁴ Để đi đến câu trả lời này, Kelly và cộng sự. (2022a) làm việc với hai công cụ toán học chính để phân tích các mô hình phi tuyến tính phức tạp (tức là học máy). Đây là hồi quy sườn với các đặc điểm phi tuyến tính được tạo ra (giống như S_i, t ở trên) và lý thuyết ma trận ngẫu nhiên để xử lý hành vi của công cụ ước tính khi P lớn so với số lượng quan sát dữ liệu huấn luyện.

³ Giả định của chúng tôi rằng các trọng số w_i và hàm phi tuyến $\sim f$ được biết theo Rahimi và Recht (2007), những người chứng minh rằng các kết quả gần đúng phổ quát có thể áp dụng ngay cả khi các trọng số được tạo ngẫu nhiên.

⁴ Điều này đúng mà không cần xác định ở chế độ có độ phức tạp cao ($P > T$) và đúng ở chế độ có độ phức tạp thấp miễn là độ co rút thích hợp được sử dụng. Kelly và cộng sự. (2022a) đưa ra lựa chọn độ co nhắm tối đa hóa hiệu suất của mô hình ngoài mẫu dự kiến.

2.1.1 Hồi quy sườn với các tính năng được tạo

Giả định mô hình đầu tiên của họ⁵ tập trung vào các mô hình dự đoán tuyến tính nhiều chiều theo (2.3), mà chúng tôi gọi là “mô hình thực nghiệm”. Giải thích (2.3) không phải là thu nhập của tài sản chịu sự tác động của nhiều động lực cơ bản tuyến tính. Thay vào đó, DGP không xác định, nhưng có thể được tính gần đúng bằng sự mở rộng phi tuyến S của một số biến điều khiển cơ bản (có thể nhỏ). Trong ngôn ngữ học máy, S là “các tính năng được tạo”^{6X}. bắt nguồn từ các tính năng thô X, chẳng hạn thông qua sự lan truyền mạng lưới thần kinh phi tuyến.

Một khía cạnh rõ ràng của vấn đề này là mô hình thực nghiệm bị xác định sai. Đặc tả đúng của (2.3) đòi hỏi phải mở rộng chuỗi vô hạn, nhưng trong thực tế chúng ta bị mắc kẹt với một số hữu hạn các số hạng, P.

Các mô hình P nhỏ ổn định vì có ít tham số để ước tính (phương sai thấp), nhưng cung cấp giá trị gần đúng kém (độ lệch lớn). Tiềm đề nền tảng của học máy là các thông số kỹ thuật mô hình linh hoạt (P lớn) có thể được tận dụng để cải thiện khả năng dự đoán. Ước tính của họ có thể nhiều (phương sai cao) nhưng cung cấp giá trị gần đúng chính xác hơn (độ lệch nhỏ).

Trước hết, không rõ ràng rằng lựa chọn nào của P là tốt nhất xét về sự cân bằng độ lệch-phương sai. Với tư cách là nhà kinh tế học, cuối cùng chúng ta tìm kiếm sự đánh đổi giữa sai lệch và phương sai để mang lại kết quả kinh tế tối ưu-chẳng hạn như mức hữu dụng cao hơn cho nhà đầu tư. Việc tìm kiếm các mô hình đạt được sự tối ưu về kinh tế đã hướng dẫn Kelly et al. (2022a) theo đuổi lý thuyết về các mô hình có độ phức tạp cao.

Giả định mô hình hóa thứ hai chọn công cụ ước tính của (2.3) để là bình phương tối thiểu được chuẩn hóa theo sườn núi:

$$\beta(z) = zI + T \sum_{t=1}^T \frac{1}{\sigma_t^2} \epsilon_t^2, \quad (2.4)$$

⁵Chúng tôi đưa ra các giả định ở đây ở mức độ cao. Chúng tôi giới thiệu đến bạn đọc quan tâm Kelly và cộng sự. (2022a) để biết các giả định kỹ thuật chi tiết.

⁶Như được đề xuất bằng cách suy ra (2.3) từ phương trình (2.1), điều này hy sinh ít tính tổng quát vì một số bài báo gần đây đã thiết lập sự tương đương giữa các mô hình tuyến tính nhiều chiều và các mô hình phức tạp hơn như mạng lưới thần kinh sâu (Jacot et al., 2018; Hastie và cộng sự, 2019; Allen-Zhu và cộng sự, 2019). Để đơn giản, Kelly và cộng sự. (2022a) tập trung vào dự báo chuỗi thời gian cho một tài sản rủi ro duy nhất (được mở rộng cho bảng đa tài sản của Didisheim và cộng sự (2023)).

2.1. Công cụ phân tích mô hình học máy

19

trong đó z là tham số co rút của sườn núi. Không phải tất cả các chi tiết của công cụ ước tính này đều là trọng tâm trong các lập luận của chúng tôi—nhưng việc chính quy hóa là rất quan trọng. Nếu không chính quy hóa thì mẫu số của (2.4) là số ít trong chế độ có độ phức tạp cao ($P > T$), mặc dù chúng ta sẽ thấy nó cũng có ý nghĩa quan trọng đối với hành vi của $\beta(z)$ trong các mô hình có độ phức tạp thấp ($P < T$).

Cuối cùng, để mô tả những hậu quả kinh tế của các mô hình có độ phức tạp cao đối với các nhà đầu tư, Kelly et al. (2022a) giả định rằng các nhà đầu tư sử dụng dự đoán để xây dựng chiến lược giao dịch được xác định là

$$R_t^{\pi} + 1 = \pi_t R_t + 1,$$

trong đó π_t là trọng số thời gian làm tăng hoặc giảm vị thế tài sản rủi ro tương ứng với dự báo lợi nhuận của mô hình. Đối với phân tích của họ, π_t được đặt bằng lợi nhuận ngoài mẫu dự kiến từ mô hình dự đoán phức tạp của họ. Và họ cho rằng phúc lợi của nhà đầu tư được đo bằng tỷ lệ Sharpe vô điều kiện, $E[R_t^{\pi} - 1]$

$$SR = \frac{E[R_t^{\pi} - 1]}{\sqrt{E[(R_t^{\pi} - 1)^2]}}. \quad (2.5)$$

Mặc dù có các chiến lược giao dịch hợp lý và tiêu chí đánh giá hiệu suất khác, nhưng đây là những lựa chọn phổ biến giữa các học giả và người hành nghề và mang lại lợi ích về tính minh bạch và khả năng điều chỉnh.

2.1.2 Lý thuyết ma trận ngẫu nhiên

Công thức hồi quy sườn ở trên mô tả các mô hình học máy giống như mạng lưới thần kinh về mặt hồi quy tuyến tính. Người ta hy vọng rằng, với cách biểu diễn này, có thể nói điều gì đó cụ thể về hành vi ngoài mẫu dự kiến của các mô hình phức tạp trong giới hạn $P \rightarrow \infty$ với $P/T = c > 0$. Các tiệm cận cần thiết cho học máy là khác với các đặc tính dành cho các đặc tính kinh tế lượng tiêu chuẩn (sử dụng các phép tính gần đúng tiệm cận cho $T \rightarrow \infty$ với P cố định). Lý thuyết ma trận ngẫu nhiên phù hợp một cách lý tưởng để mô tả hành vi hồi quy đường vân trong cài đặt P lớn. Để đơn giản hóa ký hiệu, chúng tôi loại bỏ T khỏi cuộc thảo luận bằng cách luôn đề cập đến mức độ tham số hóa mô hình liên quan đến lượng dữ liệu huấn luyện. Nghĩa là, chúng ta sẽ chỉ theo dõi tỷ lệ $c = P/T$, mà chúng ta gọi là “độ phức tạp của mô hình”.

Điểm mấu chốt của việc mô tả hành vi $\beta(z)$ khi $P \rightarrow \infty$ là ma trận hiệp phương sai mẫu $P \times P$ của các tín hiệu, $\Psi := T S t S^{-1}$. Ý thuyết ma trận ngẫu nhiên mô tả sự phân bố giới hạn của các giá trị riêng của Ψ . Kiến thức về phân phối này đủ để xác định hiệu suất dự đoán ngoài mẫu dự kiến (R2) của hồi quy sườn núi, cũng như tỷ lệ Sharpe ngoài mẫu dự kiến của chiến lược tính thời gian liên quan. Cụ thể hơn, những đại lượng này được xác định bởi

$$m(z; c) := \lim_{P \rightarrow \infty} \frac{1}{P} \text{tr} (\Psi - zI)^{-1} \quad (2.6)$$

đó là biến đổi Stieltjes giới hạn của phân bố giá trị riêng của Ψ . Từ (2.6), chúng tôi nhận ra mối liên hệ chặt chẽ với hồi quy sườn bởi vì Phép biến đổi Stieltjes bao gồm dạng ma trận sườn (Ψ^{-1} . chức năng zI) của $m(z; c)$ được biết đến từ một phiên bản tổng quát của định luật Marcenko-Pastur. Từ $m(z; c)$, chúng ta có thể tính toán rõ ràng tỷ lệ R2 và Sharpe ngoài mẫu dự kiến cũng như độ nhạy của nó đối với độ phức tạp của mô hình dự đoán (xem Phần 3 và 4 của Kelly và cộng sự (2022a) để biết chi tiết về điều này điêm).

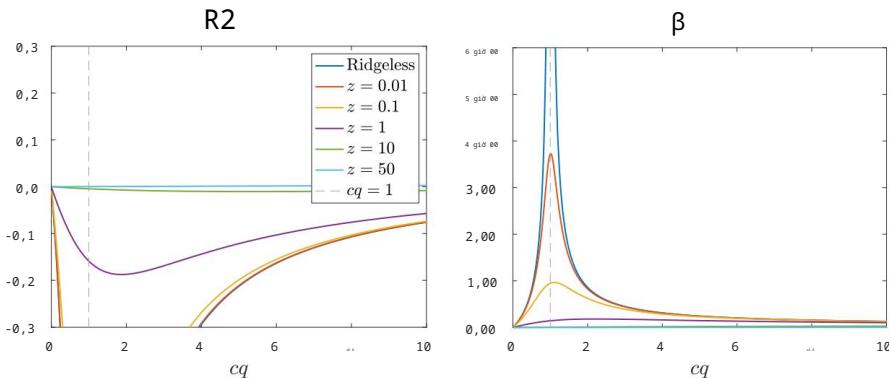
Nói cách khác, độ phức tạp của mô hình đóng một vai trò quan trọng trong việc hiểu hành vi của mô hình. Nếu T tăng với tốc độ nhanh hơn số lượng yếu tố dự đoán (tức là $c = 0$), thì các tiệm cận T lớn và P cố định truyền thống sẽ xuất hiện. Trong trường hợp này, hành vi ngoài mẫu dự kiến của một mô hình trùng với hành vi được ước tính trong -vật mẫu. Đường nhiên, đây là một kịch bản khó xảy ra và khá kém thú vị. Trường hợp thú vị tương ứng với các mô hình học máy được tham số hóa cao với dữ liệu khan hiếm, $P/T = c > 0$, và chính ở đây đã xuất hiện các hành vi mô hình ngoài mẫu đáng ngạc

2.2 Lớn hơn thường tốt hơn

Kelly và cộng sự. (2022a) đưa ra các tuyên bố lý thuyết chặt chẽ về các đặc tính của mô hình học máy có độ phức tạp cao và các chiến lược giao dịch liên quan. Phần trình bày hiện tại của chúng tôi tập trung vào các khía cạnh định tính chính của những kết quả đó dựa trên sự hiệu chỉnh của chúng đối với vấn đề dự đoán lợi nhuận thị trường. Cụ thể, họ giả định tổng biến động lợi nhuận bằng 20% mỗi năm và R2 dự đoán "đúng" không khả thi là 20% mỗi tháng (nếu dạng hàm thực và tất cả các tín hiệu đều có sẵn đầy đủ cho người dự báo). Sự phức tạp cần trở khả năng trau dồi của mô hình

2.2. Lớn hơn thường tốt hơn

21



Hình 2.1: Độ chính xác dự đoán ngoài mẫu từ các mô hình được chỉ định sai

Lưu ý: Giới hạn định mức R và β ngoài mẫu của hồi quy sườn núi là hàm của c và z từ Định luật 5 của Kelly và cộng sự. (2022a). Hiệu chuẩn giả định Ψ là ma trận đồng nhất, $b = 0,2$ và độ phức tạp của mô hình thực là 10.

trên DGP thực vì không có đủ dữ liệu để hỗ trợ quá trình tham số hóa nặng nề của mô hình, do đó, R^2 khả thi nhất mà hiệu chỉnh này ngụ ý là gần 1% mỗi tháng. Chúng tôi tập trung vào cài đặt được chỉ định sai bằng cách xem xét các mô hình thực nghiệm sử dụng các tập hợp con khác nhau của các yếu tố dự đoán thực.

Trong hiệu chuẩn này, DGP thực chưa biết được giả định là có độ phức tạp là $c = 10$. Tham số $q \in [0, 1]$ chi phối sự phức tạp của mô hình thực nghiệm so với sự thật. Chúng tôi phân tích hành vi của các mô hình thực nghiệm gần đúng có độ phức tạp từ rất đơn giản ($q \approx 0$, $cq \approx 0$ và do đó bị xác định sai nghiêm trọng) đến rất phức tạp ($q = 1$, $cq = 10$ tương ứng với mô hình gần đúng phong phú nhất và trên thực tế phục hồi đúng đặc điểm kỹ thuật). Các mô hình có độ phức tạp rất thấp là các mô hình gần đúng kém nhưng các tham số của chúng có thể được ước tính với độ chính xác cao. Khi cq tăng, mô hình thực nghiệm gần đúng với sự thật hơn, nhưng phương sai dự báo tăng (nếu không được điều chỉnh). Việc hiệu chỉnh cũng xem xét một loạt các hình phạt về sườn núi, z .

Trước tiên, hãy xem xét công cụ ước tính bình phương tối thiểu thông thường (OLS), $\beta(0)$, đây là trường hợp đặc biệt của (2.4) với $z = 0$. Khi $c \approx 0$ thì mô hình là

⁷Để đơn giản, các yếu tố dự đoán được giả định là có thể trao đổi được, do đó, chỉ có kích thước của một tập hợp con mới quan trọng chứ không phải danh tính của các yếu tố dự đoán cụ thể trong tập hợp con đó.

rất đơn giản, do đó nó không có khả năng ước chừng độ chính xác và mang lại R2 về cơ bản bằng 0. Khi P tăng và tiến gần đến T từ bên dưới, giá trị gần đúng của mô hình được cải thiện, nhưng mẫu số của ước lượng bình phương nhỏ nhất sẽ tăng lên, tạo ra sai số dự báo bùng nổ.

Hiện tượng này được minh họa ở hình 2.1. Khi P = T, mô hình phù hợp chính xác hoặc "nội suy" dữ liệu huấn luyện (vì lý do này, c = 1 được gọi là "ranh giới nội suy"), do đó cách giải thích phổ biến về hành vi bùng nổ của $\beta(0)$ là nghiêm trọng quá phù hợp mà không khái quát hóa ngoài mẫu.

Khi P vượt ra ngoài T, chúng ta bước vào chế độ được tham số hóa quá mức hoặc độ phức tạp cao. Trong chế độ này, có nhiều tham số hơn số quan sát, bài toán bình phương nhỏ nhất có nhiều nghiệm và ma trận hiệp phương sai nghịch đảo của các biến hồi quy không được xác định. Tuy nhiên, nghịch đảo giả của nó được xác định và điều này tương ứng với một giải pháp duy nhất cụ thể cho bài toán bình phương tối thiểu: $T \sum_{t=1}^T z_t^2 + \frac{1}{T} \sum_{t=StRt+1}^{StS} z_t^2$.

Trong số nhiều giải pháp phù hợp chính xác với dữ liệu huấn luyện, giải pháp này có chỉ tiêu 2 nhỏ nhất. Trên thực tế, nó tương đương với công cụ ước tính đường vân khi tham số độ co ngót tiến tới 0:

$$\beta(0+) = \lim_{z \rightarrow 0^+} \frac{1}{z} \sum_{t=1}^{StS} z_t^2 - \frac{1}{\sum_{t=StRt+1}^{StS} z_t^2}$$

Lời giải $\beta(0+)$ được gọi là hàm ước lượng hồi quy "không có đường vân" (đường màu xanh lam trong Hình 2.1). Khi $c \leq 1$, OLS là công cụ ước lượng không có đường vân, trong khi với $c > 1$ trường hợp không có đường vân được xác định bởi giới hạn $z \rightarrow 0$.

Đáng ngạc nhiên là R2 không có đường vân tăng lên khi độ phức tạp của mô hình tăng lên trên 1. Lý do là khi c trở nên lớn hơn, sẽ có không giàn giải pháp lớn hơn cho hồi quy không có đường vân để tìm kiếm và do đó nó có thể tìm thấy các beta có chuẩn 2 nhỏ hơn vẫn nội suy dữ liệu huấn luyện. Điều này hoạt động như một dạng thu hẹp, làm lệch ước tính beta về 0. Do sai lệch này, phương sai dự báo giảm xuống và R2 được cải thiện. Nói cách khác, mặc dù $z \rightarrow 0$, nghiệm không có đường vân vẫn chuẩn hóa công cụ ước lượng bình phương nhỏ nhất và càng lớn thì c càng lớn. Vào thời điểm c rất lớn, R2 ngoài mẫu dự kiến sẽ chuyển sang vùng dương. Tài sản này

⁸Hãy nhớ lại rằng giả nghịch đảo Moore-Penrose A^+ của ma trận A được xác định thông qua $A^+ = (A^T A)^{-1} A^T$ nếu AA^T khả nghịch và $A^+ = A(AA^T)^{-1}$ nếu AA^T khả nghịch.

2.2. Lớn hơn thường tốt hơn

23

bình phương tối thiểu không có đường vân là một hiện tượng mới được ghi nhận trong tài liệu thống kê và vẫn là một chủ đề nghiên cứu mới nổi.⁹ Kết quả này thách thức học thuyết kinh tế tài chính tiêu chuẩn vốn nhấn mạnh vào tính phân tích mô hình, chứng minh rằng người ta có thể cải thiện tính chính xác của dự báo lợi nhuận bằng cách đẩy kích thước của mô hình vượt xa kích thước mẫu.¹⁰ Hình 2.1 mô tả

hành vi thống kê của các mô hình có độ phức tạp cao. Hình 2.2 hướng sự chú ý đến hậu quả kinh tế của chúng. Bảng bên phải cho thấy sự biến động của chiến lược giao dịch máy học như một hàm số của độ phức tạp của mô hình. Độ biến động của chiến lược thay đổi từng cái một với định mức β và với R^2 (cả ba đại lượng đều là những cách thể hiện khác nhau của phuong sai sai số dự báo). Điểm quan trọng là khi độ phức tạp của mô hình tăng quá $c = 1$, độ biến động của chiến lược giao dịch liên tục giảm. Độ phức tạp làm tăng độ co rút tiềm ẩn của công cụ ước tính không có đường vân, làm giảm độ biến động của lợi nhuận (và $z > 0$ còn làm giảm độ biến động hơn nữa).

Phần bên trái của Hình 2.2 cho thấy hành vi kinh tế quan trọng của các mô hình có độ phức tạp cao - lợi nhuận kỳ vọng ngoài mẫu của chiến lược tính thời gian. Lợi nhuận kỳ vọng thấp đối với các chiến lược đơn giản. Một lần nữa, điều này là do các mô hình đơn giản đưa ra xấp xỉ DGP kém. Độ phức tạp của mô hình ngày càng tăng giúp bạn đến gần hơn với sự thật và tăng lợi nhuận kỳ vọng của chiến lược giao dịch một cách đơn điệu.¹¹

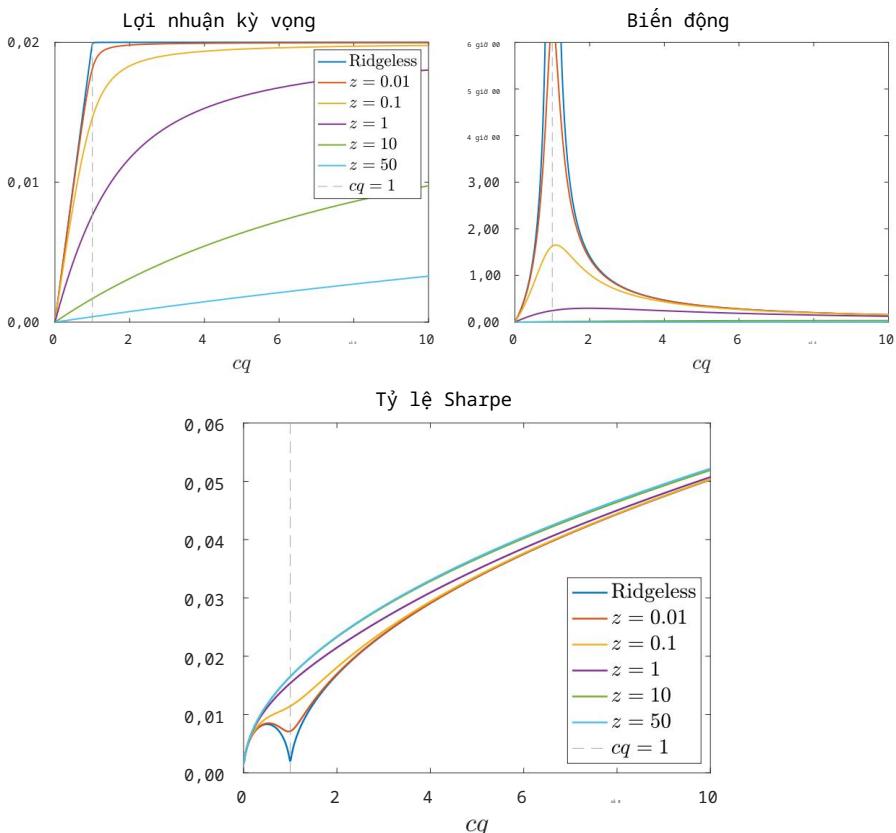
Điều này có ý nghĩa gì đối với sự an toàn của nhà đầu tư? Bảng dưới cùng trong Hình 2.2 cho thấy tiện ích về mặt tỷ lệ Sharpe ngoài mẫu dự kiến.¹² Tỷ lệ Sharpe ngoài mẫu dẫn đến sự cân bằng sai lệch-phương sai cổ điển. Lợi nhuận kỳ vọng hoàn toàn nắm bắt được các hiệu ứng sai lệch. Đối với độ phức tạp thấp

⁹Xem Spigler và cộng sự. (2019), Belkin và cộng sự. (2018), Belkin và cộng sự. (2019), Belkin và cộng sự. (2020), và Hastie và cộng sự. (2019).

¹⁰Các đường cong còn lại trong Hình 2.1 cho thấy R ngoài mẫu do độ co rút không tầm thường² bị ảnh hưởng của đường vân. Các mẫu R cơ bản giống như trường hợp không² có gờ, nhưng việc cho phép $z > 0$ có thể cải thiện R hơn nữa.

¹¹Trong trường hợp không có đường gờ, lợi ích của độ phức tạp tăng thêm đạt mức tối đa khi $c = 1$. Lợi nhuận kỳ vọng không có đường gờ bằng phẳng với $c \geq 1$ vì những cải tiến giá tăng trong phép tính gần đúng DGP được bù đắp chính xác bởi độ lệch tăng dần của độ co rút không có đường gờ.

¹²Trong quá trình hiệu chỉnh, chiến lược giao dịch mua và giữ được chuẩn hóa để có tỷ lệ Sharpe bằng 0. Do đó, tỷ lệ Sharpe trong Hình 2.2 trên thực tế là tỷ lệ Sharpe thu được từ thời điểm dựa trên dự báo mô hình, liên quan đến nhà đầu tư mua và nắm giữ.



Hình 2.2: Chiến lược đánh đổi thời gian giữa rủi ro và lợi nhuận ngoài mẫu dự kiến

Lưu ý: Hạn chế lợi nhuận kỳ vọng, độ biến động và tỷ lệ Sharpe ngoài mẫu của chiến lược định thời gian như là một hàm của cq và z từ Dự luật 5 của Kelly và cộng sự. (2022a). Hiệu chuẩn giả định Ψ là ma trận đồng nhất, $b = 0,2$ và độ phức tạp của mô hình đúng là $c = 10$.

mô hình, sự thiên vị xảy ra thông qua việc xác định sai mô hình nhưng không phải thông qua sự co rút. Đối với các mô hình có độ phức tạp cao, độ lệch do đặc tả sai là nhỏ nhưng độ lệch co lại trở nên lớn. Lý thuyết cho thấy rằng lợi nhuận kỳ vọng tăng lên cùng với sự phức tạp, cho thấy rằng việc xác định sai sai lệch tôn kém hơn so với sai lệch thu hẹp khi nói đến dự kiến trở lại. Trong khi đó, sự biến động của chiến lược hoàn toàn phụ thuộc vào dự báo hiệu ứng biến dạng. Cả hai mô hình đơn giản ($c \approx 0$) hoặc mô hình rất phức tạp

(c 1) tạo ra phương sai thấp. Với những mẫu này trong phương sai sai lệch cân bằng, thì tỷ lệ Sharpe ngoài mẫu cũng tăng lên phức tạp như trong Hình 2.2.

Thật thú vị khi so sánh những phát hiện này với hiện tượng về "giảm dần kép" hoặc thực tế là MSE ngoài mẫu có mẫu không đơn điệu về độ phức tạp của mô hình khi z gần bằng 0 (Belkin và cộng sự, 2018; Hastie và cộng sự, 2019). Hình ảnh phản chiếu của dòng dõi kép trong MSE là "tăng gấp đôi" của tỷ lệ Sharpe không có gờ. Kelly và cộng sự. (2022a) chứng minh rằng tỷ lệ Sharpe không có gờ giảm xuống ở $c = 1$ là giả độ co rút không đủ. Với mức độ co rút phù hợp (rõ ràng đặc trưng bởi Kelly et al. (2022a)), sự phức tạp thậm chí còn trở thành một đức tính tốt ở chế độ có độ phức tạp thấp: bướu biến mất và "đi lên gấp đôi" chuyển thành "đi lên vĩ nh viễn."

Tóm lại, những kết quả này thách thức giáo điều về sự tiết kiệm được thảo luận ở đầu phần này. Họ chứng minh rằng, trong thực tế Trong trường hợp các mô hình thực nghiệm được xác định sai, thì sự phức tạp là một ưu điểm. Đây là đúng không chỉ về mặt hiệu suất thống kê ngoài mẫu (như được hiển thị của Belkin và cộng sự, 2019; Hastie và cộng sự, 2019 và những người khác) mà còn trong điều kiện kinh tế của tiện ích nhà đầu tư ngoài mẫu. Ngược lại với hiểu biết thông thường, hiệu suất của danh mục học máy có thể về mặt lý thuyết được cải thiện bằng cách đẩy tham số hóa mô hình vượt xa số lượng quan sát đào tạo.

Kelly và cộng sự. (2022a) kết luận với khuyến nghị về cách thực hành tốt nhất Những hạn chế trong việc sử dụng các mô hình phức tạp:

"Kết quả của chúng tôi không phải là giấy phép để thêm các yếu tố dự đoán tùy ý đến một mô hình. Thay vào đó, chúng tôi khuyến khích i) bao gồm tất cả những điều hợp lý các yếu tố dự đoán có liên quan và ii) sử dụng các mô hình phi tuyến phong phú thay vì hơn các thông số kỹ thuật tuyến tính đơn giản. Làm như vậy sẽ đưa ra dự đoán và lợi ích danh mục đầu tư, ngay cả khi dữ liệu đào tạo khan hiếm, và đặc biệt khi đi kèm với sự thu hẹp một cách thận trọng."

Để rút ra kết quả trên, Kelly et al. (2022a) áp đặt một giả định rằng khả năng dự đoán được phân bổ đồng đều trên các tín hiệu. Tại Thoạt nhìn, điều này có vẻ quá hạn chế vì nhiều yếu tố dự đoán tiêu chuẩn sẽ không thỏa mãn giả định này. Tuy nhiên, giả định là phù hợp với (và thực sự được thúc đẩy bởi) mạng lưới thần kinh tiêu chuẩn

các mô hình trong đó các tính năng thô được trộn lẫn và truyền phi tuyến thành các tính năng được tạo cuối cùng, như trong phương trình (2.2). Thứ tự của các tính năng S được tạo về cơ bản là ngẫu nhiên theo bước khởi tạo của quá trình huấn luyện mạng. Hơn nữa, trong công việc thực nghiệm của họ, Kelly et al. (2022a), Kelly và cộng sự, 2022b, và Didisheim và cộng sự. (2023) sử dụng công thức mạng nơ-ron được gọi là hồi quy đặc trưng ngẫu nhiên để đảm bảo thỏa mãn giả định này.

2.3 Cái nêm phức tạp

Didisheim và cộng sự. (2023) mở rộng phân tích của Kelly et al. (2022a) theo một số cách và đưa ra "mức độ phức tạp", được định nghĩa là sự khác biệt dự kiến giữa hiệu suất trong mẫu và ngoài mẫu.

Để đơn giản, hãy xem xét một mô hình thực nghiệm được xác định chính xác. Trong môi trường có độ phức tạp thấp với $c \approx 0$, luật số lớn được áp dụng và kết quả là các ước tính trong mẫu hội tụ về mô hình thực.

Do sự hội tụ này, hiệu suất trong mẫu của mô hình biểu thị chính xác hiệu suất ngoài mẫu dự kiến của nó. Nghĩa là, không có sự phức tạp, không có ranh giới giữa hành vi trong mẫu và ngoài mẫu.

Nhưng khi $c > 0$, một cái nêm phức tạp xuất hiện, bao gồm hai thành phần. Độ phức tạp làm tăng khả năng dự đoán trong mẫu của mô hình đã huấn luyện so với khả năng dự đoán của mô hình thực. Đây là định nghĩa truyền thống về mức độ phù hợp quá mức và nó là thành phần nêm đầu tiên. Nhưng độ phức tạp cao cũng có nghĩa là mô hình thực nghiệm không có đủ dữ liệu (liên quan đến việc tham số hóa của nó) để khôi phục mô hình thực sự - đây là sự thất bại của quy luật số lớn do tính phức tạp. Điều này làm phát sinh thành phần nêm thứ hai, là sự thiếu hụt về hiệu suất ngoài mẫu so với mô hình thực. Sự thiếu hụt này có thể được coi là "giới hạn trong việc học" do tính phức tạp của mô hình. Mức độ phức tạp – sự khác biệt được mong đợi giữa hiệu suất trong mẫu và ngoài mẫu – là tổng của mức độ phù hợp quá mức và các giới hạn đối với việc học.

Cái nêm phức tạp có một số ý nghĩa hấp dẫn đối với định giá tài sản. Đưa ra một dự đoán đã được hiện thực hóa , người ta có thể sử dụng (khả thi) lý thuyết ma trận ngẫu nhiên R2 để chứng minh mức độ có thể dự đoán được trong mô hình "đúng" (nhưng không khả thi). Một số nghiên cứu đã ghi nhận

2.3. Cái nêm phức tạp

27

dự đoán lợi nhuận ngoài mẫu tích cực đáng kể bằng cách sử dụng mô hình học máy, ở mức khoảng 1% mỗi tháng đối với từng cổ phiếu riêng lẻ. Thực tế này, kết hợp với suy luận lý thuyết về các giới hạn đối với việc học, cho thấy R2 dự đoán thực sự không khả thi phải cao hơn nhiều. Liên quan đến vấn đề này, ngay cả khi có những cơ hội kinh doanh chênh lệch giá (hoặc đơn giản là tỷ lệ Sharpe rất cao) được ngụ ý bởi mô hình thực tế, những hạn chế trong việc học hỏi khiến các nhà đầu tư trong thế giới thực không thể tiếp cận được những cơ hội này. Trong bối cảnh thực nghiệm thực tế, Didisheim et al. (2023) gợi ý rằng các tỷ lệ Sharpe có thể đạt được bị suy giảm gần một bậc độ lớn so với quá trình tạo dữ liệu thực sự do không thể ước tính chính xác các mối quan hệ thống kê phức tạp.

Đà và cộng sự. (2022) xem xét một môi trường kinh tế riêng biệt trong đó các tác nhân, cụ thể là nhà kinh doanh chênh lệch giá, áp dụng các chiến lược chênh lệch giá thông kê nhằm nỗ lực tối đa hóa tỷ lệ Sharpe ngoài mẫu của họ. Những nhà kinh doanh chênh lệch giá này cũng phải đối mặt với những trở ngại về mặt thống kê (chẳng hạn như "sự phức tạp" ở đây) trong việc tìm hiểu DGP của các bảng chữ cái. Đà và cộng sự. (2022) cho thấy rằng bất kể các nhà kinh doanh chênh lệch giá sử dụng phương pháp học máy nào, họ cũng không thể nhận ra tỷ lệ Sharpe không khả thi tối ưu trong một số môi trường có tỷ lệ tín hiệu trên nhiều thấp nhất định. Hơn nữa, ngay cả khi các nhà kinh doanh chênh lệch giá áp dụng một chiến lược giao dịch khả thi tối ưu thì vẫn có một khoảng cách đáng kể giữa tỷ lệ Sharpe của họ và tỷ lệ không khả thi tối ưu. Chúng tôi thảo luận chi tiết hơn về từng bài viết này trong Chương 4.6.

3

Dự đoán trả về

Định giá tài sản theo kinh nghiệm là đo lường phần bù rủi ro tài sản. Phép đo này có hai dạng: một dạng tìm cách mô tả và hiểu sự khác biệt về phần bù rủi ro giữa các tài sản, dạng kia tập trung vào động lực chuỗi thời gian của phần bù rủi ro. Là khoảnh khắc đầu tiên quay trở lại, đây là điểm khởi đầu tự nhiên để khảo sát tài liệu thực nghiệm về học máy tài chính. Khoảnh khắc đầu tiên ngay lập tức tóm tắt i) mức chiết khấu mà các nhà đầu tư coi là khoản đền bù hợp lý cho việc nắm giữ tài sản rủi ro (có khả năng bị bóp méo bởi những xung đột dẫn đến yếu tố định giá sai) và ii) khía cạnh hàng đầu của các cơ hội đầu tư sẵn có đối với một nhà đầu tư tiềm năng.

Theo định nghĩa, dự đoán lợi nhuận là phép đo lợi nhuận vượt quá dự kiến có điều kiện của một tài sản :1

$$R_{i,t+1} = E_t [R_{i,t+1}] + \epsilon_{i,t+1}, \quad (3.1)$$

Liên quan đến phương trình (1.2), $E_t [R_{i,t+1}] = E[R_{i,t+1}|I_t]$ thể hiện một dự báo có điều kiện trên tập thông tin. Nó thu thập tất cả các biến động không thể đoán trước trong quá khứ và không thể đoán trước. Khi xây dựng các mô hình kê dữ liệu thị trường, cần lùi lại một bước và nhận ra rằng

¹R_{i,t+1} là lợi nhuận của một tài sản vượt quá lãi suất phi rủi ro với các tài sản được lập chỉ mục theo i = 1, ..., N_t và các ngày theo t = 1, ..., T.

dữ liệu được tạo ra từ một quá trình cực kỳ phức tạp. Một số lượng lớn các nhà đầu tư có sở thích và bộ thông tin cá nhân khác nhau sẽ tương tác và trao đổi chứng khoán để tối đa hóa lợi ích của họ. Các nhà giao dịch thường xuyên quyết định về giá và trình tự thay đổi giá (và các dòng tiền không thường xuyên như cổ tức bằng cổ phiếu hoặc phiếu giảm giá trái phiếu) tạo ra một chuỗi lợi nhuận. Tập thông tin ngũ ý bởi kỳ vọng có điều kiện trong (3.1) phản ánh tất cả thông tin - công khai, riêng tư, rõ ràng, tinh vi, rõ ràng hoặc không rõ ràng - theo một cách nào đó ảnh hưởng đến giá cả được quyết định tại thời điểm t. Chúng tôi nhấn mạnh nguồn gốc phức tạp của kỳ vọng có điều kiện này bởi vì tiếp theo chúng tôi thực hiện bước nhảy vọt trong việc xác định một hàm cụ thể để mô tả hành vi của nó:

$$Et [Ri,t+1] = g (zi,t). \quad (3.2)$$

Mục tiêu của chúng tôi là biểu diễn $Et [Ri,t+1]$ dưới dạng hàm g bất biến nhưng nói chung là của các biến dự đoán P-chiều zi,t có sẵn cho các nhà nghiên cứu như chúng tôi. Nghĩa là, chúng tôi hy vọng tách ra được một hàm mà khi chúng tôi đặt điều kiện vào zi,t , sẽ giải thích đầy đủ tính không đồng nhất trong lợi nhuận kỳ vọng 2 trên tất cả các tài sản và trong mọi thời gian, vì g không phụ thuộc vào i hay t.

Bằng cách duy trì cùng một biểu mẫu theo thời gian và giữa các nội dung, ước tính có thể tận dụng thông tin từ toàn bộ bảng điều khiển, mang lại sự ổn định cho ước tính lợi nhuận kỳ vọng cho bất kỳ nội dung riêng lẻ nào. Điều này trái ngược với một số phương pháp định giá tài sản tiêu chuẩn ước tính lại mô hình cắt ngang trong từng khoảng thời gian hoặc ước tính độc lập các mô hình chuỗi thời gian cho từng tài sản (xem Giglio và cộng sự, 2022a). Do quy trình giao dịch phức tạp tạo ra dữ liệu lợi nhuận, giả định "phổ quát" này là đầy tham vọng. Đầu tiên, thật khó để tưởng tượng rằng các nhà nghiên cứu có thể đặt điều kiện dựa trên cùng một tập hợp thông tin như thị trường.

người tham gia (phê bình kinh điển của Hansen-Richard). Thứ hai, do bối cảnh văn hóa và công nghệ xung quanh các thị trường không ngừng phát triển, chưa kể đến những ý tưởng bất chợt của con người có thể ảnh hưởng đến giá cả, khái niệm về một g (·) phổ quát dường như rất xa vời. Vì vậy, trong khi một số nhà kinh tế có thể xem khuôn khổ này là quá linh hoạt (vì nó cho phép sử dụng một hàm dự đoán và hàm tùy ý), có vẻ như (3.2) đặt

2 Ngoài ra, g (·) chỉ phụ thuộc z thông qua zi,t . Trong hầu hết các phân tích của chúng tôi, các dự đoán sẽ không sử dụng thông tin từ lịch sử trước thời điểm t hoặc từ các cổ phiếu riêng lẻ ngoài thời kỳ thứ i, mặc dù điều này được khái quát hóa trong một số phân tích mà chúng tôi tham khảo sau.

những hạn chế khó khăn đối với mô hình lợi nhuận kỳ vọng. Đối với những người coi nó là hạn chế một cách vô lý, thì bất kỳ khả năng nào của khuôn khổ này trong việc mô tả mạnh mẽ các hành vi khác nhau của lợi nhuận tài sản-theo nội dung, thời gian và cụ thể trên cơ sở ngoài mẫu-có thể được coi là một thành công khó có thể xảy ra.

Các phương trình (3.1) và (3.2) đưa ra một khuôn mẫu tổ chức cho tài liệu học máy về dự đoán lợi nhuận. Phần này được tổ chức theo dạng hàm $g(z_i, t)$ được sử dụng trong mỗi bài báo để tính gần đúng hàm dự đoán thực $g(z_i, t)$ trong mẫu của chúng tôi. Chúng tôi nhấn mạnh những phát hiện thực nghiệm chính của mỗi bài báo, đồng thời nêu bật những đóng góp mới về phương pháp luận và phân biệt các kết quả thực nghiệm của từng bài báo. Chúng tôi tránh thảo luận về những khác biệt chi tiết về các biến điều kiện trong các bài viết khác nhau, nhưng người đọc nên lưu ý rằng những khác biệt này cũng là nguyên nhân dẫn đến sự khác biệt trong kết quả thực nghiệm (trên và ngoài những khác biệt về dạng hàm số).

Có hai dòng văn học riêng biệt gắn liền với chuỗi thời gian và chương trình nghiên cứu chéo đã thảo luận ở trên. Tài liệu về các mô hình học máy theo chuỗi thời gian cho tài sản tổng hợp (ví dụ: danh mục đầu tư chỉ số vốn cổ phần hoặc trái phiếu) được phát triển trước đó nhưng là tài liệu nhỏ hơn. Những nỗ lực trong việc dự đoán lợi nhuận chuỗi thời gian đã được thực hiện trong tài liệu tiếp theo của Shiller (1981) về câu đố biến động quá mức vào những năm 1980. Khi các nhà nghiên cứu bắt đầu định lượng mức độ biến động quá mức trên thị trường, đã xuất hiện một cuộc đua kinh tế lượng về tỷ lệ chiết khấu thay đổi theo thời gian thông qua hồi quy dự đoán. Sự phát triển của tài liệu dự đoán chuỗi thời gian xảy ra sớm hơn so với tài liệu nghiên cứu chéo do có sẵn dữ liệu sớm hơn về lợi nhuận tổng hợp của danh mục đầu tư và do tính đơn giản so sánh của việc dự báo một chuỗi thời gian duy nhất. Kích thước nhỏ của tập dữ liệu lợi nhuận thị trường cũng là một lý do hạn chế

phân tích học máy của dữ liệu này. Với rất ít quan sát (vài trăm kết quả hàng tháng), sau một vài nỗ lực nghiên cứu, người ta sẽ cạn kiệt tính hợp lý rằng các suy luận "ngoài mẫu" thực sự là ngoài mẫu.

Tài liệu về phương pháp học máy để dự đoán lợi nhuận trong bảng của nhiều tài sản riêng lẻ mới hơn, lớn hơn nhiều và tiếp tục phát triển. Nó được xếp vào danh mục "mặt cắt ngang lợi nhuận" vì các nghiên cứu ban đầu về các cổ phiếu có tên riêng lẻ đã tìm cách giải thích sự khác biệt về lợi nhuận trung bình vô điều kiện giữa các cổ phiếu - tức là dữ liệu được rút gọn lại.

một mặt cắt ngang duy nhất của lợi nhuận trung bình. Tuy nhiên, trong phiên bản hiện đại của nó, cái gọi là chương trình nghiên cứu "mặt cắt ngang" có dạng một bài toán dự đoán bảng thực sự. Mục tiêu là để giải thích cả sự khác biệt theo thời gian và sự khác biệt chéo trong lợi nhuận kỳ vọng có điều kiện. Lý do chính cho sự mở rộng nhanh chóng và liên tục của tài liệu này là sự phong phú của dữ liệu. Kích thước mặt cắt ngang của bảng điều khiển có thể nhân số lượng quan sát chuỗi thời gian với hệ số vài nghìn hoặc nhiều hơn (ví dụ: trong trường hợp cổ phiếu, trái phiếu hoặc quyền chọn một tên).

Mặc dù có mối tương quan chéo đáng chú ý trong những quan sát này, nhưng hầu hết sự khác biệt về lợi nhuận của tài sản có tên riêng lẻ đều có đặc điểm riêng. Hơn nữa, các mô hình lợi nhuận đường như rất không đồng nhất. Nói cách khác, khía cạnh bảng của vấn đề đưa ra vô số hiện tượng để các nhà nghiên cứu thực nghiệm khám phá, ghi lại và cố gắng hiểu trong khuôn khổ kinh tế.

Chúng tôi đã chọn sơ đồ tổ chức cho phần này để phân loại tài liệu theo các phương pháp học máy được sử dụng và trong mỗi phần, chúng tôi thảo luận về cả chuỗi thời gian và ứng dụng bảng. Với sự phổ biến của các nghiên cứu cắt ngang, chủ đề này nhận được phần lớn sự chú ý của chúng tôi.

3.1 Dữ liệu

Phần lớn tài liệu học máy tài chính nghiên cứu một tập dữ liệu chuẩn bao gồm lợi nhuận hàng tháng của chứng khoán Mỹ và các tín hiệu cấp độ chứng khoán đi kèm được xây dựng chủ yếu từ dữ liệu CRSP-Compustat.

Cho đến gần đây, có rất ít lựa chọn được tiêu chuẩn hóa để xây dựng bảng dữ liệu về tháng chứng khoán này. Các nhà nghiên cứu khác nhau sử dụng các bộ dự đoán mức tồn kho khác nhau (ví dụ: Lewellen (2015) sử dụng 15 tín hiệu, Freyberger và cộng sự (2020) sử dụng 36, Gu và cộng sự (2020b) sử dụng 94) và áp đặt các bộ lọc quan sát khác nhau (ví dụ: loại trừ quan sát các cổ phiếu có giá cổ phiếu danh nghĩa dưới 5 USD, không bao gồm một số ngành nhất định như tài chính hoặc tiện ích, v.v.). Và các bài viết khác nhau thường sử dụng các cấu trúc khác nhau cho các tín hiệu có cùng tên và cơ sở kinh tế (ví dụ: các công thức khác nhau về "giá trị", "chất lượng", v.v.).

May mắn thay, nghiên cứu gần đây đã đạt được tiến bộ trong việc hợp lý hóa các quyết định về dữ liệu này bằng cách phát hành công khai dữ liệu và mã cho một hệ thống được tiêu chuẩn hóa.

bảng điều khiển cấp độ chứng khoán, có sẵn trực tiếp từ máy chủ Dịch vụ Dữ liệu Nghiên cứu Wharton. Jensen và cộng sự. (2021) xây dựng 153 tín hiệu chứng khoán, đồng thời cung cấp mã nguồn và tài liệu để người dùng có thể dễ dàng kiểm tra, phân tích và sửa đổi các lựa chọn theo kinh nghiệm. Hơn nữa, dữ liệu của họ không chỉ có sẵn cho chứng khoán Hoa Kỳ mà còn cho chứng khoán ở 93 quốc gia trên thế giới và được cập nhật thường xuyên để phản ánh các bản phát hành dữ liệu CRSP-Compustat hàng năm. Những tài nguyên này có thể được truy cập tại jkpfactors.com. Jensen và cộng sự. (2021) nhấn mạnh cách tiếp cận đồng nhất để xây dựng tín hiệu bằng cách cố gắng đưa ra quyết định nhất quán về cách sử dụng các mục dữ liệu CRSP-Compustat khác nhau trong các tín hiệu khác nhau. Chen và Zimmermann (2021) cũng đăng mã và dữ liệu về chứng khoán Hoa Kỳ tại openassetpricing.com.

Về mặt dữ liệu được tiêu chuẩn hóa để dự đoán lợi nhuận tổng hợp của thị trường, Welch và Goyal (2008) đăng các biến số dự báo và lợi nhuận hàng tháng và hàng quý được cập nhật cho thị trường chứng khoán Hoa Kỳ tổng hợp.3 Rapach và Chu (2022) cung cấp đánh giá mới nhất về các yếu tố dự báo thị trường bổ sung.

3.2 Thiết kế thí nghiệm

Quá trình ước tính và lựa chọn giữa nhiều mô hình là trọng tâm của định nghĩa học máy được đưa ra ở trên. Đương nhiên, việc chọn một mô hình hoạt động tốt nhất theo độ phù hợp trong mẫu (hoặc mẫu huấn luyện) sẽ phỏng đại hiệu suất của mô hình vì việc tăng tham số mô hình sẽ cải thiện độ phù hợp trong mẫu một cách cơ học. Các mô hình đủ lớn sẽ phù hợp chính xác với dữ liệu huấn luyện. Khi việc lựa chọn mô hình trở thành một phần của quá trình nghiên cứu, chúng ta không còn có thể dựa vào hiệu suất trong mẫu để đánh giá các mô hình nữa.

Các phương pháp phổ biến để lựa chọn mô hình dựa trên tiêu chí thông tin hoặc xác thực chéo. Một tiêu chí thông tin như Akaike (AIC) hoặc Bayes/Schwarz (BIC) cho phép các nhà nghiên cứu lựa chọn trong số các mô hình dựa trên hiệu suất của mẫu huấn luyện bằng cách đưa ra mức phạt hiệu suất theo lý thuyết xác suất liên quan đến số lượng tham số mô hình. Điều này đóng vai trò như một đối trọng đối với sự cải thiện về mặt cơ học về độ vừa vặn do tham số hóa nặng hơn. Tiêu chí thông tin nhằm mục đích chọn một mô hình từ tập ứng cử viên có khả năng có mẫu ngoài mẫu tốt nhất

³Có sẵn tại [Sites.google.com/view/agoyal145](https://sites.google.com/view/agoyal145).

hiệu suất dự đoán theo một số số liệu.

Xác thực chéo có cùng mục tiêu với AIC và BIC, nhưng tiếp cận vấn đề theo cách dựa trên dữ liệu nhiều hơn. Nó so sánh các mô hình dựa trên hiệu suất ngoài mẫu "giả" của chúng. Xác thực chéo sẽ tách các quan sát thành một tập được sử dụng để huấn luyện và một tập khác (các quan sát giả ngoài mẫu) để đánh giá hiệu suất. Bằng cách tách mẫu huấn luyện khỏi mẫu đánh giá, việc xác thực chéo sẽ tránh được hiệu suất vượt trội về mặt cơ học của các mô hình lớn hơn bằng cách mô phỏng hiệu suất của mô hình ngoài mẫu. Xác thực chéo sẽ chọn các mô hình dựa trên hiệu suất dự đoán của chúng trong dữ liệu giả ngoài mẫu.

Mỗi tiêu chí thông tin và xác nhận chéo đều có những ưu điểm và nhược điểm. Trong một số trường hợp, AIC và xác nhận chéo đưa ra các lựa chọn mô hình tương đương tiệm cận (Stone, 1977; Nishii, 1984). Nhược điểm của tiêu chí thông tin là chúng bắt nguồn từ những giả định lý thuyết nhất định, vì vậy nếu dữ liệu hoặc mô hình vi phạm các giả định này thì tiêu chí lý thuyết phải được sửa đổi và điều này có thể khó khăn hoặc không khả thi. Việc triển khai xác thực chéo thường dễ dàng điều chỉnh hơn để giải quyết các thuộc tính dữ liệu đầy thách thức như sự phụ thuộc nối tiếp hoặc các giá trị cực trị và có thể được áp dụng cho hầu hết mọi thuật toán học máy (Arlot và Celisse, 2010). Mặt khác, do tính chất ngẫu nhiên và lặp đi lặp lại của nó, việc xác thực chéo có thể tạo ra các đánh giá mô hình ôn ào hơn và có thể kém về mặt tính toán. Theo thời gian, tài liệu về học máy đã chuyển sang phụ thuộc nhiều vào xác thực chéo vì tính thích ứng và tính phổ biến rõ ràng của nó, đồng thời nhờ giảm chi phí trong điện toán hiệu năng cao.

Để cung cấp một góc nhìn cụ thể về việc lựa chọn mô hình với xác nhận chéo và cách nó phù hợp hơn với các thiết kế thực nghiệm học máy, chúng tôi phác thảo các thiết kế mẫu được Gu et al áp dụng . (2020b) và một số nghiên cứu tiếp theo.

3.2.1 Ví dụ: Thiết kế cố định

Giả sử mẫu nghiên cứu đầy đủ bao gồm các quan sát chuỗi thời gian được lập chỉ mục $t = 1, \dots, T$. Chúng ta bắt đầu với một ví dụ về sơ đồ phân tách mẫu cố định. Mẫu đầy đủ của các quan sát T được chia thành ba mẫu con riêng biệt. Đầu tiên là mẫu "huấn luyện" được sử dụng để ước tính tất cả các Ứng viên

các mô hình. Mẫu huấn luyện bao gồm các quan sát T_{train} từ $t = 1, \dots, T_{train}$. Tập hợp các mô hình ứng cử viên thường được điều chỉnh bởi một tập hợp các "siêu tham số" (còn thường được gọi là "tham số điều chỉnh"). Một ví dụ về siêu tham số xác định tính liên tục của các mô hình ứng cử viên là tham số độ co trong hồi quy sườn, trong khi ví dụ về tham số điều chỉnh mô tả một tập hợp mô hình rời rạc là số lượng thành phần được chọn từ PCA.

Mẫu thứ hai, hay mẫu "xác nhận", bao gồm các quan sát giả ngoài mẫu. Dự báo điểm dữ liệu trong mẫu xác thực được xây dựng dựa trên các mô hình ước tính từ mẫu huấn luyện.

Hiệu suất của mô hình (thường được xác định theo hàm mục tiêu của công cụ ước tính mô hình) trên mẫu xác thực sẽ xác định giá trị siêu tham số cụ thể nào (tức là mô hình cụ thể nào) được chọn.

Mẫu xác nhận bao gồm các quan sát $T_{validate} = T_{train} + T_{train} + T_{validate}$. Điểm Lưu ý rằng mặc dù mô hình ban đầu được ước tính là $1, \dots$, chỉ từ các dữ liệu T_{train} đầu tiên, sau khi đặc tả mô hình được chọn, các tham số của nó thường được ước tính lại bằng cách sử dụng các quan sát $T_{train} + T_{validate}$ đầy đủ để khai thác toàn bộ hiệu quả của mô hình trong -dữ liệu mẫu khi xây dựng dự báo ngoài mẫu.

Tất nhiên, các mẫu xác thực phù hợp không thực sự nằm ngoài mẫu vì chúng được sử dụng để điều chỉnh, do đó, bản thân hiệu suất xác thực có thể bị sai lệch khi lựa chọn. Do đó, mẫu "thử nghiệm" thứ ba (không được sử dụng để ước tính hay điều chỉnh) được sử dụng để đánh giá cuối cùng về hiệu quả dự đoán của phương pháp. Mẫu kiểm định bao gồm các quan sát $T_{test} = T_{train} + T_{validate} + 1, \dots, T_{train} + T_{validate} + T_{test}$.

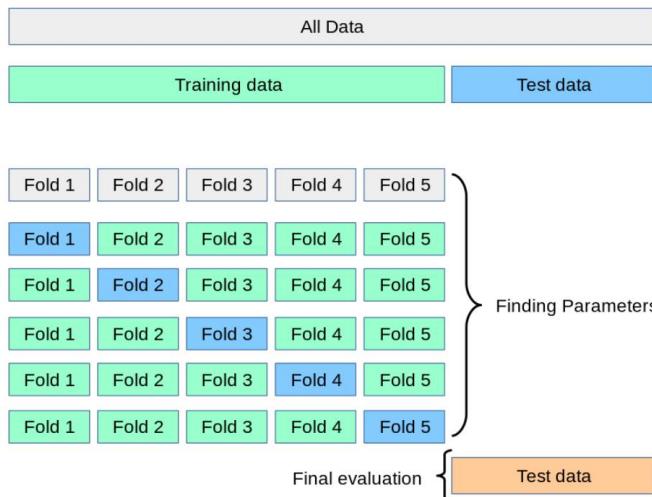
Hai điểm đáng chú ý trong ví dụ thiết kế đơn giản hóa này.

Đầu tiên, mô hình cuối cùng trong ví dụ này được ước tính một lần và mãi mãi bằng cách sử dụng dữ liệu thông qua $T_{train} + T_{validate}$. Tuy nhiên, nếu mô hình được ước tính lại theo cách đệ quy trong suốt thời gian thử nghiệm, nhà nghiên cứu có thể đưa ra các dự báo ngoài mẫu hiệu quả hơn. Lý do để dựa vào cách phân chia mẫu cố định là vì các mô hình ứng viên cần đào tạo rất chuyên sâu về mặt tính toán, do đó, việc đào tạo lại chúng có thể không khả thi hoặc phải gánh chịu chi phí tính toán lớn (ví dụ: xem phân tích CNN của Jiang và cộng sự, 2022).).

Thứ hai, việc phân chia mẫu trong ví dụ này tôn trọng thứ tự chuỗi thời gian của dữ liệu. Động lực của thiết kế này là để tránh rò rỉ thông tin vô ý ngược thời gian. Điều này đi xa hơn, nó

3.2. Thiết kế thử nghiệm

35



Hình 3.1: Minh họa xác thực chéo K-Fold tiêu chuẩn

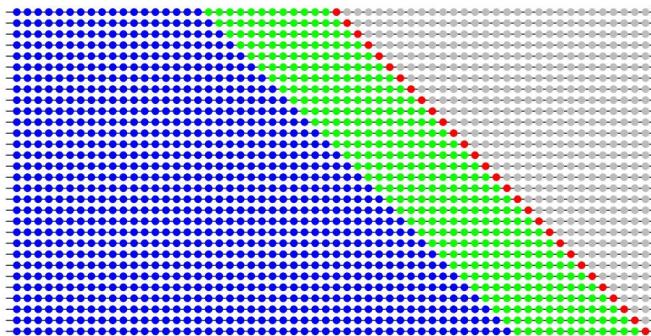
Nguồn: <https://scikit-learn.org/>

Điều phổ biến trong các ứng dụng chuỗi thời gian là đưa mẫu cảm biến giữa mẫu huấn luyện và mẫu xác nhận để mối tương quan nối tiếp giữa các mẫu không làm sai lệch xác thực. Để có mối tương quan chuỗi mạnh mẽ hơn, các lệnh cảm biến dài hơn là phù hợp.

Thứ tự tạm thời của các mẫu huấn luyện và xác nhận là không thực sự cần thiết và có thể khiến việc sử dụng dữ liệu không hiệu quả cho quyết định lựa chọn mô hình. Ví dụ: một biến thể về thứ tự thời gian trong thiết kế này sẽ thay thế mẫu xác thực cố định bằng sơ đồ xác thực chéo K-Fold truyền thống hơn. Trong trường hợp này, các quan sát $T_{train} + T_{validate}$ đầu tiên có thể được sử dụng để tạo ra K mẫu xác thực khác nhau, từ đó có thể đưa ra lựa chọn có khả năng sáng suốt hơn.

Điều này sẽ phù hợp, ví dụ, với dữ liệu không tương quan nối tiếp.

Hình 3.1 minh họa sơ đồ xác thực chéo K-Fold, tạo ra K mẫu xác thực qua đó hiệu suất được tính trung bình để đưa ra quyết định lựa chọn mô hình.



Hình 3.2: Minh họa xác thực chéo theo thứ tự thời gian để quy. Lưu ý: Các chấm màu xanh biểu thị các quan sát huấn luyện, các chấm màu xanh lá cây biểu thị các quan sát xác thực và các chấm màu đỏ biểu thị các quan sát kiểm tra. Mỗi hàng đại diện cho một bước trong thiết kế để quy. Hình minh họa này tương ứng với trường hợp cửa sổ đào tạo mở rộng (thay vì cuộn).

3.2.2 Ví dụ: Thiết kế để quy

Khi xây dựng một dự báo ngoài mẫu cho lợi nhuận đạt được tại thời điểm t , nhà phân tích thường mong muốn sử dụng mẫu cập nhật nhất để ước tính mô hình và đưa ra các dự báo ngoài mẫu. Trong trường hợp này, các mẫu huấn luyện và xác nhận dựa trên các quan sát tại $t-1$. Ví dụ: để chia 50/50 thành các mẫu huấn luyện/xác thực, 1, ..., thiết kế có định ở trên sẽ được điều chỉnh để tạo huấn luyện trên 1, ... và xác thực trên $t - 1$. Sau đó, mô hình được chọn sẽ được tái-estimated + 1, ..., sử dụng tất cả dữ liệu đến $t - 1$ và dự báo ngoài mẫu cho t được tạo ra.

$$\frac{t-1}{2}$$

Tại thời điểm $t+1$, toàn bộ quá trình đào tạo/xác nhận/kiểm tra được lặp lại một lần nữa. Quá trình đào tạo sử dụng các quan sát $t_{\text{khắc nhẫn}} + 1, \dots, t$, và mô hình đã chọn được ước tính lại thông qua t để đưa ra dự báo ngoài mẫu. Quá trình để quy này lặp lại cho đến khi dự báo ngoài mẫu cuối cùng được tạo cho quan sát T . Lưu ý rằng do quá trình xác thực được tiến hành lại mỗi kỳ nên mô hình đã chọn có thể thay đổi trong suốt quá trình để quy. Hình 3.2 minh họa sơ đồ xác thực chéo để quy này.

Một biến thể phổ biến của thiết kế này là sử dụng cửa sổ huấn luyện cuộn thay vì cửa sổ mở rộng. Điều này có lợi nếu có nghi ngờ về sự mất ổn định về cấu trúc trong dữ liệu hoặc nếu có mô hình hóa hoặc

3.3. Điểm chuẩn: Mô hình tuyến tính đơn giản

37

lợi ích thử nghiệm trong việc duy trì kích thước mẫu đào tạo bằng nhau trong suốt quá trình độ quy.

3.3 Điểm chuẩn: Mô hình tuyến tính đơn giản

Mô hình bảng cơ bản về lợi nhuận chứng khoán mà bất kỳ phương pháp học máy nào cũng nên so sánh là mô hình tuyến tính đơn giản. Đối với một tập hợp các tính năng dự đoán mức tồn kho nhất định $z_{i,t}$, mô hình bảng tuyến tính cố định hàm dự đoán là $g(z_{i,t}) = \beta z_{i,t}$:

$$R_{i,t+1} = \beta z_{i,t} + i_{i,t+1}. \quad (3.3)$$

Có nhiều công cụ ước tính khác nhau cho mô hình này phù hợp với các giả định khác nhau về cấu trúc của ma trận hiệp phương sai lỗi.

Trong nghiên cứu tài chính thực nghiệm, phổ biến nhất là hồi quy Fama và Macbeth (1973). Petersen (2008) phân tích các đặc tính kinh tế lượng của Fama-MacBeth và so sánh nó với các công cụ ước tính bảng khác.

Haugen và Baker (1996) và Lewellen (2015) là tiền thân của tài liệu về học máy để phân tích lợi nhuận. Đầu tiên, họ sử dụng số lượng tín hiệu tương đối lớn: Haugen và Baker (1996) sử dụng khoảng 40 biến liên tục và các biến giả ngành, và Lewellen (2015) sử dụng 15 biến liên tục. Thứ hai, họ huấn luyện đệ quy mô hình tuyến tính bảng trong (3.3) và nhấn mạnh hiệu suất ngoài mẫu của các mô hình được huấn luyện của họ. Điều này phân biệt phân tích của họ với một chương trình thực nghiệm phổ biến trong tài liệu sắp xếp cổ phiếu vào các nhóm dựa trên một hoặc hai đặc điểm - về cơ bản là mô hình dự đoán lợi nhuận "không tham số" xác nhận dạng hàm cho $g(z_{i,t})$ mà không thực hiện ước tính.⁴ Cả Haugen và Baker (1996) và Lewellen (2015) đều đánh giá hiệu suất của mô hình ngoài mẫu về mặt kinh tế của hiệu suất chiến lược giao dịch. Haugen và Baker (1996) phân tích bổ sung dữ liệu quốc tế, trong khi Lewellen (2015) phân tích bổ sung độ chính xác của mô hình về mặt dự đoán R2.

⁴Theo hiểu biết tốt nhất của chúng tôi, Basu (1977) là người đầu tiên thực hiện việc sắp xếp danh mục đầu tư dựa trên đặc điểm. Anh ta thực hiện các phép phân loại theo tỷ lệ giá trên thu nhập ở cấp độ cổ phiếu. Việc áp dụng cách tiếp cận này của Fama và French (1992) đã thiết lập các loại danh mục đầu tư như một phương pháp chính thống để phân tích hiệu quả của một công cụ dự đoán lợi nhuận cổ phiếu ứng viên.

Bảng 3.1: Tỷ suất sinh lợi trung bình hàng tháng của Haugen và Baker (1996)

Nhân tố	1979/01 đến 1986/06 Giá trị t trung bình -0,97% 0,52% Khối lượng	1986/07 đến 1993/12 Chi số t trung bình
Tiền lãi vượt quá một tháng	giao dịch/ -17.04	-0,72% -11.04
Tiền lãi vượt quá mười hai tháng	vốn hóa 7,09	0,52% 7.09
thị trường -0,35% Lợi nhuận vượt quá hai tháng -0,20% 0,27% Thu nhập trên giá 0,24% Lợi nhuận trên vốn chủ sở hữu Giá trị số	-5,28 -4,97	-0,20% -0,11% -2,33 -2,37
0,35% -0,10% Xu hướng khối	sách 4,56	0,26% 4,42
trái trong sáu tháng Dòng	lượng giao dịch 0,24% Lợi nhuận vượt	4,34 0,13% 2.06
tiền theo giá 0,13% Sự thay đổi của dòng tiền theo giá -0,11%	3,90	0,39% 6,72
-3,17	-0,09%	-2,58
3,01	0,19%	2,55
2,64	0,26%	4,42
-2,55	-0,15%	-3,38

Nguồn: Haugen và Baker (1996).

Haugen và Baker (1996) cho thấy rằng danh mục đầu tư được tối ưu hóa được xây dựng từ dự báo lợi nhuận của mô hình tuyến tính tốt hơn thị trường tổng hợp, và làm như vậy ở mỗi quốc gia trong số năm quốc gia mà họ nghiên cứu. Một trong những điều tuyệt vời nhất của họ phát hiện thú vị là sự ổn định trong các mẫu dự đoán, được tái tạo trong Bảng 3.1 (dựa trên Bảng 1 trong bài báo của họ). Các hệ số trên những yếu tố dự đoán quan trọng trong nửa đầu mẫu của họ không chỉ cùng dấu nhưng độ lớn và thống kê tương tự nhau một cách đáng kinh ngạc trong nửa sau của mẫu của họ.

Lewellen (2015) cho thấy rằng, trong mặt cắt ngang của tất cả các chứng khoán Mỹ, mô hình tuyến tính bảng có R2 ngoài mẫu rất có ý nghĩa trong khoảng 1% mỗi tháng, thể hiện khả năng điều chỉnh về mặt định lượng mức độ dự báo lợi nhuận với lợi nhuận thực hiện. Điều này tạo nên sự khác biệt mô hình tuyến tính từ phương pháp sắp xếp, thường được đánh giá vào khả năng phân biệt đáng kể lợi nhuận kỳ vọng cao và thấp cổ phiếu-tức là khả năng so sánh tương đối mà không nhất thiết phải có độ lớn phù hợp. Hơn nữa, mô hình tuyến tính 15 biến chuyển thành hiệu suất chiến lược giao dịch án tượng, như thể hiện trong Bảng 3.2 (được lập lại từ Bảng 6A của Lewellen (2015)). Trọng lượng bằng nhau chiến lược dài hạn mua lợi nhuận kỳ vọng dựa trên mô hình cao nhất thập phân vị và bán không mức thấp nhất kiểm được tỷ lệ Sharpe hàng năm là 1,72 trên cơ sở ngoài mẫu (0,82 cho thập phân giá trị trọng số). Nói chung, các

3.3. Điểm chuẩn: Mô hình tuyến tính đơn giản

39

Bảng 3.2: Tỷ suất sinh lợi nhân tố trung bình hàng tháng và tỷ lệ Sharpe hàng năm từ Lewellen (2015)

	Trọng số bằng nhau						Trọng số giá trị					
	Dự đoán	t-stat	trung	bình	Std	Shp	Dự đoán	t-stat	trung	bình	Std	Shp
Bảng A: Tất cả cổ phiếu												
Thấp (L) 2	-0,90	-0,32	7,19	-0,84	-0,15	-0,11	-0,76	0,11	6,01	0,37	0,06	
	0,40	5,84	1,30	0,21	0,60	0,24	-0,10	0,45	4,77	1,89	0,32	
	5,46	2,06	0,44	0,78	5,28	0,38	0,21	0,65	4,65	2,84	0,49	
3	2,74	0,64	0,81	5,36	2,82	0,51	0,44	0,69	4,67	2,97	0,51	
	0,83	1,04	5,36	3,62	1,02	0,52	0,63	0,81	5,01	3,34	0,56	
	1,12	5,55	3,68	1,25	1,31	0,67	0,82	0,88	5,22	3,28	0,58	
4	5,97	4,04	1,66	6,76	4,38	0,70	1,01	1,04	5,67	3,46	0,64	
	1,55	2,29	2,17	7,97	4,82	0,76	1,24	1,15	6,03	3,62	0,66	
5 6 7 8 9		3,20	2,49	5,02		0,85	1,54	1,34	6,68	3,80	0,69	
Cao (H)	10,00					0,94	2,19	1,66	8,28	3,73	0,70	
H - L						1,72	2,94	1,55	6,56	4,51	0,82	

Nguồn: Lewellen (2015).

bằng chứng của Haugen và Baker (1996) và Lewellen (2015) chứng minh rằng các mô hình bảng tuyến tính đơn giản có thể ước tính các kết hợp trong thời gian thực của nhiều yếu tố dự đoán có hiệu quả trong việc dự báo lợi nhuận và xây dựng chiến lược giao dịch.

Chuyển sang phân tích chuỗi thời gian, câu đó về độ biến động quá mức của Shiller (1981) đã thúc đẩy một lượng lớn tài liệu tìm cách lượng hóa mức độ biến đổi theo thời gian của tỷ lệ chiết khấu, cũng như một hướng nghiên cứu lý thuyết hiệu quả.

hợp lý hóa hành vi năng động của tỷ lệ chiết khấu (ví dụ Campbell và Cochrane, 1999; Bansal và Yaron, 2004; Gabaix, 2012; Wachter, 2013).

Công cụ được lựa chọn trong việc theo đuổi sự thay đổi tỷ lệ chiết khấu theo kinh nghiệm là hồi quy chuỗi thời gian tuyến tính. Như đã lưu ý trong Rapach và Chu (2013) khảo sát về khả năng dự đoán của thị trường chứng khoán, công cụ dự đoán phổ biến nhất là tỷ lệ giá cổ tức tổng hợp (ví dụ Campbell và Shiller, 1988), mặc dù hàng chục yếu tố dự đoán khác đã được nghiên cứu. Nhìn chung, tài liệu tập trung vào các mô hình dự đoán đơn biến hoặc đa biến nhỏ, đôi khi kết hợp với những hạn chế về kinh tế như tính không tiêu cực những hạn chế về dự báo lợi nhuận thị trường (Campbell và Thompson, 2008) và áp đặt các hạn chế về phương trình chéo trong giá trị hiện tại bản sắc (Cochrane, 2008; Van Binsbergen và Kojien, 2010). trong một phê bình có ảnh hưởng, Welch và Goyal (2008) cho rằng sự phong phú bằng chứng trong mẫu về khả năng dự đoán lợi nhuận thị trường từ tuyến tính đơn giản

các mô hình không thể khái quát hóa ngoài mẫu. Tuy nhiên, Rapach và cộng sự. (2010) cho thấy rằng các kỹ thuật kết hợp dự báo tạo ra những dự báo thị trường ngoài mẫu đáng tin cậy.

3.4 Mô hình tuyển tính bị phạt

Mượn một cụm từ từ sinh học tiến hóa, các mô hình tuyển tính của Haugen và Baker (1996) và Lewellen (2015) là một “loài chuyển tiếp” trong tài liệu tài chính. Giống như các phương pháp kinh tế lượng truyền thống, các nhà nghiên cứu sửa đổi các thông số kỹ thuật mô hình của họ một cách tiên nghiệm. Nhưng giống như học máy, họ xem xét một tập hợp các yếu tố dự đoán lớn hơn nhiều so với những người tiên nghiệm và nhán mạnh đến hiệu suất dự báo ngoài mẫu.

Trong khi các bài báo này nghiên cứu hàng chục yếu tố dự đoán lợi nhuận, thì danh sách các đặc điểm dự đoán được phân tích trong số lượng tài liệu lên tới hàng trăm (Harvey và cộng sự, 2016; Hou và cộng sự, 2018; Jensen và cộng sự, 2021).Thêm vào đó là mối quan tâm đến việc mở rộng mô hình để kết hợp sự phụ thuộc trạng thái vào các mối quan hệ tiềm đoán (ví dụ: Schaller và Norden, 1997; Cujean và Hasler, 2017), và quy mô tham số hóa của mô hình tuyển tính nhanh chóng tăng lên đến hàng nghìn tham số. Gu và cộng sự. (2020b) xem xét đặc điểm tuyển tính cơ sở để dự đoán bằng lợi nhuận hàng tháng của chứng khoán. Họ sử dụng khoảng 1.000 công cụ dự đoán là các tương tác nhân lén của khoảng 100 đặc điểm cổ phiếu với khả năng dự báo đã được chứng minh đối với lợi nhuận của từng cổ phiếu và 10 công cụ dự đoán tài chính vĩ mô tổng hợp đã được chứng minh thành công trong việc dự đoán lợi nhuận thị trường. Mặc dù thực tế là những yếu tố dự đoán này đã cho thấy nhiều hứa hẹn trong nghiên cứu trước đó, Gu và cộng sự. (2020b) cho thấy OLS không thể đạt được sự phù hợp ổn định của một mô hình với nhiều tham số như vậy cùng một lúc, dẫn đến hiệu suất ngoài mẫu thảm hại. R2 dự đoán là

35% mỗi tháng và chiến lược giao dịch dựa trên những dự đoán này sẽ kém hiệu quả hơn

Hầu như không có gì đáng ngạc nhiên khi các ước tính OLS thất bại với rất nhiều yếu tố tiên đoán. Khi số lượng yếu tố dự đoán P tiến gần đến số lượng quan sát NT, mô hình tuyển tính trở nên kém hiệu quả hoặc thậm chí không nhất quán. Nó bắt đầu xử lý nhiều hơn là trích xuất tín hiệu. Điều này đặc biệt rắc rối đối với vấn đề dự đoán phản hồi trong đó tỷ lệ tín hiệu trên tạp âm rất thấp.

Một kết luận trọng tâm từ cuộc thảo luận về các mô hình “phức tạp” như vậy

trong Phần 2 như sau: Điều quan trọng để tránh tình trạng quá phù hợp là hạn chế mô hình bằng cách chuẩn hóa công cụ ước tính. Điều này có thể được thực hiện bằng cách đầy đủ phức tạp (được định nghĩa là c trong Phần 2) lên trên một (điều này ngầm chuẩn hóa công cụ ước tính bình phương nhỏ nhất) hoặc bằng cách áp đặt hình phạt rõ ràng thông qua sườn núi hoặc sự rút gọn khác. Mô hình tuyến tính đơn giản không làm được điều đó. Độ phức tạp của nó nằm trong vùng phương sai cao đến mức khó chịu (cao hơn 0, nhưng không cao hơn 1) và nó không sử dụng chính quy hóa rõ ràng.

Hàm dự đoán cho mô hình tuyến tính bị phạt giống như mô hình tuyến tính đơn giản trong phương trình (3.3). Nghĩa là, nó tiếp tục chỉ xem xét các yếu tố dự đoán cơ bản, chưa được chuyển đổi. Các phương pháp bị phạt khác nhau bằng cách thêm một hình phạt vào hàm tổn thất ban đầu, chẳng hạn như hình phạt "lưới đàn hồi" phỏ biến, dẫn đến việc mất đi bị phạt

$$L(\beta; \rho, \lambda) = \sum_{t=1}^T \left[R_i, t+1 - \beta z_i, t+1 + \lambda(1 - \rho) \sum_{j=1}^P |\beta_j| + \frac{\lambda \rho}{2} \sum_{j=1}^P \beta_j^2 \right]$$
(3.4)

Mạng đàn hồi bao gồm hai siêu tham số không âm, λ và ρ , và bao gồm hai bộ điều chỉnh nói tiếng là trường hợp đặc biệt. Trường hợp $\rho = 1$ tương ứng với hồi quy sườn, sử dụng hình phạt 2 tham số, làm cho tất cả các ước tính hệ số gần bằng 0 nhưng không áp đặt các số 0 chính xác ở bất kỳ đâu. Ridge là một phương pháp rút gọn giúp ngăn các hệ số trở nên lớn quá mức. Trường hợp $\rho = 0$ tương ứng với Lasso và sử dụng giá trị tuyệt đối hoặc "1", hình phạt tham số. Hình học của Lasso đặt các hệ số trên một tập hợp con của các hiệp phương sai về chính xác bằng 0 nếu hình phạt đủ lớn. Theo nghĩa này, lasso áp đặt sự thua thót đối với thông số kỹ thuật và do đó có thể được coi vừa là một lựa chọn có thể thay đổi vừa là một thiết bị thu nhỏ. Đối với các giá trị trung gian của ρ , lưới đàn hồi khuyến khích các mô hình đơn giản với sự kết hợp khác nhau giữa hiệu ứng đưỡng gờ và dây kéo.

Gu và cộng sự. (2020b) cho thấy rằng sự thất bại của công cụ ước tính OLS trong bảng dự đoán lợi nhuận theo tháng chứng khoán được khắc phục bằng cách đưa ra hình phạt rộng co giãn. Dự đoán ngoài mẫu R2 trở nên dương và chênh lệch thập phân vị dài ngắn có trọng số bằng nhau dựa trên các dự đoán rộng đàn hồi trả về tỷ lệ Sharpe ngoài mẫu là 1,33 mỗi năm. Nói cách khác, đó không phải là điểm yếu của thông tin dự đoán được thể hiện trong 1.000 yếu tố dự đoán mà là chi phí thống kê (quá phù hợp và kém hiệu quả) của

gánh nặng tham số nặng nề, điều đó gây bất lợi cho hiệu suất của OLS. Anh ấy và cộng sự. (2022a) phân tích các tập hợp lưỡi đòn hồi và nhận thấy rằng các mô hình lưỡi đòn hồi riêng lẻ tốt nhất thay đổi khá nhanh theo thời gian và các tập hợp đó hoạt động tốt trong việc nắm bắt những thay đổi này. Rapach và cộng sự. (2013) áp dụng (3.4) để dự báo lợi nhuận thị trường ở các quốc gia.

Phương pháp hồi quy bị phạt là một số công cụ học máy được sử dụng thường xuyên nhất trong tài chính, phần lớn nhờ vào khả năng xử lý khái niệm và tính toán của chúng. Ví dụ, công cụ ước tính hồi quy sườn có sẵn ở dạng đóng nên nó có tính đơn giản về mặt tính toán giống như OLS. Không có biểu diễn dạng đóng chung của công cụ ước tính Lasso, vì vậy nó phải được tính toán bằng số, nhưng các thuật toán hiệu quả để tính toán Lasso có mặt khắp nơi trong các gói phần mềm thống kê (bao gồm Stata, Matlab và các thư viện Python khác nhau).

Freyberger và cộng sự. (2020) kết hợp hồi quy bị phạt với mô hình cộng tính tổng quát (GAM) để dự đoán bằng lợi nhuận hàng tháng của cổ phiếu. Trong ứng dụng của họ, hàm $p_k(z_i, t)$ là một phép biến đổi phi tuyến theo phần tử của các biến P trong z_i, t :

$$g(z_i, t) = \sum_{k=1}^K \beta_k p_k(z_i, t). \quad (3.5)$$

Mô hình của họ mở rộng tập hợp các yếu tố dự đoán bằng cách sử dụng $k = 1, \dots, K$ các phép biến đổi phi tuyến như vậy, với mỗi phép biến đổi k có vectơ $Q \times 1$ riêng của hệ số hồi quy tuyến tính β_k . Freyberger và cộng sự. (2020) sử dụng công thức spline bậc hai cho các hàm cơ sở của chúng, mặc dù khả năng của hàm cơ sở là vô tận và các phép biến đổi phi tuyến có thể được áp dụng chung cho nhiều yếu tố dự đoán (ngược lại với từng phần tử).

Bởi vì các thuật ngữ phi tuyến tính được đưa vào một cách bổ sung, nên việc dự báo bằng GAM có thể được tiếp cận bằng các công cụ ước tính giống như bất kỳ mô hình tuyến tính nào. Nhưng khái niệm mở rộng chuỗi làm nền tảng cho GAM nhanh chóng nhân số lượng tham số mô hình, do đó, việc phạt kiểm soát mức độ tự do có xu hướng mang lại lợi ích cho hiệu suất ngoài mău. Freyberger và cộng sự. (2020) áp dụng hàm phạt được gọi là Lasso nhóm

(Huang và cộng sự, 2010), có dạng

$$\lambda \sum_{j=1}^Q \beta_j^{2k, j1/2}, \quad (3.6)$$

trong đó $\beta_j^{2k, j}$ là hệ số trên hàm cơ sở k áp dụng cho tín hiệu chứng khoán j . Hình phạt này đặc biệt phù hợp với cài đặt mở rộng spline. Như tên gọi của nó, nhóm Lasso chọn tất cả K số hạng spline liên kết với một đặc tính j cho trước, hoặc không chọn số nào trong số đó.

Freyberger và cộng sự. Kết quả nhóm Lasso của (2020) cho thấy chưa đến một nửa số tín hiệu chứng khoán được nghiên cứu phổ biến trong tài liệu có khả năng dự đoán độc lập về lợi nhuận. Họ cũng ghi lại tầm quan trọng của tính phi tuyến, cho thấy rằng đặc tính phi tuyến đầy đủ của chúng chiếm ưu thế so với đặc tính tuyến tính lồng nhau xét về hiệu suất chiến lược giao dịch ngoài mẫu.

Chinco và cộng sự. (2019) cũng sử dụng Lasso để nghiên cứu khả năng dự đoán lợi nhuận. Nghiên cứu của họ có một số khía cạnh đặc đáo. Đầu tiên là việc họ sử dụng dữ liệu tần suất cao - họ dự báo lợi nhuận cổ phiếu trước một phút đã được đào tạo về hồi quy luân phiên trong 30 phút. Thứ hai, họ sử dụng các mô hình hoàn toàn riêng biệt cho từng cổ phiếu, khiến cho phân tích của họ trở thành một tập hợp lớn các hồi quy tuyến tính chuỗi thời gian. Và, thay vì sử dụng các đặc điểm cấp độ chứng khoán tiêu chuẩn làm công cụ dự đoán, bộ tính năng của chúng bao gồm ba độ trễ lợi nhuận một phút cho tất cả các cổ phiếu trong mặt cắt ngang của NYSE. Có lẽ khía cạnh thú vị nhất của mô hình này là khả năng điều chỉnh các hiệu ứng dự đoán giữa các cổ phiếu. Đặc tả hồi quy cuối cùng của họ là

$$R_{i,t} = a_i + \beta_{i,1} R_{t-1} + \beta_{i,2} R_{t-2} + \beta_{i,3} R_{t-3} + \epsilon_{i,t}, \quad i = 1, \dots, N \quad (3.7)$$

trong đó R_t là vectơ bao gồm lợi nhuận của tất cả các cổ phiếu ở phút t . Các tác giả ước tính mô hình này cho từng cổ phiếu i có Lasso bằng cách sử dụng tham số hình phạt dành riêng cho cổ phiếu được chọn với xác thực chéo 10 lần. Họ phát hiện ra rằng các yếu tố dự đoán chính thay đổi khá đáng kể theo từng thời kỳ và có xu hướng là lợi nhuận của những cổ phiếu đang đưa tin tức cơ bản. Những hiểu biết sâu sắc về kinh tế từ các mô hình này chưa được đưa ra đầy đủ, nhưng sức mạnh của bằng chứng dự đoán cho thấy các phương pháp học máy áp dụng cho lợi nhuận tần số cao có khả năng tiết lộ những hiện tượng mới và thú vị liên quan đến luồng thông tin và động lực chung mà chúng gây ra giữa các tài sản. .

Avramov và cộng sự. (2022b) nghiên cứu mức độ năng động của các yếu tố cơ bản ở cấp độ công ty liên quan đến sự biến động tiếp theo của giá cổ phiếu của một công ty. Họ có quan điểm bắt khả tri về các chi tiết của động lực cơ bản. Cách tiếp cận dựa trên dữ liệu của họ xem xét độ lệch của tất cả các mục dữ liệu Compustat hàng quý so với giá trị trung bình của chúng trong ba quý gần đây nhất, thay vì ưu tiên chọn thủ công các nguyên tắc cơ bản cụ thể. Tập hợp sai lệch lớn này được tổng hợp thành một chỉ số dự đoán lợi nhuận duy nhất thông qua học tập có giám sát.

Cụ thể, họ ước tính các phép hồi quy lasso bằng tổng hợp để dự báo lợi nhuận trên cổ phiếu i bằng cách sử dụng tất cả các độ lệch Compustat của cổ phiếu i và gọi giá trị phù hợp từ các hồi quy này là Chỉ số độ lệch cơ bản hoặc FDI. Chiến lược chênh lệch thập phân trọng số giá trị là cổ phiếu FDI dài hạn cao nhất và cổ phiếu FDI ngắn hạn thấp nhất kiểm được tỷ lệ thông tin ngoài mẫu hàng năm là 0,8 so với mô hình bốn yếu tố Fama-Pháp-Carhart.

3.5 Giảm kích thước

Khía cạnh chính quy hóa của học máy nói chung có lợi cho các vấn đề dự đoán nhiều chiều vì nó làm giảm bậc tự do. Có nhiều cách có thể để đạt được điều này. Các mô hình tuyến tính bị phạt làm giảm mức độ tự do bằng cách thu hẹp các hệ số về 0 và/hoặc buộc các hệ số về 0 trên một tập hợp con các yếu tố dự đoán. Nhưng điều này có thể tạo ra những dự báo dưới mức tối ưu khi các yếu tố dự báo có mối tương quan cao.

Hãy tưởng tượng một cài đặt trong đó mỗi yếu tố dự đoán bằng với mục tiêu dự báo cộng với một số thuật ngữ nhiều iid. Trong tình huống này, giải pháp dự báo hợp lý chỉ đơn giản là sử dụng giá trị trung bình của các yếu tố dự đoán trong hồi quy dự đoán đơn biến.

Ý tưởng về tính trung bình dự đoán này là bản chất của việc giảm kích thước. Việc hình thành các tổ hợp tuyến tính của các bộ dự đoán giúp giảm thiểu để cách ly tín hiệu tốt hơn. Trước tiên, chúng tôi thảo luận về hai kỹ thuật giảm kích thước cổ điển, hồi quy thành phần chính (PCR) và bình phương nhỏ nhất một phần (PLS), tiếp theo là hai phần mở rộng của PCA, PCA tỷ lệ và PCA được giám sát, được thiết kế cho cài đặt tín hiệu trên nhiều tháp. Những phương pháp này xuất hiện trong tài liệu như những công cụ giảm kích thước để dự đoán chuỗi thời gian về lợi nhuận thị trường hoặc các biến số kinh tế vĩ mô. Tiếp theo, chúng tôi mở rộng việc sử dụng chúng sang cài đặt bảng để dự đoán mặt cắt ngang của lợi nhuận và sau đó

giới thiệu một phương pháp tập trung vào tài chính gần đây hơn được gọi là phân tích danh mục đầu tư chính phù hợp với vấn đề này. Trong phần này, chúng tôi tập trung vào các ứng dụng giảm kích thước để dự đoán. Việc giảm kích thước đóng một vai trò quan trọng trong việc định giá tài sản ngoài dự đoán và chúng tôi sẽ nghiên cứu những điều này trong các phần tiếp theo. Ví dụ: nhiều phương pháp dựa trên PCA là trọng tâm của phân tích nhân tố tiềm ẩn và được xử lý dưới phạm vi mô hình định giá nhân tố học máy trong Phần 4.

3.5.1 Thành phần chính và bình phương nhỏ nhất từng phần

Chúng tôi chính thức hóa cuộc thảo luận về hai phương pháp này trong cài đặt hồi quy dự đoán chung:

$$yt+h = x t \theta + t+h, \quad (3.8)$$

trong đó y có thể đề cập đến lợi nhuận thị trường hoặc các biến số kinh tế vĩ mô như tăng trưởng GDP, thất nghiệp và lạm phát, xt là vectơ $P \times 1$ của các yếu tố dự đoán và h là phạm vi dự đoán.

Ý tưởng của việc giảm kích thước là thay thế các yếu tố dự đoán có chiều cao bằng một tập hợp các "yếu tố" có chiều thấp, ft , tóm tắt thông tin hữu ích trong xt . Mỗi quan hệ của họ thường được đưa vào mô hình nhân tố tiêu chuẩn:

$$xt = \beta ft + ut. \quad (3.9)$$

Dựa vào (3.9), ta có thể viết lại (3.8) dưới dạng:

$$yt+h = f ta + ^t+h. \quad (3.10)$$

Các phương trình (3.9) và (3.10) có thể được biểu diễn dưới dạng ma trận như sau:

$$X = \beta F + U, \quad Y = F a + E, \quad (3.11)$$

trong đó X là ma trận $P \times T$, F là $K \times T$, F là ma trận $K \times (T - h)$ với h cột cuối cùng được loại bỏ khỏi F , $Y = (yh+1, yh+2, \dots, yT)$, $E = (^h+1, ^h+2, \dots, ^T)$.

Hồi quy thành phần chính (PCR) là một quy trình gồm hai bước. Đầu tiên, nó kết hợp các yếu tố dự đoán thành một số tổ hợp tuyến tính, $ft = \Omega K xt$,

và tìm các trọng số kết hợp QK theo cách đệ quy. Sự kết hợp j giải quyết th giao tuyển tuyển tính

$$w_j = \arg \max_w \text{Var}(x_{tw}),$$

$$\text{st } w_1 = 1, \text{Cov}(x_{tw}, x_{tw1}) = 0, l = 1, 2, \dots, j - 1. \quad (3.12)$$

Rõ ràng, việc lựa chọn các thành phần hoàn toàn không dựa trên mục tiêu dự báo mà nhằm mục đích duy trì tốt nhất cấu trúc hiệp phương sai giữa các yếu tố dự đoán.⁵ Trong bước thứ hai, PCR sử dụng các thành phần ước tính ft trong hồi quy dự đoán tuyển tính tiêu chuẩn (3.10).

Stock và Watson (2002) đề xuất dự báo các biến số kinh tế vĩ mô trên cơ sở PCR. Chúng chứng minh tính nhất quán của việc phục hồi các yếu tố ở giai đoạn đầu (lên đến một phép quay) và dự đoán ở giai đoạn thứ hai, khi cả số lượng yếu tố dự đoán P và cỡ mẫu T đều tăng.

Trong số những ứng dụng sớm nhất của dự báo PC để dự đoán lợi nhuận chứng khoán là Ludvigson và Ng (2007). Mục tiêu dự báo của họ là lợi nhuận hàng quý của chỉ số trọng số giá trị CRSP hoặc mức độ biến động hàng quý của chỉ số đó. Họ xem xét hai bộ thành phần chính, tương ứng với các lựa chọn khác nhau cho các yếu tố dự đoán bao gồm X. Đầu tiên là một tập hợp lớn các chỉ số bao trùm tất cả các khía cạnh của nền kinh tế vĩ mô (sản lượng, việc làm, nhà ở, mức giá, v.v.). Thứ hai là một tập hợp lớn các dữ liệu tài chính, bao gồm chủ yếu là dữ liệu tổng hợp về giá thị trường và cổ tức của Hoa Kỳ, lãi suất tín dụng và trái phiếu chính phủ cũng như lợi nhuận của các danh mục đầu tư được sắp xếp theo đặc điểm và ngành khác nhau.

Tổng cộng, họ kết hợp 381 yếu tố dự đoán riêng lẻ trong phân tích của mình. Họ sử dụng BIC để lựa chọn các thông số kỹ thuật của mô hình bao gồm nhiều sự kết hợp khác nhau của các thành phần ước tính. Các thành phần được trích xuất từ các công cụ dự báo tài chính có khả năng dự báo ngoài mẫu đáng kể về lợi nhuận và biến động của thị trường, trong khi các chỉ số kinh tế vĩ mô thì không. Các dự báo trung bình và biến động phù hợp thể hiện mối liên hệ tích cực cung cấp bằng chứng về sự cân bằng giữa rủi ro và lợi nhuận cổ điển ở cấp độ thị trường tổng hợp. Để giúp giải thích những dự đoán của họ, các tác giả chỉ ra rằng "Hai yếu tố nổi bật là đặc biệt quan trọng đối với lợi nhuận vượt trội hàng quý: yếu tố biến động có mối tương quan cao với lợi nhuận bình thường".

⁵Vấn đề tối ưu hóa này được giải quyết một cách hiệu quả thông qua việc phân rã giá trị số ít của X, ma trận P × T với các cột trong {x1, ..., xT}.

và yếu tố phần bù rủi ro có mối tương quan cao với các yếu tố rủi ro đã được thiết lập rõ ràng để giải thích mặt cắt ngang của lợi nhuận kỳ vọng."

Mở rộng khuôn khổ thị trường vốn cổ phần trước đây của họ, Ludvigson và Ng (2010) sử dụng các thành phần chính của các yếu tố dự đoán tài chính và kinh tế vĩ mô để dự báo lợi nhuận vượt trội của trái phiếu kho bạc. Họ ghi nhận hiệu suất dự đoán lợi nhuận trái phiếu đáng tin cậy vượt xa tỷ giá kỳ hạn (Fama và Bliss, 1987; Cochrane và Piazzesi, 2005).

Họ nhấn mạnh i) sự không nhất quán của kết quả này với các mô hình cấu trúc thuật ngữ affine hàng đầu (trong đó về mặt lý thuyết, lãi suất kỳ hạn bao trùm tất cả các biến thể có thể dự đoán được trong lợi nhuận trái phiếu trong tương lai) và ii) rằng lợi nhuận kỳ vọng có điều kiện ước tính của họ rõ ràng là nghịch chu kỳ, giải quyết tính chu kỳ khó hiểu của rủi ro trái phiếu phần bù phát sinh khi các thành phần kinh tế vĩ mô bị loại khỏi mô hình dự báo.

Jurado và cộng sự. (2015) sử dụng ứng dụng PCR thông minh để ước tính rủi ro kinh tế vĩ mô. Tiền đề của lập luận của họ là một thước đo phương sai có điều kiện tốt phải điều chỉnh một cách hiệu quả cho các điều kiện có điều kiện. có nghĩa là,

$$\text{Var}(y_{t+1}|I_t) = E(y_{t+1} - E[y_{t+1}|I_t])^2 |I_t| .$$

Nếu mức độ dự đoán trung bình bị đánh giá thấp, rủi ro có điều kiện sẽ được đánh giá quá cao. Các tác giả xây dựng dựa trên ý tưởng của Ludvigson và Ng (2007) và Ludvigson và Ng (2010) để bao hòa các dự đoán về lợi nhuận thị trường (và các chuỗi kinh tế vĩ mô khác) với thông tin chứa trong các thành phần chính của các biến dự báo kinh tế vĩ mô. Những cải thiện trong ước tính rủi ro kinh tế vĩ mô có những hậu quả kinh tế thú vị. Đầu tiên, các giai đoạn có mức độ không chắc chắn tăng cao ít hơn và cách nhau xa hơn so với những gì người ta tin trước đây.

Thứ hai, các ước tính được cải thiện cho thấy mối liên hệ chặt chẽ hơn giữa rủi ro gia tăng và hoạt động kinh tế vĩ mô suy thoái.

PCR xây dựng các yếu tố dự đoán chỉ dựa trên sự cộng hưởng giữa các yếu tố dự đoán. Ý tưởng là để điều chỉnh càng nhiều sự khác biệt tổng thể giữa các yếu tố dự đoán càng tốt bằng cách sử dụng một số lượng kích thước tương đối nhỏ. Điều này xảy ra trước và độc lập với bước dự báo.

Điều này tự nhiên cho thấy một thiếu sót tiềm tàng của PCR - đó là nó không xem xét được mục tiêu dự báo cuối cùng trong cách tiến hành giảm kích thước.

Bình phương tối thiểu một phần (PLS) là một giải pháp thay thế cho PCR giúp giảm tính chiêu bằng cách khai thác trực tiếp hiệp phương sai giữa các yếu tố dự đoán và mục tiêu dự báo. Không giống như PCR, PLS tìm kiếm các thành phần của X để tối đa hóa mối tương quan mang tính dự đoán với mục tiêu dự báo, do đó có trọng số th trong thành phần PLS là

$$w_j = \arg \max_w \text{Cov}(y_t + h, x_{tw})^2,$$

$$\text{st } w_w = 1, \text{Cov}(x_{tw}, x_{tw}) = 0, \quad l = 1, 2, \dots, j - 1. \quad (3.13)$$

Cốt lõi của PLS là tập hợp các mô hình đơn biến ("một phần") dự báo $y_t + h$ một yếu tố dự đoán tại một thời điểm. Sau đó, nó xây dựng một tổ hợp tuyến tính của các yếu tố dự đoán được đánh giá bằng khả năng dự đoán đơn biến của chúng. Để hình thành nhiều thành phần PLS, mục tiêu và tất cả các yếu tố dự đoán được trực giao với các thành phần được xây dựng trước đó và quy trình được lặp lại trên tập dữ liệu trực giao.

Kelly và Pruitt (2015) phân tích các đặc tính kinh tế lượng của các mô hình dự đoán PLS (và một khái quát hóa được gọi là bộ lọc hồi quy ba lần) và lưu ý khả năng phục hồi của nó khi bộ dự đoán chứa các yếu tố chi phối không liên quan đến dự đoán, đó là một tình huống hạn chế - đó là hiệu quả của PCR. Liên quan đến vấn đề này, Kelly và Pruitt (2013) phân tích khả năng dự đoán lợi nhuận tổng hợp của thị trường bằng cách sử dụng giá trị hiện tại. Họ lưu ý rằng các phép hồi quy giá trị hiện tại truyền thống (ví dụ: Campbell và Shiller, 1988) phải đối mặt với vấn đề sai số trong các biến và đề xuất giải pháp hồi quy nhiều chiều nhằm khai thác điểm mạnh của PLS. Trong phân tích của họ, Z bao gồm các tỷ lệ định giá (giá trị số sách trên thị trường hoặc cổ tức trên giá) cho một mặt cắt ngang lớn của tài sản. Áp đặt các hạn chế kinh tế, họ rút ra một hệ thống giá trị hiện tại liên quan đến lợi nhuận thị trường với mặt cắt ngang của các tỷ lệ định giá cấp độ tài sản. Làm sao-

Chưa bao giờ, các yếu tố dự đoán cũng được thúc đẩy bởi các yếu tố chung về tăng trưởng cổ tức tổng hợp dự kiến. Việc áp dụng PCR vào mặt cắt ngang của các tỷ lệ định giá gấp phải vấn đề là các yếu tố thúc đẩy tăng trưởng cổ tức không đặc biệt hữu ích cho việc dự báo lợi nhuận thị trường. Công cụ ước tính PLS học cách bỏ qua các thành phần có phương sai cao nhưng có khả năng dự đoán thấp này để ưu tiên các thành phần có khả năng dự đoán lợi nhuận cao hơn. Dự báo dựa trên PLS của họ đạt được R2 ngoài mẫu là 13% cho lợi nhuận hàng năm. Điều này mang lại lợi ích kinh tế lớn cho các nhà đầu tư

3.5. Giảm kích thước

49

sẵn sàng điều chỉnh thời điểm thị trường, tăng tỷ lệ Sharpe lên hơn một phần ba so với nhà đầu tư mua và nắm giữ. Hơn nữa, chúng ghi lại sự biến động lớn hơn đáng kể trong tỷ lệ chiết khấu của nhà đầu tư so với các mô hình lý thuyết hàng đầu. Chatelais và cộng sự. (2023) sử dụng một khuôn khổ tương tự để dự báo hoạt động kinh tế vĩ mô bằng cách sử dụng mặt cắt ngang của giá tài sản, về cơ bản là thực hiện phiên bản phân tích dựa trên PLS của phân tích Fama (1990) chứng minh rằng giá tài sản dẫn đến kinh tế vĩ mô kết quả.

Baker và Wurgler (2006) và Baker và Wurgler (2007) sử dụng PCR để dự báo lợi nhuận thị trường dựa trên tập hợp các chỉ số tâm lý thị trường. Hoàng và cộng sự. (2014) mở rộng phân tích này với PLS và cho thấy các chỉ số tâm lý PLS có lợi ích dự đoán đáng kể so với PCR. Họ lập luận rằng PLS tránh được một yếu tố phổ biến nhưng không liên quan đến nhiều do lưỡng trong các biến đại diện cảm tính Baker-Wurgler. Bằng cách giảm tác động gây nhiễu của tiếng ôn này, Huang et al. (2014) nhận thấy rằng tâm lý là động lực rất quan trọng của lợi nhuận kỳ vọng và khả năng dự đoán này phần lớn không được phát hiện bằng PCR. Tương tự, Chen và cộng sự. (2022b) kết hợp nhiều đại diện thu hút sự chú ý của nhà đầu tư vào một công cụ dự báo lợi nhuận thị trường dựa trên PLS thành công.

Ahn và Bae (2022) tiến hành phân tích tiềm cận của công cụ ước tính PLS và nhận thấy rằng số lượng yếu tố PLS tối ưu để dự báo có thể nhỏ hơn nhiều so với số lượng các yếu tố chung trong các biến dự đoán ban đầu. Hơn nữa, việc bao gồm quá nhiều yếu tố PLS sẽ gây bất lợi cho hiệu suất ngoài mẫu của bộ dự đoán PLS.

3.5.2 PCA được chia tỷ lệ và PCA được giám sát

Một giả định thuận tiện và quan trọng trong phân tích PCR và PLS là các yếu tố này có tính phổ biến. Các mô hình nhân tố lan tỏa rất phổ biến trong các tài liệu. Ví dụ, Bai (2003) nghiên cứu tính chất tiềm cận của công cụ ước lượng PCA của các hệ số trong trường hợp $\lambda K(\beta \beta) P$.

Giglio và cộng sự. (2022b) cho thấy hiệu suất của bộ dự đoán PCA và PLS phụ thuộc vào tỷ lệ tín hiệu trên nhiễu, được xác định bởi $P/(T \lambda K(\beta \beta))$, trong đó $\lambda K(\beta \beta)$ là giá trị riêng lớn thứ K. Khi $P/(T \lambda K(\beta \beta)) = 0$, nhìn chung không có phương pháp dự đoán nào là nhất quán. Đây được gọi là

vấn đề yếu tố yếu.

Huang et al. (2021) đề xuất quy trình PCA theo tỷ lệ, gán trọng số cho các biến dựa trên mối tương quan của chúng với mục tiêu dự đoán, trước khi áp dụng PCA. Sơ đồ trọng số tăng cường tỷ lệ tín hiệu trên tạp âm và do đó giúp phục hồi hệ số.

Tương tự như vậy, Giglio et al. (2022b) đề xuất một giải pháp thay thế PCA (SPCA) có giám sát cho phép các yếu tố dọc theo phạm vi rộng của cường độ yếu tố. Họ lưu ý rằng sức mạnh của các yếu tố phụ thuộc vào tập hợp các biến mà PCA được áp dụng. SPCA bao gồm bước sàng lọc cận biên để chọn một tập hợp con các yếu tố dự đoán trong đó có ít nhất một yếu tố mạnh. Sau đó, nó trích xuất yếu tố đầu tiên từ tập hợp con bằng PCA, chiếu mục tiêu và tất cả các yếu tố dự đoán (bao gồm cả những yếu tố không được chọn) vào yếu tố đầu tiên và xây dựng phần dư. Sau đó, nó lặp lại bước lựa chọn, PCA và bước chiếu với phần dư, trích xuất từng yếu tố một cho đến khi mối tương quan giữa phần dư và mục tiêu biến mất. Giglio và cộng sự. (2022b) chứng minh rằng cả PLS và SPCA đều có thể phục hồi các yếu tố yếu có tương quan với mục tiêu dự đoán và kết quả dự đoán đạt được tính nhất quán. Trong cài đặt mục tiêu đa biến, nếu tất cả các yếu tố tương quan với ít nhất một trong các mục tiêu dự đoán thì quy trình PLS có thể khôi phục số lượng yếu tố yếu và toàn bộ không gian yếu tố một cách nhất quán.

3.5.3 Từ dự đoán chuỗi thời gian đến dự đoán bảng

Các ứng dụng giảm kích thước được tham chiếu ở trên chủ yếu tập trung vào việc kết hợp nhiều yếu tố dự đoán để dự báo chuỗi thời gian một biến. Để dự đoán mặt cắt ngang của lợi nhuận, chúng ta cần khái quát hóa việc giảm kích thước thành cài đặt dự đoán bảng như trong (3.3). Tương tự như (3.9) và (3.10), ta có thể viết

$$R_{t+1} = Fta + Et+1, \quad Z_t = Fty + Ut,$$

trong đó R_{t+1} là vectơ lợi nhuận $N \times 1$ được quan sát vào ngày $t + 1$, Ft là ma trận $N \times K$, a là $K \times 1$, $Et+1$ là vectơ $N \times 1$ của phần dư, Z_t là 6Bai và Ng (2021) mở rộng phân

tích của họ sang các yếu tố vừa phải, tức là $P/(T \lambda K(\beta)) > 0$ và nhận thấy PCA vẫn nhất quán. Công trình trước đây về các mô hình nhân tố yếu bao gồm Onatski (2009), Onatski (2010) và Onatski (2012), những người xem xét việc thiết lập nhân tố cực kỳ yếu trong đó các nhân tố không thể phục hồi một cách nhất quán.

3.5. Giảm kích thước

51

Ma trận đặc tính $N \times P$, y là $K \times P$, và U_t là ma trận $N \times P$ của phần dư.

Sau đó chúng ta có thể xếp chồng $\{R_{t+1}\}$, $\{E_{t+1}\}$, $\{Z_t\}$, $\{F_t\}$, $\{U_t\}$ thành $NT \times 1$, Các ma trận $NT \times 1$, $NT \times P$, $NT \times K$, $NT \times P$, R , E , Z , F và U tương ứng sao cho

$$\bar{R} = F_a + E, \quad \bar{Z} = F_y + U.$$

Các phương trình này có dạng giống hệt như (3.11), do đó áp dụng được quy trình giảm kích thước nói trên.

Ánh sáng và cộng sự. (2017) áp dụng PLS bằng tổng hợp trong dự đoán lợi nhuận cổ phiếu và Gu et al. (2020b) thực hiện PCA và PLS bằng tổng hợp để dự đoán lợi nhuận của từng cổ phiếu. Gu và cộng sự. (2020b) các danh mục đầu tư dài hạn có trọng số bằng nhau dựa trên PCA và PLS có tỷ lệ Sharp lần lượt là 1,89 và 1,47 mỗi năm, vượt trội so với danh mục đầu tư dài hạn dựa trên mạng co giãn.

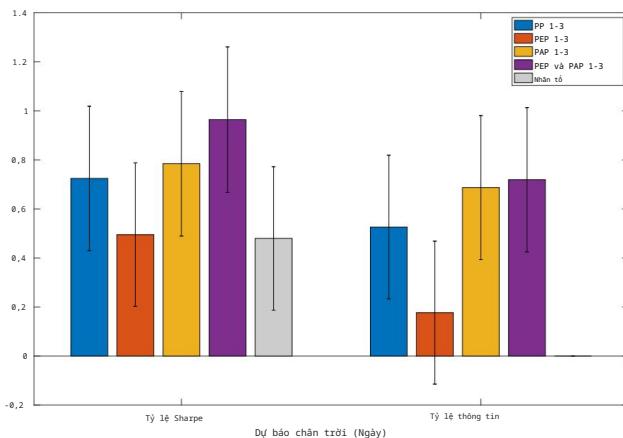
3.5.4 Danh mục đầu tư chính

Kelly và cộng sự. (2020a) đề xuất một phương pháp giảm kích thước khác để dự đoán lợi nhuận và tối ưu hóa danh mục đầu tư được gọi là "phân tích danh mục đầu tư chính" (PPA). Trong khuôn khổ của phần trước, tính đa chiều cao đến từ việc mỗi tài sản riêng lẻ có nhiều yếu tố dự báo tiềm năng. Hầu hết các mô hình định giá tài sản theo kinh nghiệm đều tập trung vào các tín hiệu dự đoán tài sản riêng ; tức là mối liên hệ giữa S_i, t và lợi nhuận của chỉ một tài sản i , $R_{i,t+1}$. PPA được thúc đẩy bởi mong muốn khai thác thông tin dự đoán chung cho nhiều tài sản cùng một lúc. Nó tận dụng các tín hiệu dự đoán của tất cả các tài sản để dự báo lợi nhuận trên tất cả các tài sản khác. Trong trường hợp này, tính chất nhiều chiều của vấn đề xuất phát từ số lượng lớn các mối quan hệ dự đoán chéo tiềm năng.

Để đơn giản, giả sử chúng ta có một tín hiệu duy nhất, S_i, t , cho mỗi nội dung (được xếp chồng thành một vectơ N, St). PPA bắt đầu từ ma trận hiệp phương sai của lợi nhuận trong tương lai của tất cả tài sản với tín hiệu của tất cả tài sản:

$$\Pi = E(R_{t+1}S_i^T) \quad R^{N \times N}.$$

Kelly và cộng sự. (2020a) gọi Π là "ma trận dự đoán". Phần đường chéo của ma trận dự đoán theo dõi các hiệu ứng dự đoán tín hiệu riêng,



Hình 3.3: Tỷ lệ hiệu quả hoạt động của danh mục đầu tư chính.

Nguồn: Hình 3 của KMP.

đó là trọng tâm của các mô hình dự đoán lợi nhuận truyền thống. Đường chéo theo dõi hiện tượng có thể dự đoán chéo.

PPA áp dụng SVD cho ma trận dự đoán. Kelly và cộng sự. (2020a) chứng minh rằng các vectơ số ít của Π – những vectơ chiếm phần lớn sự cộng biến giữa tín hiệu và lợi nhuận trong tương lai – là một tập hợp các danh mục đầu tư được chuẩn hóa, được sắp xếp từ những danh mục có thể dự đoán được nhiều nhất bới S đến những danh mục ít có thể dự đoán được nhất.

Các vectơ số ít dẫn đầu của Π là “danh mục đầu tư chính”. Kelly và cộng sự. (2020a) cho thấy danh mục đầu tư chính được hiểu trực tiếp là danh mục đầu tư tối ưu. Cụ thể, chúng là những danh mục đầu tư “có thời gian” nhất dựa trên tín hiệu S và chúng mang lại lợi nhuận trung bình cao nhất cho nhà đầu tư gặp phải hạn chế về đòn bẩy (tức là nhà đầu tư không thể nắm giữ các vị thế lớn tùy ý).

Kelly và cộng sự. (2020a) cũng chỉ ra rằng thông tin hữu ích về mô hình định giá tài sản và lỗi định giá được mã hóa trong Π . Để giải thích khả năng dự đoán lợi nhuận liên quan đến beta so với alpha, chúng ta có thể phân tách Π thành phần đối xứng (Π_s) và phần phản đối xứng (Π_a):

$$\Pi = \frac{1}{2} \frac{\Pi + \Pi_s}{\Pi_s} + \frac{1}{2} \frac{\Pi - \Pi_s}{\Pi_s} . \quad (3.14)$$

Kelly và cộng sự. (2020a) chứng minh rằng các vectơ đơn hàng đầu của Π_α ("danh mục đầu tư alpha chính") có cách hiểu là chiến lược alpha thuần túy trong khi các vectơ Π có cách hiểu là danh mục đầu tư beta thuần túy ("danh mục đầu tư rủi ro chính"). Vì vậy, Kelly và cộng sự. (2020a) đề xuất một thử nghiệm mới về mô hình định giá tài sản dựa trên lợi nhuận trung bình của danh mục đầu tư alpha chính (như một phương pháp thay thế cho các thử nghiệm như Gibbons và cộng sự, 1989). Trong khi Kelly và cộng sự. (2020a) tập trung vào trường hợp một tín hiệu duy nhất cho mỗi tài sản, He et al. (2022b) chỉ ra cách mở rộng điều này thành nhiều tín hiệu một cách dễ dàng. Goulet Coulombe và Göbel, 2023 tiếp cận một vấn đề tương tự như Kelly et al. (2020a)—xây dựng danh mục đầu tư có khả năng dự đoán tối đa—sử dụng sự kết hợp giữa rừng ngẫu nhiên và hồi quy suôn núi bị ràng buộc.

Kelly và cộng sự. (2020a) áp dụng PPA cho một số tập dữ liệu trên các loại tài sản. Hình 3.3 trình bày một ví dụ từ phân tích của họ. Tập hợp tài sản $Rt+1$ là danh mục đầu tư theo quy mô Fama-Pháp và giá trị số sách trên thị trường, và các tín hiệu S_t là động lượng chuỗi thời gian cho từng tài sản. Hình này cho thấy tỷ lệ Sharpe ngoài mẫu lớn của các danh mục đầu tư chính thu được và cho thấy rằng chúng tạo ra hệ số alpha đáng kể so với mô hình chuẩn bao gồm năm yếu tố Fama-Pháp cộng với hệ số động lượng chuỗi thời gian tiêu chuẩn.

3.6 Cây quyết định

Các mô hình định giá tài sản hiện đại (với sự tồn tại của thói quen, rủi ro dài hạn hoặc rủi ro thảm họa thay đổi theo thời gian) có mức độ phụ thuộc cao của nhà nước vào hành vi thị trường tài chính, cho thấy rằng các hiệu ứng tương tác có thể quan trọng để đưa vào các mô hình thực nghiệm. Ví dụ, Hồng và cộng sự. (2000) xây dựng một lý thuyết thông tin trong đó hiệu ứng động lượng được điều chỉnh bởi quy mô của công ty và phạm vi phân tích của nó, nhấn mạnh rằng lợi nhuận cổ phiếu kỳ vọng thay đổi theo sự tương tác giữa các đặc điểm của công ty. Về mặt khái niệm, thật đơn giản để kết hợp những hiệu ứng như vậy vào các mô hình tuyến tính bằng cách đưa ra các tương tác biến đổi, giống như phương trình (3.5) đưa ra các phép biến đổi phi tuyến. Vấn đề là do thiếu các giả định tiên nghiệm về các tương tác có liên quan nên cách tiếp cận bổ sung tổng quát này nhanh chóng gặp phải các giới hạn tính toán vì các tương tác đa chiều làm tăng số lượng tham số.

combinatorially.⁷

Cây hồi quy cung cấp một cách để kết hợp các tương tác dự báo đa chiều với chi phí tính toán thấp hơn nhiều. Cây phân chia các quan sát dữ liệu thành các nhóm có chung các tương tác đặc trưng chung. Logic là, bằng cách tìm các nhóm quan sát đồng nhất, người ta có thể sử dụng dữ liệu trong quá khứ của một nhóm nhất định để dự báo hành vi của một quan sát mới xuất hiện trong nhóm. Hình 3.4 cho thấy một ví dụ với hai yếu tố dự đoán, "size" và "b/m". Bảng điều khiển bên trái mô tả cách cây gán từng quan sát cho một phân vùng dựa trên các giá trị dự đoán của nó.

Đầu tiên, các quan sát được sắp xếp theo kích thước. Những phần trên điểm dừng 0,5 được xếp vào Loại 3. Những phần có kích thước nhỏ sau đó được sắp xếp thêm theo b/m. Các quan sát có kích thước nhỏ và b/m dưới 0,3 sẽ được xếp vào Loại 1, trong khi những quan sát có b/m trên 0,3 sẽ được xếp vào Loại 2. Cuối cùng, dự báo cho các quan sát trong mỗi phân vùng được xác định là giá trị trung bình đơn giản của biên kết quả giữa các quan sát trong phân vùng đó.

Hàm dự đoán tổng quát gắn với cây K "lá"
(nút đầu cuối) và độ sâu L là

$$g(z_i, t; \theta, K, L) = \sum_{k=1}^K \theta_k \mathbb{1}_{\{z_i, t \in C_k(L)\}}, \quad (3.15)$$

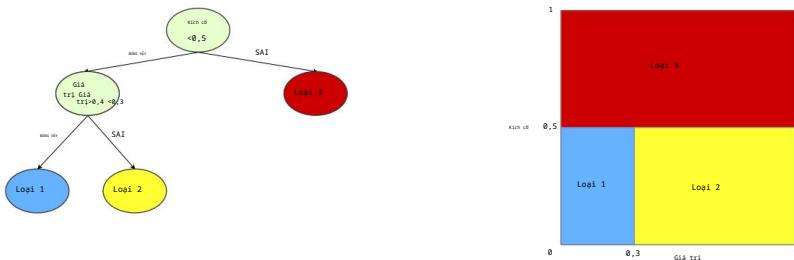
trong đó $C_k(L)$ là một trong các phân vùng K. Mỗi phân vùng là một sản phẩm của tối đa L chức năng chỉ báo. Tham số không đổi θ_k liên kết với phân vùng k được ước tính là giá trị trung bình mẫu của các kết quả trong phân vùng.

Sự phỏ biến của cây quyết định ít bắt nguồn từ cấu trúc của chúng mà nhiều hơn ở các thuật toán "tham lam" có thể có lập một cách hiệu quả các phân vùng có tính dự đoán cao với chi phí tính toán thấp. Trong khi biến dự đoán cụ thể mà một nhánh được phân chia (và giá trị cụ thể nơi xảy ra sự phân chia) được chọn để giảm thiểu sai số dự báo, khoảng trống của

⁷Như Gu và cộng sự. (2020b) lưu ý, "Việc xử phạt tham số không giải quyết được khó khăn trong việc ước tính mô hình tuyến tính khi số lượng yếu tố dự đoán lớn hơn theo cấp số nhân so với số lượng quan sát. Thay vào đó, người ta phải chuyển sang các thuật toán tối ưu hóa heuristic như hồi quy từng bước (thêm/bớt các biến liên tục cho đến khi thỏa mãn một số quy tắc dừng), sàng lọc biến (giữ lại các yếu tố dự đoán có mối tương quan đơn biến với mục tiêu dự đoán vượt quá một giá trị nhất định) hoặc các thuật toán khác."

3.6. Cây quyết định

55



Hình 3.4: Ví dụ về cây hồi quy

Nguồn: Gu và cộng sự. (2020b).

Lưu ý: Hình này trình bày sơ đồ của cây hồi quy (trái) và biểu diễn tương đương của nó (phải) trong không gian của hai đặc điểm (kích thước và giá trị). Các nút cuối của cây lần lượt có màu xanh lam, vàng và đỏ. Dựa trên giá trị của hai đặc điểm này, mẫu cổ phiếu riêng lẻ được chia thành ba loại.

sự phân chia có thể mở rộng đến mức cây không thể được tối ưu hóa toàn cục. Thay vào đó, sự phân chia được xác định một cách cận thị và cây ước tính là một xấp xỉ thô của mô hình cây tốt nhất không khả thi.

Cây có thể điều chỉnh các tương tác một cách linh hoạt (cây có độ sâu L có thể nắm bắt được các tương tác theo chiều ($L - 1$)) nhưng dễ bị quá khớp. Để chống lại điều này, cây thường được sử dụng trong các "tập hợp" được quy chuẩn hóa. Một phương pháp tổng hợp phổ biến là "tăng cường" (cây hồi quy tăng cường độ dốc, "GBRT"), kết hợp để quy các dự báo từ nhiều cây là các yếu tố dự đoán nông và yếu riêng lẻ nhưng kết hợp thành một yếu tố dự đoán mạnh duy nhất (xem Schapire, 1990; Friedman, 2001). Quy trình tăng cường bắt đầu bằng cách lắp một cây nông (ví dụ: có độ sâu $L = 1$). Sau đó, cây đơn giản thứ hai phù hợp với phần dư dự đoán từ cây đầu tiên. Dự báo từ cây thứ hai được thu nhỏ theo hệ số v ($0, 1$) sau đó được thêm vào dự báo từ cây thứ nhất. Sự co rút giúp ngăn chặn mô hình khớp quá mức các phần dư của cây trước đó. Điều này được lặp lại thành một tập hợp phụ gia gồm các cây nông B . Sự tăng cường

do đó tập hợp có ba tham số điều chỉnh (L, v, B).

Rossi và Timmermann (2015) điều tra ICAPM của Merton (1973) để đánh giá liệu phần bù rủi ro vốn cổ phần có điều kiện có thay đổi theo mức độ tiếp xúc của thị trường với các biến số trạng thái kinh tế hay không. Điểm mấu chốt của cuộc điều tra này là phép đo chính xác về điều kiện ICAPM

hiệp phương sai. Để giải quyết vấn đề này, các tác giả sử dụng cây tăng cường để dự đoán hiệp phương sai thực tế giữa chỉ số hoạt động kinh tế tổng hợp hàng ngày và lợi nhuận hàng ngày của danh mục thị trường. Các yếu tố dự đoán bao gồm một tập hợp các chuỗi dữ liệu tài chính vĩ mô từ Welch và Goyal (2008) cũng như các hiệp phương sai đã thực hiện trong quá khứ. Ngược lại với các tài liệu trước đây, các ước tính hiệp phương sai có điều kiện này có mối liên hệ tích cực đáng kể giữa chuỗi thời gian với phần bù rủi ro vốn có phần có điều kiện và hàm ý mức độ hợp lý về mặt kinh tế của tâm lý lo ngại rủi ro của nhà đầu tư. Các tác giả cho thấy rằng các mô hình tuyến tính cho hiệp phương sai có điều kiện bị xác định sai nghiêm trọng và điều này có thể gây ra kết quả hỗn hợp trong các thử nghiệm trước đây của ICAPM. Họ cho rằng kết luận tích cực của họ về ICAPM là nhờ phương pháp cây được tăng cường, phương pháp này có tính linh hoạt phi tuyến tính làm giảm sự xác định sai của hàm hiệp phương sai có điều kiện.

Trong nghiên cứu liên quan, Rossi (2018) sử dụng cây hồi quy tăng cường với các yếu tố dự đoán tài chính vĩ mô để dự báo trực tiếp lợi nhuận tổng hợp (và biến động) theo tần suất hàng tháng, nhưng không áp đặt hạn chế của ICAPM rằng khả năng dự đoán có được thông qua phương sai có điều kiện và hiệp phương sai với kinh tế. các biến trạng thái. Ông cho thấy rằng dự báo cây được tăng cường tạo ra dự đoán lợi nhuận hàng tháng R2 là 0,3% mỗi tháng ngoài mẫu (so với

0,7% khi sử dụng dự báo lợi nhuận trung bình lịch sử), với độ chính xác định hướng là 57,3% mỗi tháng. Điều này dẫn đến một hệ số alpha ngoài mẫu đáng kể so với chiến lược mua và nắm giữ trên thị trường và mức lợi ích hữu dụng tương ứng là 30% đối với một nhà đầu tư có phương sai trung bình có ác cảm rủi ro bằng hai.

Công cụ điều chỉnh cây phổi biến thứ hai là mô hình “rừng ngẫu nhiên”, giống như việc tăng cường, tạo ra một tập hợp các dự báo từ nhiều cây nồng. Theo quy trình “đóng gói” (tổng hợp bootstrap) tổng quát hơn của Breiman (2001), rừng ngẫu nhiên rút ra B mẫu bootstrap của dữ liệu, khớp một cây hồi quy riêng cho từng mẫu, sau đó tính trung bình các dự báo của chúng. Ngoài việc ngẫu nhiên hóa các mẫu ước tính thông qua bootstrap, Random Forest còn ngẫu nhiên hóa tập hợp các yếu tố dự đoán có sẵn để xây dựng cây (một cách tiếp cận được gọi là “bỏ học”). Cả hai thành phần đóng bao và bỏ học của rừng ngẫu nhiên đều điều chỉnh dự báo của nó. Độ sâu L của từng cây, số lượng mẫu bootstrap B và tỷ lệ bỏ học là các tham số điều chỉnh.

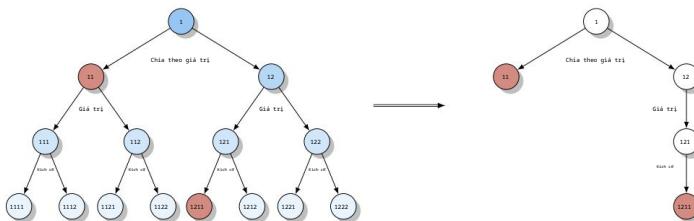
Dự đoán lợi nhuận dựa trên cây là đẳng cấu với điều kiện (chuỗi-

Tial) sắp xếp tương tự như những gì được sử dụng trong tài liệu định giá tài sản. Hãy xem xét một cách sắp xếp thứ ba tuần tự trên các tín hiệu giá trị và quy mô cổ phiếu để tạo ra chính danh mục đầu tư có quy mô và giá trị được sắp xếp kép. Đây là cây quyết định bậc ba hai cấp với các điểm phân chia lớp thứ nhất bằng với các điểm phân chia kích thước và các điểm phân chia lớp thứ hai tại các điểm phân chia giá trị. Mỗi lá j của cây có thể được hiểu là một danh mục đầu tư có lợi nhuận theo thời gian t là trung bình (có trọng số) của lợi nhuận cho tất cả các cổ phiếu được phân bổ cho lá j trong thời gian t. Tương tự, dự báo cho mỗi cổ phiếu ở lá j là lợi nhuận trung bình giữa các cổ phiếu được phân bổ cho j trong tập huấn luyện trước t.

Được thúc đẩy bởi sự đắt cầu này, Moritz và Zimmermann (2016) tiến hành phân loại danh mục đầu tư có điều kiện bằng cách ước tính cây hồi quy từ một tập hợp lớn các đặc điểm chứng khoán, thay vì sử dụng các biến sắp xếp và điểm dừng được xác định trước. Trong khi các loại truyền thống có thể đáp ứng tối đa các tương tác hai hoặc ba chiều giữa các tín hiệu chứng khoán, thì cây có thể tìm kiếm rộng hơn các tương tác đa chiều có tính dự đoán nhất. Thay vì tiến hành phân loại dựa trên cây đơn lẻ, họ sử dụng rồng ngẫu nhiên để đưa ra dự báo lợi nhuận tổng thể từ 200 cây. Các tác giả báo cáo rằng chiến lược trải rộng thập phân vị dài ngắn ngắn có trọng số bằng nhau dựa trên dự báo rồng ngẫu nhiên trước một tháng kiểm được lợi nhuận hàng tháng là 2,3% (mặc dù các tác giả cũng cho thấy rằng hiệu suất này bị ảnh hưởng nặng nề bởi chi phí giao dịch vốn kém. tín hiệu đào chiều một tháng).

Dựa trên Moritz và Zimmermann (2016), Bryzgalova et al. (2020) sử dụng một phương pháp mà họ gọi là "cây AP" để tiến hành phân loại danh mục đầu tư. Trong khi Moritz và Zimmermann (2016) nhấn mạnh sức mạnh dự đoán lợi nhuận của các loại dựa trên cây, Bryzgalova et al. (2020) nhấn mạnh tính hữu ích của chính danh mục đầu tư được sắp xếp như tài sản thử nghiệm để đánh giá các mô hình định giá tài sản. Cây AP khác với các mô hình dựa trên cây truyền thống ở chỗ nó

không tìm hiểu cấu trúc cây mà thay vào đó tạo ra nó bằng cách sử dụng các phân chia trung vị với thứ tự tín hiệu được chọn trước. Điều này được minh họa ở phía bên trái của Hình 3.5 cho một cây ba cấp được chỉ định (không ước tính) để phân chia đầu tiên theo giá trị trung bình, sau đó lại theo giá trị trung bình (nghĩa là tạo ra sự phân chia từ phân vị) và cuối cùng theo quy mô công ty trung bình. Cây AP giới thiệu việc "cắt tia" theo tiêu chí tỷ lệ Sharpe toàn cầu. Đặc biệt, từ cấu trúc cây được xác định trước, mỗi nút trong cây tạo thành một danh mục đầu tư. Thủ tục cây AP tìm giá trị trung bình



Hình 3: Minh họa việc cắt tia với nhiều đặc điểm. Hình này thể hiện các cây mẫu, nguyên bản và được cắt tia cho danh mục có độ sâu 3, và được xứng đáng dựa trên quy mô và giá trị trên thi trường là những đặc điểm duy nhất. Bộ danh mục đầu tư được cắt tia dựa dû dựa trên tam cây, trong đó hình bên phải minh họa kết quả tiềm năng cho một cây.

sự phân chia và độ sâu của chúng. Đáng chú ý, một tập hợp danh mục đầu tư tối ưu cũng có thể bị đây là sự kết hợp thông thường của các đặc điểm (Bryzgalova et al., 2020).

AP-Pruning được mong muốn về mặt thống kê. Vấn đề lựa chọn phân chia cơ bản phản ánh sự giao dịch thiên vị-phương sai-o . Danh mục cây ở các nút cao hơn đa dạng hơn, sự kết hợp hiển thị rõ ràng và có thể gây ra nhiều rủi ro phân tán hơn cho chất lượng số kết hợp. Cuối cùng, quan trọng nhất là việc thêm vào phân chia tiếp theo Phân chia AP được chọn trước chúng tạo nên chất lượng bắt đầu không đổi sẽ nhận được số bằng 0 trong các nút tối ưu hóa được xác thực chéo, trừ khi các nút con thêm thông tin về giá tài sản lớn hơn mức tăng trong khía cạnh danh mục con lığı lại sử dụng làm tài sản thử nghiệm trong phương pháp phụ trợ.12 phân tích định giá tài sản. Các tác giả lập lại phân tích này cho nhiều bộ ba. Điều quan trọng là việc lựa chọn một tập hợp danh mục cây thỏa thuận bao hàm sự phân loại các tín hiệu, xây dựng nhiều loại cây AP được cắt tia để sử dụng làm thử phiếu riêng biệt. Do đó, việc chia nhỏ AP có thể giúp cải thiện chất lượng kết quả.

tài sản. Bryzgalova và cộng sự. (2020) nêu bật khả năng tồn tại của việc tối ưu hóa của họ AP-Pruning ước tính ánh xạ giữa các cổ phiếu riêng lẻ và đặc điểm của chúng bằng cách kết hợp danh mục đầu tư như mô hình số chiết khấu ngẫu nhiên ứng cử viên, liên kết nó với tài liệu trong Phần 5.5.

kết hợp các tài sản tương ứng danh mục đầu tư lớn hơn. Trong khía cây cho phép chúng ta mô hình hóa vấn đề như một vấn đề lựa chọn danh mục đầu tư, sự thỏa thuận của chúng tối ưu hoạt động trực tiếp trên việc phân bổ các cổ phiếu riêng lẻ

Công và cộng sự (2022) lý thuyết dưới đây giới hạn của các phương pháp cây định

giá tài sản bằng cách giới thiệu mô hình Cây P-Tree, phân chia lợi

nhuận cổ phiếu dựa trên đặc điểm của Công ty để phát triển một mô hình

Mô hình định giá tài sản có diện tích Tính toán AP-Pruning. Mô hình sẽ chọn một

thành công độ tối ưu giao tiếp với bảng thông tin nhỏ, giúp giảm sai lệch.

Ứng cử viên quy tắc phân tách (chênh lệch kích thước < 0,2) từ nhóm đặc điểm (được chuẩn hóa thành phạm vi 1 đến 1) và một tập hợp các ngưỡng phân tách được xác định trước (ví dụ: -0,6, -0,2, 0,2, 0,6) để tạo thành các lá con, mỗi đại diện cho một danh mục đầu tư lợi nhuận. Họ đề xuất một tiêu chí phân chia hấp dẫn dựa trên mục tiêu định giá tài sản. Thuật toán P-tree kết thúc khi kích thước lá tối thiểu ¹⁴ được xác định trước hoặc số lượng lá tối đa là

Bản sao điện tử có tại: <https://ssrn.com/abstract=3493458>

⁸Giglio và cộng sự. (2021b) chỉ ra rằng việc lựa chọn tài sản thử nghiệm là một biện pháp khắc phục các yếu tố yếu kém. Vì vậy, việc tạo cây AP mà không cần cắt tia cũng là một công thức hợp lý để xây dựng nhiều danh mục đầu tư đa dạng hóa tốt làm tài sản thử nghiệm, có thể lựa chọn xây dựng danh mục phòng ngừa rủi ro cho các yếu tố yếu.

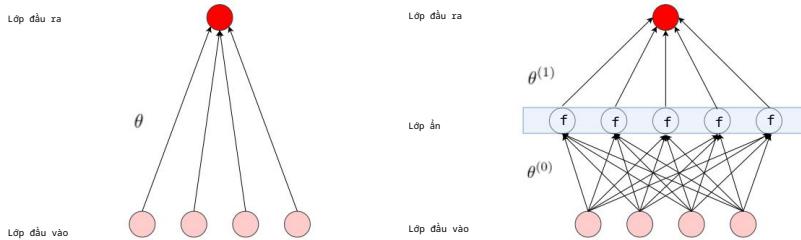
gặp. Quy trình cây P mang lại mô hình định giá tài sản có điều kiện một yếu tố, duy trì khả năng diễn giải và linh hoạt thông qua cấu trúc dựa trên cây duy nhất.

Phương pháp cây được sử dụng trong một số nhiệm vụ dự đoán tài chính ngoài dự đoán lợi nhuận. Chúng tôi thảo luận ngắn gọn về ba ví dụ: dự đoán rủi ro tín dụng, dự đoán thanh khoản và dự đoán biến động. Correia và cộng sự. (2018) sử dụng cây phân loại cơ bản để dự báo các sự kiện tín dụng (phá sản và vỡ nợ). Họ thấy rằng độ chính xác của dự đoán được hưởng lợi từ một tập hợp rộng rãi các đặc điểm rủi ro ở cấp độ công ty và họ chứng minh độ chính xác phân loại ngoài mẫu vượt trội từ mô hình của họ so với các mô hình rủi ro tín dụng truyền thống (ví dụ Altman, 1968; Ohlson, 1980). Easley và cộng sự. (2020) sử dụng mô hình rừng ngẫu nhiên để nghiên cứu động lực tiềm số cao của tính thanh khoản và rủi ro trong môi trường dữ liệu lớn thực sự-dữ liệu đánh dấu cho 87 thị trường tương lai. Họ cho thấy rằng nhiều thước đo thanh khoản thị trường (chẳng hạn như lambda của Kyle, thước đo Amihud và xác suất giao dịch có hiểu biết) có sức mạnh dự đoán đáng kể cho những động thái định hướng tiếp theo trong kết quả rủi ro và thanh khoản (bao gồm chênh lệch giá chào mua, biến động lợi nhuận và trả lại độ lệch). Mittnik và cộng sự. (2015) sử dụng cây tăng cường để dự báo biến động thị trường chứng khoán hàng tháng. Họ sử dụng chuỗi thời gian tài chính vĩ mô 84 làm công cụ dự đoán và sử dụng tính năng tăng cường để xây dựng đệ quy một tập hợp các cây dự đoán biến động nhằm tối ưu hóa khả năng ghi nhận ký Gaussian cho lợi nhuận chứng khoán hàng tháng. Các tác giả nhận thấy mức giảm lớn và có ý nghĩa thống kê về lỗi dự đoán từ các dự báo biến động dựa trên cây so với các mô hình chuẩn GARCH và EGARCH.

3.7 Mạng lưới thần kinh Vanilla

Mạng lưới thần kinh có lẽ là mô hình phổ biến nhất và thành công nhất trong học máy. Chúng có nền tảng lý thuyết là "các công cụ xấp xỉ phổ quát" cho bất kỳ chức năng dự đoán trơn tru nào (Hornik và cộng sự, 1989; Cybenko, 1989). Tuy nhiên, họ phải chịu đựng sự thiếu minh bạch và khả năng giải thích.

Gu và cộng sự. (2020b) phân tích dự đoán từ các mạng "chuyển tiếp nguồn cấp dữ liệu". Chúng tôi thảo luận chi tiết về cấu trúc này vì nó đặt nền tảng cho các kiến trúc phức tạp hơn như mạng hồi quy và mạng tích chập.



Hình 3.6: Mạng nơ-ron

Lưu ý: Hình này cung cấp sơ đồ của hai mạng thần kinh đơn giản có (phải) hoặc không có (trái) lớp ẩn. Vòng tròn màu hồng biểu thị lớp đầu vào và vòng tròn màu đỏ đậm biểu thị lớp đầu ra. Mỗi mũi tên được liên kết với một tham số trọng số. Trong mạng có lớp ẩn, hàm kích hoạt phi tuyến f biến đổi đầu vào trước khi chuyển chúng sang đầu ra.

(sẽ thảo luận sau trong bài đánh giá này). Chúng bao gồm "lớp đầu vào" của các yếu tố dự đoán thô, một hoặc nhiều "lớp ẩn" tương tác và biến đổi phi tuyến tính các yếu tố dự đoán và "lớp đầu ra" tổng hợp các lớp ẩn thành dự đoán kết quả cuối cùng.

Hình 3.6 cho thấy hai mạng ví dụ. Bảng bên trái hiển thị một mạng rất đơn giản không có lớp ẩn. Các yếu tố dự đoán (ký hiệu là z_1, \dots, z_4) được tính trọng số bởi một vectơ tham số (θ) bao gồm một điểm chặn và một tham số trọng số cho mỗi yếu tố dự đoán. Các trọng số này tổng hợp các tín hiệu thành dự báo $\theta_0 + z_k \theta_k$. Như ví dụ đã cho thấy rõ, nếu không có các nút trung gian, mạng lưới thần kinh là một mô hình hồi quy tuyến tính.

Bảng bên phải giới thiệu một lớp ẩn gồm năm nơ-ron. Một nơ-ron nhận được sự kết hợp tuyến tính của các yếu tố dự đoán và cung cấp nó thông qua một "hàm kích hoạt" phi tuyến tính f , sau đó đầu ra này được chuyển đến lớp tiếp theo. (1) (0)

Ví dụ: đầu ra từ nơ-ron thứ hai là $x = f(\theta_0 + 2, \theta_2, j) . \quad 2 \quad (0) \sum_{j=1}^4 z_j \theta_j$

Trong ví dụ này, kết quả từ mỗi nơ-ron được tổng hợp tuyến tính thành dự báo đầu ra cuối cùng:

$$g(z; \theta) = \theta_0 \theta^{(1)} + \sum_{j=1}^5 \theta_j^{(1)} x_j$$

Có nhiều lựa chọn cho hàm kích hoạt phi tuyến (như

như sigmoid, sin, hyperbol, softmax, ReLU, v.v.). Mạng chuyển tiếp nguồn cấp dữ liệu “sâu” giới thiệu các lớp ẩn bổ sung trong đó các phép biến đổi phi tuyến tính từ lớp ẩn trước đó được kết hợp thành tổ hợp tuyến tính và được chuyển đổi một lần nữa thông qua kích hoạt phi tuyến.

Sự ổn định phi tuyến lập đi lập lại này tạo ra các mô hình gần đúng có tham số hóa phong phú hơn và cao hơn.

Như trong Gu và cộng sự. (2020b), mạng nơ-ron chuyển tiếp sâu có công thức chung như sau. Gọi $K(l)$ biểu thị số lượng nơ-ron trong mỗi lớp $l = 1, \dots, L$. Xác định đầu ra của nơ-ron k trong lớp l là k . Tiếp theo, xác định vectơ đầu ra cho lớp này (được tăng cường $x^{(1)}$ thành $(1) (1) \dots (1)$ là x). Để khởi tạo mạng bao gồm một hàng số $x(0)$, hãy xác định tương z_N . Khi đó, công thức $\underset{\theta}{\underset{0}{\text{tự lớp đầu}}}$ vào bằng cách $\underset{\theta}{\underset{0}{\text{tự lớp đầu}}} = \underset{\theta}{\underset{0}{\text{tự lớp đầu}}}(\underset{\theta}{\underset{0}{\text{tự lớp đầu}}}(x_0), \underset{\theta}{\underset{0}{\text{tự lớp đầu}}}(x_1, z_1), \dots,$ đầu ra để quy cho mạng nơ-ron tại mỗi nơ-ron ở lớp $l > 0$ là $=$

$$x_k^{(1)} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \theta = f_k x_k \quad , \quad (3.16)$$

với đầu ra cuối cùng

$$q(z; \theta) = x^{(L-1)} \theta^{(L-1)}. \quad (3.17)$$

Số lượng tham số trọng số trong mỗi lớp ẩn 1 là $K(l) (1 + K(l-1))$, cộng với các trọng số $1 + K(L-1)$ khác cho lớp đầu ra. Mạng năm lớp của Gu et al. (2020b) chẳng hạn, có 30.185 tham số.

Gu và cộng sự. (2020b) ước tính các mô hình dự đoán bằng mức tồn kho hàng tháng cho mẫu CRSP từ năm 1957 đến năm 2016. Các đặc điểm thô của chúng bao gồm 94 đặc điểm cổ phiếu được chuẩn hóa theo thứ hạng tương tác với 8 chuỗi thời gian tài chính vĩ mô cũng như 74 chỉ báo ngành cho tổng số 920 đặc điểm. Họ suy ra sự đánh đổi giữa độ sâu mạng trong bài toán dự báo lợi nhuận bằng cách phân tích hiệu suất của các mạng có từ một đến năm lớp ẩn (ký hiệu là NN1 đến NN5). Dự đoán ngoài mẫu hàng tháng R2 lần lượt là 0,33%, 0,40% và 0,36% cho các mô hình NN1, NN3 và NN5. Điều này so sánh với R2 là 0,16% từ mô hình tuyến tính ba tín hiệu chuẩn (kích thước, giá trị và động lượng) do Lewellen (2015) ủng hộ. Đối với mô hình NN3, R2 cao hơn đáng kể đối với các cổ phiếu lớn ở mức 0,70%, cho thấy rằng học máy không chỉ đơn thuần khắc phục sự kém hiệu quả ở quy mô nhỏ do tính thanh khoản kém. Các

R2 ngoài mẫu tăng lên 3,40% khi dự báo lợi nhuận hàng năm thay vì
hơn hàng tháng, minh họa rằng mạng lưới thần kinh cũng có thể
cô lập các mô hình có thể dự đoán được tồn tại theo tần suất của chu kỳ kinh doanh.

Bảng 3.3: Hiệu suất của danh mục máy học

Bảng A: Trọng lượng bằng nhau		
NN1	NN3	NN5
Dự kiến Std SR trung bình	Dự kiến Std SR trung bình	Dự kiến Std SR trung bình
L -0,45 -0,78 7,43 -0,36 0,15 0,22 6,24 0,12 0,43 0,47 5,55 0,29 0,64 0,64 5,00 0,45 0,80 0,80 4,76 0,58 0,95 0,85 4,63 0,63 1. 12 0,84 4,66 0,62 1,32 0,88 4,95 0,62 1,63 1,17 5,62 0,72 2,13 7,34 1,00	-0,31 -0,92 7,94 -0,40 0,22 0,16 6,46 0,09 0,45 0,44 5,40 0,28 0,60 0,66 4,83 0,48 0,73 0,77 4,58 0,58 0,85 0,81 4,47 0,63 0. 97 0,86 4,62 0,64 1,12 0,93 4,82 0,67 1,38 1,18 5,51 0,74 2,28 2,35 8,11 1,00	-0,08 -0,83 7,92 -0,36 0,33 0,24 6,64 0,12 0,51 0,53 5,65 0,32 0,62 0,59 4,91 0,41 0,71 0,68 4,56 0,51 0,80 0,76 4,43 0,60 0,88 0,88 4,60 0,66 1,01 0,95 4,90 0,67 1,25 1,17 5,60 0,73 2,08 2,27 7,95 0,99
2 3 4 5 6 7 8 9 H 2,43		
HL 2,89	2,91 4,72 2,13	2,58 3,27 4,80 2,36
Bảng B: Trọng số giá trị		
NN1	NN3	NN5
Dự kiến Std SR trung bình	Dự kiến Std SR trung bình	Dự kiến Std SR trung bình
L -0,38 -0,29 7,02 -0,14 2 0,16 0,41 5,89 0,24 3 0,44 0,51 5,07 0,35 4 0,64 0,70 4,56 0,53 5 0,80 0,77 4,37 0,61 0,78 4,39 0,6 2 6 0,95 7 1,11 0,81 4,40 0,64 8 1,31 1,58 0,96 0,75 4,86 0,54 9 5,22 0,64 H 2,19 1,52 6,79 0,77	-0,03 -0,43 7,73 -0,19 0,34 0,30 6,38 0,16 0,51 0,57 5,27 0,37 0,63 0,66 4,69 0,49 0,71 0,69 4,41 0,55 0,76 4,46 0,59 0,79 0. 88 1,09 0,99 4,77 0,72 1,00 5,47 0,69 1,21 1,25 5,94 0,73 1,83 1,69 7,29 0,80	-0,23 -0,51 7,69 -0,23 0,23 0,31 6,10 0,17 0,45 0,54 5,02 0,37 0,60 0,67 4,47 0,52 0,73 0,77 4,32 0,62 0,85 0,86 4,35 0,68 0,96 0,88 4,76 0,64 1,11 0,94 5,17 0,63 1,34 1,02 6,02 0,58 1,99 1,46 7,40 0,68
HL 2,57	1,81 5,34 1,17	1,86 2,12 6,13 1,20

Lưu ý: Trong bảng này, chúng tôi báo cáo hiệu suất của các danh mục đầu tư được sắp xếp theo dự đoán trên giai đoạn thử nghiệm 1987-2016 (được đào tạo về dữ liệu từ 1957-1974 và xác nhận trên dữ liệu từ 1975-1986). Tất cả các cổ phiếu được sắp xếp thành thập phân vị dựa trên lợi nhuận dự đoán của chúng cho tháng tiếp theo. Cột "Pred", "Avg", "Std" và "SR" cung cấp kết quả dự đoán lợi nhuận hàng tháng cho mỗi thập phân vị, lợi nhuận trung bình thực hiện hàng tháng, tiêu chuẩn của họ độ lệch và tỷ số Sharpe tương ứng.

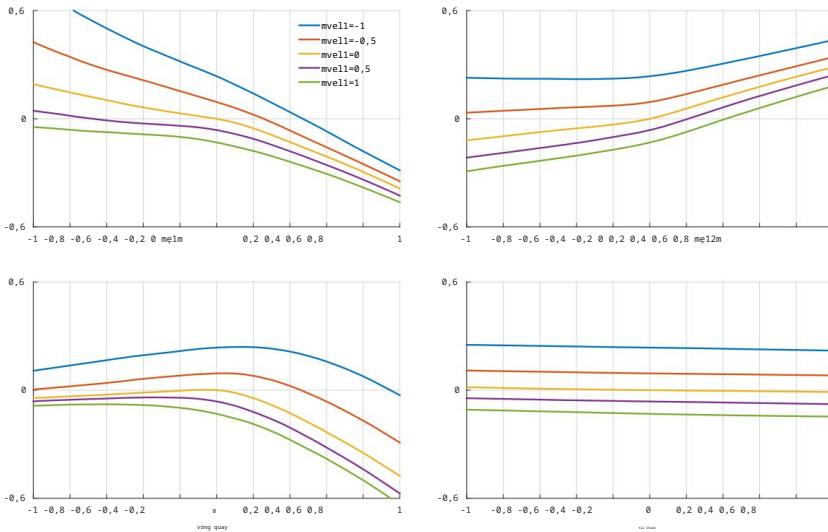
Tiếp theo, Gu và cộng sự. (2020b) báo cáo phân loại danh mục đầu tư dựa trên thần kinh dự đoán lợi nhuận hàng tháng của mạng, được tái tạo trong Bảng 3.3. Bảng A báo cáo lợi nhuận trung bình có tỷ trọng bằng nhau và Bảng B báo cáo tỷ trọng giá trị trả lại. Lợi nhuận của danh mục đầu tư ngoài mẫu tăng đều đặn trên

thập phân. Sự kết hợp định lượng giữa lợi nhuận dự đoán và lợi nhuận thực tế trung bình sử dụng mạng lưới thần kinh là rất án tượng. Danh mục đầu tư có mức chênh lệch thập phân vị dài hạn ngắn hạn kiểm được tỷ lệ Sharpe trọng số bằng nhau hàng năm lần lượt là 2,1, 2,4 và 2,2 cho NN1, NN3 và NNS. Cả ba đều có tỷ lệ Sharpe theo trọng số giá trị là 1,2. Vì vậy, trong khi dự đoán lợi nhuận R2 cao hơn đối với các cổ phiếu lớn thì nội dung dự đoán lợi nhuận của các mô hình mạng thần kinh đặc biệt mang lại lợi nhuận khi được sử dụng để giao dịch các cổ phiếu nhỏ. Avramov và cộng sự. (2022a) chứng thực phát hiện này bằng một cuộc điều tra kỹ lưỡng hơn về giới hạn chênh lệch giá tác động đến các chiến lược giao dịch học máy. Họ chứng minh rằng các dự đoán của mạng lưới thần kinh thành công nhất trong số các cổ phiếu khó định giá và khó kinh doanh chênh lệch giá. Họ nhận thấy rằng, sau khi điều chỉnh chi phí giao dịch và các cân nhắc thực tế khác, các chiến lược dựa trên mạng lưới thần kinh vẫn mang lại lợi ích đáng kể so với các tiêu chuẩn thông thường. Chúng mang lại lợi nhuận cao (đặc biệt là ở các vị thế mua), ít rủi ro giảm giá hơn và tiếp tục hoạt động tốt trong dữ liệu gần đây.

Là một hệ quy chiếu, Gu et al. (2020b) sao chép mô hình tuyến tính ba tín hiệu của Lewellen (2015) thu được tỷ lệ Sharpe có trọng số bằng nhau là 0,8 (0,6 giá trị-trọng số). Bản thân điều này đã rất án tượng, nhưng sự cải thiện lớn từ các dự đoán của mạng lưới thần kinh nhấn mạnh vai trò quan trọng của tính phi tuyến và sự tương tác trong các mô hình lợi nhuận kỳ vọng. Hình 3.7 từ Gu et al. (2020b) minh họa thực tế này bằng cách vẽ đồ thị tác động của một số đặc điểm trong mô hình của họ. Nó cho thấy lợi nhuận kỳ vọng thay đổi như thế nào khi các cặp đặc điểm thay đổi trên mức hỗ trợ [-1,1] của chúng trong khi giữ tất cả các biến khác cố định ở giá trị trung bình của chúng. Các tác động được báo cáo là sự tương tác của quy mô cổ phiếu (mvel1) với sự đảo chiều ngắn hạn (mom1m), động lượng (mom12m), tổng biến động (retvol) và tích lũy (acc). Ví dụ, hình phía trên bên trái cho thấy hiệu ứng đảo chiều ngắn hạn là mạnh nhất và về cơ bản là tuyến tính giữa các cổ phiếu nhỏ (đường màu xanh).

Nhưng trong số các cổ phiếu vốn hóa lớn (đường màu xanh lá cây), hiệu ứng đảo chiều là lõm, thể hiện rõ nhất khi lợi nhuận của các cổ phiếu vốn hóa lớn trong quá khứ rất dương .

Các mô hình trong Gu et al. (2020b) áp dụng cho bảng mức tồn kho. Phong và cộng sự. (2018) sử dụng mạng chuyển tiếp chuỗi thời gian thuận túy để dự báo lợi nhuận tổng hợp của thị trường bằng cách sử dụng các công cụ dự đoán tài chính vĩ mô Welch-Goyal và tìm ra mức tăng R2 ngoài mẫu đáng kể so với mô hình tuyến tính



Hình 3.7: Lợi nhuận kỳ vọng và tương tác đặc trưng (NN3)

Lưu ý: Độ nhạy của tỷ lệ phần trăm lợi nhuận hàng tháng dự kiến (trục tung) đối với các tương tác hiệu ứng cho mve11 với mom1m, mom12m, retvol, acc trong model NN3 (giữ tất cả các đồng biến khác cố định ở giá trị trung bình của chúng).

(kể cả những trường hợp bị phạt, giảm kích thước).

3.8 Phân tích so sánh

Một số bài báo gần đây tiến hành so sánh machine learning mô hình dự đoán trả về trong các tập dữ liệu khác nhau. Trong nghiên cứu đầu tiên như vậy, Gu và cộng sự. (2020b) thực hiện phân tích so sánh máy chính phương pháp học tập được nêu ở trên trong dự đoán bằng tháng chứng khoán cài đặt. Bảng 3.4 tái tạo kết quả chính cho bảng ngoài mẫu trả về dự đoán R2 trên các mô hình. Việc phân tích so sánh này giúp thiết lập một số sự kiện thực nghiệm mới cho việc học máy tài chính. Đầu tiên, mô hình tuyến tính đơn giản với nhiều yếu tố dự đoán ("OLS" trong bảng) gặp khó khăn về độ chính xác của dự báo, không vượt trội hơn so với bảng dự báo ngày thơ dự báo bằng không. Gu và cộng sự. (2020b) xác định R2 dự đoán của họ so với một dự báo ngày thơ về số 0 chứ không phải là tiêu chuẩn chuẩn hơn của

3.8. Phân tích so sánh

65

	OLS OLS-3 PLS PCR ENet GLM RF GBRT NN1 NN2 NN3 NN4 NN5									
	+H +H		+H +H				+H			
Tất cả	-3,46	0,16	0,27	0,26	0,11	0,19	0,33	0,34	0,33	0,39
1000 hàng đầu	-11,28	0,31	-0,14	0,06	0,25	0,14	0,63	0,52	0,49	0,62
Đầy 1000	-1,30	0,17	0,42	0,34	0,20	0,30	0,35	0,32	0,38	0,46

Bảng 3.4: Hiệu suất dự đoán mức tồn kho ngoài mẫu hàng tháng (Tỷ lệ phần trăm $R^2_{\text{đối}}$)

Lưu ý: Trong bảng này, Gu et al. (2020b) báo cáo hàng tháng $R^2_{\text{đối}}$ cho toàn bộ bảng điều khiển cỗ phiếu sử dụng OLS với tất cả các biến (OLS), OLS chỉ sử dụng quy mô, giá trị số sách trên thị trường và động lượng (OLS-3), PLS, PCR, mạng đòn hồi (ENet), mô hình tuyến tính tổng quát (GLM), rặng ngẫu nhiên (RF), cây hồi quy tăng cường độ dốc (GBRT) và mạng lưới thần kinh với một đến năm lớp (NN1-NN5). "+H" biểu thị việc sử dụng mắt Huber thay thế của tần số 12. Roos cũng được báo cáo trong các mẫu phụ chỉ bao gồm phần trên cùng 1.000 cỗ phiếu hoặc 1.000 cỗ phiếu đáy theo giá trị thị trường.

lợi nhuận trung bình của mẫu lịch sử, lưu ý:

Việc dự đoán lợi nhuận vượt trội của cỗ phiếu trong tương lai với mức trung bình trong quá khứ thường kém hiệu quả hơn một dự báo ngây thơ bằng 0. biên độ lớn. Nghĩa là, lợi nhuận cỗ phiếu trung bình lịch sử là như vậy ồn ào đến mức nó hạ thấp tiêu chuẩn dự báo "tốt" một cách giả tạo hiệu suất. Chúng tôi tránh được cảm thấy này bằng cách đo điểm chuẩn cho R^2 của chúng tôi so với giá trị dự báo bằng 0. Để đưa ra một dấu hiệu của tầm quan trọng của sự lựa chọn này khi chúng tôi đánh giá mô hình so với lợi nhuận cỗ phiếu trung bình trong quá khứ, R^2 hàng tháng ngoài mẫu của tất cả các phương pháp đều tăng khoảng ba điểm phần trăm.⁹

Việc điều chỉnh bằng cách giảm kích thước hoặc thu nhỏ sẽ cải thiện R^2 của mô hình tuyến tính xuống khoảng 0,3% mỗi tháng. mô hình phi tuyến, đặc biệt là mạng lưới thần kinh, thậm chí còn giúp ích nhiều hơn nữa, đặc biệt là trong số các mạng lưới lớn cỗ phiếu vốn hóa. Sự phi tuyến tốt hơn mô hình tuyến tính ít chi tiêu chỉ sử dụng ba tín hiệu dự đoán được lựa chọn kỹ càng từ tài liệu ("OLS-3" trong bảng).

Các mô hình phi tuyến cũng mang lại lợi ích gia tăng lớn về mặt kinh tế. Bảng 3.5 tái hiện lại Bảng 8 của Gu et al. (2020b), hiển thị

⁹Anh ấy và cộng sự. (2022a) đề xuất một mặt cắt ngang thay thế ngoài mẫu R^2 cài đặt sử dụng lợi nhuận trung bình mặt cắt ngang làm dự báo ngây thơ.

	OLS-3	PLS	PCR	ENet	GLM	RF	GBRT	NN1	NN2	NN3	NN4	NN5	+H	+H +H	+H		
Hiệu suất được điều chỉnh theo rủi ro (Giá trị có trọng số)																	
Có nghĩa là Ret.	0,94	1,02	1,22	0,60	1,06	1,81	1,921	0,22	-0,28	0,938	1,20	1,33	1,97	2,26	2,12		
FF5+Mẹ a	0,39	0,24						1,20		0,66			1,52	1,76	1,43		
ul trung bình số %	2,76	1,09	2,89	-0,89	1,68			3,95	3,11	4,68	4,74	4,92	6 giò 88		4,71		
R ²	78,60	34,95	39,11	28,04	30,78	13,43	20,68	27,67	25,81	20,84	20,47	18,23					
IR	0,54	0,21	0,57	-0,17	0,33			0,77	0,61	0,92	0,93	0,96	1,18	0,92			
Hiệu suất được điều chỉnh theo rủi ro (Có trọng số bằng nhau)																	
Có nghĩa là Ret.	1,34	2,08	2,45	2,11	2,31	2,91	2,38	1,79	3,20	2,14			3,27	3,33	3,09		
FF5+Mẹ a t(a)	0,83	1,40	1,95	1,32				1,88	1,87	2,60			3,02	3,08	2,78		
R ²	6,64	5,90	9,92	4,77	8,09			6,66	8,19	10,51	11,66	11,70	12,28	10,68			
IR	84,26	26,27	40,50	20,89	21,25	19,91				11,19	13,98	10,60	9,63	11,57	14,54		
	1h30	1,15	1,94	0,93	1,58	1h30			1,60	2,06	2,28	2,29	2,40	2,09			

Bảng 3.5: Hiệu suất, mức giảm và doanh thu được điều chỉnh theo rủi ro của danh mục học máy

Lưu ý: Các bảng báo cáo lợi nhuận trung bình hàng tháng theo phần trăm cũng như bảng chữ cái, tỷ lệ thông tin (IR) và R² đối với mô hình 5 yếu tố Fama-Pháp được tăng cường để bao gồm hệ số động lượng.

hiệu suất ngoài mǎu của các danh mục đầu tư có mức chênh lệch thập phân vị dài hạn ngắn hạn được sắp xếp về dự báo lợi nhuận từ mỗi mô hình. Đầu tiên, điều thú vị cần lưu ý rằng các mô hình có R2 ngoài mǎu tương đối nhỏ tạo ra đáng kể lợi nhuận giao dịch, xét về tỷ lệ alpha và thông tin so với Mô hình sáu yếu tố Fama-Pháp (bao gồm cả yếu tố động lượng). Đây là phù hợp với Kelly et al. (2022a) cho rằng R2 là không đáng tin cậy chẩn đoán giá trị kinh tế của dự đoán lợi nhuận; thay vào đó họ khuyên bạn nên đánh giá các phương pháp học máy tài chính dựa trên kinh tế tiêu chí (chẳng hạn như chiến lược giao dịch tỷ lệ Sharpe). Mô hình phi tuyển trong Bảng 3.5 mang lại giá trị kinh tế tốt nhất về mặt chiến lược giao dịch hiệu suất.

Các bài báo tiếp theo tiến hành phân tích so sánh tương tự ở các tài sản của lớp. Choi và cộng sự. (2022) phân tích các mô hình tương tự như Gu et al. (2020b) trên thị trường chứng khoán quốc tế. Họ đạt được kết luận tương tự rằng các mô hình hoạt động tốt nhất là phi tuyển tính. Điều thú vị là chúng chứng minh tính khả thi của việc học chuyển tiếp. Đặc biệt, một mô hình được đào tạo trên dữ liệu của Hoa Kỳ mang lại hiệu suất ngoài mǎu đáng kể khi được sử dụng để dự báo lợi nhuận chứng khoán quốc tế. Có liên quan, Jiang et al. (2018) tìm mô hình tương tự nhau giữa lợi nhuận chứng khoán và đặc điểm công ty

	OLS	PCA	PLS	Lasso	Ridge	ENet	RF	FFN	LSTM	Kết hợp	
R ²	3,36	2,07	2,03		1,85	1,89	1,87	2,19	2,37		2,28
Trung bình	0,16	0,51	0,63		0,39	0,33	0,43	0,79	0,75		0,79

Bảng 3.6: Dự đoán lợi nhuận trái phiếu so sánh bằng học máy (Bali và cộng sự, 2020)

Lưu ý: Hàng đầu tiên báo cáo ngoài mẫu R² theo tỷ lệ phần trăm cho toàn bộ bảng điều khiển trái phiếu doanh nghiệp sử dụng 43 đặc điểm trái phiếu (từ Bảng 2 của Bali và cộng sự (2020)). Hàng thứ hai báo cáo tỷ lệ phần trăm lợi nhuận ngoài mẫu trung bình hàng tháng của danh mục đầu tư trái phiếu có mức chênh lệch thập phân giá trị giá quyền được sắp xếp dựa trên lợi nhuận học máy dự báo (từ Bảng 3 của Bali và cộng sự (2020)).

Ở Trung Quốc sử dụng PCR và PLS. Gần đây, Leippold và cộng sự. (2022) so sánh mô hình học máy để dự đoán lợi nhuận vốn cổ phần của Trung Quốc và nêu bật tính thanh khoản, sự tham gia của nhà đầu tư bán lẻ và sở hữu nhà nước doanh nghiệp đóng một vai trò rõ rệt trong hành vi thị trường Trung Quốc.

Ait-Sahalia và cộng sự. (2022) nghiên cứu khả năng dự đoán của tần số cao trả về (và các đại lượng khác như thời lượng và âm lượng) bằng máy phương pháp học tập. Họ xây dựng 13 yếu tố dự đoán trong 9 thời điểm khác nhau windows, dẫn đến tổng cộng 117 biến. Họ thử nghiệm với thành phần chỉ số S&P 100 trong một mẫu kéo dài hai năm và thấy rằng tất cả các phương thức có hiệu suất rất giống nhau, ngoại trừ OLS. Những cải tiến từ phương pháp phi tuyến tính như xung ngẫu nhiên hoặc cây được tăng cường là hạn chế trong hầu hết các trường hợp khi so sánh với lasso.

Bali và cộng sự. (2020) và He et al. (2021) tiến hành phân tích so sánh về các phương pháp học máy để dự đoán lợi nhuận trái phiếu doanh nghiệp Hoa Kỳ. Các tín hiệu dự đoán được sử dụng bởi Bali et al. (2020) bao gồm một bộ lớn gồm 43 đặc điểm của trái phiếu như quy mô phát hành, xếp hạng tín dụng, thời hạn và vân vân. Các mô hình họ nghiên cứu cũng giống như Gu et al. (2020b) cộng mạng LSTM (chúng ta sẽ thảo luận về LSTM trong tiêu mục tiếp theo). Bàn 3.6 báo cáo kết quả của Bali et al. (2020). Sự so sánh của họ về máy mô hình học tập về R2 dự đoán và danh mục đầu tư trải rộng thập phân lợi nhuận phần lớn chứng thực kết luận của Gu et al. (2020b). Các mô hình tuyến tính không chính quy là mô hình hoạt động kém nhất. Hình phạt và giảm kích thước cải thiện đáng kể hiệu suất mô hình tuyến tính. Và các mô hình phi tuyến tính nhìn chung có hiệu suất tốt nhất.

Ngoài ra, Bali và cộng sự. (2020) điều tra dự đoán học máy

khuôn khổ tôn trọng ý nghĩa phi chênh lệch giá để đảm bảo tính nhất quán giữa giá vốn cổ phần và giá trị của công ty. Điều này cho phép họ hình thành các dự đoán về cơ cấu vốn của một công ty-tận dụng thông tin vốn chủ sở hữu để dự đoán lợi nhuận trái phiếu. Họ thấy rằng

"một khi chúng tôi áp dụng cấu trúc mô hình Merton (1974), các đặc điểm vốn chủ sở hữu mang lại sự cải thiện đáng kể trên và ngoài các đặc điểm trái phiếu đối với lợi nhuận trái phiếu trong tương lai, trong khi sức mạnh gia tăng của các đặc tính vốn chủ sở hữu để dự đoán lợi nhuận trái phiếu khá hạn chế trong cách tiếp cận dạng rút gọn khi các đặc tính kinh tế như vậy cấu trúc không được áp đặt."

Đây là một ví dụ điển hình về sự bổ sung giữa học máy và cấu trúc kinh tế, lặp lại lập luận của Israel và cộng sự. (2020).

Cuối cùng, Bianchi và cộng sự. (2021) tiến hành phân tích so sánh các mô hình học máy để dự đoán lợi nhuận trái phiếu chính phủ Hoa Kỳ. Đây là môi trường chuỗi thời gian thuận túy (xem các phân tích so sánh đã thảo luận ở trên về dữ liệu trả về của nhóm nghiên cứu). Tuy nhiên, Bianchi et al. (2021) đưa ra kết luận tương tự về giá trị tương đối của các mô hình tuyến tính bị phạt và giảm kích thước so với các mô hình tuyến tính không bị ràng buộc và của các mô hình phi tuyến so với các mô hình tuyến tính.

3.9 Mạng lưới thần kinh tinh vi hơn

Mạng thần kinh tái phát (RNN) là các mô hình phổ biến để thu thập các động lực phức tạp trong dữ liệu chuỗi. Về bản chất, chúng là các mô hình không gian trạng thái phi tuyến có tham số cao, khiến chúng trở thành ứng cử viên thú vị một cách tự nhiên cho các vấn đề dự đoán chuỗi thời gian. Một ứng dụng đầy hứa hẹn của RNN là mở rộng đặc tả mô hình hạn chế được đưa ra bởi (3.2) để nắm bắt sự phụ thuộc trong phạm vi dài hơn giữa kết quả trả về và đặc tính. Cụ thể, chúng ta xem xét mô hình hồi quy tổng quát cho lợi nhuận kỳ vọng của cổ phiếu i tại thời điểm t:

$E_t [R_{i,t+1}] = g(h_{i,t}), \quad (3.18)$ trong đó $h_{i,t}$ là vectơ của các biến trạng thái ẩn phụ thuộc vào $z_{i,t}$ và lịch sử quá khứ của nó.

RNN kinh điển giả định rằng

$$g(h_{i,t}) = \sigma(c + V h_{i,t}), \quad \text{chào}_t = \tanh(b + W h_{i,t} + U z_{i,t}),$$

3.9. Mạng lưới thần kinh tinh vi hơn

69

trong đó b , c , U , V , và W là các tham số chưa biết và $\sigma(\cdot)$ là một sigmoid chức năng. Phương trình trên chỉ bao gồm một lớp trạng thái ẩn.

Thật đơn giản để xếp chồng nhiều lớp trạng thái ẩn lại với nhau để xây dựng một RNN sâu có thể chứa các chuỗi phức tạp hơn.

Ví dụ, chúng ta có thể viết $h_{\text{Nó}}^{(0)} = z_i, t$, và với $1 \leq l \leq L$, ta có

$$g(h_{\text{Nó}}^{(L)}) = \sigma(c + V h_{\text{Nó}}^{(l-1)}), \quad h_{\text{Nó}}^{(l)} = \text{tanh}(b + W h_{\text{Nó}}^{(l-1)} + \tilde{O}(l)_{\text{Nó}}^{(l)}).$$

RNN chuẩn đầu tranh để nắm bắt sự phụ thuộc tầm xa. Của nó cấu trúc kính thiên văn ngũ ý trọng lượng giảm dần theo cấp số nhân của độ trễ trạng thái ở trạng thái hiện tại nên dòng gradient tầm xa biến mất nhanh chóng trong quá trình học tập.

Hochreiter và Schmidhuber (1997) đề xuất một dạng RNN đặc biệt được gọi là mô hình bộ nhớ dài hạn (LSTM). Nó chứa sự kết hợp giữa sự phụ thuộc tầm ngắn và tầm xa thông qua một chuỗi của các hàm cồng điều khiển luồng thông tin từ ht 1 đến ht (chúng ta viết tắt sự phụ thuộc vào i để đơn giản về ký hiệu):

$$ht = ot \tanh(ct), \quad ct = ft ct_{t-1} + it \quad (3.19)$$

trong đó biểu thị phép nhân theo phần tử, ct được gọi là trạng thái ô, N_o , và ft là các "cồng" bản thân chúng là hàm sigmoid của zt và ht 1:

$$\text{tại} = \sigma(Wazt + Uah_t + ba), \quad (3.20)$$

trong đó a có thể là công i , o hoặc f và Wa , Ua và ba là các tham số.

Trạng thái ô bao hoạt động giống như một băng chuyền, đại diện cho "bộ nhớ" đơn vị của mô hình. Các giá trị quá khứ của ct 1 góp phần bổ sung vào ct (trừ từ một số điều chỉnh bằng ft). Cơ chế này cho phép mạng ghi nhớ thông tin tầm xa. Trong khi đó, ft kiểm soát bao nhiêu thông tin từ quá khứ bị lãng quên nên gọi là công quên. Thông tin mới từ ht 1 và zt đến qua c^t :

$$c^t = \tanh(Wczt + Ucht + ba), \quad (3.21)$$

được đưa vào ô nhớ ct . Công đầu vào, nó , kiểm soát nội dung của c^t cần ghi nhớ. Cuối cùng, ot xác định thông tin nào được chuyển sang trạng thái ẩn tiếp theo ht , và do đó được gọi là công đầu ra.

Theo trực giác, không phải tất cả thông tin được ghi nhớ bởi ct đều cần được trích xuất để dự đoán (dựa trên ht).

Mặc dù hữu ích cho việc lập mô hình chuỗi thời gian, LSTM và các phương pháp liên quan (như đơn vị định kỳ của Cho và cộng sự, 2014) ít được ứng dụng trong tài liệu tài chính thực nghiệm. Chúng tôi đã thảo luận về một ngoại lệ đáng chú ý của Bali et al. (2020) ở trên để dự đoán lợi nhuận trái phiếu doanh nghiệp. (Guizarro-Ordonez và cộng sự, 2022) sử dụng kiến trúc RNN để dự đoán lợi nhuận chứng khoán hàng ngày. Một ví dụ khác là Công et al. (2020), người so sánh các mạng chuyển tiếp nguồn cấp dữ liệu đơn giản với LSTM và các mạng thần kinh tái phát khác trong dự đoán lợi nhuận chứng khoán hàng tháng. Họ nhận thấy rằng các dự đoán từ đặc tả LSTM của họ tốt hơn so với các dự đoán từ thông số chuyển tiếp, mặc dù việc mô tả các thông số kỹ thuật của họ còn hạn chế và phân tích của họ được tiến hành phù hợp hơn với các tiêu chuẩn của tài liệu khoa học máy tính. Thật vậy, có một lượng tài liệu khoa học máy tính khá phong phú sử dụng nhiều loại mạng thần kinh để dự đoán lợi nhuận cổ phiếu (ví dụ: Sezer và cộng sự, 2020). Phân tích thực nghiệm tiêu chuẩn trong tài liệu này nhằm mục đích chứng minh bằng chứng cơ bản về khái niệm thông qua các thí nghiệm minh họa quy mô nhỏ và có xu hướng tập trung vào tầm số cao (hàng ngày hoặc trong ngày, thay vì tầm suất tháng hoặc năm có lẽ thú vị hơn về mặt kinh tế) hoặc có xu hướng phân tích một hoặc một vài tài sản tại một thời điểm (ngược lại với một mặt cắt tài sản mang tính đại diện hơn).¹⁰ Điều này trái ngược với các phân tích thực nghiệm sâu rộng hơn phò biến trong tài liệu tài chính và kinh tế có xu hướng phân tích dữ liệu hàng tháng hoặc hàng năm cho các bộ sưu tập lớn của tài sản.

3.10 Mô hình dự đoán trả về cho dữ liệu “thay thế”

Dữ liệu thay thế (hoặc thông tục hơn là “alt”) đã trở thành chủ đề phổ biến trong ngành quản lý tài sản và nghiên cứu gần đây đã đạt được những bước tiến trong việc phát triển mô hình học máy cho một số loại dữ liệu thay thế. Chúng tôi thảo luận về hai ví dụ trong phần này, dữ liệu văn bản và dữ liệu hình ảnh cũng như một số mô hình máy học có giám sát được tùy chỉnh cho các dữ liệu thay thế này nguồn.

¹⁰Ví dụ bao gồm Rather et al. (2015), Singh và Srivastava (2017), Chong và cộng sự. (2017), Bảo và cộng sự. (2017).

3.10.1 Phân tích văn bản

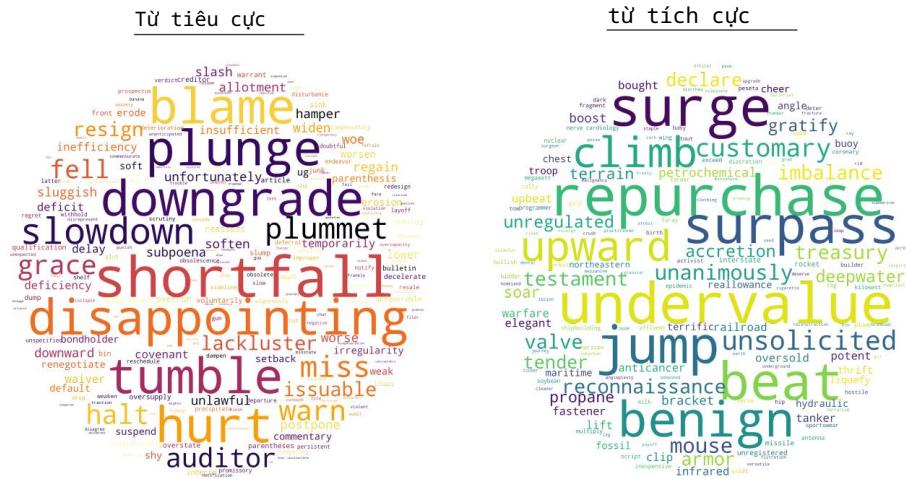
Phân tích văn bản là một trong những lĩnh vực thú vị nhất và phát triển nhanh nhất trong nghiên cứu tài chính và kinh tế. Tài liệu ban đầu đã phát triển từ việc đọc thủ công "gần gũi" của các nhà nghiên cứu (ví dụ Cowles, 1933) sang phương pháp chấm điểm cảm tính dựa trên từ điển (ví dụ Tetlock, 2007). Tài liệu này được khảo sát bởi Das et al. (2014), Gentzkow và cộng sự. (2019), và Loughran và McDonald (2020). Trong phần này, chúng tôi tập trung vào việc học có giám sát dựa trên văn bản với ứng dụng vào dự đoán tài chính.

Jegadeesh và Wu (2013), Ke và cộng sự. (2019), và Garcia và cộng sự. (2022) là ví dụ về các mô hình học có giám sát được tùy chỉnh cho vấn đề dự đoán lợi nhuận bằng cách sử dụng cách trình bày số thuật ngữ hoặc "túi từ" (BoW) của tài liệu văn bản. Ké và cộng sự. (2019) mô tả một mô hình xác suất chung để tạo ra một bài báo về một cổ phiếu và lợi nhuận sau đó của cổ phiếu đó. Một bài viết được lập chỉ mục bởi một tham số "tình cảm" cơ bản duy nhất xác định xu hướng của bài viết về tin tốt hay tin xấu về cổ phiếu. Thông số tương tự này dự đoán hướng sinh lợi trong tương lai của cổ phiếu. Từ mẫu đào tạo các bài báo và kết quả trả về có liên quan, Ke et al. (2019) ước tính tập hợp các thuật ngữ mang lại cảm xúc cao nhất (tức là có khả năng dự đoán lợi nhuận cao nhất) và các giá trị cảm tính liên quan của chúng (tức là hệ số dự đoán của chúng).

Về bản chất, chúng cung cấp một phương pháp dựa trên dữ liệu để xây dựng các từ điển cảm tính được tùy chỉnh cho các nhiệm vụ học tập có giám sát cụ thể.

Phương pháp của họ, được gọi là "SESTM" (Trích xuất tình cảm thông qua sàng lọc và mô hình hóa chủ đề), có ba thành phần trung tâm. Bước đầu tiên tách biệt các thuật ngữ phù hợp nhất khỏi vốn từ vựng rất lớn của các thuật ngữ thông qua sàng lọc tương quan dự đoán. Bước thứ hai chỉ định trọng số cảm tính theo từng thuật ngữ cụ thể bằng cách sử dụng mô hình chủ đề được giám sát. Bước thứ ba sử dụng mô hình chủ đề ước tính để chỉ định điểm cảm tính ở cấp độ bài viết thông qua khả năng bị phạt tối đa.

SESTM dễ dàng và rẻ tiền để tính toán. Bản thân mô hình này có thể xử lý được về mặt phân tích và công cụ ước tính rút gọn thành hai phương trình cốt lõi. Cách tiếp cận mô hình hóa của họ nhấn mạnh đến tính đơn giản và khả năng diễn giải. Do đó, nó là "hộp trắng" và dễ kiểm tra và diễn giải, trái ngược với nhiều mô hình NLP tiên tiến được xây dựng xung quanh hệ thống thần kinh mạnh mẽ nhưng không rõ ràng.



Hình 3.8: Từ ngữ mang tính tình cảm

Lưu ý: Hình này báo cáo danh sách các từ trong tập S. Kích thước phông chữ của một từ tỷ lệ thuận với tông màu trung bình của nó trên tất cả 17 mẫu đào tạo.

thông số kỹ thuật nhúng mạng. Như một minh họa, Hình 3.8 báo cáo hệ số ước tính trên các token là yếu tố dự báo lợi nhuận mạnh nhất dưới dạng đám mây từ. Đám mây được chia thành các mảng thông báo có hệ số dương hoặc âm và kích thước của mỗi mảng thông báo tỷ lệ thuận với độ lớn của hệ số dự đoán ước tính của nó.

Ké và cộng sự. (2019) nghĩ ra một loạt chiến lược giao dịch để chứng minh sức mạnh dự đoán lợi nhuận tiềm năng của SESTM. Khi so sánh trực tiếp, SESTM vượt trội đáng kể so với RavenPack (nhà cung cấp dịch vụ chăm sóc cảm tính thương mại hàng đầu được các nhà quản lý tài sản lớn sử dụng) và các phương pháp dựa trên từ điển như Loughran và McDonald (2011).

Một số bài báo áp dụng mô hình học có giám sát vào dữ liệu văn bản BoW để dự đoán các kết quả tài chính khác. Manela và Moreira (2017) sử dụng hồi quy vectơ hỗ trợ để dự đoán biến động thị trường. Davis và cộng sự. (2020) sử dụng việc công bố hệ số rủi ro 10-K để hiểu phản ứng lợi nhuận khác biệt của các công ty đối với đại dịch COVID-19, tận dụng phương pháp hồi quy nghịch đảo đa thức của Taddy (2013). Kelly và cộng sự. (2018) giới thiệu một phương pháp gọi là hồi quy đa thức phân tán rào cản (HDMR) để

cải thiện thông số kỹ thuật của mô hình đểm và sử dụng nó để xây dựng chỉ số dựa trên văn bản đo lường sức khỏe của khu vực trung gian tài chính.

Những phân tích này tiến hành theo hai bước chung. Bước 1 quyết định cách biểu diễn bằng số của dữ liệu văn bản. Bước 2 sử dụng các biểu diễn dưới dạng dữ liệu trong mô hình kinh tế lượng để mô tả một số hiện tượng kinh tế (ví dụ: lợi nhuận tài sản, sự biến động và các nguyên tắc cơ bản về kinh tế vĩ mô trong các tài liệu tham khảo ở trên).

Các trình bày văn bản tài chính được tham chiếu ở trên có một số hạn chế. Đầu tiên, tất cả các ví dụ này đều bắt đầu từ cách biểu diễn BoW, quá đơn giản và chỉ truy cập thông tin ở dạng văn bản có thể truyền tải theo tần suất sử dụng thuật ngữ. Nó hy sinh gần như toàn bộ thông tin được truyền tải thông qua trật tự từ hoặc mối quan hệ ngữ cảnh giữa các thuật ngữ. Thứ hai, tính chiều cực cao của các biểu diễn BoW dẫn đến thiếu hiệu quả về mặt thống kê—Các mô hình kinh tế lượng Bước 2 phải bao gồm nhiều tham số để xử lý tất cả các thuật ngữ này mặc dù nhiều thuật ngữ truyền tải thông tin không đáng kể. Việc giảm kích thước như LDA và sàng lọc tương quan có lợi vì chúng giảm thiểu tính kém hiệu quả của BoW. Tuy nhiên, chúng có nguồn gốc từ BoW và do đó không tránh được việc mất thông tin do dựa vào số lượng thuật ngữ ngay từ đầu. Thứ ba, và tinh tế hơn, các biểu diễn giảm kích thước mang tính cụ thể cho kho ngữ liệu. Ví dụ, khi Bybee et al. (2020) xây dựng mô hình chủ đề của họ, các chủ đề chỉ được ước tính từ The Wall Street Journal, mặc dù thực tế là nhiều chủ đề là cấu trúc ngôn ngữ chung và có thể được suy luận tốt hơn bằng cách sử dụng văn bản bổ sung bên ngoài mẫu của chúng.

Giang và cộng sự. (2023) đưa tài liệu tiến thêm một bước nữa bằng cách xây dựng cách trình bày văn bản tin tức tinh tế bắt nguồn từ cái gọi là “mô hình ngôn ngữ lớn” (LLM). Sau đó, họ sử dụng những cách trình bày này để cải thiện các mô hình về lợi nhuận cổ phiếu kỳ vọng. LLM được đào tạo trên các tập dữ liệu văn bản lớn trải rộng trên nhiều nguồn và chủ đề. Khóa đào tạo này được thực hiện bởi các nhóm nghiên cứu chuyên ngành thực hiện kỳ công Herculean trong việc ước tính mô hình ngôn ngữ có mục đích chung với tham số hóa thiên văn trên dữ liệu văn bản thực sự lớn. LLM có hàng tỷ tham số (hoặc nhiều hơn) và được đào tạo trên hàng tỷ ví dụ văn bản (bao gồm cả kho sách hoàn chỉnh khổng lồ và phần lớn Internet). Nhưng đối với mỗi LLM, kỳ tích ước tính này được thực hiện một lần, sau đó mô hình ước tính sẽ được cung cấp để phân phối cho các nhà nghiên cứu không chuyên triển khai.

trong các nhiệm vụ hạ nguồn.

Nói cách khác, LLM ủy quyền Bước 1 của quy trình trên cho một số chuyên gia trên thế giới có thể thực hiện nó tốt nhất. Sau đó, mô hình kinh tế lượng Bước 2 có thể được xây dựng dựa trên đầu ra LLM. Giống như LDA (hoặc thậm chí BoW), đầu ra của mô hình nền tảng là biểu diễn vectơ số (hoặc "nhúng") của một tài liệu. Một nhà nghiên cứu không chuyên có được kết quả này bằng cách cung cấp tài liệu quan tâm thông qua phần mềm (trong nhiều trường hợp là nguồn mở). Lợi ích chính của LLMS ở Bước 1 là nó cung cấp các cách trình bày văn bản phức tạp và được đào tạo bài bản hơn so với những gì được sử dụng trong tài liệu được tham chiếu ở trên. Lợi ích này đến từ tính biểu cảm của các tham số hóa mô hình phi tuyến nặng và từ việc đào tạo các ví dụ ngôn ngữ mở rộng trên nhiều lĩnh vực, trong suốt lịch sử loài người và bằng nhiều ngôn ngữ khác nhau. Khả năng chuyển giao của LLM làm cho quy mô kiến thức chưa từng có này có sẵn cho nghiên cứu tài chính.

Giang và cộng sự. (2023) phân tích các dự đoán lợi nhuận dựa trên văn bản tin tức được xử lý thông qua một số LLM, bao gồm cả Biểu diễn bộ mã hóa hai chiều từ Transformer (BERT) của Devlin et al. (2018), Máy biến áp được đào tạo trước (GPT) của Radford và cộng sự. (2019) và Máy biến áp được huấn luyện trước mở (OPT) của Zhang và cộng sự. (2022). Họ nhận thấy rằng các dự đoán từ các phần nhúng LLM được đào tạo trước vượt trội hơn các dự đoán lợi nhuận học máy dựa trên văn bản phổ biến về hiệu suất chiến lược giao dịch ngoài mẫu và hiệu suất vượt trội của LLM bắt nguồn từ thực tế là chúng có thể nắm bắt thành công hơn ý nghĩa ngữ cảnh trong các tài liệu.

3.10.2 Phân tích hình ảnh

Phần lớn công nghệ học máy hiện đại đã phát triển xung quanh nhiệm vụ phân tích hình ảnh và thị giác máy tính, với lợi ích lớn trong các nhiệm vụ liên quan đến hình ảnh bắt nguồn từ sự phát triển của các mô hình mạng thần kinh tích chập (CNN). Giang và cộng sự. (2022) giới thiệu kỹ thuật phân tích ảnh CNN cho bài toán dự đoán kết quả.

Một tài liệu tài chính lớn nghiên cứu các mô hình giá trong quá khứ dự báo lợi nhuận trong tương lai như thế nào. Quan điểm triết học làm nền tảng cho những phân tích này phổ biến nhất là quan điểm kiểm tra giả thuyết. Các nhà nghiên cứu

xây dựng một mô hình dự đoán lợi nhuận dựa trên xu hướng giá-chẳng hạn như sự hồi quy của lợi nhuận một tháng trước đối với lợi nhuận trung bình trong 12 tháng trước đó-như một thử nghiệm về giả thuyết vô hiệu của thị trường hiệu quả dạng yếu. Tuy nhiên, thật khó để tìm thấy trong tài liệu một giả thuyết thay thế cụ thể nào. Nói cách khác, các công cụ dự đoán lợi nhuận dựa trên giá được nghiên cứu trong tài liệu nói chung là đặc biệt và được phát hiện thông qua quá trình học tập thống kê chuyên sâu của con người diễn ra đằng sau bức màn của quá trình nghiên cứu học thuật. Giang và cộng sự. (2022) xem xét lại ý tưởng về khả năng dự đoán lợi nhuận dựa trên giá từ một quan điểm triết học khác dựa trên học máy. Với những bước tiến gần đây trong việc tìm hiểu hành vi của con người ảnh hưởng như thế nào đến mô hình giá cả (ví dụ như Barberis và Thaler, 2003; Barberis, 2018), thật hợp lý khi cho rằng giá cả chưa đựng những mô hình tinh vi và phức tạp mà không có thể phát triển các giả thuyết cụ thể có thể kiểm chứng được. Giang và cộng sự. (2022) nghĩ ra một phương pháp học máy có hệ thống để gợi ra các mô hình dự đoán lợi nhuận dựa trên dữ liệu định giá, thay vì thử nghiệm các giả thuyết đặc biệt cụ thể.

Thách thức đối với việc khám phá như vậy là cân bằng các mô hình linh hoạt có thể phát hiện các mẫu có khả năng tinh vi với mong muốn duy trì khả năng điều khiển và khả năng diễn giải của các mô hình đó. Để điều hướng sự cân bằng này, Jiang et al. (2022) thể hiện giá lịch sử dưới dạng hình ảnh và sử dụng máy CNN được phát triển tốt để phân tích hình ảnh nhằm tìm kiếm các mô hình dự đoán. Hình ảnh của chúng bao gồm giá mở cửa hàng ngày, giá cao nhất, giá thấp nhất và giá đóng cửa (thường được gọi là biểu đồ "OHLC") được phủ lên bởi đường trung bình động nhiều ngày của giá đóng cửa và biểu đồ thanh cho khối lượng giao dịch hàng ngày (xem Hình 3.9 để biết thông tin liên quan). ví dụ từ Yahoo Finance).

CNN được thiết kế để tự động trích xuất các đặc điểm từ hình ảnh có tính dự đoán cho các nhãn giám sát (là lợi nhuận thực tế trong tương lai trong trường hợp của Jiang và cộng sự, 2022). Dữ liệu thô bao gồm các mảng giá trị pixel. CNN thường có một số khối xây dựng cốt lõi giúp chuyển đổi dữ liệu pixel thành các tính năng dự đoán. Các khối xây dựng được xếp chồng lên nhau theo nhiều cấu hình ống lồng khác nhau tùy thuộc vào ứng dụng hiện tại. Chúng làm mịn nội dung hình ảnh về mặt không gian để giảm nhiễu và làm nổi bật các đường viền hình dạng nhằm tối đa hóa mối tương quan giữa hình ảnh với nhãn của chúng. Các thông số của các khối xây dựng được học như một phần



Hình 3.9: Biểu đồ Tesla OHLC từ Yahoo! Tài chính

Lưu ý: Biểu đồ OHLC cho cổ phiếu Tesla với đường giá trung bình động 20 ngày và thanh khối lượng hàng ngày. Dữ liệu hàng ngày từ ngày 1 tháng 1 năm 2020 đến ngày 18 tháng 8 năm 2020.

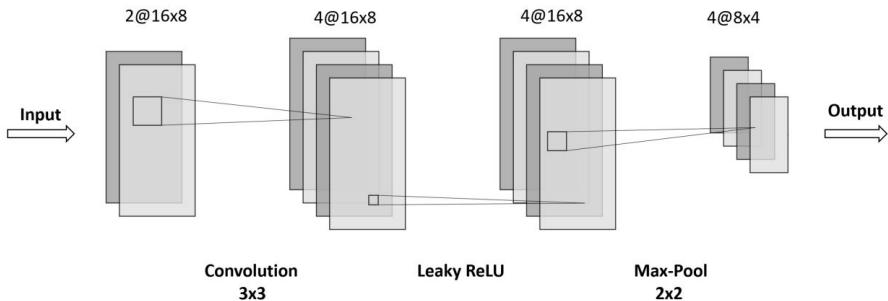
của quá trình ước lượng mô hình.

Mỗi khối xây dựng bao gồm ba hoạt động: tích chập, kích hoạt và gộp. "Convolution" là một dạng tương tự về mặt không gian của việc làm mịn hạt nhân cho chuỗi thời gian. Convolution quét qua hình ảnh và đối với mỗi phần tử trong ma trận hình ảnh, tạo ra một bản tóm tắt nội dung hình ảnh trong khu vực xung quanh ngay lập tức. Phép tích chập hoạt động thông qua một tập hợp các "bộ lọc" đã học, là các ma trận trọng số hạt nhân có kích thước thấp lấy trung bình các phần tử ma trận lân cận.

Hoạt động thứ hai trong khối xây dựng, "kích hoạt", là một phép biến đổi phi tuyến được áp dụng theo từng phần tử cho đầu ra của bộ lọc tích chập. Ví dụ: kích hoạt "Leaky ReLU" sử dụng hàm tuyến tính lồi từng phần, có thể được coi là làm sắc nét độ phân giải của đầu ra bộ lọc tích chập nhất định.

Hoạt động cuối cùng trong một khối xây dựng là "tổng hợp tối đa". Thao tác này sử dụng một bộ lọc nhỏ quét qua ma trận đầu vào và trả về giá trị lớn nhất của các phần tử đi vào bộ lọc tại mỗi vị trí trong ảnh. Max-pooling đóng vai trò vừa là một thiết bị giảm kích thước vừa là một công cụ khử nhiễu.

Hình 3.10 minh họa cách kết hợp tích chập, kích hoạt và gộp tối đa để tạo thành khối xây dựng cơ bản cho mô hình CNN. Bằng cách xếp chồng nhiều khối này lại với nhau, trước tiên mạng sẽ tạo ra các biểu diễn



Hình 3.10: Sơ đồ khái niệm

Lưu ý: Khối xây dựng của mô hình CNN bao gồm lớp tích chập với bộ lọc 3×3 , lớp ReLU bị rò rỉ và lớp tổng hợp tối đa 2×2 . Trong ví dụ về đồ chơi này, đầu vào có kích thước 16×8 với 2 kênh. Để tăng gấp đôi độ sâu của đầu vào, 4 bộ lọc được áp dụng, tạo ra đầu ra với 4 kênh. Lớp gộp tối đa thu hai chiều đầu tiên (chiều cao và chiều rộng) của đầu vào xuống một nửa và giữ nguyên độ sâu. ReLU bị rò rỉ giữ nguyên kích thước của đầu vào trước đó. Nói chung, với đầu vào có kích thước $h \times w \times d$, đầu ra có kích thước $h/2 \times w/2 \times 2d$. Một ngoại lệ là khái niệm đầu tiên của mỗi mô hình CNN lấy hình ảnh thang độ xám làm đầu vào: đầu vào có độ sâu 1 và số lượng bộ lọc CNN là 32, tăng độ sâu của đầu ra lên 32.

các thành phần nhỏ của ảnh sau đó dần dần tập hợp chúng thành các biểu diễn của các vùng lớn hơn. Đầu ra từ khái niệm cuối cùng được làm phẳng thành một vectơ và mỗi phần tử được coi như một tính năng trong lớp chuyển tiếp nguồn cấp dữ liệu được kết nối đầy đủ, tiêu chuẩn cho bước dự đoán cuối cùng.¹¹ Jiang et al. (2022) đào tạo một mô hình CNN để dự đoán hướng lợi nhuận

của cổ phiếu trong tương lai bằng cách sử dụng dữ liệu chứng khoán hàng ngày của Hoa Kỳ từ năm 1993 đến năm 2019. Chiến lược giao dịch chênh lệch giá theo thập phân vị dài hạn ngắn hạn được cân bằng lại hàng tuần dựa trên xác suất thay đổi giá dương dựa trên CNN sẽ mang lại lợi nhuận Tỷ lệ Sharpe hàng năm ngoài mẫu là 7,2 trên cơ sở trọng số bằng nhau và 1,7 trên cơ sở trọng số giá trị. Điều này vượt trội hơn các chiến lược xu hướng giá nổi tiếng – các dạng động lượng và đảo chiều khác nhau – với một mức chênh lệch lớn và đáng kể. Họ cho thấy rằng việc biểu diễn hình ảnh là động lực chính cho sự thành công của mô hình, vì các thuật toán mạng thần kinh chuỗi thời gian khác gặp khó khăn trong việc khớp với hiệu suất của hình ảnh CNN. Giang

¹¹Ch. 9 của Goodfellow và cộng sự. (2016) cung cấp phần giới thiệu tổng quát hơn về CNN.

et al. (2022) lưu ý rằng chiến lược của họ có thể gần đúng một phần bằng cách thay thế dự báo CNN bằng tín hiệu đơn giản hơn. Đặc biệt, một cổ phiếu có giá đóng cửa gần nhất ở gần đây của phạm vi cao-thấp trong những ngày gần đây có xu hướng tăng giá trong tuần tiếp theo. Mô hình này, trước đây chưa được nghiên cứu trong tài liệu, được CNN phát hiện từ dữ liệu hình ảnh và ổn định theo thời gian cũng như trên các phân khúc quy mô thị trường và giữa các quốc gia.

Glaeser và cộng sự. (2018) nghiên cứu thị trường nhà ở sử dụng hình ảnh bất động sản nhà ở. Họ sử dụng mô hình CNN được đào tạo trước (Resnet-101 từ He và cộng sự, 2016) để chuyển đổi hình ảnh thành một vector đặc trưng cho từng thuộc tính, sau đó cô đọng thêm bằng cách sử dụng các thành phần chính và cuối cùng các thành phần được thêm vào một tiêu chuẩn khác mô hình định giá nhà theo chủ nghĩa khoái lạc. Glaeser và cộng sự. (2018) nhận thấy rằng dữ liệu đặc tính thu được từ hình ảnh giúp cải thiện mức độ phù hợp ngoài mẫu của mô hình he-donic. Aubry và cộng sự. (2022) dự đoán giá đầu giá tác phẩm nghệ thuật thông qua mạng thần kinh bằng cách sử dụng hình ảnh tác phẩm nghệ thuật đi kèm với đặc điểm tác phẩm nghệ thuật không trực quan và sử dụng mô hình của họ để ghi lại một số thông tin thiểu số quả trên thị trường nghệ thuật. Obaid và Pukthuanthong (2022) áp dụng CNN để phân loại ảnh trên Wall Street Journal và xây dựng chỉ số tâm lý nhà đầu tư hàng ngày, dự đoán sự đảo chiều lợi nhuận của thị trường và khôi phục giao dịch. Mỗi liên hệ này mạnh nhất trong số các cổ phiếu có giới hạn chênh lệch giá nghiêm ngặt hơn và trong thời kỳ rủi ro cao. Họ cũng nhận thấy rằng các bức ảnh truyền tải thông tin thay thế cho văn bản tin tức.¹²

¹²Liên quan, Đặng và cộng sự. (2022) đề xuất một mô hình lý thuyết để hợp lý hóa cách thức nội dung đồ họa của báo cáo 10-K ảnh hưởng đến lợi nhuận chứng khoán.

4

Sự đánh đổi rủi ro-lợi nhuận

Phần trước chủ yếu tập trung vào các mô hình dự đoán có giám sát, không có quan điểm về sự đánh đổi rủi ro-lợi nhuận và do đó không cấu thành các mô hình định giá tài sản. Trong phần này, chúng tôi phát triển các mô hình định giá nhân tố bằng cách sử dụng các phương pháp học không giám sát và bán giám sát để mô hình hóa sự cân bằng giữa rủi ro và lợi nhuận một cách rõ ràng.

4.1 Nền tảng APT

Lý thuyết định giá chênh lệch giá (APT) của Ross (1976) đặt nền tảng cho việc phân tích máy học dựa trên dữ liệu của các mô hình định giá theo yếu tố. Nó chứng minh rằng với đặc tả mô hình một phần-về cơ bản chỉ yêu cầu cấu trúc nhân tố tuyến tính, một số nhân tố cố định và giả định kinh tế tối thiểu về không có chênh lệch giá-chúng ta có thể tìm hiểu mô hình định giá tài sản đơn giản bằng cách nghiên cứu danh mục nhân tố và hiểu những thành phần nào. Lợi nhuận có thể đa dạng hóa được và cái nào không. Nói cách khác, APT cung cấp một kế hoạch chi tiết để phân tích thực nghiệm sự đánh đổi giữa rủi ro và lợi nhuận mà không yêu cầu bất kỳ kiến thức nào về cơ chế làm phát sinh các yếu tố định giá tài sản. Do đó, các phương pháp học máy để phân tích nhân tố tiềm ẩn có thể được tận dụng để tiến hành các phân tích mới và mạnh mẽ về hiện tượng định giá tài sản theo kinh nghiệm.

Chúng tôi cũng giới thiệu đến độc giả bản khảo sát về mô hình hệ số sinh lợi của Giglio et al. (2022a). Họ sắp xếp tài liệu thành các danh mục dựa trên việc các yếu tố có thể quan sát được, beta có thể quan sát được hay không thể quan sát được. Các yếu tố hoặc beta có thể quan sát được làm phát sinh các phương pháp hồi quy chuỗi thời gian và mặt cắt ngang. Trong khảo sát này, chúng tôi tập trung vào trường hợp khó khăn hơn trong đó các yếu tố và beta tiềm ẩn (hoặc tốt nhất là có thể quan sát được một phần), điều này cho phép chúng tôi đi sâu hơn vào một số chi tiết về kỹ thuật mô hình hóa yếu tố học máy.

4.2 Mô hình nhân tố vô điều kiện

Tiền đề của APT [của Ross \(1976\)](#) là hệ số thống kê sau người mẫu:

$$R_t = \alpha + \beta F_t + \epsilon_t, \quad (4.1)$$

trong đó R_t thu thập tài sản dư thừa trả về R_i, t thành vectơ $N \times 1$, β là ma trận $N \times K$ các tài sản vectơ tiềm ẩn $K \times 1$ có thể có giá trị trung bình khác 0 $y = E(F_t)$ được hiểu là yếu tố, , là một vectơ $N \times 1$ của phần dư trung bình bằng 0. phần bù rủi ro yếu tố F_t , và

Vectơ chặn $N \times 1$ α biểu thị lỗi về giá. Chúng là các thành phần của lợi nhuận tài sản kỳ vọng không được giải thích bằng các yếu tố rủi ro . APT [của Ross \(1976\)](#) và các hậu duệ của nó (ví dụ Huberman, 1982; Ingersoll, 1984; Chamberlain và Rothschild, 1983) xác định rằng kinh doanh chênh lệch giá không gần tương đương với

$$\alpha \text{ Var}(\epsilon_t) = 1 \alpha \leq 1, \text{ dưới dạng } N \rightarrow \infty. \quad (4.2)$$

Nghĩ α là, khoản bù đắp cho việc gánh chịu rủi ro đặc thù không bùng nổ khi vũ trụ đầu tư mở rộng.

4.2.1 Ước tính các yếu tố thông qua PCA

Được thúc đẩy bởi APT, Chamberlain và Rothschild (1983), Connor và Korajczyk (1986), Connor và Korajczyk (1988) ủng hộ PCA như một công cụ ước tính mô hình nhân tố khi các yếu tố và beta tiềm ẩn. MỘT

¹Chúng ta sử dụng p để biểu thị $a = OP(b)$, và sử dụng P để nếu $a \in P$ và b

$P(a)$.

Cách tiếp cận thuận tiện hơn nhưng tương đương hơn là tiến hành phân tách giá trị số ít (SVD) của các kết quả trả về không có ý nghĩa a $R^- = R(\frac{1}{T} - Tt = Rt) \perp T$:

$$R^- = \sum_{j=1}^K \sigma_j \xi_j + \text{Bạn}, \quad (4.3)$$

trong đó $\{\sigma_j\}$, $\{\xi_j\}$ tương ứng với các giá trị số ít đầu tiên của K , vectơ số ít bên trái và bên phải của R^- , K là bất kỳ công cụ ước tính nhất quán nào (ví dụ: Bai và Ng, 2002) của số phần tử trong Rt là ma trận phần dư và T là vectơ $T \times 1$ của phần tử. Sự phân tách này mang lại ước tính về sự đổi mới của yếu tố $Vt = Ft = E(Ft)$ và mức độ rủi ro β như

$$V = T^{1/2} (\xi_1, \xi_2, \dots, \xi_K), \quad \beta = T^{-1/2} (\sigma_1 \zeta_1, \sigma_2 \zeta_2, \dots, \sigma_K \zeta_K). \quad (4.4)$$

Các ước lượng hệ số được chuẩn hóa sao cho chúng thỏa mãn $VV = TIK$. Ngoài ra, chúng ta có thể chuẩn hóa β sao cho $\beta \beta = NIK$. Có sự không xác định cơ bản trong các mô hình nhân tố tiềm ẩn, tức là các nhân tố quay và tải trọng quay ngược chiều dẫn đến không có thay đổi trong quá trình tạo dữ liệu. Do đó, các hệ số và tải của chúng chỉ có thể xác định được khi có một phép biến đổi tuyến tính khả nghịch (tức là phép quay). Do đó, các cách chuẩn hóa khác nhau mang lại các ước tính tương đương về các hệ số và tải trọng, trong đó chúng chỉ khác nhau bởi một số phép quay. Bai (2003) chứng minh tính nhất quán của các ước tính PCA này và rút ra phân bố tiệm cận của chúng theo giả định rằng tất cả các yếu tố đều có sức lan tỏa, tức là, $\lambda K(\beta \beta) \rightarrow N$.

P Connor và Korajczyk (1988) trước tiên điều tra hiệu suất của mô hình nhân tố tiềm ẩn bằng cách sử dụng một mặt cắt ngang lớn của khoảng 1.500 cổ phiếu. Họ nhận thấy rằng, mặc dù mô hình nhân tố dựa trên PCA hoạt động tốt hơn CAPM trong việc giải thích sự cân bằng giữa rủi ro và lợi nhuận trong mẫu của họ, nhưng nó vẫn thừa nhận những lỗi định giá lớn và đáng kể. Nói chung, các mô hình nhân tố vô điều kiện gặp khó khăn trong việc mô tả dữ liệu ở mức tồn kho. Dựa trên nghiên cứu này và các nghiên cứu liên quan, các mô hình nhân tố tiềm ẩn vô điều kiện (và ước tính của chúng thông qua PCA) phần lớn không còn được ưa chuộng trong giai đoạn sau Connor và Korajczyk (1988). Kelly và cộng sự. (2020b) chứng thực những phát hiện đó bằng dữ liệu gần đây hơn. Họ cho thấy rằng, trong bảng dữ liệu về cổ phiếu CRSP từ năm 1962-2014, PCA cực kỳ không đáng tin cậy khi mô tả phần bù rủi ro của từng cổ phiếu riêng lẻ.

Gần đây có sự hồi sinh của PCA đối với mô hình hệ số sinh lợi. Điều này phần lớn xuất phát từ thực tế là, mặc dù PCA gặp khó khăn khi mô tả các bảng cổ phiếu riêng lẻ, nhưng nó lại thành công hơn nhiều trong việc mô hình hóa các bảng danh mục đầu tư. Ví dụ, Kelly và cộng sự. (2020b), Kozak và cộng sự. (2018), và Pukthuanthong et al. (2019) chứng minh rằng các mô hình nhân tố được ước tính từ một nhóm danh mục đầu tư bất thường có thể định giá những danh mục đầu tư đó với sai số định giá nhỏ về mặt kinh tế. Những phân tích này được xây dựng dựa trên công trình trước đó của Geweke và Chu (1996), những người sử dụng phương pháp lấy mẫu Gibbs để trích xuất các yếu tố tiềm ẩn từ dữ liệu cấp danh mục đầu tư.

Một nhược điểm tiềm ẩn của cách tiếp cận nhân tố tiềm ẩn nằm ở chỗ khó diễn giải các nhân tố ước tính do tính không xác định của các mô hình nhân tố tiềm ẩn. Tuy nhiên, sẽ có lợi khi áp dụng cách tiếp cận hệ số tiềm ẩn bắt cứ khi nào đối tượng quan tâm là bất biến đối với phép quay. Chúng tôi cung cấp một ví dụ như vậy tiếp theo.

4.2.2 Công cụ ước tính rủi ro ba lần

Phí bảo hiểm rủi ro yếu tố mô tả mức độ rủi ro mà nhà đầu tư yêu cầu để nắm giữ rủi ro liên quan đến yếu tố đó. Nhiều mô hình kinh tế lý thuyết được phát triển trên cơ sở một số yếu tố phi thương mại (bản thân chúng không phải là danh mục đầu tư) như tiêu dùng, tăng trưởng GDP, lạm phát, thanh khoản và rủi ro khí hậu. Để ước tính phần bù rủi ro của một yếu tố phi thương mại, chúng ta cần xây dựng một danh mục đầu tư tương tự yếu tố và ước tính lợi nhuận kỳ vọng của nó. Để khắc phục ý tưởng, giả sử rằng yếu tố không thể giao dịch này, Gt liên quan đến mặt cắt ngang của tài sản ở dạng sau:

$$Gt = \xi + \eta Vt + Zt , \quad (4.5)$$

trong đó Zt là sai số đo và $Vt = Ft - E(Ft)$. Theo mô hình này, phần bù rủi ro của Gt được tính bằng η . Vì Gt đôi khi được thúc đẩy từ một số lý thuyết kinh tế nên danh mục đầu tư và phần bù rủi ro bắt chước của nó có thể được giải thích về mặt kinh tế. Hơn nữa, ηVt và η là bất biến phép quay và có thể nhận dạng được mặc dù Vt , η và y chỉ có thể nhận dạng được trong một số phép quay. Để thấy điều này, Bai và Ng (2002) đã biết rằng tồn tại một ma trận H sao cho V bất kỳ t . Nếu chúng ta viết $t^P HVT$, vì lại DGP của Rt và Gt đối với HVT , sau đó

Phần bù rủi ro của V_t là H_t , và gánh nặng của G_t lên H_{Vt} trở thành η_H 1 , tuy nhiên sự đổi mới của danh mục đầu tư bắt chước và phần bù rủi ro của G_t vẫn là η_H $1H_{Vt} = \eta_{Vt}$ và η_H $1H_t = \eta_t$.

Giglio và Xiu (2021) đề xuất một công cụ ước tính ba bước để suy luận về η_t , kết hợp hồi quy Fama-MacBeth với PCA. Bước đầu tiên là tiến hành PCA và ước tính các hệ số và tải theo (4.4). Bước thứ hai phục hồi phần bù rủi ro của các yếu tố tiềm ẩn thông qua hồi quy Fama-MacBeth:

$$\gamma = \frac{1}{T} \left(\beta \beta \right)^{-1} \beta R_t . \quad (4.6)$$

Bước thứ ba phục hồi tải G_t trên các hệ số tiềm ẩn ước tính:

$$\eta = \frac{1}{T} G_t V_t .$$

Công cụ ước tính phần bù rủi ro do đó trở thành: η_t .

Giglio và Xiu (2021) thiết lập tính chất tiềm cận của công cụ ước tính kết quả. Phân tích tiềm cận của họ về Fama-MacBeth trên kết quả đầu ra PCA mở đường cho suy luận thống kê về mức độ quan tâm trong việc định giá tài sản với các mô hình yếu tố tiềm ẩn, bao gồm phần bù rủi ro, hệ số chiết khấu ngẫu nhiên (Giglio và cộng sự, 2021b) và alpha (Giglio et al. . , 2021a).

Công cụ ước tính ba lần có liên quan chặt chẽ đến danh mục đầu tư bắt chước dựa trên hồi quy PCA. Quy trình này giống như hồi quy G_t trên PC của R_t để xây dựng danh mục bắt chước nhân tố của nó và tính toán lợi nhuận trung bình của nó để đạt được phần bù rủi ro của G_t . Việc sử dụng PC thay vì tài sản ban đầu trong R_t là một hình thức chính quy hóa. Quan điểm này khuyến khích việc áp dụng các phương pháp chính quy hóa khác trong học máy, như sườn núi và dây thông lọng, khi tạo danh mục đầu tư bắt chước.

Về mặt thực nghiệm, Bảng 4.1 thu thập các ước tính phần bù rủi ro bằng các phương pháp khác nhau cho một số yếu tố phi thương mại, bao gồm những đổi mới AR(1) trong tăng trưởng sản xuất công nghiệp (IP), những đổi mới VAR(1) trong ba thành phần chính đầu tiên của 279 các biến tài chính vĩ mô của Ludvigson và Ng (2010) , hệ số thanh khoản của Pástor và Stambaugh (2003) , hệ số vốn trung gian của He et al. (2017) , bốn yếu tố từ Novy-Marx (2014) (nhiệt độ tháng cao ở

Manhattan, độ thường nhiệt độ bề mặt đất toàn cầu, độ thường nhiệt độ gần như chu kỳ ở Thái Bình Dương hay “El Niño” và số lượng vết đen mặt trời), và yếu tố dựa trên mức tiêu thụ tổng hợp từ Malloy et al. (2009).

Bảng này nêu bật hai vấn đề của hồi quy hai bước thông thường : độ lệch biến bị bỏ qua và độ lệch sai số đo. Ước tính hai lần phụ thuộc vào các yếu tố chuẩn mà các nhà nghiên cứu chọn làm đối chứng. Tuy nhiên, lý thuyết kinh tế thường không đưa ra hướng dẫn nào về những yếu tố nào có thể đóng vai trò kiểm soát. Việc bỏ qua các yếu tố kiểm soát nói chung sẽ làm sai lệch ước tính phần bù rủi ro. Lấy yếu tố thanh khoản và vốn trung gian làm ví dụ. Ước tính phần bù rủi ro cho thay đổi trước đây từ 226 bp mỗi tháng dựa trên hồi quy hai bước đơn biến thành 57 bp dựa trên cùng một phương pháp nhưng với các yếu tố FF3 làm đối chứng.

Tương tự, ước tính cho lần thay đổi sau từ 101bp thành 43bp.

Phương trình (4.5) cũng cho phép các yếu tố nhiễu ($\eta = 0$) và yếu (η nhỏ) như các trường hợp đặc biệt, thường sẽ làm sai lệch suy luận về phần bù rủi ro trong hồi quy hai bước (một hiện tượng được Kan và Zhang lần đầu tiên ghi lại, 1999). Ví dụ, bốn yếu tố từ Novy-Marx (2014) là ví dụ về các biến có vẻ dự đoán lợi nhuận trong hồi quy dự đoán tiêu chuẩn, nhưng mối liên hệ kinh tế của chúng với thị trường chứng khoán có vẻ yếu. Tuy nhiên, ba trong số bốn yếu tố này có phần bù rủi ro đáng kể dựa trên hồi quy hai lần với các yếu tố FF3. Các yếu tố vĩ mô (như PC hay tăng trưởng tiêu dùng) cũng yếu. Phương pháp tiếp cận ba bước giải quyết cả độ lệch biến bị bỏ qua và lỗi đo lường vì nó ước tính các yếu tố tiềm ẩn trong lần đầu tiên, sử dụng chúng làm các biện pháp kiểm soát trong hồi quy cắt ngang lần thứ hai và áp dụng một hồi quy chuỗi thời gian khác trong lần thứ ba để loại bỏ phép đo. lỗi. Nhờ tính chắc chắn này, các ước tính từ hai cột cuối của Bảng 4.1 có vẻ hợp lý hơn về mặt kinh tế.

4.2.3 Phần mở rộng PCA

Mặc dù PCA là phương pháp phổ biến nhất để phục hồi yếu tố nhưng vẫn có những lựa chọn thay thế với các tính năng độc đáo. Ví dụ, Giglio và cộng sự. (2021a) áp dụng việc hoàn thiện ma trận để ước tính mô hình nhân tố có thể đối phó với bảng lợi nhuận không cân bằng. Giả sử Ω là ma trận $N \times T$

4.2. Mô hình nhân tố vô điều kiện

85

Các yếu tố	Điều khiển hai lượt không có	Hai lượt với FF3	Hồi quy ba lần
thanh khoản	2,26 (0,90) 1,01	0,57 (0,68)	0,37 (0,16)
Interterm. Mù lười trai.	(0,45) -319,01 (255,73)	0,43 (0,45) -277,96	0,60 (0,31)
nhiệt độ NY.	(4,85) 56,85 (17,42)	(124,08) -3,33 (2,07)	-0,69 (13,90)
Nhiệt độ toàn cầu	-6,65 (937,73)	(7,11) -15,34	0,05 (0,21)
El Niño	(0,14) 84,90 (24,76)	882,89	0,41 (0,82)
Vết đen Mặt	-409,37 .93)	(405,40) -0,14 (0,05)	4,01 (35,63)
trời Tăng	-0,36 (14:30)	6 (13,57)	-0,01 (0,00)
trường IP Macro	(0,16)	23,91 (8,97)	3,26 (1,58)
PC1 Macro PC2	9h35	-31,24 (9,74)	-0,88 (1,27)
Macro PC3 Nhược	-5,94	0,07 (0,05)	-1,25 (1,51)
điểm. sự phát triển	0,26		0,00 (0,01)

Bảng 4.1: Hồi quy ba lần: Kết quả thực nghiệm

Lưu ý: Đối với mỗi yếu tố, bảng báo cáo ước tính phần bù rủi ro theo điểm phần trăm mỗi tháng bằng cách phương pháp khác nhau, với hạn chế là tỷ lệ beta bằng 0 bằng với tỷ lệ hóa đơn T được quan sát: hai phiên bản của mặt cắt ngang hai lần hồi quy, không sử dụng yếu tố kiểm soát trong mô hình và sử dụng ba mô hình Fama-Pháp các yếu tố tương ứng; công cụ ước tính ba lần sử dụng 7 yếu tố tiềm ẩn. Tài sản thử nghiệm bao gồm 647 danh mục đầu tư được sắp xếp theo các đặc điểm khác nhau (từ Bảng B1 của Giglio và Xiu, 2021).

với phần tử (i, t) bằng 1 khi và chỉ nếu R_{it} không bị thiếu. Các Thuật toán hoàn thiện ma trận giải quyết tối ưu lồi sau vấn đề:

$$X = \arg \min_{X} \|R - X\|_F^2 + \lambda X_n,$$

trong đó biểu thị tích Hadamard, $X_n = \prod_{i=1}^{N,T} \psi_i(X)$ với $\psi_i(X)$ là giá trị số ít lớn thứ i của X , và λ là giá trị điều chỉnh tham số. Bằng cách trừ phạt chuẩn 1 của các giá trị số ít của X , thuật toán cố gắng tìm giá trị gần đúng thứ hạng thấp của lợi nhuận ma trận R chỉ sử dụng các mục được quan sát của nó. X ước tính là hoàn thành ma trận của R có cấu trúc cấp thấp, từ đó các yếu tố tiềm ẩn và tải do đó có thể được phục hồi thông qua SVD.

Việc triển khai tiêu chuẩn của PCA áp dụng SVD cho ma trận trả về vượt quá giới hạn R^- . Khi làm như vậy, nó ước tính các yếu tố tiềm ẩn và beta chỉ từ thời điểm trả về thứ hai tập trung vào mẫu. Lettau và Pelger (2020b) chỉ ra rằng việc PCA chỉ dựa vào thông tin thời điểm thứ hai dẫn đến ước tính mô hình nhân tố không hiệu quả. Lý thuyết định giá tài sản ngụ ý một hạn chế giữa phương tiện tài sản và beta yếu tố (đặc biệt, thông qua phiên bản vô điều kiện của phương trình Euler (1.2)). Vì thế họ

lập luận rằng việc dựa nhiều hơn vào thông tin tải yếu tố có trong dữ liệu thời điểm trả về đầu tiên có thể cải thiện hiệu suất tổng thể của các công cụ ước tính PCA. Dựa trên hiểu biết sâu sắc này, họ phát triển "PCA bù rủi ro" (RP-PCA) công cụ ước tính, áp dụng PCA cho mô-đun thứ hai không tập trung để trả về, $T \lambda > 1^T t=1 R t R t+\lambda(T^T t=1 R t)(T^T t=1 R t)$, ở đâu đóng vai trò là tham số điều chỉnh. Connor và Korajczyk (1988) cũng sử dụng PCA không tập trung, nhưng tuân theo trường hợp $\lambda = 0$, trong khi PCA tiêu chuẩn tương ứng với $\lambda = 1$.

Lettau và Pelger (2020a) thiết lập lý thuyết tiệm cận của RP-PCA và chỉ ra rằng nó hiệu quả hơn PCA khi các yếu tố

phổ biến trong trường hợp không có sai sót về giá. Mặc dù cách tiếp cận PCA tiêu chuẩn có hiệu quả đối với sự tồn tại của sai số về giá, thuật ngữ sai số này có thể làm sai lệch công cụ ước tính RP-PCA ở chỗ lợi nhuận kỳ vọng không còn song song với hệ số tải. Chúng tôi phỏng đoán rằng độ lệch như vậy là không đáng kể về mặt tiệm cận nếu áp đặt hạn chế kinh tế của việc không cạnh tranh chênh lệch giá (4.2), trong trường hợp đó độ lớn của α đủ nhỏ để nó không làm sai lệch các ước tính của các yếu tố và tải của chúng một cách tiệm

cận. Giglio và cộng sự. (2021b) chỉ ra rằng sức mạnh của một yếu tố phụ thuộc vào về việc lựa chọn nội dung thử nghiệm. Ngay cả yếu tố thị trường cũng có thể trở nên yếu đi nếu tất cả tài sản thử nghiệm là danh mục đầu tư dài hạn không có tác động đến yếu tố này. Để giải quyết vấn đề yếu tố yếu này trong ước tính phần bù rủi ro, họ đề xuất quy trình chọn tài sản thử nghiệm dựa trên PCA được giám sát, như đã thảo luận trong Phần 3.5.2. Hơn nữa, quy trình này có thể được áp dụng để phát hiện các yếu tố còn thiếu trong mô hình hệ số chiết khấu ngẫu nhiên.

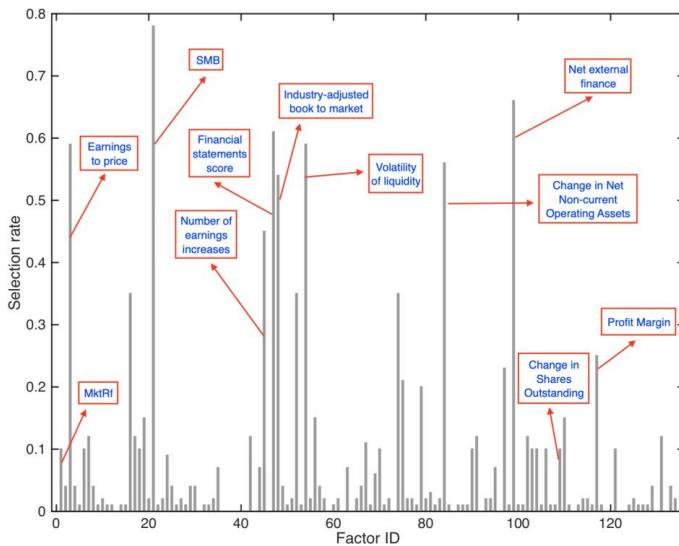
4.2.4 Những yếu tố nào?

Việc tìm kiếm các yếu tố giải thích mặt cắt ngang của lợi nhuận cổ phiếu kỳ vọng đã tạo ra hàng trăm ứng cử viên tiềm năng. Nhiều trong số đó là dư thừa, không có khả năng giải thích cho việc định giá tài sản do một số yếu tố khác. Một số hoàn toàn vô dụng và không có lời giải thích
quyền lực.

Phương pháp học máy có thể giải quyết thách thức về các yếu tố dư thừa và vô dụng thông qua việc giảm kích thước và lựa chọn biến.

²Bryzgalova và cộng sự. (2023) giới thiệu các mục tiêu có động cơ kinh tế khác để xác định các yếu tố định giá sau này.

Ví dụ: hồi quy Lasso của lợi nhuận trung bình dựa trên hiệp phương sai các yếu tố có thể giúp xác định một tập hợp các yếu tố chi tiết định giá mặt cắt ngang của tài sản. Đồng thời, sai sót trong lựa chọn là không thể tránh khỏi: việc trang bị quá mức có thể dẫn đến việc lựa chọn các biến vô ích; các biến có khả năng giải thích tương đối yếu có thể bị bỏ qua; các biến dư thừa có thể được chọn thay cho các biến thực. Hình 4.1 của Feng et al. (2020) cho thấy rằng các biến được chọn bởi lasso khác nhau đáng kể giữa các hạt ngẫu nhiên khác nhau được áp dụng trong xác thực chéo, khiến mẫu được chia ngẫu nhiên thành nhiều nếp gấp (xem Hình 3.1).



Hình 4.1: Tỷ lệ lựa chọn nhân tố

Ghi chú: Nguồn: Feng et al. (2020). Hình vẽ cho thấy, đối với mỗi yếu tố được xác định bởi ID yếu tố (trên trục X), tỷ lệ hạt giống ngẫu nhiên mà mỗi yếu tố được chọn bằng phương pháp hồi quy lasso của lợi nhuận trung bình đối với hiệp phương sai của yếu tố thông qua 200 lần xác thực chéo ngẫu nhiên.

Phong và cộng sự. (2020) đề xuất một phương pháp kết hợp khung học máy kép của Chernozhukov và cộng sự. (2018) với hồi quy cắt ngang hai bước để xác định một tập hợp các yếu tố mạch lạc và phát triển phân bố tiệm cận của công cụ ước tính của chúng, nhờ đó họ có thể đưa ra suy luận về các yếu tố.

Theo kinh nghiệm, Feng et al. (2020) áp dụng quy trình suy luận để quy định phân biệt các yếu tố hữu ích với các yếu tố vô ích và dư thừa khi chúng được giới thiệu trong tài liệu.

Những phát hiện thực nghiệm của họ cho thấy rằng nếu cách tiếp cận của họ được áp dụng hàng năm bắt đầu từ năm 1994, chỉ có 17 yếu tố trong số hơn 120 yếu tố ứng cử viên được coi là hữu ích, với phần lớn được xác định là dư thừa hoặc vô dụng.

Một tài liệu khác xem xét việc lựa chọn mô hình nhân tố từ góc độ không chắc chắn của mô hình và tính trung bình của mô hình, với công việc liên quan bao gồm Avramov et al. (2021) và Chib và cộng sự. (2023). Avramov và cộng sự. (2021) cho thấy rằng các quan điểm trước đây về mức độ lớn của tỷ lệ Sharpe có thể có ý nghĩa đối với việc đưa vào cả hai yếu tố và yếu tố dự đoán. Nhìn chung, sự không chắc chắn của mô hình Bayes là một chủ đề thú vị để phát triển hơn nữa trong lĩnh vực học máy tài chính.

4.3 Mô hình nhân tố có điều kiện

Phần trước tập trung vào phiên bản vô điều kiện của phương trình Euler (1.2) với beta tinh và giá rủi ro (được rút ra bằng cách thay thế λ bằng tập rỗng, hay "điều hòa giảm"). Nói chung, khi chúng ta thay đổi tập điều kiện, cách biểu diễn hệ số cũng thay đổi do thời điểm điều kiện của nội dung thay đổi. Có thể xảy ra trường hợp khi chúng ta điều chỉnh hệ số beta của tài sản và lợi nhuận kỳ vọng không thay đổi - trong trường hợp đặc biệt này, các mô hình có điều kiện và không điều kiện đều giống nhau.

Tuy nhiên, nghiên cứu thực nghiệm chứng minh rằng hiệp phương sai của tài sản có thể dự đoán được cao trong chuỗi thời gian và nhiều bằng chứng cho thấy rằng phương tiện tài sản cũng có thể dự đoán được. Nói cách khác, chúng ta có thể loại trừ trường hợp đặc biệt "tinh có điều kiện".

Điều kiện hay không điều kiện? Đó là câu hỏi khi xây dựng mô hình hệ số sinh lợi. Quan điểm của chúng tôi là, bất cứ khi nào có thể, nhà nghiên cứu nên mong muốn xây dựng một mô hình có điều kiện hiệu quả. Các mô hình có điều kiện đầy tham vọng - chúng mô tả tính chất phụ thuộc vào nhà nước của giá tài sản, và do đó nắm bắt được hành vi của thị trường ở độ phân giải tốt hơn.

Tuy nhiên, các mô hình có điều kiện cũng đòi hỏi khắt khe hơn, đòi hỏi nhà nghiên cứu phải cung cấp dữ liệu liên quan để tóm tắt các điều kiện phổ biến.

Thông tin điều kiện như vậy có thể mở rộng và có thể yêu cầu các mô hình được tham số hóa phong phú hơn để nắm bắt các đặc điểm có điều kiện.

hành vi cư xử. Khi không có thông tin điều kiện liên quan, một mô hình vô điều kiện cho phép nhà nghiên cứu hiểu được tài sản cơ bản

hành vi mà không nhất thiết phải hiểu động lực thị trường chi tiết và với các mô hình có thể đơn giản hơn. Vì lý do này, phần lớn tài liệu ban đầu về phân tích hệ số sinh lợi đều theo đuổi các thông số kỹ thuật vô điều kiện (như trong phần trước). Trong phần này chúng ta tập trung vào việc xây dựng mô hình có điều kiện.

Tương tự với (4.1), mô hình nhân tố tiềm ẩn có điều kiện được véc tơ hóa là

$$R_{t+1} = \alpha_t + \beta_t F_{t+1} + \epsilon_{t+1}, \quad (4.7)$$

trong đó các lỗi về hệ số và giá cả hiện thay đổi theo tập thông tin điều hòa I_t .

4.3.1 IPCA

Nếu không có các hạn chế bổ sung, về phái của (4.7) chứa quá nhiều bậc tự do và không thể xác định được mô hình. Phân tích các thành phần chính được thiết kế (IPCA) của Kelly và cộng sự. (2020b) đạt được tiến bộ bằng cách đưa ra các hạn chế liên kết beta (và alpha) của nội dung với các yếu tố có thể quan sát được. Mô hình IPCA có dạng

$$R_{t+1} = \underbrace{Z_t \Gamma \alpha}_{\alpha_t} + \underbrace{Z_t \Gamma \beta}_{\beta_t} F_{t+1} + \epsilon_{t+1}, \quad (4.8)$$

trong đó Z_t là ma trận $N \times L$ sắp xếp dữ liệu về các đặc điểm L có thể quan sát được (hoặc "công cụ") cho mỗi nội dung.³ F_{t+1} lại là vectơ $K \times 1$ của các yếu tố tiềm ẩn. Harvey và Ferson (1999), và gần đây là Gagliardini et al. (2016), cũng mô hình hóa hệ số tải như các hàm thay đổi theo thời gian của các đối tượng có thể quan sát được, nhưng các hệ số của chúng hoàn toàn có thể quan sát được.

Cốt lõi của mô hình IPCA là đặc điểm kỹ thuật của β_t . Đầu tiên và quan trọng nhất, các công cụ thay đổi theo thời gian kết hợp trực tiếp động lực học vào hệ số tải có điều kiện. Về cơ bản hơn, việc kết hợp các công cụ cho phép dữ liệu bổ sung để định hình mô hình nhân tố, khác với các kỹ thuật nhân tố tiềm ẩn vô điều kiện như PCA vốn chỉ ước tính cấu trúc nhân tố từ dữ liệu trả về. Và neo tải vào các vật có thể quan sát được

³Thông thường, một trong những đặc điểm của Z là một số hạng không đổi.

xác định mô hình bằng cách thay thế một phần các tham số không xác định bằng dữ liệu.

Mô trộn $L \times K \Gamma\beta$ xác định ánh xạ từ một số lượng lớn các đặc điểm (L) đến một số lượng nhỏ mức độ phơi nhiễm yếu tố rủi ro (K). Khi chúng tôi ước tính $\Gamma\beta$, chúng tôi đang tìm kiếm một số kết hợp tuyến tính của các đặc điểm ứng cử viên mô tả tốt nhất cấu trúc tài sản tự tiêm ẩn. Để đánh giá cao điểm này, trước tiên hãy tưởng tượng một trường hợp trong đó $\Gamma\beta$ là ma trận nhận dạng L chiều. Trong trường hợp này, dễ dàng chứng minh rằng Ft bao gồm L các yếu tố tiềm ẩn tỷ lệ thuận với nhau và các đặc tính xác định hệ số beta của các yếu tố này.

$R_{t+1|t}$,

Điều này gợi nhớ đến Rosenberg (1974) và MSCI Barra (một mô hình nhân tố rủi ro thường được các nhà thực hành sử dụng). Mô hình của Barra bao gồm hàng chục đặc điểm và chỉ số ngành trong Z_t . Khi số lượng đặc điểm công ty L lớn, số lượng tham số tự do bằng với số lượng hiện thực hóa hệ số $\{Ft\}$ hoặc $L \times T$, thường lớn so với cỡ mẫu. Có sự dư thừa đáng kể trong các yếu tố Barra (một số lượng nhỏ các thành phần chính nắm bắt hầu hết các biến thể chung của chúng), điều này cho thấy nó có thể được tham số hóa quá mức

và không hiệu quả.

IPCA giải quyết vấn đề này bằng cách giảm kích thước trong không gian đặc trưng. Nếu có nhiều đặc điểm cung cấp tín hiệu nhiều nhưng mang tính thông tin về mức độ rủi ro của cổ phiếu, thì việc tổng hợp các đặc điểm thành các kết hợp tuyến tính sẽ tách tín hiệu của chúng và tính trung bình nhiều của chúng.

Thách thức của việc di chuyển tài sản, chẳng hạn như cổ phiếu phát triển từ nhỏ đến lớn hoặc từ tăng trưởng sang giá trị, đặt ra vấn đề trong việc mô hình hóa lợi nhuận kỳ vọng có điều kiện ở cấp độ cổ phiếu bằng các phương pháp chuỗi thời gian đơn giản.

Giải pháp tiêu chuẩn là tạo ra các danh mục đầu tư có các giá trị đặc trưng trung bình ổn định theo thời gian, nhưng cách tiếp cận này trở nên không thực tế nếu cần nhiều đặc điểm để mô tả chính xác danh tính của một tài sản. Giải pháp IPCA tham số hóa beta với các đặc điểm xác định rủi ro và lợi nhuận của cổ phiếu. IPCA theo dõi quá trình di chuyển danh tính của nội dung thông qua các phiên bản beta của nó, lần lượt được xác định bởi các đặc điểm của nội dung. Điều này giúp nhà nghiên cứu không cần phải nhóm tài sản vào danh mục đầu tư theo cách thủ công vì mô hình theo dõi rõ ràng danh tính của tài sản theo đặc điểm của chúng. Như một

4.3. Mô hình nhân tố có điều kiện

91

Thống kê tài sản thử nghiệm		K				
		1	3	4	5	6
Bảng A: IPCA						
Cổ phiếu	Tổng R ²	14,9	17,6	19	18,2	18,7
	Dự đoán. R ²	0,36	0,43	0,70	0,43	0,70
	Np	636	1908	2544	3180	3816
Danh mục đầu tư	Tổng R ²	90,3	97,1	98,0	98,4	98,8
	Dự đoán. R ²	2,01	2,10	2,13	2,41	2,39
	Np	636	1908	2544	3180	3816
Phản B: Các yếu tố có thể quan sát được						
Cổ phiếu	Tổng R ²	11,9	18,9	20,9	21,9	23,7
	2 Dự đoán. R	0,31	0,29	0,28	0,29	0,23
	Np	11452	34356	45808	57268	68712
Danh mục đầu tư	Tổng R ²	65,6	85,1	87,5	86,4	88,6
	Dự đoán. R ²	1,67	2,07	1,98	2,06	1,96
	Np	37	111	148	185	222

Bảng 4.2: So sánh IPCA với các mô hình nhân tố khác

Lưu ý: Bảng báo cáo tổng và R dự đoán² theo phần trăm và số lượng ước tính tham số (Np) cho mô hình IPCA bị hạn chế ($\Gamma\alpha = 0$) (Bảng A) và cho mô hình có thể quan sát được mô hình hệ số với tải trọng tinh (Bảng B). Thông số mô hình nhân tố quan sát được lần lượt là CAPM, FF3, FFC4, FF5 và FFC6 trong các cột K = 1, 3, 4, 5, 6 (từ Bảng 2 của Kelly và cộng sự, 2020b).

Kết quả là, mô hình chứa một hệ thống tài sản có chiều cao (ví dụ: cổ phiếu riêng lẻ) mà không cần hình thành danh mục đầu tư đặc biệt.

Cuối cùng, đặc điểm IPCA trong (4.7) cũng cho phép khả năng những đặc điểm đó có thể ủy quyền cho alpha thay vì beta. Thông thường Các mô hình định giá tài sản giả định rằng sự khác biệt về lợi nhuận kỳ vọng giữa tài sản chỉ là do sự khác biệt trong mức độ rủi ro. Tuy nhiên, khi Vectơ hệ số L \times 1 $\Gamma\alpha$ khác 0, các đặc điểm cấp tồn kho có thể dự đoán lợi nhuận theo cách không phù hợp với sự đánh đổi rủi ro-lợi nhuận. IPCA giải quyết vấn đề này bằng cách ước tính alpha dưới dạng kết hợp tuyến tính các đặc điểm (được xác định bởi $\Gamma\alpha$) giải thích tốt nhất các điều kiện lợi nhuận kỳ vọng trong khi kiểm soát vai trò của các đặc tính trong yếu tố phơi nhiễm rủi ro. Nếu sự liên kết các đặc điểm với lượng hàng tồn kho trung bình lợi nhuận khác với sự liên kết của các đặc điểm với yếu tố rủi ro tài, IPCA sẽ ước tính $\Gamma\alpha$ khác 0, từ đó xác định việc định giá sai (tức là bồi thường cho việc nắm giữ tài sản không liên quan đến tài sản đó) rủi ro hệ thống).

Bảng 4.2 so sánh IPCA (Bảng A) với nhiều số la-

yếu tố lèu cho các mô hình hàng đầu khác trong văn học. Bộ mô hình so sánh đầu tiên bao gồm các yếu tố có thể quan sát được xác định trước, được ước tính bằng cách sử dụng phương pháp truyền thống là hồi quy chuỗi thời gian theo từng tài sản. Mô hình $K = 1$ là CAPM, $K = 3$ là mô hình ba yếu tố Fama-Pháp (1993) bao gồm thị trường, SMB và HML ("FF3" từ nay về sau), $K = 4$ là Carhart (1997, "FFC4") ") bổ sung MOM vào mô hình FF3, $K = 5$ là mô hình 5 yếu tố Fama-Pháp (2015, "FF5") bổ sung RMW và CMA vào các yếu tố FF3, $K = 6$ ("FFC6") bao gồm MOM bên cạnh các yếu tố FF5. Tất cả các mô hình trong Bảng 4.2 đều được ước tính với giới hạn điểm chặn bằng 0 bằng cách áp đặt $\Gamma\alpha = 0$ trong IPCA hoặc bằng cách bỏ qua điểm chặn trong hồi quy chuỗi thời gian.

Bảng 4.2 báo cáo tổng R2 dựa trên việc hiện thực hóa hệ số đương thời, R2 dự đoán (thay thế hiện thực hóa hệ số bằng phần bù rủi ro trung bình của nó), cũng như số lượng tham số ước tính (N_p) cho mỗi mô hình. Chúng tôi báo cáo sự phù hợp ở cấp độ cổ phiếu riêng lẻ và ở cấp độ danh mục đầu tư được quản lý theo đặc điểm. Các số liệu thống kê này được tính toán trong mẫu.4 Đối với từng cổ phiếu riêng lẻ, các mô hình nhân tố có thể quan sát được tạo ra tổng R2 cao hơn một chút so với IPCA. Tuy nhiên, để thực hiện được điều này, các yếu tố có thể quan sát được dựa vào nhiều thông số hơn IPCA.

Trong mẫu gồm 11.452 cổ phiếu với 37 công cụ trong 599 tháng này, các mô hình nhân tố quan sát được ước tính số lượng thông số nhiều hơn ($\approx 11452/(37 + 599)$) gấp 18 lần ($\approx 11452/(37 + 599)$) so với IPCA! Nói tóm lại, IPCA cung cấp một mô tả tương tự về rủi ro hệ thống trong lợi nhuận chứng khoán như những yếu tố hàng đầu có thể quan sát được trong khi sử dụng các tham số ít hơn gần 95%. Đồng thời, IPCA cung cấp mô tả chính xác hơn đáng kể về mức đền bù rủi ro của cổ phiếu so với các mô hình nhân tố có thể quan sát được, được chứng minh bằng R2 dự đoán. Đối với danh mục đầu tư được quản lý theo đặc điểm, tổng R2 và R2 dự đoán của mô hình nhân tố có thể quan sát đều bị ảnh hưởng so với IPCA.

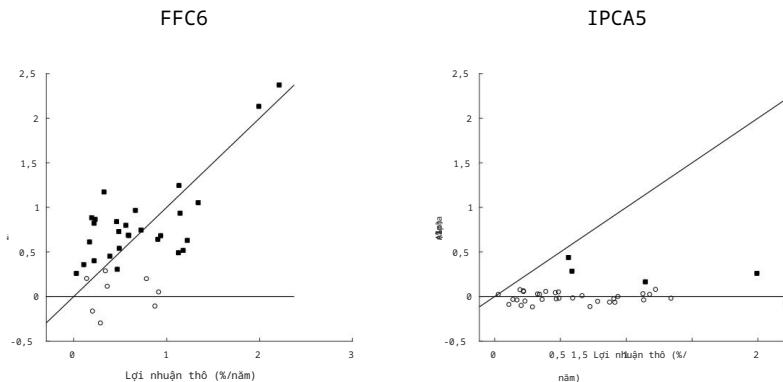
Hình 4.2 so sánh các mô hình về sai số định giá trung bình của 37 danh mục đầu tư "bất thường" được quản lý theo đặc tính.5 Biểu đồ bên trái hiển thị bảng chữ cái của danh mục đầu tư từ mô hình FFC6 so với mức trung bình thô của chúng

⁴Kelly và cộng sự. (2020b) cho thấy IPCA có độ bền ngoài mẫu rất cao, trong khi các mô hình khác như những mô hình sử dụng các yếu tố có thể quan sát được hoặc PCA có xu hướng bị suy giảm ngoài mẫu nghiêm trọng hơn.

⁵Để so sánh, danh mục đầu tư được ký kết lại và chia tỷ lệ để có phương tiện tích cực và độ biến động hàng năm là 10%.

4.3. Mô hình nhân tố có điều kiện

93



Hình 4.2: Bảng chữ cái của danh mục đầu tư được quản lý theo đặc điểm

Lưu ý: Bảng bên trái và bên phải báo cáo bảng chữ cái cho danh mục đầu tư được quản lý theo đặc điểm liên quan đến mô hình FFC6 và mô hình 5 yếu tố IPCA. Alpha được vẽ dựa trên lợi nhuận vượt trội trung bình thô của danh mục đầu tư. Các bảng chữ cái có thống kê t vượt quá 2,0 được hiển thị bằng các ô vuông được lấp đầy, trong khi các bảng chữ cái không đáng kể được hiển thị bằng các vòng tròn không được lấp đầy (từ Hình 1 của Kelly và cộng sự, 2020b).

lợi nhuận vượt quá, phủ lên đường 45 độ. Các bảng chữ cái có thống kê t vượt quá 2,0 được mô tả bằng các ô vuông được lấp đầy, trong khi các bảng chữ cái không đáng kể được hiển thị bằng các vòng tròn không được lấp đầy. 29 danh mục đầu tư được quản lý theo đặc điểm có mức alpha đáng kể so với mô hình FFC6. Các bảng chữ cái tập trung xung quanh đường 45 độ, cho thấy rằng lợi nhuận trung bình của chúng về cơ bản không thể giải thích được bằng các yếu tố có thể quan sát được. Biểu đồ bên phải hiển thị mức trung bình chuỗi thời gian của các bảng chữ cái có điều kiện từ đặc tả IPCA năm yếu tố. Bốn danh mục đầu tư có hệ số alpha có điều kiện khác biệt đáng kể so với 0 nhưng có quy mô nhỏ về mặt kinh tế. Hình 4.2 ủng hộ kết luận rằng việc xây dựng hệ số tiềm ẩn của IPCA thành công hơn trong việc định giá một loạt danh mục đầu tư vốn cổ phần so với các mô hình định giá tài sản với các hệ số có thể quan sát được.

Khung IPCA đã được sử dụng để nghiên cứu việc định giá tài sản chéo ở nhiều thị trường khác nhau, bao gồm cả phiếu quốc tế (Langlois, 2021; Windmueller, 2022), trái phiếu doanh nghiệp (Kelly và cộng sự, Forthcoming), quyền chọn chỉ số vốn cổ phần (Büchner và Kelly, 2022), quyền chọn vốn sở hữu một tên (Goyal và Saretto, 2022) và tiền tệ (Bybee et al., 2023). Nó cũng được sử dụng để hiểu lợi nhuận của các tín hiệu xu hướng giá

(Kelly và cộng sự, 2021) và nền tảng tường thuật của các mô hình định giá tài sản (Bybee và cộng sự, 2023b).

4.4 Mô hình nhân tố phức tạp

Một số bài báo nghiên cứu khái quát hóa đặc tả (4.8) cho các phiên bản beta được thiết bị trong các mô hình nhân tố điều kiện tiềm ẩn. IPCA có thể được xem như một phép tính gần đúng tuyến tính đối với mức độ rủi ro dựa trên dữ liệu đặc điểm có thể quan sát được. Trong khi nhiều mô hình định giá tài sản dự đoán mối liên hệ phi tuyến tính giữa lợi nhuận kỳ vọng và các biến trạng thái, thì tài liệu lý thuyết lại đưa ra rất ít hướng dẫn để sàng lọc danh sách các biến điều kiện và dạng hàm số. Sự ra đời của học máy cho phép chúng ta giải quyết sự mơ hồ về dạng hàm này bằng một loạt các mô hình phi tuyến.

Trong công việc ban đầu, Connor et al. (2012) và Fan và cộng sự. (2016b) cho phép đặc tả beta phi tuyến tính bằng cách coi beta là hàm phi tham số của các đặc tính điều hòa (nhưng, không giống như IPCA, đặc tính này được cố định theo thời gian để có khả năng điều khiển). Kim và cộng sự. (2020) áp dụng khuôn khổ này để nghiên cứu hành vi của các danh mục đầu tư "chênh lệch giá" nhằm phòng ngừa rủi ro nhân tố.

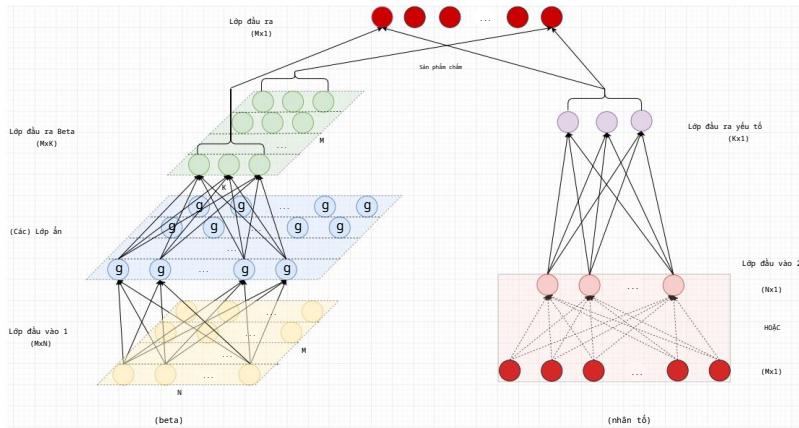
Gu và cộng sự. (2020a) mở rộng mô hình IPCA bằng cách cho phép beta trở thành một hàm mạng thần kinh có các đặc điểm. Hình 4.3 sơ đồ mô hình "bộ mã hóa tự động có điều kiện" (CA) của họ. Hình 4.3 minh họa cấu trúc cơ bản của nó, khác với (4.8) bằng cách truyền dữ liệu đầu vào (công cụ Zt) thông qua các hàm kích hoạt phi tuyến. CA là mô hình học sâu đầu tiên về lợi nhuận vốn chủ sở hữu giải thích rõ ràng sự cân bằng giữa rủi ro và lợi nhuận. Gu và cộng sự. (2020a) cho thấy rằng tổng R2 của CA tương tự như của IPCA, nhưng nó cải thiện đáng kể so với IPCA về mặt R2 dự đoán. Nói cách khác, CA cung cấp mô tả chính xác hơn về khả năng bù đắp có điều kiện của tài sản đối với rủi ro yếu tố.

Gu và cộng sự. (2020a) là một mô hình có độ phức tạp cao và hiệu suất thực nghiệm mạnh mẽ của nó gợi ý về lợi ích của độ phức tạp trong các mô hình nhân tố tương tự như mô hình được nghiên cứu trong các mô hình dự đoán của Kelly và cộng sự. (2022a).

Didisheim và cộng sự. (2023) chính thức hóa ý tưởng này và chứng minh tính phức tạp trong việc định giá theo yếu tố. Họ xây dựng phân tích của mình xung quanh hệ số chiết khấu ngẫu nhiên có điều kiện (SDF), hệ số này thường có thể được viết dưới dạng danh mục đầu tư.

4.4. Mô hình nhân tố phức tạp

95



Hình 4.3: Mô hình bộ mã hóa tự động có điều kiện

Lưu ý: Hình này trình bày sơ đồ của mô hình bộ mã hóa tự động có điều kiện, trong mà bộ mã hóa tự động được tăng cường để kết hợp các hiệp phương sai trong quá trình tải hệ số sự chỉ rõ. Phía bên trái mô tả cách hệ số tải (màu xanh lá cây) phụ thuộc vào đặc điểm chắc chắn (màu vàng) của lớp đầu vào 1 thông qua hàm kích hoạt g trên các nơ-ron của lớp ẩn. Mỗi hàng nơ-ron màu vàng đại diện cho vector đặc điểm của một mã chứng khoán. Phía bên phải mô tả tương ứng các nhân tố. Các nút (màu tím) là sự kết hợp có trọng số của các nơ-ron của lớp đầu vào 2, có thể là danh mục đầu tư được quản lý theo đặc điểm (màu hồng) hoặc tài sản riêng lẻ trả về (màu đỏ). Trong trường hợp trước, chúng tôi tiếp tục sử dụng mũi tên đứt để chỉ ra rằng danh mục đầu tư được quản lý theo đặc điểm dựa vào các tài sản riêng lẻ thông qua các danh mục được xác định trước. trọng lượng (không được ước tính). Trong cả hai trường hợp, đầu vào hiệu quả có thể được coi là lợi nhuận của tài sản riêng lẻ, chính xác là những gì lớp đầu ra (màu đỏ) hướng tới gần đúng, do đó mô hình này có chung tính thắn như một bộ mã hóa tự động tiêu chuẩn. Nguồn: Gu và cộng sự. (2020a).

tài sản rủi ro:

$$Mt+1 = 1 \quad w(Xt) \quad Rt+1, \quad (4.9)$$

trong đó $Rt+1$ là vector lợi nhuận vượt trội trên N tài sản. Vector N $w(Xt)$ chứa các trọng số danh mục đầu tư có điều kiện của SDF, với Xt biểu diễn các biến điều hòa kéo dài trong tập thông tin t .

Nếu không biết về dạng hàm của nó, $w()$ có thể xấp xỉ bằng mô hình học máy:

$$w(Xt) \approx \sum_{p=1}^P \lambda_p s_p(Xt)$$

trong đó $Sp(Xt)$ là một hàm cơ sở phi tuyến nào đó của Xt và số lượng tham số P trong phép tính gần đúng là lớn. Mô hình học máy của SDF có thể được hiểu là mô hình định giá nhân tố với các yếu tố P :

$$Mt+1 \approx 1 \quad \underset{P}{\lambda p Fp, t+1}, \quad (4.10)$$

trong đó mỗi "yếu tố" $Fp, t+1$ là một danh mục tài sản rủi ro được quản lý bằng cách sử dụng các "đặc điểm" tài sản phi tuyến $Sp(Xt)$ làm trọng số. Kết quả lý thuyết chính của Didisheim et al. (2023) cho thấy rằng càng có nhiều yếu tố trong mô hình định giá tài sản thì càng tốt. Trong thiết lập này, việc thêm các yếu tố có nghĩa là sử dụng cách trình bày thông tin phong phú hơn có trong Xt để đạt được xấp xỉ tốt hơn về SDF thực sự. Sự cải thiện về độ chính xác gần đúng giúp giảm chi phí thông kê khi phải ước tính nhiều tham số. Kết quả là, hệ số alpha ngoài mẫu dự kiến sẽ giảm khi số lượng yếu tố tăng lên. Việc giải thích ưu điểm của sự phức tạp này là một thách thức đối với quan điểm APT truyền thống cho rằng một số lượng nhỏ các yếu tố rủi ro cung cấp mô tả đầy đủ về sự đánh đổi giữa rủi ro và lợi nhuận đối với bất kỳ tài sản có thể giao dịch nào. Điều này có hàm ý rằng, ngay cả khi không có hoạt động kinh doanh chênh lệch giá và SDF tồn tại, thì thực tế là SDF phải được ước tính ngụ ý rằng có thể (trên thực tế là dự kiến) liên tục tìm ra các yếu tố "rủi ro" thực nghiệm mới mà người khác không định giá được và rằng việc thêm các yếu tố này vào mô hình định giá sẽ liên tục cải thiện hiệu suất ngoài mẫu của nó.

4.5 Mô hình tần số cao

Sự săn có ngày càng tăng của dữ liệu cấp độ giao dịch tần số cao về mặt cắt ngang ngày càng tăng của các tài sản có thể giao dịch mang lại cơ hội duy nhất trong việc ước tính rủi ro của từng tài sản và sự phụ thuộc lẫn nhau của chúng. Các thước đo phi tham số đơn giản về độ biến động và hiệp phương sai (Andersen và Bollerslev, 1998; Andersen và cộng sự, 2001; Barndorff-Nielsen và Shephard, 2002) đưa ra minh chứng ban đầu về cách khai thác dữ liệu giá cả phong phú và kịp thời trong ngày để hiểu rõ hơn về biến động của thị trường tài sản. Việc sử dụng các biện pháp tần số cao giúp giải quyết một số thách thức khi nghiên cứu chuỗi thời gian tần số thấp. Ví dụ, nó giúp nhà nghiên cứu điều chỉnh các phá vỡ cấu trúc và các tham số thay đổi theo thời gian

4.5. Mô hình tần số cao

97

với những giả định tối thiểu. Hơn nữa, nhiều giả định tiêu chuẩn về tính tuyến tính, tính dừng, sự phụ thuộc và tính không đồng nhất trong chuỗi thời gian có diễn thường không cần thiết cho việc mô hình hóa dữ liệu trong ngày.

Chúng tôi xác định hai luồng tài liệu gần đây áp dụng kỹ thuật học máy để ước tính hiệp phương sai nhiều chiều và cải thiện khả năng dự báo biến động với dữ liệu tần số cao.

Ước tính hiệp phương sai chính xác là rất quan trọng để xây dựng danh mục đầu tư thành công. Nhưng việc ước tính các ma trận hiệp phương sai lớn là một vấn đề thống kê đầy thách thức do lỗi nguyên của chiều. Một số phương pháp dựa vào các hình thức chính quy hóa khác nhau (Bickel và Levina, 2008a; Bickel và Levina, 2008b; Cai và Liu, 2011; Ledoit và Wolf, 2012; Ledoit và Wolf, 2004) để cải thiện các ước tính. Lấy cảm hứng từ APT, Fan et al. (2008) để xuất các công cụ ước tính ma trận hiệp phương sai dựa trên mô hình nhân tố trong trường hợp mô hình nhân tố nghiêm ngặt với các yếu tố có thể quan sát được, và Fan et al. (2013) đưa ra một cách tiếp cận với hệ số gần đúng

cấu trúc với các yếu tố tiềm ẩn.

Cấu trúc nhân tố cũng cần thiết ở tần số cao khi kích thước của bảng điều khiển đạt đến kích thước mẫu. Tuy nhiên, các kỹ thuật kinh tế lượng về cơ bản là khác nhau trong môi trường lấy mẫu tần số thấp và tần số cao. Cái sau thường được áp dụng trong cài đặt thời gian liên tục dựa trên mô hình bán tổng hợp thời gian liên tục chung, cho phép biến đổi ngẫu nhiên và nhảy trong động lực quay trở lại. Ait-Sahalia và Xiu (2019) phát triển lý thuyết tiệm cận của PCA phi tham số để xử lý dữ liệu trong ngày, mở đường cho việc ứng dụng mô hình nhân tố trong thời gian liên tục. Ngoài ra, Fan và cộng sự. (2016a) và Ait-Sahalia và Xiu (2017) phát triển các công cụ ước tính ma trận hiệp phương sai lớn sử dụng dữ liệu tần số cao cho từng cổ phiếu trên cơ sở mô hình hệ số thời gian liên tục.

Một chương trình nghị sự đầy hứa hẹn là kết hợp tài liệu về đo lường rủi ro tần suất cao với tài liệu về mặt cắt ngang của lợi nhuận kỳ vọng, tận dụng thông tin rủi ro phong phú hơn để hiểu rõ hơn về sự cân bằng giữa rủi ro và lợi nhuận. Một số nghiên cứu có liên quan theo hướng này bao gồm Bollerslev et al. (2016), người tính toán beta cổ phiếu riêng lẻ liên quan đến các thành phần liên tục và tăng vọt của một yếu tố thị trường duy nhất trong bối cảnh thời gian liên tục, nhưng liên kết các ước tính này với mặt cắt ngang của lợi nhuận trong thiết lập thời gian rời rạc. Ait-Sahalia và cộng sự. (2021)

cung cấp suy luận về phần bù rủi ro trong một khung thời gian liên tục thống nhất, đồng thời cho phép nhiều yếu tố và beta ngẫu nhiên trong giai đoạn đầu tiên, đồng thời coi beta trong lần thứ hai là các thành phần được ước tính trong lần đầu tiên, khái quát hóa phương pháp suy luận cổ điển của Shanken (1992a). Về mặt thực nghiệm, họ kiểm tra các mô hình nhân tố của lợi nhuận trong ngày bằng cách sử dụng Fama- French và các nhân tố động lượng được lấy mẫu cứ sau 15 phút do Ait-Sahalia et al xây dựng. (2020).

Ý tưởng đo lường mức độ biến động bằng cách sử dụng dữ liệu tần số cao cũng thúc đẩy một chương trình đầy hứa hẹn trong việc dự báo mức độ biến động. Mô hình tự hồi quy không đồng nhất (HAR) của các thước đo biến động được thực hiện trong quá khứ (Corsi, 2009) đã nổi lên như mô hình dự báo biến động hàng đầu trong nghiên cứu học thuật và thực tiễn ngành. Một số bài báo gần đây đã xem xét các chiến lược học máy để dự báo biến động, bao gồm Li và Tang (2022) và Bollerslev et al. (2022). Nhưng không giống như phân tích dự đoán lợi nhuận trong đó dự đoán học máy trực tiếp chuyển thành tỷ lệ Sharpe cao hơn, không rõ mức độ mà dự báo học máy vượt trội hơn các mô hình HAR hiện tại về mặt kinh tế. Đây là một câu hỏi mở thú vị trong văn học.

4.6 Chữ cái

Phản này thảo luận về tài liệu về thử nghiệm alpha và học máy. Alpha là phần lợi nhuận kỳ vọng chưa được tính đến bởi yếu tố beta và do đó là một đối tượng phụ thuộc vào mô hình. Bởi vì lý thuyết kinh tế thường quá cách điệu để xác định danh tính của tất cả các yếu tố và vì dữ liệu có thể không đủ phong phú để suy ra các yếu tố thực sự theo cách dựa trên dữ liệu, nên việc xác định sai mô hình là thách thức kéo dài trong việc phân biệt alpha với dồn bù “công bằng”. đối với mức độ rủi ro của yếu tố.

Ví dụ, có thể các hệ số alpha ước tính là biểu hiện của hệ số beta yếu tố yếu, gợi nhớ đến bài toán biến bị bỏ sót trong hồi quy. Nói cách khác, alpha của người này chính là beta của người khác. Trong mô hình nhân tố tiềm ẩn, alpha và beta cuối cùng được phân biệt bằng điểm giới hạn cường độ nhân tố giúp phân biệt các yếu tố với nhiều đặc ứng.

Chúng tôi tập trung phân tích alpha từ góc độ của các mô hình nhân tố tiềm ẩn vô điều kiện. Sự nhấn mạnh của chúng tôi vào vô điều kiện thay vì

hơn bảng chữ cái có điều kiện được thúc đẩy bởi trọng tâm trong tài liệu.

Sự nhấn mạnh của chúng tôi vào các mô hình nhân tố tiềm ẩn được thúc đẩy bởi quan điểm của chúng tôi rằng những lo ngại về việc xác định sai sẽ ít nghiêm trọng hơn đối với các mô hình nhân tố tiềm ẩn.

4.6.1 Thủ nghiệm Alpha và tầm quan trọng kinh tế

Trọng tâm lâu dài của việc định giá tài sản theo kinh nghiệm là giả thuyết không cho rằng tất cả các giá trị alpha đều bằng 0, $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_N = 0$. Đây là một giả thuyết duy nhất khác với vấn đề thử nghiệm alpha nhiều giả thuyết mà chúng ta sẽ thảo luận sau. Việc bác bỏ H_0 được hiểu là bằng chứng của việc xác định sai mô hình định giá tài sản hoặc định giá sai các tài sản thử nghiệm (và có lẽ do nhầm lẫn là vi phạm APT của Ross (1976)).

Kiểm định GRS nổi tiếng (Gibbons và cộng sự, 1989) của H_0 là kiểm định Chi bình phương được thiết kế cho các mô hình nhân tố có chiều thấp với các nhân tố có thể quan sát được. Fan và cộng sự. (2015) và Pesaran và Yamagata (2017) đề xuất các thử nghiệm có cùng giá trị rỗng nhưng ở cài đặt nhiều chiều. Đây là một bước tiến quan trọng vì nó loại bỏ hạn chế ($T > N + K$) trong thử nghiệm GRS ban đầu và cải thiện sức mạnh của thử nghiệm khi N lớn.

Mặc dù các phương pháp này ban đầu được đề xuất cho các mô hình có các yếu tố có thể quan sát được nhưng có thể mở rộng chúng sang các mô hình yếu tố tiềm ẩn. Trên thực tế, Giglio và cộng sự. (2021a) xây dựng công cụ ước tính của α bằng cách sử dụng hệ số tải ước tính và phần bù rủi ro tương ứng được đưa ra bởi (4.4) và (4.6) :

$$\alpha = \frac{1}{T} \sum_{t=1}^T \beta y_t$$

Họ cũng rút ra sự khai triển tiềm cận cần thiết của α , mở đường cho việc xây dựng các thử nghiệm về alpha trong các mô hình nhân tố tiềm ẩn.

Thống kê kiểm tra GRS được xây dựng dựa trên $(S^2)^{-1} = \alpha \sum \alpha$, có thể được hiểu là tỷ lệ Sharpe bình phương tối ưu của một danh mục đầu tư không có sự tiếp xúc với các yếu tố. Ước tính tỷ lệ Sharpe này là một chuyện, thực hiện chiến lược giao dịch để hiện thực hóa nó lại là một chuyện khác. Nghĩa là, việc bác bỏ giả thuyết zero alpha trong một thử nghiệm giống GRS không nhất thiết có nghĩa là việc bác bỏ đó là quan trọng về mặt kinh tế. Bất kỳ sự định lượng có ý nghĩa nào về tầm quan trọng kinh tế của bảng chữ cái đều phải xem xét tính khả thi của chiến lược giao dịch của một nhà kinh doanh chênh lệch giá. Đánh giá các từ chối thống kê về mặt kinh tế vừa có giá trị hơn đối với tài sản

nghiên cứu về giá cả và phù hợp hơn với những người thực

hành.6 APT giả định rằng các nhà kinh doanh chênh lệch giá biết các thông số thực sự trong quá trình tạo ra lợi nhuận. Một giả định như vậy có thể là tính với điều kiện là cỡ mẫu đủ lớn, trong trường hợp đó các tham số được bộc lộ một cách tiệm cận và các nhà kinh doanh chênh lệch giá hành xử (xấp xỉ) như thể họ biết các tham số thực. Điều đáng chú ý là, trong cài đặt APT, các nhà kinh doanh chênh lệch giá phải biết số lượng bảng chữ cái ngày càng tăng, do đó kích thước mặt cắt ngang lớn so với kích thước mẫu thông thường. Do đó, thật vô lý khi cho rằng các nhà kinh doanh chênh lệch giá có thể học được bảng chữ cái ngay cả trong giới hạn T lớn .

Đà và cộng sự. (2022) xem lại APT và nới lỏng giả định về các tham số đã biết. Trong bối cảnh của họ, các nhà kinh doanh chênh lệch giá phải sử dụng chiến lược giao dịch khả thi dựa trên dữ liệu lịch sử với cỡ mẫu T . Đối với bất kỳ chiến lược khả thi nào tại thời điểm t , họ xác định tỷ lệ Sharpe có điều kiện trong giai đoạn tiếp theo của chiến lược này là:

$$S(w) := E(w_{Rt+1|It}) / \text{Var}(w_{Rt+1|It})^{1/2},$$

trong đó $\hat{\sigma}$ là thông tin được đặt tại t . Chúng chứng tỏ rằng $S(w)$ tuân theo

$$S(w) \leq S(G)^2 + \gamma \sum_v \gamma^{1/2} + oP(1), \quad (4.12)$$

trong đó $S(G)^2 := E(\alpha|G) \Sigma \text{ma}^1 E(\alpha|G)$, với $N = \infty$, Σ_v là hiệp phương sai-trậnance của các thừa số và G là tập thông tin được sinh ra bởi $\{(R_s, \beta, V_s, \Sigma) : t \leq T + 1 \text{ s} \leq t\}$.

Đáng chú ý, $\gamma \sum_v \gamma^{1/2}$ là tỷ số Sharpe của danh mục nhân tố tối ưu. Do đó, đối với bất kỳ chiến lược trung hòa yếu tố nào w , tức là $w \beta = 0$,

$$S(w) \leq S(G) + oP(1). \quad (4.13)$$

Kết quả này gợi ý rằng chính ước tính hậu nghiệm của α sẽ xác định tỷ lệ Sharpe khả thi tối ưu và áp đặt giới hạn trên cho

⁶Như Shanken (1992b) trình bày

... nội dung thực tế được đưa ra cho khái niệm 'kinh doanh chênh lệch giá gần đúng', bằng cách mô tả các cơ hội đầu tư sẵn có do độ lệch lợi nhuận kỳ vọng quan sát được ... Tôi tin rằng sẽ học được nhiều hơn nữa bằng cách xem xét mức độ mà chúng ta có thể ước chừng một sự chênh lệch giá với các tài sản hiện có.

lợi nhuận của kinh doanh chênh lệch giá thống kê. Bất kỳ chiến lược học máy nào, đơn giản hay phức tạp, đều cần tuân theo giới hạn tỷ lệ Sharpe khả thi này.

Nói chung, $E(S(G)) \leq E^2(S)$ trong đó $E^2(S)$ là kết quả của chiến lược khả thi và S có thể được coi là tỷ lệ Sharpe tối ưu không khả thi.

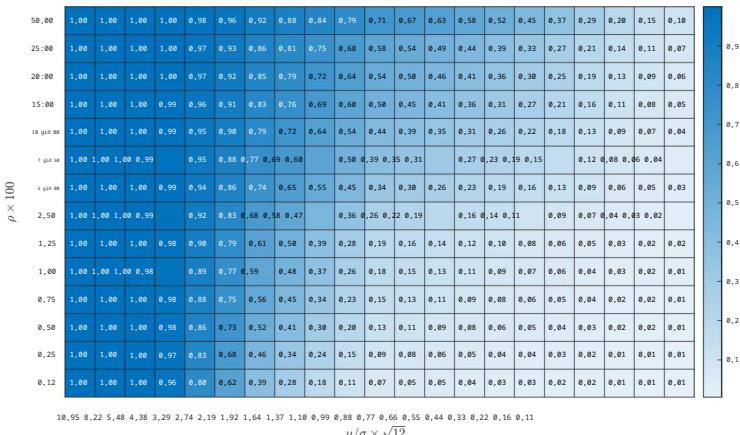
"Khoảng cách tỷ lệ Sharpe" giữa các chiến lược khả thi và không khả thi mô tả sự khó khăn của việc học thống kê. Nếu học khó thì chênh lệch lớn. Hình 4.4 trình bày tỷ lệ giũa hai tỷ lệ Sharpe này bằng số cho một quá trình tạo ra lợi nhuận giả định trong đó α nhận giá trị bằng 0 với xác suất $1 - \rho$ và một giá trị μ hoặc μ mỗi giá trị có xác suất $\rho/2$. Phần dư có ma trận hiệp phương sai Σ

là một đường chéo có phương sai σ đặc². Trong lớp DGP này, μ / σ trưng cho cường độ của α (so với nhiều), trong khi ρ đặc trưng cho độ hiém của nó. Việc thay đổi μ / σ và ρ giúp tìm ra vai trò của các quá trình tạo ra khác nhau trong hoạt động kinh doanh chênh lệch giá, được thể hiện trong Hình 4.4. Khi μ / σ tăng lên, alpha đủ mạnh và dễ học, đồng thời khoảng cách tỷ lệ Sharpe nhỏ hơn. Độ hiém của alpha đóng vai trò ít nổi bật hơn, mặc dù alpha phổ biến hơn cũng dẫn đến khoảng cách nhỏ hơn.

Đà và cộng sự. (2022) chỉ ra cách định lượng khoảng cách giữa tỷ lệ Sharpe không khả thi và khả thi. Họ đánh giá APT trên cơ sở mô hình 27 yếu tố, trong đó 16 đặc điểm và 11 biến giả ngành GICS được sử dụng làm phiên bản beta có thể quan sát được. Tỷ lệ Sharpe không khả thi ước tính là trên 2,5 đối với một mẫu thử nghiệm từ tháng 1 năm 1975 đến tháng 12 năm 2020 và con số này cao gấp bốn lần so với tỷ lệ Sharpe khả thi là khoảng 0,5 mà các chiến lược học máy đạt được. Thực tế là tỷ lệ Sharpe khả thi (trước chi phí giao dịch) dưới 0,5 cho thấy APT trên thực tế hoạt động khá tốt về mặt thực nghiệm. Về lý thuyết, khoảng cách giũa tỷ lệ Sharpe khả thi và không khả thi sẽ tăng thêm nếu các nhà kinh doanh chênh lệch giá gặp nhiều trở ngại về mặt thống kê hơn, chẳng hạn như mô hình xác định sai, ma trận hiệp phương sai dư không thua thớt, v.v.

4.6.2 Nhiều thử nghiệm

Kể từ khi CAPM ra đời, cộng đồng kinh tế tài chính đã cùng nhau tìm kiếm những "điểm bất thường"; tức là danh mục đầu tư có alpha đến CAPM. Một số trong số này, như kích thước, giá trị và một số thứ khác, có

Hình 4.4: Tỷ số giữa $S(G)$ và S

Lưu ý: Hình vẽ trình bày tỷ lệ Sharpe tối ưu giữa danh mục đầu tư chênh lệch giá khả thi và không khả thi. Cài đặt mô phỏng dựa trên một mô hình đơn giản, trong đó chỉ 100 × p% tài sản có bằng chữ cái khác 0, với mỗi mục nhập tương ứng với tỷ lệ Sharpe hàng năm $\mu / \sigma \times \sqrt{12}$. Nguồn: Da et al. (2022).

được tiếp thu vào các mô hình chuẩn (Fama và French, 1993; Fama và French, 2015). Sau đó, những bất thường mới được đề xuất khi các nhà nghiên cứu phát hiện ra một alpha so với chuẩn mục hiện hành. Harvey và cộng sự. (2016) khảo sát tài liệu và đối chiếu danh sách hơn ba trăm điểm bất thường được ghi nhận. Họ đưa ra những lời chỉ trích quan trọng rằng nỗ lực tìm kiếm sự bất thường đã không giải thích chính xác việc kiểm tra nhiều giả thuyết khi đánh giá tầm quan trọng của những bất thường mới.⁷ Thử nghiệm nhiều lần trong bối cảnh này để cập nhật việc kiểm tra đồng thời: $a_i = 0$, với $i = 1, 2, \dots, N$. Cái này

một tập hợp các giả thuyết không: Bài 0

toán Hi về cơ bản khác với việc kiểm tra giả thuyết không duy nhất $H_0 : a_1 = a_2 = \dots = a_N = 0$ đã thảo luận trước đó. Nhiều thử nghiệm dễ xảy ra sự cố phát hiện sai vì một phần của các thử nghiệm alpha riêng lẻ chắc chắn sẽ xuất hiện đáng kể chỉ do tình cờ và do đó

⁷Nói rộng hơn, Harvey (2017) trong số những người khác nhấn mạnh xu hướng các nhà nghiên cứu tài chính đưa ra kết luận sai sót từ các phân tích của họ bằng cách không tính đến các thử nghiệm không được báo cáo, không tính đến nhiều thử nghiệm và khuyến khích nghề nghiệp thúc đẩy "p-hack".

giả thuyết vô hiệu của họ bị bác bỏ một cách không chính xác.

Giả sử ti là một thông kê kiểm tra cho giá trị null H_0 . Giả sử H_1 từ chối bất cứ khi nào $|t_i| > c_i$ đối với một số giá trị tối hạn được xác định trước ci. Đặt $H_0 \{1, \dots, N\}$ biểu thị tập hợp các chỉ số mà các giả thuyết không tương ứng là đúng. Ngoài ra, gọi R là tổng số lần từ chối trong một mẫu và gọi F là số lần từ chối sai trong mẫu đó:

$$F = \sum_{t \leq i=1}^{N} 1\{i \leq N : |t_i| > c_i \text{ và } i \in H_0\}, R = \sum_{t \leq i=1}^{N} 1\{i \leq N : |t_i| > c_i\}.$$

Cả F và R đều là các biến ngẫu nhiên, nhưng R có thể quan sát được còn F thì không.

Đối với bất kỳ mức xác định trước nào τ ($0, 1$), giả sử là 5%, các thử nghiệm riêng lẻ đảm bảo rằng tỷ lệ lỗi trên mỗi thử nghiệm được giới hạn dưới đây bởi τ : $E(F)/N \leq \tau$. Nói cách khác, số lượng từ chối sai dự kiến có thể lớn bằng $N\tau$. Để hạn chế số lượng từ chối sai, một đề xuất thay thế là chọn giá trị tối hạn lớn hơn để kiểm soát tỷ lệ lỗi thông minh theo dòng (FWER): $P(F \geq 1) \leq \tau$. Thật không may, đề xuất thứ hai lại quá bảo thủ trong thực tế. Đề xuất thứ ba, có từ thời Stewamini và Hochberg (1995), là kiểm soát trực tiếp tỷ lệ phát hiện sai: $FDR \leq \tau$ trong đó tỷ lệ phát hiện sai (FDP) và kỳ vọng của nó, FDR, được định nghĩa là $FDP = F/\max\{R, 1\}$ và $FDR = E(FDP)$.

Trong khi tài liệu về định giá tài sản từ lâu đã nhận thức được vấn đề thu thập dữ liệu chung (Lo và MacKinlay, 1990; Sullivan và cộng sự, 1999), những đề xuất ban đầu đề xuất các giả thuyết null đơn lẻ thay thế, chẳng hạn như $H_0 : \max_i \alpha_i = 0$ hoặc $H_0 : E(\alpha_i) = 0$ (xem ví dụ White, 2000; Kosowski và cộng sự, 2006; Fama và French, 2010). Barras và cộng sự. (2010), Bajgrowicz và Scaillet (2012), và Harvey và cộng sự. (2016) là một trong những người đầu tiên áp dụng các phương pháp kiểm soát FDR hoặc FWER trong bối cảnh định giá tài sản để hạn chế nhiều thử nghiệm. Harvey và Liu (2020) đề xuất phương pháp double-bootstrap để kiểm soát FDR, đồng thời tính đến tỷ lệ âm tính giả và tỷ lệ chênh lệch.8 Giglio et al. (2021a) đề xuất một cách tiếp cận suy luận chặt chẽ để kiểm soát FDR trên alpha trong mô hình nhân tố tiềm ẩn, đồng thời giải quyết sai lệch biến bị bỏ sót, th

⁸Chen và cộng sự. (2023) nghiên cứu hàng trăm yếu tố và chỉ tìm ra 2 yếu tố bất thường độc lập sau khi kiểm soát FDR.

dữ liệu và tính chiều cao trong số lượng bài kiểm tra. Jensen và cộng sự. (2021) đề xuất mô hình phân cấp Bayes để thực hiện hiệu chỉnh nhiều thử nghiệm của họ, mô hình này tận dụng hành vi chung và trước đó của các yếu tố bằng θ -alpha , cho phép ước tính alpha của các yếu tố thu hẹp về phía trước và vay mượn sức mạnh của nhau.

Cuối cùng, thử nghiệm nhiều lần về bản chất là một vấn đề thống kê. Các phương pháp thống kê nói trên thường đáp ứng các tiêu chí của một bài kiểm tra thống kê tốt, chẳng hạn như kiểm soát lỗi Loại I, tỷ lệ phát hiện sai, v.v. Tuy nhiên, hiệu quả kinh tế mới là điều mà các đại lý quan tâm nhất. Hai mục tiêu này thường xung đột với nhau. Jensen và cộng sự. (2021) và Đà và cộng sự. (2022) chỉ ra rằng thử nghiệm nhiều lần như một công cụ để lựa chọn alpha thường dẫn đến các chiến lược giao dịch cực kỳ thận trọng, mặc dù nó bảo vệ FDR một cách hoàn hảo. Jensen và cộng sự. (2021) chứng minh rằng một nhà nghiên cứu bao gồm các yếu tố trong danh mục đầu tư của họ dựa trên phương pháp thử nghiệm đa cấp bậc Bayesian sẽ đạt được sự cải thiện lớn và đáng kể so với nhà đầu tư sử dụng phương pháp kiểm soát FDR thận trọng hơn .

5

Danh mục đầu tư tối ưu

Trong phần này, chúng tôi thảo luận và phân tích các phương pháp học máy để lựa chọn danh mục đầu tư. Vấn đề lựa chọn danh mục đầu tư nằm ở trung tâm của tài chính. Nó nhằm mục đích phân bổ hiệu quả các nguồn lực của nhà đầu tư để đạt được

tiết kiệm tối ưu cho tăng trưởng và tất cả các vấn đề chính trong việc định giá tài sản đều có mối liên hệ mật thiết với nó. Dưới các giả định kinh tế yếu (chẳng hạn như không có hoạt động kinh doanh chênh lệch giá), danh mục đầu tư hiệu quả có phương sai trung bình (MVE) trong một nền kinh tế là một đại diện có thể giao dịch được của hệ số chiết khấu ngẫu nhiên (SDF), và do đó tóm tắt cách những người tham gia thị trường đánh đổi rủi ro và lợi nhuận. đạt được mức giá cân bằng (Hansen và Richard, 1987). Tương tự như vậy, mức độ tiếp xúc của các tài sản riêng lẻ với danh mục đầu tư MVE ánh xạ từng cái một thành lợi nhuận kỳ vọng của tài sản, có nghĩa là MVE tương đương với mô hình định giá beta một yếu tố giải thích sự khác biệt giữa các bộ phận trong lợi nhuận trung bình (Roll, 1977). Và, trong vấn đề dự đoán lợi nhuận được phân tích ở Phần 3, hiệu suất dự đoán được đánh giá một cách phổ biến về mặt lợi ích mà nó mang lại trong danh mục đầu tư tối ưu.

Có nhiều phương pháp thống kê có sẵn để theo đuổi danh mục đầu tư tối ưu. Tất cả các phương pháp tiếp cận đều kết hợp các khía cạnh của đặc tính phân phối tài sản (rủi ro và lợi nhuận của chúng) và sở thích của nhà đầu tư đối với sự đánh đổi giữa rủi ro và lợi nhuận. Vấn đề quan trọng của Markowitz (1952) mang lại

các nhà đầu tư có kiến thức về phân phôi lợi nhuận.¹ Nếu không cần ước tính phân phôi này, việc lựa chọn danh mục đầu tư là bài toán một bước về tối đa hóa hữu dụng. Trích dẫn câu nói của Hayek trong phần giới thiệu, "nếu chúng ta có được kiến thức đầy đủ về các phương tiện sẵn có, thì vấn đề còn lại hoàn toàn là logic." Một lần nữa, đây rõ ràng không phải là vấn đề kinh tế mà nhà đầu tư phải đối mặt. Vấn đề lựa chọn danh mục đầu tư của nhà đầu tư gắn bó chặt chẽ với việc ước tính, điều này cần thiết để đổi mới với sự thiếu hiểu biết của họ về phân phôi lợi nhuận. Nhà đầu tư đưa ra quyết định danh mục đầu tư với thông tin không hoàn hảo.

Kể từ Markowitz, rất nhiều tài liệu đề xuất các công cụ phương pháp luận để giải quyết vấn đề ước tính và đạt được danh mục đầu tư có lợi. Đó là một vấn đề cứng đầu mà các sắc thái của nó thường cản trở những đổi mới sáng tạo về phương pháp luận. Thành công thực tế phụ thuộc vào khả năng thông báo mục tiêu tiện ích với dữ liệu hạn chế. Nói cách khác, vấn đề về danh mục đầu tư đã chín muồi cho các giải pháp học máy.

Một tài liệu mới nổi đề xuất các giải pháp mới bắt nguồn từ các ý tưởng học máy về thu gọn, lựa chọn mô hình và tham số hóa linh hoạt. Đầu tiên, dự báo học máy, xuất phát từ những cân nhắc thống kê thuần túy (chẳng hạn như mô hình dự đoán lợi nhuận hoặc rủi ro) có thể được đưa vào vấn đề lựa chọn danh mục đầu tư và được xử lý như số lượng đã biết. Tuy nhiên, một nhà đầu tư xem xét các mục tiêu ước tính và danh mục đầu tư một cách tách biệt sẽ hành xử dưới mức tối ưu. Sự nhiễu loạn của ước tính đóng vai trò như một nguồn rủi ro cho danh mục đầu tư của nhà đầu tư, do đó, việc tách biệt hoàn toàn vấn đề ước tính sẽ làm mất đi tính hữu dụng một cách không cần thiết. Theo trực giác, nhà đầu tư có thể làm tốt hơn bằng cách hiểu các đặc tính của độ nhiễu ước tính và tính toán nó trong các tính toán đánh đổi rủi ro-lợi nhuận của họ. Điều này đòi hỏi phải xem xét tổng hợp việc ước tính và tối ưu hóa tiện ích, một lần nữa dẫn đến các công thức vấn đề được giải quyết một cách hấp dẫn bằng các phương pháp học máy. Chúng tôi bắt đầu cuộc thảo luận của mình bằng cách minh họa những hạn chế của các giải pháp ngây thơ tách biệt các vấn đề ước tính và tối đa hóa tiện ích. Điều này cung cấp một sự hiểu b

¹Có một tài liệu lý thuyết đáng kể rộng vấn đề Markowitz với các dạng ưu đãi phức tạp, phân phôi lợi nhuận, hạn chế danh mục đầu tư và xung đột thị trường, trong khi vẫn duy trì giả định rằng các nhà đầu tư có tất cả kiến thức cần thiết về phân phôi lợi nhuận. Brandt (2010) và Fabozzi và cộng sự. (2010) khảo sát các phần của tài liệu này.

trên đó chúng tôi xây dựng cuộc thảo luận về các công cụ học máy nhằm cải thiện các giải pháp lựa chọn danh mục đầu tư.

5.1 Danh mục "plug-in"

Chúng tôi bắt đầu với một dẫn chứng rõ ràng từ Kan và Chu (2007), những người đã phân tích tác động của việc ước lượng tham số trong vấn đề lựa chọn danh mục đầu tư để thực hiện. Lợi nhuận một kỳ của N tài sản vượt quá lãi suất phi rủi ro, R_t , là một vectơ bình thường iid với khoảnh khắc thứ nhất và thứ hai $\mu \in \mathbb{R}^N$ và $\Sigma \in \mathbb{R}^{N \times N}$. Thái độ của nhà đầu tư được tóm tắt bằng phương trình bậc hai R hữu dụng với ác cảm rủi ro γ , và nhà đầu tư tìm kiếm danh mục đầu tư rủi ro w (kết hợp với tài sản phi rủi ro) để tối đa hóa tài sản $w^\top R$. Tiện ích kỳ vọng trong một kỳ:

$$E[U(w)] = w^\top \mu - \frac{1}{2} w^\top \Sigma w. \quad (5.1)$$

Khi biết được sự phân phối lợi nhuận, giải pháp danh mục đầu tư tối đa hóa tiện ích là

$$w = \frac{1}{\gamma} \Sigma^{-1} \mu. \quad (5.2)$$

Khi chúng ta tính đến thực tế kiến thức không hoàn hảo về phân bổ lợi nhuận, nhà đầu tư phải quyết định cách sử dụng thông tin sẵn có để tối ưu hóa lợi ích mà họ thu được từ danh mục đầu tư của mình. Chúng tôi giả định rằng các nhà đầu tư có sẵn mâu quan sát T được rút ra từ phân bổ lợi nhuận nói trên.

Một cách tiếp cận đơn giản và phổ biến đối với việc lựa chọn danh mục đầu tư xem xét mục tiêu thống kê của việc ước tính phân phối lợi nhuận một cách riêng biệt. Đầu tiên, nhà đầu tư suy ra giá trị trung bình và hiệp phương sai của lợi nhuận mà không quan tâm đến đặc tính hữu dụng. Thứ hai, nhà đầu tư coi các ước tính μ và Σ là mômen lợi nhuận thực sự và dựa trên những ước tính này, nhà đầu tư sẽ lựa chọn trọng số danh mục đầu tư để tối ưu hóa tiện ích. Điều này được gọi một cách thông tục là công cụ ước tính "plug-in", bởi vì nó dẫn đến một giải pháp thay thế đầu vào cho giải pháp (5.2) bằng các đối tác ước tính

$$w = \frac{1}{\gamma} \Sigma^{-1} \mu. \quad (5.3)$$

Động lực của giải pháp plug-in xuất phát từ thực tế là, nếu μ và Σ là các ước tính nhất quán thì w cũng nhất quán với w . Nó

cung cấp cách sử dụng đơn giản và trực tiếp các dự báo học máy để đạt được mục tiêu kinh tế mong muốn. Trong khi tính nhất quán của w có thể là một tính chất lý thuyết hấp dẫn, hiếm khi chúng ta có đủ dữ liệu để phát huy tính chất hội tụ. Khi N lớn, tính nhất quán sẽ gặp nguy hiểm. Do đó, thật không may, giải pháp danh mục đầu tư bổ sung có xu hướng hoạt động kém trong thực tế và gây tai hại trong bối cảnh mà số lượng tài sản bắt đầu gần bằng số lượng quan sát đào tạo (Jobson và Korkie, 1980; Michaud, 1989).

Kan và Chu (2007) kết nối vấn đề ước tính danh mục đầu tư với lý thuyết quyết định thống kê bằng cách lưu ý rằng sự khác biệt về lợi ích của nhà đầu tư phát sinh từ danh mục đầu tư tối ưu "thực sự" (w^*) so với danh mục đầu tư bổ sung (w) có thể được coi là một vấn đề kinh tế. hàm mất mát với dự kiến giá trị

$$E[L(w^*, w) | \mu, \Sigma] = U(w^*) - E[U(w) | \mu, \Sigma]. \quad (5.4)$$

Đây là "hàm rủi ro" thống kê nổi tiếng, mà trong môi trường hiện tại có thể được xem là chi phí tương đương chắc chắn của việc sử dụng trọng số danh mục đầu tư dưới mức tối ưu.

Giả sử nhà đầu tư ước lượng μ và Σ sử dụng mẫu đầu số liệu thống kê

$$\mu = \frac{1}{T} \sum_{t=1}^T R_t, \quad \Sigma = \frac{1}{T} \sum_{t=1}^T (R_t - \mu)(R_t - \mu)^T,$$

thì Kan và Zhou (2007) cho thấy hàm rủi ro của danh mục đầu tư plug-in là

$$E[L(w^* - SR_2, w) | \mu, \Sigma] = -a_1 + a_2 2\gamma \quad (5.5)$$

trong đó SR_2 là tỷ lệ Sharpe bình phương của chiến lược tối ưu thực sự w và (a_1, a_2) là các hằng số chỉ phụ thuộc vào số lượng mẫu quan sát T và số lượng tài sản N .

Công thức cho thấy rằng

tồn thắt dự kiến bằng 0 đối với w và hoàn toàn dương đối với bất kỳ trọng lượng nào khác. Giữ N cố định, cả a_1 và a_2 đều tiến tới 0 khi T tăng, cho thấy rằng sự mất mát là không quan trọng khi T lớn so với

2Các công thức rõ ràng được cho bởi $a_1 = \frac{1}{T-N-2} - 2 \cdot \frac{\frac{T}{(T-N-1)(T-N-2)} \cdot \frac{a_2}{N-4}}$,
 $\frac{NT(T-2)}{(T-N-1)(T-N-2)(T-N-4)}$.

N. Giữ T cố định, cả a1 và a2 đều tăng trong N, nghĩ a là tổn thất trả nên nghiêm trọng hơn khi có nhiều tham số hơn (nhiều tài sản hơn), tất cả các yếu tố khác đều bằng nhau. Và nhà đầu tư chịu tổn thất lớn hơn khi tỷ lệ Sharpe cao hơn hoặc tâm lý lo ngại rủi ro thấp hơn vì cả hai điều này đều khiến nhà đầu tư nghiêng về các tài sản rủi ro, từ đó làm tăng khả năng nhà đầu tư gấp phải sự không chắc chắn trong ước tính.

Phương trình (5.5) cho phép định lượng chi phí tiện ích của công cụ ước tính plug-in trong các hiệu chỉnh hợp lý. Những chi phí này gợi ý sự thiếu hụt lớn về tiện ích đối với các nhà đầu tư phụ thuộc quá mức vào các đặc tính nhất quán của w. Nhưng ngoài thách thức về chiều cao, một nhược điểm rõ ràng của giải pháp plug-in là không thể chấp nhận được. Các quy tắc danh mục đầu tư khác, chẳng hạn như các quy tắc dựa trên phương pháp thu hẹp hoặc lý thuyết quyết định Bayes, cải thiện hiệu suất ngoài mẫu dự kiến cho bất kỳ giá trị nào của các khoảnh khắc thực μ và Σ . Công cụ ước tính plug-in theo nghĩ a này là một giải pháp danh mục đầu tư đặc biệt kém hấp dẫn. Tính không thể chấp nhận được của nó bắt nguồn từ việc không tính đến sự phụ thuộc của tiện ích sử dụng cuối vào việc ước tính tham số.

Bằng cách đặt phân tích của họ vào khung lý thuyết quyết định, Kan và Chu (2007) tiếp theo đưa ra một minh chứng rõ ràng rằng các nhà đầu tư có sẵn các công cụ để giảm thiểu tổn thất tiện ích ngoài mẫu và dễ dàng cải thiện danh mục đầu tư bổ trợ. Cụ thể, nhà đầu tư có thể hiểu trước mức độ không chắc chắn trong ước tính tác động như thế nào đến tiện ích ngoài mẫu của họ và cách cô ấy có thể nội hóa tác động này bằng cách quy tắc danh mục đầu tư đã sửa đổi nhằm giảm tác động có hại của nhiều ước tính. Kan và Chu (2007) chứng minh tính không thể chấp nhận được của danh mục đầu tư plug-in bằng cách chứng minh rằng những điều chỉnh danh mục đầu tư đơn giản (như nghiêng giải pháp plug-in theo hướng có trọng số nặng hơn trong tài sản phi rủi ro hoặc trộn nó với danh mục phương sai tối thiểu của plug-in) tạo ra kết quả cao hơn. Hỗn hợp kỳ vọng đối với bất kỳ giá trị nào của μ và Σ . Cái nhìn sâu sắc này là bản chất của lý thuyết quyết định thống kê - kết quả được cải thiện bằng cách tích hợp các khía cạnh của mục tiêu hữu dụng của nhà đầu tư vào bài toán thống kê về ước lượng trọng số, thay vì coi việc ước lượng và tối đa hóa hữu dụng là những vấn đề riêng biệt

3Có rất nhiều tài liệu phân tích lựa chọn danh mục đầu tư thông qua lăng kính lý thuyết quyết định kinh tế và Bayes (xem khảo sát của Avramov và Chu (2010)). Những ví dụ ban đầu như Jorion (1986) và Frost và Savarino (1986) chứng minh lợi ích tiện ích do giải pháp co ngót Bayes-Stein mang lại. Những đóng góp khác trong lĩnh vực này sử dụng thông tin và động cơ kinh tế của các linh mục trước đó (Black và Litterman,

Nhưng cách tiếp cận lý thuyết về quyết định cũng làm này sinh một nghịch lý. Việc xây dựng bài toán trên cực kỳ đơn giản. Việc trả lại là bình thường. Tùy chọn là bậc hai. Không có khía cạnh nào của vấn đề là mơ hồ hoặc không xác định được. Khi đó, có vẻ như bất kỳ giải pháp nào cho vấn đề này (quyết định về mặt lý thuyết hoặc cách khác) đều phải rõ ràng và rõ ràng tương tự. Nhưng đây không phải là trường hợp! Để thấy được điều này, chúng ta hãy xem xét những gì mà mỗi giả định mang lại cho chúng ta. Giả định về tính chuẩn iid ngụ ý rằng phương tiện mẫu và hiệp phương sai là số liệu thống kê đầy đủ cho tập hợp thông tin của nhà đầu tư (mẫu quan sát lợi nhuận T). Từ đó, chúng ta biết rằng bất kỳ giải pháp hiệu quả nào cho bài toán tiện ích bậc hai đều phải phụ thuộc vào dữ liệu thông qua hai thống kê này chứ không phụ thuộc vào dữ liệu nào khác, vì vậy chúng ta biết hạn chế các giải pháp ở dạng

$$w = f(\mu, \Sigma). \quad (5.6)$$

Tiếp theo, hữu dụng bậc hai cộng với tính chuẩn iid mang lại cho chúng ta khả năng phân tích dễ dàng về tổn thất dự kiến trong (5.5).⁴ Chúng ta tìm kiếm một quy tắc danh mục đầu tư giúp cực tiểu hóa $E[L(w, w)]$ μ, Σ . Tuy nhiên, nếu không có hướng dẫn thêm về hàm f thì vấn đề giảm thiểu tổn thất dự kiến sẽ không được giải quyết. Đây chính là nghịch lý. Công thức đường như hoàn chỉnh của bài toán Markowitz chỉ mang lại một giải pháp rõ ràng khi biết các tham số. Khi tính đến độ không chắc chắn của ước tính, vấn đề sẽ cung cấp không đủ hướng dẫn để ước tính các quy tắc danh mục đầu tư nhằm tối đa hóa tiện ích ngoài mẫu dự kiến.

Một số đạt được tiến bộ bằng cách áp đặt một hình thức cụ thể của f . Ví dụ, để chứng minh tính không thể chấp nhận được của giải pháp trình cắm thêm, Kan và Chu (2007) giới hạn f ở một tập hợp các hàm tuyến tính và chỉ ra rằng trong tập hợp này có các giải pháp danh mục thống trị thống nhất quy tắc trình cắm thêm. Trong khi đó, họ lưu ý rằng những lựa chọn tốt cho w "có thể là một hàm phi tuyến rất phức tạp của μ và Σ và có vô số cách để xây dựng nó. Tuy nhiên, việc xác định các

1992; Mục sư, 2000; Mục sư và Stambaugh, 2000; Tú và Chu, 2010). Mặc dù đây không phải là các phương pháp học máy nhưng chúng đã mở đường cho việc phát triển các giải pháp tối ưu hóa tiện ích và ước tính tích hợp hiện đã trở thành tiêu chuẩn trong các phương pháp danh mục học máy.

4Nhưng khả năng phân tích dễ dàng này chỉ đạt được đối với một số lựa chọn nhất định của f (ví dụ, tuyến tính trong μ, Σ 1, hoặc Σ 1 μ).

$f(\mu, \Sigma)$ tối ưu ." Tu và Chu (2010) và Kan và cộng sự. (2022) xây dựng ý tưởng này bằng cách kết hợp danh mục đầu tư thay thế và đề xuất các quy tắc danh mục đầu tư nhằm giảm thiểu tổn thất tiện ích trong điều kiện ước tính không chắc chắn. Yuan và Chu (2022) mở rộng phân tích này sang trường hợp nhiều chiều trong đó $N > T$. Thay vì hạn chế dạng chức năng của chiến lược danh mục đầu tư f , Da et al. (2022) đã đạt được tiến bộ trong vấn đề này bằng cách áp đặt các hạn chế đối với quá trình tạo dữ liệu trả về.

5.2 Ước tính và tối ưu hóa tích hợp

Vậy nghịch lý về khả năng không thể chấp nhận được sẽ đưa chúng ta đến đâu một cách tổng quát hơn? Chúng ta có một hàm số chưa xác định xuất hiện trong một bài toán kinh tế. Đây là cơ hội lý tưởng để tận dụng sức mạnh của học máy. Trong cài đặt hiện tại, tính năng này có thể hoạt động như thế nào? Có lẽ chúng ta có thể chọn một mô hình linh hoạt như mạng nơ-ron để tham số hóa f và tìm kiếm các giá trị tham số hàm nhằm tối đa hóa tiện ích ngoài mẫu dự kiến (hoặc, tương đương, giảm thiểu hàm rủi ro $E[L(w, w) | \mu, \Sigma]$)? Vấn đề là $E[L(w, w) | \mu, \Sigma]$ chỉ có thể được suy ra trong một số trường hợp đặc biệt nhất định như công cụ ước tính plug-in. Đối với f tổng quát, tiện ích ngoài mẫu dự kiến không có sẵn ở dạng đóng và nếu có thì nó phụ thuộc vào các tham số thực chưa biết μ và Σ .

Một cách tiếp cận khác là chọn quy tắc danh mục đầu tư f để tối ưu hóa tiện ích trong mẫu và thường xuyên hóa công cụ ước tính để khuyến khích hiệu suất danh mục đầu tư ngoài mẫu ổn định (ví dụ: thông qua xác thực chéo). Điều này nằm gọn trong quy trình học máy gồm hai bước diễn hình:

Bước 1. Chọn một lớp hàm được lập chỉ mục theo tham số điều chỉnh (ví dụ: lớp có thể là các mô hình đường gờ tuyến tính, được lập chỉ mục bởi hình phạt đường gờ z) và sử dụng mẫu huấn luyện để ước tính các tham số mô hình (một bộ ước tính cho mỗi hàm giá trị của z). Hãy lưu ý xem mô hình Markowitz đơn giản, mang tính phân tích ngay lập tức biến thành một vấn đề về học máy như thế nào. Khi các tham số thực được biết đến, mô trường Markowitz được chỉ định chặt chẽ sẽ mang lại giải pháp danh mục đầu tư dạng đóng đơn giản và trực quan. Tuy nhiên, chỉ với một chút độ không chắc chắn trong ước tính, vấn đề lý thuyết quyết định chỉ được xác định một phần và quy tắc danh mục đầu tư.

không thể xác định được việc giảm thiểu mất mău ngoài mău. Điều này dẫn đến việc chúng ta sử dụng mục tiêu kinh tế làm mục tiêu ước lượng. Nói cách khác, cấu trúc lý thuyết quyết định của vấn đề buộc chúng ta phải tích hợp các mục tiêu thống kê và kinh tế cũng như xem xét các thông số kỹ thuật học máy có tư duy cởi mở.

Bước 2. Trong mău xác thực, chọn các tham số điều chỉnh (ví dụ: hình phạt sườn z cụ thể) để tối ưu hóa hiệu suất ngoài mău dự kiến . Nền tảng lý thuyết của việc chính quy hóa bắt nguồn từ việc giảm thiểu rủi ro ước tính. Trong khuôn khổ Kan và Chu (2007), những nền tảng này rất rõ ràng và việc chính quy hóa được thực hiện thông qua các tính toán phân tích. Quá trình hành động tiêu chuẩn hơn thay thế các tính toán lý thuyết bằng kinh nghiệm.

Tiện ích ngoài mău dự kiến được tính gần đúng bằng tiện ích thực hiện được trong mău xác thực và giá trị z tối đa hóa tiện ích xác thực sẽ được chọn. Việc xác nhận chéo theo kinh nghiệm này loại bỏ nhu cầu đơn giản hóa lý thuyết cần thiết để tối ưu hóa tiện ích dự kiến một cách trực tiếp.

Khi vấn đề lựa chọn danh mục đầu tư được đưa vào khung học máy , người ta có thể bắt đầu khai quát hóa bất kỳ giả định hạn chế nào khác.

Ví dụ: phân phối lợi nhuận không nhất thiết phải chuẩn tắc cũng như không phải iid, vì chúng ta không cần tính toán rõ ràng độ thỏa dụng ngoài mău dự kiến.

Do đó chúng ta có thể viết quy tắc danh mục đầu tư chung như

$$w = f(XT), \quad (5.7)$$

trong đó XT đối chiếu tất cả dữ liệu liên quan đến quá trình ra quyết định. có thể bao gồm các biến {Rt} mău trả về điều T_t = Bản thân nó cũng như mọi điều hòa chỉnh phân phối trả về có điều kiện. Chúng ta có thể tương tự như vậy, hoán đổi các thông số kỹ thuật ưu tiên để giải quyết các mối lo ngại về rủi ro cuối, các mục tiêu đa chiều hoặc các vấn đề phức tạp khác.

5.3 Hồi quy tỷ lệ Sharpe tối đa

Các phương pháp học máy biên giới để lựa chọn danh mục đầu tư trực tiếp xem xét việc tối ưu hóa tiện ích khi ước tính các quy tắc danh mục đầu tư.

5.3. Hồi quy tỷ lệ Sharpe tối đa

113

Tiến bộ ban đầu quan trọng trong việc lựa chọn danh mục học máy đến từ Ait-Sahalia và Brandt (2001), Brandt và Santa-Clara (2006), và Brandt et al. (2009). Đóng góp trọng tâm của họ là chính thức hóa vấn đề danh mục đầu tư như một quy trình một bước tích hợp tối đa hóa tiện ích vào vấn đề thông kê về ước lượng hàm trọng số. Cách tiếp cận này không đưa ra giả định về phân phối lợi nhuận. Thay vào đó, nó chỉ định mức hữu dụng của nhà đầu tư và một dạng hàm rõ ràng cho hàm trọng số danh mục đầu tư của nhà đầu tư dưới dạng các hiệp phương sai có thể quan sát được. Các tham số của hàm trọng số được ước tính bằng cách tối đa hóa tiện ích trung bình trong mẫu. Theo Brandt (1999), chúng tôi gọi đây là cách tiếp cận “trọng lượng danh mục đầu tư theo tham số”.

Để giữ cho phần trình bày của chúng tôi cụ thể, chúng tôi tiếp tục đi theo tinh thần của Markowitz và chuyên môn hóa phân tích của chúng tôi vào khuôn khổ tiện ích phương sai trung bình. Động lực chính của chúng tôi để làm như vậy là chúng tôi có thể sử dụng “trọng số danh mục tham số” dưới dạng hồi quy OLS. Để làm như vậy, chúng tôi dựa vào Định lý 1 của Britten-Jones (1999), người đã chỉ ra rằng giải pháp plug-in Markowitz trong (5.3) tỷ lệ thuận với hệ số OLS trong hồi quy của vectơ không đổi trên lợi nhuận tài sản mẫu. Đặc biệt, hồi quy OLS

$$1 = w R_t + u_t \quad (5.8)$$

cung cấp hệ số 5

$$\text{với } \overset{\text{OLS}}{\Sigma} = 1\mu .$$

Theo trực giác, hồi quy (5.8) tìm kiếm sự kết hợp giữa lợi nhuận vượt quá rủi ro R_t hoạt động gần nhất có thể với hằng số dương.

Điều này tương đương với việc tìm ra sự kết hợp giữa các tài sản rủi ro với tỷ lệ Sharpe trong mẫu cao nhất có thể. Mặc dù điều này hoàn toàn tỷ lệ với việc xác định danh mục tiếp tuyến trong mẫu của các tài sản rủi ro, nhưng chúng ta sẽ thấy rằng công thức hồi quy rất hấp dẫn khi kết hợp các phương pháp học máy vào các vấn đề danh mục đầu tư được tham số hóa. Trong tương lai, chúng tôi gọi hồi quy có dạng (5.8) là “hồi quy tỷ lệ Sharpe tối đa” hoặc MSRR.

Brandt và cộng sự. (2009) tập trung vào việc tham số hóa trọng số danh mục đầu tư

$$w_{i,t} = -w_{i,t} + s_{i,t}\beta \quad (5.9)$$

$$\text{Để rút ra kết quả này, hãy lưu ý rằng } w^{\text{OLS}} = (\Sigma + \mu\mu)^{-1}\mu = \Sigma^{-1}\mu(1 + \mu\Sigma^{-1}\mu)^{-1} .$$

trong đó w^i_t là trọng số chuẩn đã biết của tài sản i tại thời điểm t và s_i , t là tập hợp K tín hiệu đi tuyến tính vào quy tắc danh mục đầu tư với vectơ hệ số K chiều β . Điều này giới thiệu phạm vi cho trọng số danh mục đầu tư linh hoạt thay đổi theo các biến điều kiện. Để đơn giản về mặt ký hiệu, chúng tôi chuẩn hóa w^i_t về 0 và xếp các trọng số và tín hiệu vào phương trình vectơ

$$w_t = S\beta, \quad (5.10)$$

trong đó w_t là $N \times 1$ và S là $N \times K$. Thay thế vào (5.8), MSRR trở thành

$$1 = \beta (S^T 1R_t) + u_t = \beta F_t + u_t. \quad (5.11)$$

Chúng ta có thể xem w_t là tập hợp các trọng số động trên N tài sản cơ sở R_t . Hoặc, tương đương, β là tập hợp các trọng số tinh trên K "các yếu tố" (danh mục đầu tư được quản lý theo đặc điểm):

$$F_t = S^T 1R_t. \quad (5.12)$$

Nói cách khác, hạn chế áp dụng cùng một tổ hợp tín hiệu cho tất cả các tài sản sẽ ngay lập tức chuyển chiến lược này thành chiến lược giao dịch cắt ngang (vì F đại diện cho danh mục đầu tư được hình thành dưới dạng hiệp phương sai cắt ngang của lợi nhuận với từng tín hiệu riêng lẻ). Hệ số công danh mục ước tính là $\beta = \text{Cov}(F) / \text{Var}(F)$. Theo ngôn ngữ của Brandt và Santa-Clara (2006), việc tham số hóa trọng số MSRR trong (5.10) "tăng không gian tài sản," từ không gian cơ sở trở về không gian các thừa số. Với phần mở rộng này, chúng ta quay trở lại bài toán cơ bản trong (5.8) ước tính trọng số cố định của danh mục đầu tư theo các yếu tố. Lợi ích của công thức MSRR là chúng ta có sẵn số liệu thống kê kiểm tra OLS đơn giản để đánh giá sự đóng góp của từng tín hiệu (như đề xuất ban đầu của Britten-Jones, 1999).

Brandt (1999) tóm tắt lợi ích của các quy tắc danh mục đầu tư tham số, lưu ý

"bước lập mô hình lợi nhuận chắc chắn là gót chân Achilles của phương pháp kinh tế lượng truyền thống.... Tập trung trực tiếp

⁶Brandt và cộng sự. (2009) cũng lưu ý sự dễ dàng của việc cung cấp tính trung lập của ngành hoặc các lựa chọn tập trung vào trọng số khác cũng như xử lý các kích thước mặt cắt thay đổi theo thời gian (trong trường hợp đó họ khuyến nghị $\frac{1}{N_t}$ chuẩn hóa s_i, t, β trong (5.9)), trong đó N_t là số lượng tài sản ở thời kỳ t). Chúng tôi tóm tắt những cân nhắc này để đơn giản về mặt ký hiệu.

do đó, về tỷ trọng danh mục đầu tư tối ưu sẽ giảm đáng kể khả năng xảy ra lỗi ước tính và xác định sai mô hình."

Họ nhấn mạnh rằng việc kiểm soát lỗi ước tính bắt nguồn từ thứ nguyên tham số được loại bỏ nhiều (liên quan đến việc ước tính phương tiện và hiệp phương sai của tất cả tài sản).

Brandt và cộng sự. (2009) Ứng dụng thực nghiệm nghiên cứu các quy tắc danh mục đầu tư cho hàng nghìn cổ phiếu chỉ sử dụng ba tham số (nghĩa là chỉ dựa trên quy mô, giá trị và đặc điểm đảo chiều ngắn hạn của cổ phiếu). Công thức bình phương tối thiểu cho các quy tắc danh mục đầu tư được tham số hóa thúc đẩy việc suy nghĩ về các điều chỉnh để kết hợp các cấu trúc học máy. Chúng tôi ngay lập tức muốn xem xét các hồi quy K lớn để kết hợp một tập hợp thông tin điều hòa phong phú, bao gồm cả tập hợp đầy đủ các tín hiệu yếu tố và có lẽ là tương tác của chúng với các biến trạng thái kinh tế vĩ mô, cùng với các chiến lược chính quy hóa máy học thông thường để thu hẹp và lựa chọn mô hình. Ý tưởng về danh mục đầu tư tham số được chuẩn hóa theo dây đã được đề xuất bởi Brandt và Santa-Clara (2006) và Brandt et al. (2009), và được điều tra kỹ lưỡng bởi DeMiguel et al.

(2020).⁷ Điểm hay của MSRR là phần mềm hiệu quả cho hồi quy lasso và lưới đàn hồi có thể được triển khai ngay lập tức cho bài toán lựa chọn danh mục đầu tư. Về cơ bản, điều này bao gồm tất cả các loại hồi quy bị phạt sử dụng hình phạt 1 hoặc 2 tham số, bao gồm một loạt các cân nhắc danh mục đầu tư thực tế như kiểm soát đòn bẩy (thường được coi là hình phạt 1) hoặc điều chỉnh chi phí giao dịch (thường được mô hình hóa dưới dạng hình phạt 2). Trong trường hợp đặc biệt của hồi quy sườn núi với hình phạt z, mục tiêu MSRR bị phạt là

$$\frac{\text{phút}}{w} \quad (1 - \beta F_t)^2 + z\beta\beta \quad (5.13)$$

với giải pháp

$$\beta = (\text{Cov}(F) + zI)^{-1}E(F), \quad (5.14)$$

điều chỉnh danh mục các yếu tố tiếp tuyến mẫu chuẩn bằng cách thu hẹp hiệu phương sai mẫu về mức đồng nhất. Điều này kết nối một cách tự nhiên

⁷Ao và cộng sự. (2018) cũng nghiên cứu một vấn đề liên quan đến việc lựa chọn danh mục đầu tư được điều chỉnh theo dây mà họ gọi là MAXSER.

với chủ đề rộng hơn về thu hẹp hiệp phương sai để tối ưu hóa danh mục đầu tư được thảo luận bởi Ledoit và Wolf (2004) và Ledoit và Wolf (2012).

Cùng với việc dễ dàng thao tác các thông số kỹ thuật MSRR, khả năng tồn tại của các công thức tuyển tính bị phạt sẽ làm tăng triển vọng thú vị cho thiết kế danh mục đầu tư tham số. Ví dụ: chúng ta có thể phá vỡ giới hạn các hệ số giống hệt nhau cho tất cả nội dung trong (5.10) bằng cách sửa đổi

$$w_{i,t} = s_i t \beta_i \quad (5.15)$$

và biểu diễn hồi quy liên quan

$$1 = \text{vec}(B) \cdot \{\text{vec}(S_t - 1) (R_t - 1K)\} + u_t \quad , \quad (5.16)$$

trong đó $B = [\beta_1, \dots, \beta_N]$. Hồi quy này giải thích việc tìm kiếm danh mục đầu tư tiếp tuyến giữa các chiến lược “được quản lý” bởi NK, với mỗi chiến lược riêng lẻ là sự tương tác giữa lợi nhuận trên tài sản i với một trong các tín hiệu của nó. Một sửa đổi nhỏ của vấn đề này giúp có thể điều chỉnh tài sản với các bộ dự đoán khác nhau. Như một bước ngoặt trong khả năng áp dụng MSRR bị phạt, một cách tiếp cận thú vị là phạt β^- , cho phép có sự không đồng nhất giữa các tài sản nhưng thu hẹp quy tắc trọng số của tất cả tài sản về quy tắc trung bình, theo tinh thần của Bayes theo kinh nghiệm.

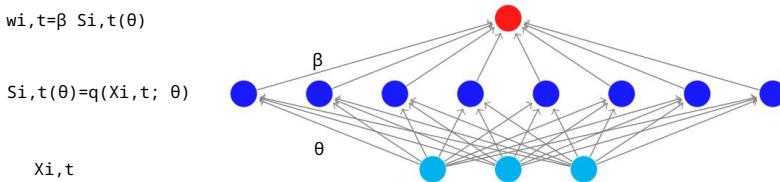
Brandt và cộng sự. (2009) thảo luận về một số phần mở rộng và cải tiến cho vấn đề danh mục tham số của họ, bao gồm các phép biến đổi không âm về trọng số đối với các chiến lược chỉ dài hạn, cân nhắc chi phí giao dịch và tầm nhìn đầu tư nhiều giai đoạn. Hầu hết các phần mở rộng này có thể được xếp vào khung MSRR được nêu ở đây miễn là mục tiêu phương sai trung bình được giữ lại. Đối với các hàm tiện ích tổng quát hơn, việc ước tính phải bước ra khỏi khung hồi quy OLS và yêu cầu tối ưu hóa bằng số. Điều này khó khăn hơn về mặt tính toán đối với việc tham số hóa nhiều chiều nhưng lại thêm rất ít độ phức tạp về mặt khái niệm.

5.4 MSRR có độ phức tạp cao

MSRR có khả năng thích ứng với các mô hình học máy phức tạp như mạng thần kinh. Một ví dụ đơn giản về kiến trúc mạng MSRR tiềm năng được thể hiện trong Hình 5.1. Một tập hợp các tín hiệu L “thô” cho chứng khoán

5.4. MSRR có độ phức tạp cao

117



Hình 5.1: Mạng thần kinh MSRR

Lưu ý: Minh họa một kiến trúc thần kinh đơn giản ánh xạ các biến điều hòa $X_{i,t}$ thành các trọng số danh mục đầu tư tối ưu $w_{i,t}$.

i, ký hiệu là $X_{i,t}$, là đầu vào của mạng. Chúng lan truyền thông qua một chuỗi các phép biến đổi phi tuyến (được tham số hóa bởi các hệ số mạng θ) để tạo ra một vectơ gồm K tín hiệu đầu ra, $S_{i,t}(\theta) = q(X_{i,t}; \theta)$ R K. Sau trong đó q R L đó, một công thức MSRR tiêu chuẩn khác sử dụng các đặc tính được chuyển đổi để rút ra quy tắc danh mục đầu tư hiệu quả có phương sai trung bình:

$$1 = \beta S_{i,t}(\theta) R_t + u_t . \quad (5.17)$$

Ước tính các tham số (β, θ) là một bài toán bình phương nhỏ nhất phi tuyến, dễ thực hiện bằng các gói phần mềm mạng thần kinh sẵn có. Giải thích ước tính này là nó tìm kiếm các phép biến đổi phi tuyến tính linh hoạt của các tín hiệu ban đầu được thiết lập để thiết kế các đặc điểm tài sản mới có thể chủ động quản lý các vị trí trong tài sản cơ sở và tạo ra các chiến lược tỷ lệ Sharpe cao.

Simon và cộng sự. (2022) theo đuổi khuôn khổ này bằng cách sử dụng mạng nơ-ron chuyển tiếp nguồn cấp dữ liệu trong hàm trọng số danh mục đầu tư với ba lớp ẩn lần lượt gồm 32, 16 và 8 nơ-ron trong mỗi lớp. Họ cho thấy rằng mạng lưới thần kinh nâng cao hiệu suất danh mục đầu tư tham số. Đặc tả danh mục đầu tư tuyến tính với cùng các yếu tố dự đoán mang lại tỷ lệ Sharpe ngoài mẫu hàng năm là 1,8, trong khi danh mục mạng thần kinh đạt được tỷ lệ Sharpe là 2,5 hoặc cải thiện 40%.

Didisheim và cộng sự. (2023) chứng minh về mặt lý thuyết rằng "tính chất phức tạp" cũng có tác dụng trong việc xây dựng danh mục đầu tư hiệu quả với phương sai trung bình. Theo các giả định yếu, tỷ lệ Sharpe của danh mục học máy đang tăng lên trong quá trình tham số hóa mô hình. Theo kinh nghiệm, họ trình bày một đặc tả thần kinh đặc biệt thuận tiện có nguồn gốc từ khuôn khổ

của Phần 2. Đặc biệt, họ đề xuất một kiến trúc thằn kinh sử dụng hồi quy sườn đặc trưng ngẫu nhiên. Công thức giống hệt với (5.17) ngoại trừ hệ số α ron 0 được tạo ngẫu nhiên thay vì ước tính. Theo dự đoán của lý thuyết, tỷ lệ Sharpe ngoài mǎu nhận ra trong phân tích thực nghiệm của họ đang tăng lên cùng với số lượng tham số mô hình và bằng phẳng khoảng 4 khi sử dụng 30.000 tham số.

5.5 Ước tính SDF và lựa chọn danh mục đầu tư

Sự tương đương giữa hiệu quả của danh mục đầu tư và các hạn chế về giá tài sản khác, chẳng hạn như bằng chữ cái bằng 0 trong mô hình định giá beta hoặc sự hài lòng của các phương trình Euler dựa trên SDF, ngụ ý rằng có các mục tiêu thống kê (ngoài mục tiêu tiện ích) cần áp dụng khi ước tính danh mục đầu tư tối ưu. Phần lớn tài liệu về học máy tài chính tập trung vào ước tính SDF. Kiến thức về SDF rất thú vị đối với một số câu hỏi kinh tế, chẳng hạn như suy ra sở thích của nhà đầu tư, định lượng sai sót về giá và tìm ra các nguồn rủi ro chính ảnh hưởng đến giá tài sản. Tuy nhiên, trên thực tế, tài liệu về học máy tài chính đã đặt ra những hạn chế không đầy đủ đối với vấn đề ước tính SDF để trả lời các câu hỏi về các cơ chế kinh tế cụ thể. Thay vào đó, tài liệu này có xu hướng đánh giá các kết quả ước lượng về mặt ngoại lệ.

tỷ lệ Sharpe mǎu của SDF ước tính. Nghĩa là, các phương pháp học máy để ước tính SDF tập trung vào vấn đề lựa chọn danh mục đầu tư có phương sai trung bình, được thúc đẩy bởi định lý SDF của Hansen và Richard (1987) và tương đương biên giới phương sai trung bình.

Công thức điển hình của bài toán ước lượng SDF như sau . Nếu một SDF tồn tại (ví dụ như không có kinh doanh chênh lệch giá), nó có thể được biểu diễn dưới dạng danh mục đầu tư có lợi nhuận vượt trội (trừ một số hạng không đổi)

$$M_t = 1 - w R_t . \quad (5.18)$$

Hơn nữa, SDF phải đáp ứng phương trình Euler tiêu chuẩn của nhà đầu tư ở trung tâm của lý thuyết định giá tài sản:

$$E[M_t R_t] = 0 . \quad (5.19)$$

Kết hợp những điều này chúng ta thấy:

$$E[R_t - R_t w] = 0 . \quad (5.20)$$

Phương trình này xác định một SDF có thể giao dịch. Lưu ý rằng đây cũng là điều kiện bậc nhất của bài toán MSRR Britten-Jones (1999),

$$\frac{1}{w} E(1 - w R_t)^2,$$

nó kết nối chính thức các vấn đề về ước tính SDF và tối đa hóa tỷ lệ Sharpe, và do đó nó dẫn đến danh mục tiếp tuyến dưới dạng trọng số SDF ước tính.

Kozak và cộng sự. (2020) đề xuất bài toán ước lượng SDF ở dạng có điều kiện và được tham số hóa có mối liên hệ chặt chẽ với MSRR.

Đặc biệt, họ thừa nhận một danh mục đầu tư $SDF_{wt} = St\beta$ giống như trong (5.10), và lưu ý rằng điều này dẫn đến một SDF có dạng

$$Mt = 1 - \beta F_t \quad (5.21)$$

với F_t được xác định trong (5.12).⁸ Nói cách khác, SDF định giá tài sản có điều kiện có thể được xem như một danh mục tinh gồm các yếu tố được quản lý theo đặc tính. Đóng góp chính của Kozak et al. (2020) là đưa phép chuẩn hóa chính quy vào bài toán ước lượng SDF có điều kiện. Họ chủ yếu phân tích việc chính quy hóa đường vân, ánh xạ trực tiếp tới công cụ ước tính đường vân MSRR trong phương trình (5.14). Xoay công cụ ước tính của họ vào không gian các thành phần chính của các yếu tố, Kozak et al. (2020) cho thấy sự co rút của sườn núi gây ra sự co rút nặng hơn đối với các thành phần được xếp hạng thấp hơn. Họ đưa ra quan điểm sâu sắc về cách suy luận này: "Giải thích kinh tế là chúng tôi đánh giá là không thể tin được rằng một PC có giá trị riêng thấp có thể đóng góp đáng kể vào sự biến động của SDF và do đó dẫn đến tỷ lệ Sharpe bình phương tối đa tổng thể." Kozak (2020) sử dụng phương pháp hạt nhân để ước tính SDF nhiều chiều bằng cách triển khai thông minh "thủ thuật hạt nhân" giúp loại bỏ nhu cầu ước tính số lượng tham số cắt cỗ.

Các kết quả thực nghiệm của Kozak et al. (2020) chỉ ra rằng phương pháp ước tính SDF dựa trên yếu tố với hồi quy bị phạt của họ mang lại một danh mục đầu tư có hiệu suất ngoài mẫu có hiệu quả về mặt kinh tế (và có ý nghĩa thống kê) so với điểm chuẩn tiêu chuẩn

⁸Kozak và cộng sự. (2020) chuẩn hóa SDF để có giá trị trung bình 1: $Mt = 1 - \beta (F_t - E(F_t))$. Chúng tôi áp dụng một cách chuẩn hóa khác để phù hợp với vấn đề MSRR trong Britten-Jones (1999).

các mô hình định giá tài sản Việc xây dựng SDF từ 50 danh mục yếu tố bất thường sẽ tạo ra SDF có tỷ lệ thông tin hàng năm là 0,65 so với CAPM. Nếu thay vì sử dụng 50 đặc điểm cho bộ tín hiệu của mình, họ bổ sung các đặc điểm thô bằng lũy thừa thứ hai và thứ ba của các đặc điểm đó cũng như các tương tác đặc tính theo cặp (với tổng số 1.375 tín hiệu), tỷ lệ thông tin này tăng lên 1.32.9 Giglio et al. (2021b) phân tích và so sánh các tính chất tiềm cận của Kozak et al. (2020) với các công cụ ước tính dựa trên

PCA và RP-PCA do Giglio và Xiu (2021) và Lettau và Pelger (2020a) đề xuất, trong khung mô hình nhân tố vô điều kiện chung (4.1). Với y và V được xác định bởi (4.6) và (4.4), công cụ ước tính SDF dựa trên PCA được cho bởi

$$Mt = 1 - y V^{-t}.$$

Giglio và cộng sự. (2021b) cho thấy rằng các công cụ ước tính sườn núi và dựa trên PCA là nhất quán miễn là các yếu tố có tính phổ biến.

Không giống như Kozak et al. (2020), người đã lập mô hình SDF như một danh mục các yếu tố bất thường, Chen và cộng sự. (2021) mở rộng bài toán ước tính SDF bằng cách lập mô hình trọng số trên từng cổ phiếu riêng lẻ. Mô hình của họ khác với MSRR và bài toán danh mục tham số theo một số cách thú vị. Đầu tiên, chúng cho phép tạo ra một hàm trọng số linh hoạt bao gồm thành phần mạng lưới thần kinh tái phát. Thứ hai, họ xây dựng công cụ ước tính của mình thông qua GMM với sơ đồ thiết bị đo phức tạp. Bắt đầu từ phiên bản có điều kiện của phương trình Euler cơ bản trong (5.20), Chen et al. (2021) viết lại hàm mục tiêu GMM thành

$$\frac{1}{W} \sum_{Nj=1}^N E \left[\sum_{t=1}^T w(SM,t, Si,t) R_i, t+1 R_j, t+1 g(SM,t, Sj,t) \right] \quad (5.22)$$

Hàm $w(SM,t, Si,t)$ mô tả trọng số của SDF có thể giao dịch i th cổ phiếu trong thứ t , là hàm vô hướng của các yếu tố dự báo kinh tế vĩ mô (SM,t , ảnh hưởng đến trọng số của tất cả tài sản) và các yếu tố dự báo cụ thể theo cổ phiếu (Si,t , điều này chỉ ảnh hưởng trực tiếp đến trọng lượng của cổ phiếu i). Nó sử dụng LSTM để nắm bắt hành vi động của SM,t , có đầu ra chảy vào một

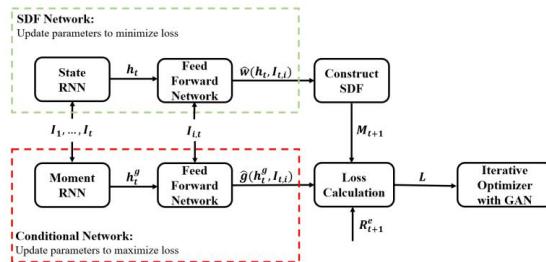
⁹Tỷ lệ thông tin được suy ra từ bảng chữ cái và sai số chuẩn được báo cáo trong Bảng 4 của Kozak et al. (2020).

Mạng có điều kiện đóng vai trò là đối thủ và cạnh tranh với mạng SDF để xác định các chiến lược tài sản và danh mục đầu tư khai thác tốt nhất. Động lực thông tin kinh tế vĩ mô được tóm tắt bằng các biến trạng thái kinh tế vĩ mô mà được Mạng thần kinh tái phát (RNN) thu được với các đơn vị Bộ nhớ ngắn dài hạn. Cấu trúc mô hình 5.5. Ước tính SDF và Lựa chọn danh mục đầu tư được tóm tắt trong

Hình 1 và từng phần khác nhau được mô tả chi tiết trong phần tiếp
tiêu mục.

121

Hình 1: Kiến trúc mô hình GAN



Hình này thể hiện kiến trúc mô hình của GAN (Mạng đối thủ sáng tạo) với RNN (Mạng thần kinh tái phát) với các ô LSTM. Mạng SDF có hai phần: (1) LSTM ước tính một số lượng nhỏ các trạng thái kinh tế vĩ mô. (2) Các trạng thái này cùng với các đặc điểm công ty được sử dụng trong FFN để xây dựng SDF ứng cử viên cho một tập hợp tài sản thử nghiệm nhất định. Mạng điều hòa cũng có hai mạng: (1) Nó tạo ra tập hợp các trạng thái kinh tế vĩ mô của riêng mình. (2) nó kết hợp với các đặc điểm công ty trong FFN để tìm ra các tài sản thử nghiệm bị định giá sai cho một SDF M nhất định. Hai mạng này cạnh tranh cho đến khi hội tụ, đó không phải là SDF cũng như nội dung thử nghiệm có thể được cải thiện.

Nguồn: [Hình ảnh](#), [Đoán Chen và cộng sự. Kế Kiến Xử Trị Mạng SDF](#) (2021) vẫn tiếp và
được gần nhau là FFN. Nó ước tính các phương tiện có điều kiện $\mu_{t,i} = \mu(I_t, I_t, i)$ bằng cách giảm thiểu
Lưu ý: Nguồn: Chen et al. (2021).

13

mạng chuyển tiếp nguồn cấp dữ liệu nơi nó được kết hợp với các công cụ dự đoán mức tồn kho
Bản sao điện tử có tại: <https://ssrn.com/abstract=350138>
để tạo ra sản lượng cuối cùng của danh mục đầu tư (xem phần trên cùng
của Hình 5.2).

Hàm $g(SM, t, S_j, t)$ là một tập hợp các biến công cụ D cung cấp hệ thống các điều kiện
mômen cần thiết để ước tính hàm trọng số danh mục đầu tư. Tuy nhiên, các biến công cụ
được lấy từ dữ liệu cơ bản thông qua hàm mạng, g. Cấu trúc của g phản ánh cấu trúc của w.
Động lực kinh tế vĩ mô được ghi lại bằng LSTM và đầu ra được kết hợp với các biến công cụ thể
của cổ phiếu S_j, t trong mạng chuyển tiếp nguồn cấp dữ liệu để tạo ra các công cụ cho lối
định giá phương trình Euler của tài sản j. Hàm lõi là chuẩn mực của các lối định giá được
tổng hợp trên tất cả các cổ phiếu, được giảm thiểu để đạt được ước tính hàm trọng số.

Có lẽ khía cạnh thú vị nhất của đặc tả này là các biến công cụ được
tạo ra theo cách đối nghịch. Trong khi công cụ ước tính tìm kiếm hàm trọng
số SDF giúp giảm thiểu sai sót về giá cho một bộ công cụ nhất định (mục
tiêu tối thiểu hóa bên ngoài trong (5.22)), nó đồng thời tìm kiếm hàm biến
công cụ khiến SDF nhất định rơi vào tình trạng xấu nhất có thể (mục tiêu
tối thiểu hóa bên ngoài trong (5.22)). mục tiêu giảm thiểu bên trong trong
(5.22)).

Được thúc đẩy bởi những nghiên cứu thực nghiệm về SDF học máy này,

Didisheim và cộng sự. (2023) phân tích về mặt lý thuyết vai trò của độ phức tạp của mô hình trong việc định hình các thuộc tính của công cụ ước tính SDF. Giống như Kelly và cộng sự. (2022a), họ tập trung vào các công cụ ước tính sườn núi có chiều cao với hai điểm khác biệt chính. Đầu tiên, họ chuyển từ cài đặt chuỗi thời gian một nội dung với một nội dung duy nhất sang cài đặt bảng với số lượng nội dung rủi ro tùy ý. Thứ hai, họ định hướng lại mục tiêu thống kê từ dự báo chuỗi thời gian sang tối ưu hóa SDF. Trong bối cảnh này, Didisheim et al. (2023) rút ra một cách rõ ràng tỷ lệ Sharpe ngoài mẫu dự kiến của SDF và các lỗi định giá theo từng mặt cắt ngang như một hàm số về độ phức tạp của nó. Kết quả trung tâm của họ là hiệu suất SDF ngoài mẫu dự kiến đang cải thiện đáng kể về độ phức tạp của mô hình SDF khi sử dụng độ co rút thích hợp. Họ báo cáo các phân tích thực nghiệm phù hợp chặt chẽ với dự đoán lý thuyết của họ. Cụ thể, họ phát hiện ra rằng các mô hình định giá tài sản thực nghiệm tốt nhất có số lượng yếu tố cực kỳ lớn (nhiều hơn số lượng quan sát đào tạo hoặc tài sản cơ sở).

5.5.1 Ước tính SDF phi thương mại

Phản trước tập trung vào việc ước tính SDF được biểu thị trong không gian có lợi nhuận vượt trội hoặc dưới dạng danh mục đầu tư có thể giao dịch. Như đã lưu ý, cấu trúc kinh tế khá tối thiểu được áp đặt trong các bài toán ước tính này gây khó khăn cho việc nghiên cứu các cơ chế kinh tế. Một tài liệu nhỏ hơn nhưng không kém phần thú vị đưa ra SDF bằng cách cân bằng giữa cấu trúc và tính linh hoạt, bao gồm việc sử dụng các thành phần mô hình phi tham số (bán) được nhúng trong khung cấu trúc một phần và tiến hành kiểm tra giả thuyết.

Thật thú vị khi nhận ra rằng ý tưởng tham số hóa các mô hình định giá tài sản bằng cách sử dụng mạng lưới thần kinh đã xuất hiện sớm trong tài liệu như Bansal và Viswanathan (1993). Các thông số kỹ thuật tiên đoán này có tất cả các thành phần cơ bản xuất hiện trong các nghiên cứu gần đây hơn, nhưng bị hạn chế bởi các tập dữ liệu nhỏ hơn và do đó nghiên cứu các thông số kỹ thuật mạng nhỏ.

Chen và Ludvigson (2009) có lẽ là ví dụ điển hình nhất cho đến nay về phương pháp này. Mô trường mô hình của họ có nguồn gốc rộng rãi từ đặc điểm ích tiêu dùng theo thói quen phi tuyến tính của Campbell và Cochrane (1999) và các mô hình liên quan của Menzly et al. (2004), Wachter (2006) và Yogo (2006). Như các tác giả đã chỉ ra một cách đúng đắn, sự đa dạng của thói quen

các thông số kỹ thuật trong tài liệu gợi ý rằng “dạng chức năng của thói quen nên được xử lý, không phải như một thứ nhất định, mà là một phần không thể thiếu của bất kỳ cuộc điều tra thực nghiệm nào,” và đến lượt họ theo đuổi mô hình thói quen bằng cách “đặt càng ít hạn chế càng tốt lên hành vi đó”. Đặc điểm kỹ thuật của thói quen và không có giới hạn tham số nào đối với quy luật chuyển động tiêu dùng.”

Mô hình của Chen và Ludvigson (2009) được tóm tắt tốt nhất bằng cách mô tả tiêu dùng theo thói quen, là yếu tố chính thúc đẩy lợi ích cận biên của nhà đầu tư. Đặc biệt, thói quen tiêu dùng là

$$X_t = CTg \quad \frac{Ct - 1}{Ct}, \dots, \frac{Ct - L}{Ct},$$

trong đó Ct là mức tiêu dùng tại thời điểm t và g là “hàm thói quen” điều chỉnh mức độ thói quen như một hàm tiêu dùng trong giai đoạn L gần đây. Thói quen tác động đến hạnh phúc của nhà đầu tư thông qua chức năng tiện ích

$$U = E^\infty \quad \delta^t \frac{(Ct - X_t) 1 - \gamma}{1 - 1 - \gamma},$$

từ đó xác định SDF (trong đó δ và γ lần lượt là hệ số chiết khấu theo thời gian và mức độ tránh rủi ro). Ngược lại với Campbell và Cochrane (1999) chỉ định dạng hàm rõ ràng cho g , Chen và Ludvigson (2009) sử dụng mạng nơ-ron chuyển tiếp nguồn cấp dữ liệu làm mô hình gần đúng chung cho hàm thói quen.

Giống như Chen và cộng sự. (2021), họ ước tính SDF bằng cách giảm thiểu định mức của các khoanh khắc có điều kiện trong phương trình Euler ngũ ý trong mô hình với dữ liệu tài chính vĩ mô làm công cụ. Một khía cạnh đặc biệt hấp dẫn của phân tích là việc cẩn thận chọn kích thước của mạng lưới thần kinh và tiến hành các bài kiểm tra thống kê phù hợp với lý thuyết tiệm cận mà các bài kiểm tra của họ dựa vào. Đầu tiên, họ mở rộng lý thuyết tiệm cận của Ai và Chen (2003) và Ai và Chen (2007) để ước tính “khoảng cách tối thiểu sàng (SMD)” của mạng lưới thần kinh để điều chỉnh mối tương quan nói tiếp trong

dữ liệu định giá tài sản. Cuối cùng, họ quyết định chọn một mô hình nhỏ theo tiêu chuẩn học máy (chỉ có 18 tham số), nhưng với mẫu chỉ có 200 quan sát, hạn chế này mang lại độ tin cậy cho các thử nghiệm của họ.

Lợi ích từ việc phân tích tiệm cận cẩn thận là khả năng tiến hành kiểm tra giả thuyết về các khía cạnh khác nhau của thói quen kiên trì trong sở thích của nhà đầu tư. Kết quả ủng hộ hàm thói quen phi tuyến tính hơn tuyến tính và

hình thành thói quen bên trong (tức là do lịch sử tiêu dùng của chính một người quy định) chứ không phải là thói quen bên ngoài (hoặc "theo kịp người Jones"). Cuối cùng, họ tiến hành so sánh các lỗi định giá từ mô hình thói quen ước tính của họ với các mô hình định giá tài sản hàng đầu khác như mô hình ba yếu tố (Fama và French, 1993) và mô hình CAPM tỷ lệ tiêu dùng-của cải logarit (Lettau và Ludvigson, 2001). và tìm ra những cải tiến đáng kể trong việc định giá quy mô và danh mục đầu tư được sắp xếp theo giá thị trường bằng cách xây dựng thói quen mạng lưới thần kinh.

5.5.2 Khoảng cách Hansen-Jagannathan

Phương pháp ước tính SDF của GMM (và SMD có liên quan) bắt nguồn từ vấn đề giảm thiểu sai số bình phương về giá. Sai số về giá được xác định theo phương trình Euler và có thể được biểu thị dưới dạng lợi nhuận vượt quá chiết khấu trung bình (phải bằng 0 theo phương trình Euler) hoặc là chênh lệch giữa lợi nhuận vượt quá trung bình chưa chiết khấu và dự đoán dựa trên mô hình của số lượng này. Mặc dù thuận tiện khi so sánh trực tiếp các mô hình về mặt sai số định giá của chúng, nhưng Hansen và Jagannathan (1997) nhấn mạnh rằng công cụ ước tính GMM hiệu quả sử dụng ma trận trọng số phụ thuộc vào mô hình để tổng hợp các sai số định giá, có nghĩa là chúng ta không thể so sánh các mô hình về mặt tối ưu của chúng. - giá trị hàm mất GMM được giảm thiểu. Để giải quyết vấn đề so sánh mô hình, Hansen và Jagannathan (1997) khuyên bạn nên đánh giá tất cả các mô hình theo hàm mất chung có ma trận trọng số độc lập với mô hình. Đặc biệt, họ đề xuất một thước đo khoảng cách có dạng

$$HJ2_{\text{tối}} = \frac{\text{phút}}{\theta_m} e(\theta_m) \Sigma^- 1 e(\theta_m), \quad (5.23)$$

trong đó $e(\theta_m)$ là vectơ lỗi định giá liên quan đến mô hình thứ m (được tham số hóa bởi θ_m). Căn bậc hai của phương trình (5.23) được gọi là "khoảng cách HJ". Có hai khía cạnh quan trọng cần lưu ý trong công thức của nó. Đầu tiên, sai số về giá được tính theo trọng số $\Sigma^- =$ độc lập với mô hình và do đó đặt tất cả các mô hình ngang nhau để so sánh. Thứ hai, các lỗi về giá dựa trên các thông số giảm thiểu khoảng cách HJ, trái ngược với các thông số do mục tiêu GMM hiệu quả quy định.

Mặc dù những khác biệt này so với hàm mục tiêu GMM tiêu chuẩn là rất nhỏ nhưng chúng trang bị cho khoảng cách HJ một số tính chất lý thuyết đặc biệt. Đầu tiên, Hansen và Jagannathan (1997) cho thấy HJm bằng sai số định giá của danh mục tài sản cơ sở bị định giá sai nhất (những tài sản tương ứng với vectơ e) phát sinh từ mô hình m. Nghĩa là, khoảng cách HJ mô tả điều tốt nhất mà mô hình m có thể làm trong việc giảm thiểu lỗi định giá trong trường hợp xấu nhất.¹⁰ Thứ hai, khoảng cách HJ mô tả khoảng cách bình thường nhỏ nhất giữa SDF (có thể bị xác định sai) của mô hình m và họ SDF định giá chính xác tất cả tài sản.

Cả hai thuộc tính này đều ngụ ý rằng khoảng cách HJ là một hàm mục tiêu đặc biệt hấp dẫn để đào tạo các mô hình học máy của SDF và danh mục đầu tư tối ưu. Một mô hình học máy giảm thiểu khoảng cách HJ sẽ mạnh mẽ (vì nó tìm cách giảm thiểu hiệu suất của mô hình trong trường hợp xấu nhất) và sẽ cung cấp một mô hình ít bị xác định sai nhất theo nghĩa a thứ 2. Huấn luyện các mô hình machine learning thông qua khoảng cách HJ phù hợp với lời khuyên của Ludvigson (2013)

“để nhấn mạnh hơn vào nghiên cứu thực nghiệm về các phương pháp tạo điều kiện thuận lợi cho việc so sánh các mô hình bị xác định sai cạnh tranh , đồng thời giảm bớt sự nhấn mạnh vào việc kiểm tra thuyết riêng lẻ về việc liệu một mô hình duy nhất có được chỉ định mà không có lỗi hay không.”

Tiến trình đã bắt đầu trong việc tích hợp khoảng cách HJ vào các bài toán học máy tài chính. Kozak và cộng sự. (2020) liên kết phương pháp ước tính SDF của họ với bài toán giảm thiểu khoảng cách HJ với hình phạt sườn núi (hoặc dây thòng lọng). Khoảng cách của chúng có thể được coi là khoảng cách HJ vô điều kiện áp dụng cho không gian của các thừa số. Gần đây, Gagliardini và Ronchetti (2019), Antoine et al. (2018), Nagel và Singleton (2011) phân tích các phương pháp so sánh mô hình thông qua khoảng cách HJ có điều kiện . Didisheim và cộng sự. (2023) rút ra hành vi lý thuyết của khoảng cách HJ trong các mô hình học máy phức tạp và chỉ ra rằng khoảng cách đang giảm dần theo độ phức tạp của mô hình. Vẫn còn phạm vi rộng lớn cho nghiên cứu trong tương lai về các mô hình học máy được đào tạo thông qua khoảng cách HJ có điều kiện hoặc vô điều kiện .

¹⁰ Danh mục đầu tư trong trường hợp xấu nhất này về cơ bản là “danh mục đầu tư Markowitz” chứa các lỗi định giá, đạt được sai số định giá cao nhất trên mỗi đơn vị biến động.

5.6 Chi phí giao dịch và học tập cung cấp

Tài liệu về máy học tài chính cung cấp một khuôn khổ linh hoạt để kết hợp một số đặc điểm thành một thước đo duy nhất về lợi nhuận kỳ vọng tổng thể (ví dụ: Gu và cộng sự, 2020b). Các tài liệu tương tự ghi lại “tầm quan trọng của đặc điểm” tương đối của các đặc điểm dự đoán lợi nhuận khác nhau (ví dụ: Chen và cộng sự, 2021). Những phát hiện này cho thấy rằng sự thành công trong dự đoán của các phương pháp học máy thường được thúc đẩy bởi các đặc điểm tồn tại trong thời gian ngắn phù hợp với các cổ phiếu nhỏ và kém thanh khoản, ví dụ: Avramov và cộng sự, 2022a, cho rằng chúng có thể ít quan trọng hơn đối với nền kinh tế thực (ví dụ Van Binsbergen và Opp, 2019). Chi phí giao dịch cao của các chiến lược danh mục đầu tư dựa trên học máy ngụ ý rằng những chiến lược này khó thực hiện trong thực tế và rộng hơn là đặt ra câu hỏi về mức độ liên quan và cách giải thích khả năng dự đoán được ghi lại trong tài liệu này. Có phài ước tính lợi nhuận kỳ vọng của máy học chỉ cho chúng ta biết về những lần định giá sai mà các nhà đầu tư không bận tâm đến việc kinh doanh chênh lệch giá vì chi phí quá lớn, việc định giá sai quá thoảng qua và thị trường bị định giá sai quá nhỏ để quan trọng?

Hoặc, liệu các dự đoán của máy học nhận biết chi phí giao dịch cũng có tác dụng đối với các cổ phiếu lớn, trong khoảng thời gian quan trọng và theo cách quan trọng đối với nhiều loại nhà đầu tư, từ đó dẫn đến những hiểu biết kinh tế mới và quan trọng ?

Trong nghiên cứu tài chính, người ta thường tách câu hỏi về khả năng triển khai danh mục đầu tư khỏi thành phần dự đoán lợi nhuận (hoặc ước tính trọng lượng danh mục đầu tư) của vấn đề. Theo cách tiếp cận này, vấn đề ước tính được loại trừ khỏi chi phí giao dịch và doanh thu, và không có gì là khi các chiến lược đầu tư tổng hợp tạo ra lợi nhuận âm sau khi trừ đi chi phí giao dịch. Có một lời giải thích tự nhiên cho kết quả này. Không phải là thị trường không biết có các mô hình dự đoán; đó là các mô hình tồn tại vì chúng quá tốn kém để giao dịch hoặc do các giới hạn đối với hoạt động chênh lệch giá khác. Đây là một thách thức quan trọng đối với các phương pháp dự đoán trong tài chính. Nếu không bằng cách nào đó nhúng các giới hạn chênh lệch giá vào mô hình dự đoán thống kê, mô hình sẽ có xu hướng tách biệt các mô hình dự đoán không thực tế để giao dịch vì đây là những mô hình hiện chưa được khai thác trên thị trường. Điều này không loại trừ khả năng có những hiện tượng có thể dự đoán được và có thể xảy ra trong thực tế.

bị bóc lột; tức là khả năng dự đoán có lợi nhuận sau khi trừ chi phí giao dịch. Điều đó chỉ có nghĩa là nếu không có hướng dẫn về chi phí, mô hình thống kê không thể phân biệt giữa các mô hình dự đoán có thể thực hiện được và không thể thực hiện được. Vì vậy, nếu trung bình các mô hình lợi nhuận có thể dự đoán nổi bật nhất là những mô hình có liên quan đến giới hạn chênh lệch giá (một kịch bản hợp lý do sự cạnh tranh không khoan nhượng trên thị trường chứng khoán), thì đây sẽ là những mô hình đầu tiên bị cô lập bởi dự đoán ML. các mô hình.

Phản này thảo luận về những nỗ lực nhằm tạo ra các danh mục máy học có thể được triển khai thực tế bởi những người tham gia thị trường có tài sản lớn được quản lý, chẳng hạn như các quỹ hưu trí lớn hoặc các nhà quản lý tài sản chuyên nghiệp khác. Nếu một chiến lược có thể thực hiện được trên quy mô lớn thì các biến dự đoán thúc đẩy nhu cầu danh mục đầu tư đó có thể mang lại nhiều thông tin về tỷ lệ chiết khấu cân bằng.

5.6.1 Chi phí giao dịch

Các tài liệu về lựa chọn danh mục đầu tư khi đối mặt với chi phí giao dịch thường giải quyết vấn đề bằng cách giả định rằng hàm chi phí giao dịch đã biết hoặc dữ liệu chi phí giao dịch có sẵn. Điều này loại bỏ sự cần thiết của bước “khám phá” trong thuật toán lựa chọn danh mục học máy, giữ vấn đề ít nhiều trong khung học tập có giám sát tiêu chuẩn và tránh máy học tăng cường. Brandt và cộng sự. (2009) áp dụng phương pháp này khi thiết lập danh mục đầu tư được tham số hóa cơ bản.

Gần đây, Jensen và cộng sự. (2022) đưa hàm chi phí giao dịch đã biết vào mục tiêu lựa chọn danh mục đầu tư, sau đó sử dụng kết hợp học máy và các hạn chế kinh tế để tìm hiểu quy tắc danh mục đầu tư tối ưu.

Chúng tôi phác thảo mô hình và giải pháp học tập của họ ở đây.

Các tài liệu tài chính từ lâu đã phải vật lộn với những thách thức kỹ thuật về xung đột chi phí giao dịch trong việc lựa chọn danh mục đầu tư (ví dụ Baldazzi và Lynch, 1999; Lynch và Baldazzi, 2000). Garleanu và Pedersen (2013) xây dựng danh mục phân tích dựa trên chi phí giao dịch. Kết quả lý thuyết quan trọng của họ là, khi có chi phí giao dịch, danh mục đầu tư tối ưu không chỉ phụ thuộc vào lợi nhuận kỳ vọng $E_t [R_{t+1}]$ mà còn phụ thuộc vào trả về trên tất cả các chân trời tương lai, $E_t [R_{t+\tau}] \quad \tau = 1, 2, \dots$ ¹¹ kỳ vọng của Do đó,

¹¹Theo trực giác, nhà đầu tư thích giao dịch theo các mô hình dự đoán dài hạn vì chi phí giao dịch có thể được khấu hao trong nhiều thời kỳ, trong khi khả năng dự đoán ngắn hạn

Ý tưởng tiếp cận vấn đề dự đoán lợi nhuận bằng học máy trong khi tính toán chi phí giao dịch đang gặp phải căng thẳng. Garleanu và Pedersen (2013) đạt được giải pháp danh mục đầu tư dễ điều chỉnh bằng cách liên kết tất cả các phạm vi với nhau bằng đặc tả tự hồi quy tuyến tính, giúp duy trì tham số hóa ở mức nhỏ và thu gọn chuỗi kỳ vọng thu gọn thành hàm tuyến tính của các giá trị tín hiệu phổ biến. Học máy thường tìm kiếm sự linh hoạt trong mô hình dự đoán lợi nhuận; nhưng nếu cần các mô hình dự đoán lợi nhuận riêng biệt cho tất cả các tầm nhìn trong tương lai thì môi trường sẽ nhanh chóng trở thành một đặc điểm kỹ thuật khó hiểu ngay cả theo các tiêu chuẩn học máy. Làm thế nào chúng ta có thể duy trì tính linh hoạt trong đặc tả mô hình mà không đưa ra các giả định hạn chế về động lực của lợi nhuận kỳ vọng? 12

Jensen và cộng sự. (2022) giải quyết căng thẳng này bằng ba đổi mới quan trọng. Đầu tiên, chúng làm giảm tính tuyến tính của lợi nhuận kỳ vọng trong các tín hiệu dự đoán và cho phép thực hiện các hàm dự đoán phi tuyến. Thứ hai, họ nới lỏng giả định về lợi nhuận kỳ vọng tự hồi quy của Garleanu và Pedersen (2013) bằng cách yêu cầu động lực chuỗi thời gian dừng nhưng không hạn chế. Từ hai giải pháp này, về mặt lý thuyết, họ có thể rút ra quy tắc danh mục đầu tư như một hàm chung của các tín hiệu điều hòa. Đóng góp thứ ba của họ là chỉ ra rằng một thành phần của quy tắc danh mục đầu tư này (chức năng "mục tiêu" mà nhà đầu tư chuyển dần dần danh mục đầu tư của mình sang) có thể được tham số hóa bằng đặc tả mạng thần kinh, giúp vượt qua khó khăn đã nói ở trên khi yêu cầu các mô hình dự đoán lợi nhuận học máy khác nhau. cho các tầm nhìn đầu tư khác nhau. 13 Giải pháp của họ là một ví dụ về mô hình kết hợp lợi ích của cơ cấu kinh tế với tính linh hoạt của học máy. Kinh tế học xuất hiện dưới dạng một giải pháp cho vấn đề tối đa hóa tiện ích của nhà đầu tư, vấn đề này áp đặt một hình thức nhất quán năng động lên quy tắc danh mục đầu tư. Chỉ sau bước này, chức năng học máy linh hoạt mới được đưa vào quy tắc danh mục đầu tư trong khi vẫn duy trì tính nhất quán về mặt kinh tế của quy tắc.

nhanh chóng bị ăn thịt do điều chỉnh danh mục đầu tư thường xuyên.

12Một căng thẳng bổ sung khác trong bối cảnh này là, nếu một giải pháp danh mục đầu tư chi phí giao dịch phải xây dựng dự báo cho nhiều phạm vi lợi nhuận, thì số lượng quan sát đào tạo sẽ giảm khi phạm vi dự báo tăng lên, do đó nhìn chung có rất ít dữ liệu để cung cấp thông tin dự báo cho phạm vi dài hạn.

13Điều này đòi hỏi giả định rằng các tín hiệu là Markovian, nhưng động lực học của chúng không cần phải xác định thêm.

5.6. Chi phí giao dịch và học tập củng cố

129

Cụ thể, nhà đầu tư tìm kiếm một danh mục đầu tư có trọng lượng để tối đa hóa hữu dụng bậc hai theo chi phí xòng:

$$\text{tiện ích} = E \mu(st) wt - \frac{\kappa t}{2} (wt - gtwt - 1) \Lambda (wt - gtwt - 1) - \frac{\gamma t}{2} w t \Sigma wt \quad (5.24)$$

trong đó $\mu(st)$, Σ và Λ tóm tắt tương ứng giá trị tài sản, hiệp phương sai và chi phí giao dịch. Chúng thay đổi theo thời gian theo những cách có thể dự đoán được liên quan đến tập biến điều hòa st , mặc dù vì mục đích ở đây chúng ta thảo luận về mô hình đơn giản hóa với hiệp phương sai tĩnh và chi phí giao dịch. wt là tài sản nhà đầu tư đang quản lý (quy mô danh mục đầu tư của nhà đầu tư là yếu tố đầu tiên được xem xét đối với chi phí giao dịch) và gt là mức tăng trưởng tài sản so với giai đoạn trước. Đề xuất 3 của Jensen et al. (2022) cho thấy quy tắc danh mục đầu tư tối ưu khi có chi phí giao dịch là

$$wt \approx mgwt - 1 + (I - m)Tại \quad (5.25)$$

trong đó nhà đầu tư giao dịch một phần từ danh mục kế thừa $wt - 1$ sang danh mục đầu tư "mục tiêu". Tại thời điểm t :

$$Tại = (I - m) \sum_{\tau=0}^{\infty} (\mu(st+\tau) - g^-(st+\tau)) \frac{1}{\Sigma Y^{st+\tau}} \cdot \quad (5.26)$$

Danh mục mục tiêu phụ thuộc vào m , hàm ma trận phi tuyến tính của chi phí giao dịch và hiệp phương sai, cũng như tốc độ tăng trưởng danh mục đầu tư trung bình $g^- = E[gt]$.

Danh mục đầu tư mục tiêu là mức trung bình được làm mịn theo cấp số nhân của các danh mục đầu tư Markowitz theo từng thời kỳ, điều này phản ánh thực tế rằng việc điều chỉnh rất tốn kém nên sẽ có lợi khi làm mịn các quan điểm danh mục đầu tư theo thời gian.

Giải pháp danh mục đầu tư có thể được viết lại dưới dạng tổng vô hạn của các danh mục mục tiêu trong quá khứ và sự tăng trưởng của chúng theo thời gian:

$$wt = \sum_{\theta=0}^{\infty} \sum_{\tau=1}^{\infty} m gt \tau+1 (I - m) A(st, \theta). \quad (5.27)$$

Giải pháp này trong (5.27) bao gồm tất cả cấu trúc kinh tế năng động có liên quan cần thiết cho một danh mục đầu tư tối ưu khi có chi phí giao dịch. Nhưng nó cũng cho thấy rằng, nếu chúng ta thay thế giải pháp này vào bài toán tối đa hóa lợi ích ban đầu thì nhiệm vụ học tập có thể bị giảm bớt.

để tìm hàm danh mục đầu tư mục tiêu $A(\cdot)$ tối đa hóa hữu dụng kỳ vọng. Dựa trên cái nhìn sâu sắc này, Jensen et al. (2022) sử dụng đặc tả mạng thần kinh có đặc điểm Fourier ngẫu nhiên (như trong Kelly và cộng sự, 2022a; Didisheim và cộng sự, 2023), nhưng thay vì xấp xỉ hàm dự đoán trả về, chúng gần đúng trực tiếp với hàm mục tiêu, $A(\cdot) = f(st)\beta$ cho f một tập hợp P hàm đặc trưng ngẫu nhiên đã biết và β là vectơ hệ số hồi quy trong R

P . Theo kinh nghiệm, thuật toán học danh mục đầu tư một bước, nhận biết chi phí giao dịch của họ mang lại hiệu suất chi phí ròng vượt trội so với các phương pháp tiếp cận hai bước tiêu chuẩn trong tài liệu trước tiên tìm hiểu danh mục đầu tư bắt khả tri về chi phí giao dịch, sau đó làm tròn danh mục đầu tư đã học theo thời gian để giảm thiểu chi phí.

5.6.2 Học tăng cường

Học tăng cường nói chung bao gồm các mô hình học máy để giải quyết vấn đề ra quyết định tuần tự trong điều kiện không chắc chắn. Chúng đặc biệt mạnh mẽ đối với các môi trường mô hình hóa trong đó một tác nhân thực hiện một hành động để tối đa hóa một số chức năng phần thưởng tích lũy nhưng

trạng thái của hệ thống và do đó việc phân phối các kết quả trong tương lai bị ảnh hưởng bởi hành động của tác nhân. Trong trường hợp này, các lựa chọn của tác nhân là các biến điều kiện quan trọng để tìm hiểu hàm hoàn trả. Nhưng vì hầu hết các cặp trạng thái/hành động không được quan sát trong quá khứ nên dữ liệu có thể được coi là phần lớn chưa được gắn nhãn và do đó mô hình không phù hợp để giám sát trực tiếp. Bởi vì các nhãn có liên quan chỉ được thúc đẩy bởi hành động của tác nhân, phần lớn việc học tăng cường sẽ phân chia nỗ lực giữa việc học bằng thử nghiệm ("khám phá") và tối ưu hóa phần thưởng mong đợi trong tương lai dựa trên những gì họ đã học được ("khai thác").

Việc tác nhân tìm cách tối ưu hóa các phần thưởng trong tương lai có nghĩa là việc học tăng cường có thể là một công cụ có giá trị để lựa chọn danh mục đầu tư. Tuy nhiên, trong bài toán lựa chọn danh mục đầu tư cơ bản đã nêu trước đó, trạng thái của hệ thống không bị ảnh hưởng bởi các quyết định của nhà đầu tư. Nhà đầu tư được coi là người chấp nhận giá, do đó, bất kỳ danh mục đầu tư nào mà cô ấy xây dựng đều không làm ảnh hưởng đến động lực của rủi ro và lợi nhuận trong tương lai. Do đó, với dữ liệu đã cho về lợi nhuận thực hiện của tài sản cơ sở, việc lựa chọn danh mục đầu tư máy học có thể được thực hiện bằng các kỹ thuật học có giám sát. Nói cách khác, lợi ích của việc học tăng cường đối với người chấp nhận giá có thể bị hạn chế.

5.6. Chi phí giao dịch và học tập củng cố

131

Tất nhiên, giả định về người chấp nhận giá là không thực tế đối với nhiều nhà đầu tư. Các trung gian tài chính như nhà quản lý tài sản và ngân hàng thường là những người tham gia tích cực nhất trên thị trường tài chính. Sự nổi bật của họ với tư cách là "nhà giao dịch cận biên" và xu hướng giao dịch khối lượng lớn có nghĩa là các quyết định về danh mục đầu tư của họ có tác động đến giá và do đó làm thay đổi trạng thái của hệ thống. Một nhà đầu tư như vậy phải tìm hiểu xem các quyết định của họ ảnh hưởng như thế nào đến động lực thị trường và điều này tạo ra động cơ khuyến khích thử nghiệm các hành động như một phần của thuật toán học tập. Nghĩa là, một khi các nhà đầu tư phải nội hóa tác động lên giá của chính họ thì các công cụ học hỏi tăng cường sẽ rất phù hợp với vấn đề lựa chọn danh mục đầu tư.

Các tài liệu tài chính chính thống có rất ít công trình nghiên cứu về học tập tăng cường để lựa chọn danh mục đầu tư,¹⁴ mặc dù có rất nhiều tài liệu khoa học máy tính về chủ đề này. Tác động giá của nhà đầu tư trong việc quyết định phần thưởng trong tương lai của họ càng nổi bật thì các phương pháp học tập củng cố càng có giá trị. Do đó, xu hướng tập trung vào các chiến lược đầu tư tần suất tương đối thấp của ngành tài chính khiến không có gì ngạc nhiên khi việc học tăng cường không được duy trì, mặc dù chúng tôi kỳ vọng điều này sẽ thay đổi trong những năm tới.

Trong khi đó, tài liệu về khoa học máy tính phần lớn đã áp dụng phương pháp học tăng cường cho các vấn đề danh mục đầu tư có tần suất cao hơn liên quan đến việc tạo lập thị trường và thực hiện giao dịch. Mặc dù chúng tôi không trình bày tài liệu này ở đây nhưng đối với những độc giả quan tâm, chúng tôi khuyên bạn nên thực hiện khảo sát của Hamblly et al. (2022).

¹⁴Cong và cộng sự. (2020) là một ví dụ.

6

Kết luận

Chúng tôi cố gắng đề cập đến các tài liệu đang phát triển về học máy tài chính. Mục tiêu chính trong công việc của chúng tôi là giúp độc giả nhận ra máy học là một công cụ không thể thiếu để phát triển sự hiểu biết của chúng ta về các hiện tượng thị trường tài chính. Chúng tôi nhấn mạnh các lĩnh vực nhận được nhiều sự quan tâm nghiên cứu nhất cho đến nay, bao gồm dự đoán lợi nhuận, mô hình nhân tố rủi ro và lợi nhuận, yếu tố chiết khấu ngẫu nhiên và lựa chọn danh mục đầu tư.

Thật không may, phạm vi của cuộc khảo sát này đã buộc chúng tôi phải hạn chế hoặc bỏ qua phạm vi đề cập đến một số chủ đề quan trọng về học máy tài chính. Một chủ đề được hưởng lợi từ phương pháp học máy bị bỏ qua là mô hình hóa rủi ro. Điều này bao gồm các mô hình phương sai và hiệp phương sai có điều kiện, đặc biệt là mô hình hóa ma trận hiệp phương sai nhiều chiều. Một lượng lớn tài liệu sử dụng các phương pháp học máy để cải thiện việc ước tính ma trận hiệp phương sai và từ đó cải thiện hiệu suất của danh mục đầu tư được tối ưu hóa. Liên quan chặt chẽ đến mô hình rủi ro là chủ đề định giá phái sinh. Trên thực tế, một số ứng dụng sớm nhất của mạng lưới thần kinh trong tài chính liên quan đến việc định giá quyền chọn và các bề mặt biến động nguy ý. Các ví dụ nổi bật về học máy hoặc mô hình phi tham số cho nghiên cứu phái sinh bao gồm Ait-Sahalia và Lo (1998) và

Ait-Sahalia và Lo (2000), Anders và cộng sự. (1998), Rosenberg và Engle (2002), Bollerslev và Todorov (2011), Israelov và Kelly (2017), cùng nhiều người khác.

Hầu hết các nỗ lực nghiên cứu cho đến nay đều liên quan đến học máy để cải thiện hiệu suất trong các nhiệm vụ dự đoán. Đây là định của tăng bằng. Một hướng quan trọng cho phân tích học máy tài chính thế hệ tiếp theo là làm sáng tỏ hơn về các cơ chế kinh tế và trạng thái cân bằng. Một cách khác là sử dụng các phương pháp học máy để giải các mô hình cấu trúc phức tạp và có tính phi tuyến cao. Riêng biệt, bản chất tiến hóa của nền kinh tế và thị trường, được định hình bởi sự thay đổi công nghệ và môi trường pháp lý thay đổi, đặt ra cho các nhà kinh tế thách thức trong việc lập mô hình thay đổi cơ cấu. Một hướng nghiên cứu tiềm năng thú vị là tận dụng các phép tính gần đúng mô hình linh hoạt mà học máy mang lại để phát hiện và thích ứng tốt hơn với những thay đổi cấu trúc.

Trong khi cuộc khảo sát này tập trung vào khía cạnh định giá tài sản của tài chính, học máy đang xâm nhập vào các lĩnh vực khác như tài chính doanh nghiệp, khởi nghiệp, tài chính hộ gia đình và bất động sản. Ví dụ : Hu và Ma (2020) sử dụng kỹ thuật máy học để xử lý các đoạn video quảng cáo của các doanh nhân đầy tham vọng nhằm kiểm tra vai trò của khả năng thuyết phục của diễn giả đối với thành công về tài chính ở giai đoạn đầu. Li và cộng sự. (2021) sử dụng phân tích văn bản về các cuộc gọi thu nhập để định lượng văn hóa doanh nghiệp và tác động của nó đến kết quả của công ty. Lyonnet và Stern (2022) sử dụng máy học để nghiên cứu cách các nhà đầu tư mạo hiểm đưa ra quyết định đầu tư. Erel và cộng sự. (2021) cung cấp bằng chứng cho thấy thuật toán học máy có thể giúp các công ty tránh được những lựa chọn giám đốc công ty tồi. Fuster và cộng sự. (2022) định lượng tác động cân bằng của thuật toán học máy trên thị trường thế chấp. Mặc dù học máy đã được sử dụng rộng rãi hơn trong việc định giá tài sản, nhưng ứng dụng sâu hơn của nó vào các vấn đề tài chính doanh nghiệp là một lĩnh vực nghiên cứu thú vị trong tương lai.

Người giới thiệu

- Ahn, SC và J. Bae. (2022). "Dự báo với bình phương nhỏ nhất một phần khi có sẵn số lượng lớn các yếu tố dự đoán". Công nghệ. trả lời. Đại học bang Arizona và Đại học Glasgow.
- Ai, C. và X. Chen. (2003). "Ước tính hiệu quả của các mô hình với các hạn chế về thời điểm có điều kiện chứa các hàm không xác định". Kinh tế lượng. 71(6): 1795–1843.
- Ai, C. và X. Chen. (2007). "Ước tính các mô hình hạn chế thời điểm có điều kiện bán tham số có thể bị xác định sai với các biến điều hòa khác nhau". Tạp chí Kinh tế lượng. 141(1): 5–43.
- Ait-Sahalia, Y. và MW Brandt. (2001). "Lựa chọn biến đổi cho lựa chọn danh mục đầu tư". Tạp chí Tài chính. 56: 1297–1351.
- Ait-Sahalia, Y., J. Fan, L. Xue và Y. Chu. (2022). "Làm thế nào và khi nào có thể dự đoán được lợi nhuận cổ phiếu có tần suất cao?" Công nghệ. trả lời. Trường Đại học Princeton .
- Ait-Sahalia, Y., J. Jacod và D. Xiu. (2021). "Hồi quy Fama- MacBeth theo thời gian liên tục". Công nghệ. trả lời. Đại học Princeton và Đại học Chicago.
- Ait-Sahalia, Y., I. Kalnina và D. Xiu. (2020). "Mô hình nhân tố tần số cao và hồi quy". Tạp chí Kinh tế lượng. 216: 86–105.
- Ait-Sahalia, Y. và AW Lo. (1998). "Ước tính phi tham số về mật độ giá nhà nước tiềm ẩn trong giá tài sản tài chính". Tạp chí tài chính. 53(2): 499–547.

- Ait-Sahalia, Y. và AW Lo. (2000). "Quản lý rủi ro phi tham số và ác cảm rủi ro ngụ ý". Tạp chí kinh tế lượng. 94(1-2): 9-51.
- Ait-Sahalia, Y. và D. Xiu. (2017). "Sử dụng phân tích thành phần chính để ước tính mô hình nhân tố thứ nguyên cao với dữ liệu tần số cao". Tạp chí Kinh tế lượng. 201: 388-399.
- Ait-Sahalia, Y. và D. Xiu. (2019). "Phân tích thành phần chính của dữ liệu tần số cao". Tạp chí của Hiệp hội Thống kê Hoa Kỳ . 114: 287-303.
- Allen-Zhu, Z., Y. Li và Z. Song. (2019). "Lý thuyết hội tụ cho học sâu thông qua quá trình tham số hóa". Trong: Hội nghị quốc tế về học máy. PMLR. 242-252.
- Altman, EI (1968). "Các tỷ lệ tài chính, phân tích phân biệt và dự đoán phá sản doanh nghiệp". Tạp chí Tài chính. 23(4): 589-609.
- Anders, U., O. Korn và C. Schmitt. (1998). "Cải thiện việc định giá các lựa chọn: Cách tiếp cận mạng lưới thần kinh". Tạp chí dự báo. 17(5-6): 369-388.
- Andersen, TG và T. Bollerslev. (1998). "Trả lời những người hoài nghi: Có, các Mô hình biến động tiêu chuẩn cung cấp những dự báo chính xác". Tạp chí kinh tế quốc tế. 39: 885-905.
- Andersen, TG, T. Bollerslev, FX Diebold và P. Labys. (2001). "Việc phân phối biến động tỷ giá hối đoái được thực hiện". Tạp chí của Hiệp hội Thống kê Hoa Kỳ. 96: 42-55.
- Antoine, B., K. Proulx và E. Renault. (2018). "SDF giả thật trong các mô hình định giá tài sản có điều kiện*". Tạp chí Kinh tế lượng tài chính . 18(4): 656-714. issn: 1479-8409. doi: [10.1093/jjfinec/nby017](https://doi.org/10.1093/jjfinec/nby017). eprint: <https://academic.oup.com/jfec/article-abstract/18/4/656/35053168/nby017.pdf>. url: <https://doi.org/10.1093/jjfinec/nby017>.
- Ao, M., L. Yingying và X. Zheng. (2018). "Tiếp cận hiệu quả phương sai trung bình cho danh mục đầu tư lớn". Tạp chí nghiên cứu tài chính. 32(7): 2890-2919.
- Arlot, S. và A. Celisse. (2010). "Khảo sát các thủ tục xác nhận chéo để lựa chọn mô hình". Khảo sát thống kê. 4: 40-79.
- Aubry, M., R. Kraussl, M. Gustavo và C. Spaenjers. (2022). "Những nhà đầu giá thiên vị". Tạp chí Tài chính, sắp xuất bản.

- Avramov, D., S. Cheng và L. Metzker. (2022a). "Học máy so với Hạn chế về kinh tế: Bằng chứng từ khả năng dự đoán lợi nhuận hàng tồn kho". Khoa học quản lý, sắp xuất bản.
- Avramov, D., S. Cheng, L. Metzker và S. Voigt. (2021). "Mô hình nhân tố tích hợp". Tạp chí Tài chính, sắp xuất bản.
- Avramov, D., G. Kaplanski và A. Subrahmanyam. (2022b). "Các yếu tố cơ bản về sự trôi dạt giá trong thị trường vốn: Quan điểm chính quy hóa hồi quy". Khoa học quản lý. 68(10): 7658-7681.
- Avramov, D. và G. Chu. (2010). "Phân tích danh mục đầu tư Bayesian". Đánh giá thường niên về kinh tế tài chính. 2(1): 25-47.
- Bài, J. (2003). "Lý thuyết suy luận cho mô hình nhân tố có kích thước lớn". Kinh tế lượng. 71(1): 135-171.
- Bài, J. và S. Ng. (2002). "Xác định số lượng nhân tố trong mô hình nhân tố gần đúng". Kinh tế lượng. 70: 191-221.
- Bài, J. và S. Ng. (2021). "Các mô hình nhân tố gần đúng với tài yếu hơn". Công nghệ. trả lời. Đại học Columbia.
- Bajgrowicz, P. và O. Scaillet. (2012). "Xem lại giao dịch kỹ thuật: Khám phá sai lầm, kiểm tra tính kiên trì và chi phí giao dịch". Tạp chí Kinh tế Tài chính. 106(3): 473-491.
- Baker, M. và J. Wurgler. (2006). "Tâm lý nhà đầu tư và mặt cắt ngang của lợi nhuận chứng khoán". Tạp chí Tài chính. 61(4): 1645-1680.
- Baker, M. và J. Wurgler. (2007). "Tâm lý nhà đầu tư trên thị trường chứng khoán". Tạp chí quan điểm kinh tế. 21(2): 129-152.
- Baldazzi, P. và AW Lynch. (1999). "Chi phí giao dịch và khả năng dự đoán trước: một số tính toán chi phí tiện ích". Tạp chí Kinh tế Tài chính. 52: 47-78.
- Bali, TG, A. Goyal, D. Huang, F. Jiang và Q. Wen. (2020). "Dự đoán lợi nhuận trái phiếu doanh nghiệp: Merton đáp ứng công nghệ học máy". Công nghệ. trả lời. Đại học Georgetown.
- Bansal, R. và S. Viswanathan. (1993). "Không có chênh lệch giá và định giá chênh lệch giá: Một cách tiếp cận mới". Tạp chí Tài chính. 48(4): 1231-1262 .
- Bansal, R. và A. Yaron. (2004). "Rủi ro trong dài hạn: Giải pháp tiềm năng cho bài toán định giá tài sản". Tạp chí Tài chính. 59(4): 1481-1509.

- Bảo, W., J. Yue và Y. Rao. (2017). "Một khung học sâu dành cho chuỗi thời gian tài chính sử dụng bộ mã hóa tự động xếp chồng và bộ nhớ ngắn hạn dài hạn ". XIN MỘT. 12(7): 1-24.
- Barberis, N. (2018). "Các mô hình giá tài sản và khôi lượng giao dịch dựa trên tâm lý học". Trong: Sổ tay kinh tế học hành vi: Ứng dụng và cơ sở 1. Tập. 1. Khác. 79-175.
- Barberis, N. và R. Thaler. (2003). "Một cuộc khảo sát về tài chính hành vi". Sổ tay Kinh tế Tài chính. 1: 1053-1128.
- Barndorff-Nielsen, O.E và N. Shephard. (2002). "Phân tích kinh tế lượng về biến động thực tế và việc sử dụng nó trong việc ước tính các mô hình biến động ngẫu nhiên". Tạp chí của Hiệp hội Thống kê Hoàng gia, B. 64: 253-280.
- Barras, L., O. Scailliet và R. Wermers. (2010). "Những khám phá sai lầm về hoạt động của quỹ tương hỗ: Đo lường may mắn bằng các bảng chữ cái ước tính". Tạp chí Tài chính. 65(1): 179-216.
- Bartlett, PL, PM Long, G. Lugosi và A. Tsigler. (2020). "Áp dụng quá mức lành tính trong hồi quy tuyến tính". Kỷ yếu của Viện Hàn lâm Khoa học Quốc gia. 117(48): 30063-30070.
- Basu, S. (1977). "Hiệu suất đầu tư của cổ phiếu phổ thông liên quan đến tỷ lệ giá trên thu nhập của chúng: Kiểm tra giả thuyết thị trường hiệu quả". Tạp chí Tài chính. 32(3): 663-682.
- Belkin, M., D. Hsu, S. Ma và S. Mandal. (2018). "Dung hòa việc học máy hiện đại và sự đánh đổi độ lệch. bản in điện tử arXiv".
- Belkin, M. (2021). "Vừa vặn mà không sợ hãi: những hiện tượng toán học đáng chú ý của học sâu qua lăng kính nội suy". Acta Số. 30: 203-248.
- Belkin, M., D. Hsu và J. Xu. (2020). "Hai mô hình giảm dần cho các tính năng yếu". Tạp chí SIAM về Toán học Khoa học Dữ liệu. 2(4): 1167-1180.
- Belkin, M., A. Rakhlin và AB Tsybakov. (2019). "Việc nội suy dữ liệu có mâu thuẫn với tính tối ưu về mặt thống kê không?" Trong: Hội nghị quốc tế lần thứ 22 về trí tuệ nhân tạo và thống kê. PMLR. 1611- 1619.

- Stewamini, Y. và Y. Hochberg. (1995). "Kiểm soát tỷ lệ phát hiện sai: một cách tiếp cận thực tế và mạnh mẽ để thử nghiệm nhiều lần". Tạp chí của Hiệp hội Thống kê Hoàng gia. Loạt B (Phương pháp luận). 57(1): 289-300.
- Bianchi, D., M. Büchner và A. Tamoni. (2021). "Phí bù rủi ro trái phiếu nhờ học máy". Tạp chí nghiên cứu tài chính. 34(2): 1046-1089.
- Bickel, PJ và E. Levina. (2008a). "Chính quy hóa hiệp phương sai bằng ngưỡng". Biên niên sử thống kê. 36(6): 2577-2604.
- Bickel, PJ và E. Levina. (2008b). "Ước tính chính quy của ma trận hiệp phương sai lớn". Biên niên sử thống kê. 36: 199-227.
- Đen, F. và R. Litterman. (1992). "Tối ưu hóa danh mục đầu tư toàn cầu". Tạp chí phân tích tài chính. 48(5): 28-43.
- Bollerslev, T., SZ Li và V. Todorov. (2016). "Tăng cường phiên bản beta: Beta liên tục và không liên tục và mặt cắt ngang của lợi nhuận cổ phiếu kỳ vọng". Tạp chí Kinh tế Tài chính. 120: 464-490 .
- Bollerslev, T., MC Medeiros, A. Patton và R. Quaedvlieg. (2022). "Từ số 0 đến anh hùng: Các phương sai (Co) từng phần được hiện thực hóa". Tạp chí Kinh tế lượng. 231: 348-360.
- Bollerslev, T. và V. Todorov. (2011). "Ước tính các bước nhảy đuôi". Kinh tế lượng. 79(6): 1727-1783.
- Box, GEP, GM Jenkins, GC Reinsel và GM Ljung. (2015). Phân tích chuỗi thời gian: Dự báo và kiểm soát. ngày 5. Wiley.
- Höp, GE và G. Jenkins. (1970). Phân tích chuỗi thời gian: Dự báo và kiểm soát. San Francisco: Ngày Holden.
- Brandt, MW (1999). "Ước tính danh mục đầu tư và lựa chọn tiêu dùng: Phương pháp tiếp cận phương trình Euler có điều kiện". Tạp chí Tài chính. 54(5): 1609-1645.
- Brandt, MW (2010). "Vấn đề lựa chọn danh mục đầu tư". Trong: Sổ tay kinh tế lượng tài chính. Ed. của Y. Ait-Sahalia và LP Hansen. Amsterdam, Hà Lan: Bắc Hà Lan. 269-336.
- Brandt, MW và P. Santa-Clara. (2006). "Lựa chọn danh mục đầu tư năng động bằng cách tăng không gian tài sản". Tạp chí Tài chính. 61(5): 2187-2217.

Brandt, MW, P. Santa-Clara và R. Valkanov. (2009). "Chính quy hóa hiệp phương sai bằng các chính sách danh mục đầu tư tham số: Khai thác các đặc điểm trong mặt cắt ngang của lợi nhuận vốn cổ phần". Đánh giá các nghiên cứu tài chính. 22: 3411-3447.

Breiman, L. (1995). "Toán học tổng quát hóa". Trong: Nhà xuất bản CRC. Chap. Suy ngẫm sau khi làm báo cáo trọng tài cho NIPS. 11-15.

Breiman, L. (2001). "Rừng ngẫu nhiên". Học máy. 45(1): 5-32.

Britten-Jones, M. (1999). "Lỗi lấy mẫu trong ước tính trọng số danh mục đầu tư hiệu quả trung bình-phương sai". Tạp chí Tài chính. 54(2): 655-671.

Brown, T., B. Mann, N. Ryder, M. Subbiah, JD Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, TÔI. Sutskever và D. Amodei. (2020). "Mô hình ngôn ngữ là những người học ít cơ hội". Trong: Những tiến bộ trong hệ thống xử lý thông tin thần kinh. Ed. của H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan và H. Lâm. Tập. 33. Hiệp hội Curran, Inc. 1877-1901.

Bryzgalova, S., V. DeMiguel, S. Li và M. Pelger. (2023). "Các yếu tố định giá tài sản với mục tiêu kinh tế". Có sẵn tại SSRN 4344837.

Bryzgalova, S., M. Pelger và J. Zhu. (2020). "Rừng xuyên rừng : Xây dựng mặt cắt ngang của lợi nhuận tài sản". Công nghệ. trả lời. Trường Kinh doanh Luân Đôn và Đại học Stanford.

Buchner, M. và BT Kelly. (2022). "Mô hình nhân tố cho lợi nhuận quyền chọn". Tạp chí Kinh tế Tài chính.

Bybee, L., L. Gomes và J. Valente. (2023a). "Các yếu tố vĩ mô cho mặt cắt ngang của lợi nhuận tiền tệ".

Bybee, L., BT Kelly, A. Manela và D. Xiu. (2020). "Cấu trúc của tin tức kinh tế". Công nghệ. trả lời. Cục nghiên cứu kinh tế quốc gia.

Bybee, L., BT Kelly và Y. Su. (2023b). "Định giá tài sản tường thuật: Các yếu tố rủi ro hệ thống có thể giải thích được từ văn bản tin tức". Đánh giá các nghiên cứu tài chính.

Cai, T. và W. Liu. (2011). "Ngưỡng thích ứng cho ước tính ma trận hiệp phương sai thưa thớt". Tạp chí của Hiệp hội Thống kê Hoa Kỳ. 106: 672-684.

- Campbell, JY và JH Cochrane. (1999). "Bằng sức ép của thói quen: Giải thích dựa trên tiêu dùng về hành vi tổng hợp của thị trường chứng khoán".
Tạp chí Kinh tế chính trị. 107(2): 205-251.
- Campbell, JY và SB Thompson. (2008). "Dự đoán lợi nhuận hàng tồn kho dựa trên từ mẫu: Có gì có thể đánh bại mức trung bình lịch sử không?"
Tạp chí nghiên cứu tài chính. 21(4): 1509-1531.
- Campbell, JY và RJ Shiller. (1988). "Giá cổ phiếu, thu nhập và cổ tức dự kiến". Tạp chí Tài chính. 43(3): 661-676.
- Chamberlain, G. và M. Rothschild. (1983). "Chênh lệch giá, cấu trúc nhân tố và phân tích phương sai trung bình trên thị trường tài sản lớn". Kinh tế - méta. 51: 1281-1304.
- Chatelais, N., A. Stalla-Bourdillon và MD Chinn. (2023). "Dự báo hoạt động thực tế bằng cách sử dụng thông tin thị trường chứng khoán liên ngành".
Tạp chí Tiền tệ và Tài chính Quốc tế. 131: 102800.
- Chen, AY và T. Zimmermann. (2021). "Định giá tài sản chéo nguồn mở". Đánh giá tài chính quan trọng, sắp ra mắt.
- Chen, B., Q. Yu và G. Chu. (2023). "Các yếu tố hữu ích ít hơn bạn nghĩ". Có sẵn tại SSRN 3723126.
- Chen, H., WW Dou và L. Kogan. (2022a). "Đo lường "Vật chất tối" trong các mô hình định giá tài sản". Tạp chí Tài chính, sắp xuất bản.
- Chen, J., G. Tang, J. Yao và G. Chu. (2022b). "Sự chú ý của nhà đầu tư và lợi nhuận cổ phiếu". Tạp chí phân tích tài chính và định lượng. 57(2): 455-484.
- Chen, L., M. Pelger và J. Zhu. (2021). "Học sâu về định giá tài sản".
SSRN.
- Chen, X. và SC Ludvigson. (2009). "Đất của người nghiên cứu? một cuộc điều tra thực nghiệm về các mô hình định giá tài sản dựa trên thói quen". Tạp chí Kinh tế lượng ứng dụng. 24(7): 1057-1093.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, WK Newey và J. Robins. (2018). "Học máy kép/không thiên vị cho các thông số cấu trúc và điều trị". Tạp chí Kinh tế lượng. 21(1): C1-C68.
- Chib, S., L. Zhao và G. Chu. (2023). "Người chiến thắng từ người chiến thắng: Một câu chuyện của các yếu tố rủi ro". Khoa học quản lý.
- Chinco, A., AD Clark-Joseph và M. Ye. (2019). "Tín hiệu thưa thớt trong mặt cắt ngang của lợi nhuận". Tạp chí Tài chính. 74(1): 449-492.

Cho, K., B. van Merriënboer, D. Bahdanau và Y. Bengio. (2014). "Về các đặc tính của dịch máy thần kinh: Các phương pháp tiếp cận bộ mã hóa-giải mã". Trong: Kỷ yếu SSST-8, Hội thảo lần thứ 8 về Cú pháp, Ngữ nghĩa và Cấu trúc trong Dịch thuật Thống kê. 103-111. doi: 10.3115/v1/W14-4012. url: <https://aclanthology.org/W14-4012>.

Choi, D., W. Jiang và C. Zhang. (2022). "Alpha đi khắp mọi nơi: Học máy và lợi nhuận chứng khoán quốc tế". Công nghệ. trả lời. Đại học Trung Quốc Hồng Kông.

Chong, E., C. Han và FC Park. (2017). "Mạng lưới học sâu để phân tích và dự đoán thị trường chứng khoán: Phương pháp, trình bày dữ liệu và nghiên cứu trường hợp". Hệ thống chuyên gia với các ứng dụng. 83: 187-205.

Cochrane, JH (2009). Định giá tài sản: Phiên bản sửa đổi. Nhà xuất bản Đại học Princeton .

Cochrane, JH (2008). "Con chó không sửa: Cách bảo vệ khả năng quay trở lại có thể đoán trước được". Tạp chí nghiên cứu tài chính. 21(4): 1533-1575.

Cochrane, JH và M. Piazzesi. (2005). "Phí bù rủi ro trái phiếu". Người Mỹ Tạp chí kinh tế. 95(1): 138-160.

Công, LW, G. Feng, J. He và X. He. (2022). "Định giá tài sản bằng cây bảng điều khiển theo tiêu chí phân chia toàn cầu". Công nghệ. trả lời. Đại học Thành phố Hồng Kông.

Công, LW, K. Tang, J. Wang và Y. Zhang. (2020). "AlphaPortfolio dành cho đầu tư và AI có thể giải thích về mặt kinh tế". Có sẵn tại SSRN.

Connor, G., M. Hagmann và O. Linton. (2012). "Ước tính bán tham số hiệu quả của mô hình Fama-Pháp và các phần mở rộng". Kinh tế - méta. 80(2): 713-754.

Connor, G. và RA Korajczyk. (1986). "Đo lường hiệu suất với lý thuyết định giá chênh lệch giá: Một khuôn khổ phân tích mới".

Tạp chí Kinh tế Tài chính. 15(3): 373-394.

Connor, G. và RA Korajczyk. (1988). "Rủi ro và lợi nhuận trong APT cân bằng: Áp dụng phương pháp thử nghiệm mới". Tạp chí kinh tế tài chính. 21(2): 255-289.

Correia, M., J. Kang và S. Richardson. (2018). "Sự biến động của tài sản". Đánh giá các nghiên cứu về kế toán. 23(1): 37-94.

- Corsi, F. (2009). "Một mô hình bộ nhớ dài gần đúng đơn giản về sự biến động đã nhận ra". Tạp chí Kinh tế lượng Tài chính. 7: 174-196.
- Cowles, A. 3. (1933). "Các nhà dự báo thị trường chứng khoán có thể dự báo được không?" Kinh tế số liệu. 1(3): 309-324.
- Cujean, J. và M. Hasler. (2017). "Tại sao khả năng dự đoán lợi nhuận lại tập trung vào thời điểm khó khăn?" Tạp chí Tài chính. 72(6): 2717-2758 .
- Cybenko, G. (1989). "Xấp xỉ bằng cách xếp chồng hàm sigmoidal". Toán học điều khiển, tín hiệu và hệ thống. 2(4): 303-314.
- Da, R., S. Nagel và D. Xiu. (2022). "Giới hạn thống kê của trọng tài". Công nghệ. trả lời. Gian hàng Chicago.
- Das, SR và cộng sự. (2014). "Văn bản và bối cảnh: Phân tích ngôn ngữ trong tài chính". Nền tảng và Xu hướng trong Tài chính. 8(3): 145-261.
- Davis, SJ, S. Hansen và C. Seminario-Amez. (2020). "Mức độ rủi ro ở cấp độ công ty và lợi nhuận cổ phiếu sau đại dịch COVID-19". Công nghệ. trả lời. Đại học Chicago.
- DeMiguel, V., A. Martin-Utrera, FJ Nogales và R. Uppal. (2020). "Quan điểm về chi phí giao dịch dựa trên nhiều đặc điểm của công ty ". Tạp chí nghiên cứu tài chính. 33(5): 2180-2222.
- Đặng, W., L. Gao, B. Hu và G. Chu. (2022). "Nhìn thấy là tin tưởng: Báo cáo hàng năm". Có sẵn tại SSRN 3723126.
- Devlin, J., M.-W. Chang, K. Lee và K. Toutanova. (2018). "Bert: Đào tạo trước các máy biến áp hai chiều sâu để hiểu ngôn ngữ ". bản in trước arXiv arXiv:1810.04805.
- Didisheim, A., S. Ke, B. Kelly và S. Malamud. (2023). "Sự phức tạp trong các mô hình định giá nhân tố". Công nghệ. trả lời. Đại học Yale.
- Easley, D., M. López de Prado, M. O'Hara và Z. Zhang. (2020). "Cấu trúc vi mô trong thời đại máy móc". Tạp chí nghiên cứu tài chính. 34(7): 3316-3363.
- Erel, I., LH Stern, C. Tan và MS Weisbach. (2021). "Lựa chọn giám đốc bằng cách sử dụng máy học". Tạp chí nghiên cứu tài chính. 34(7): 3226-3264.
- Fabozzi, FJ, D. Huang và G. Chu. (2010). "Danh mục đầu tư vững chắc: đóng góp từ nghiên cứu hoạt động và tài chính". Biên niên sử nghiên cứu hoạt động. 176(1): 191-220.

Fama, EF và KR tiếng Pháp. (1993). "Các yếu tố rủi ro phổ biến trong lợi nhuận của cổ phiếu và trái phiếu". Tạp chí kinh tế tài chính. 33(1): 3-56.

Fama, EF và KR tiếng Pháp. (2010). "May mắn và kỹ năng trong khía cạnh lợi nhuận của quỹ tương hỗ". Tạp chí Tài chính. 65(5): 1915-1947.

Fama, EF và KR tiếng Pháp. (2015). "Mô hình định giá tài sản năm yếu tố". Tạp chí kinh tế tài chính. 116(1): 1-22.

Fama, EF (1970). "Thị trường vốn hiệu quả: Đánh giá lý thuyết và thực nghiệm". Tạp chí Tài chính. 25(2): 383-417.

Fama, EF (1990). "Lợi nhuận cổ phiếu, lợi nhuận kỳ vọng và hoạt động thực tế". Tạp chí Tài chính. 45(4): 1089-1108.

Fama, EF và RR Hạnh phúc. (1987). "Thông tin về tỷ giá kỳ hạn dài hạn". Tạp chí kinh tế Mỹ. 77(4): 680-692.

Fama, EF và KR tiếng Pháp. (1992). "Mặt cắt ngang của lợi nhuận cổ phiếu kỳ vọng". Tạp chí Tài chính. 47: 427-465.

Fama, EF và JD Macbeth. (1973). "Rủi ro, lợi nhuận và cân bằng: Thủ nghiệm thực nghiệm". Tạp chí Kinh tế Chính trị. 81(3): 607-636.

Fan, J., Y. Fan và J. Lv. (2008). "Ước tính ma trận hiệp phương sai thử nguyên cao bằng cách sử dụng mô hình nhân tố". Tạp chí Kinh tế lượng. 147: 186-197.

Fan, J., A. Furger và D. Xiu. (2016a). "Kết hợp tiêu chuẩn phân loại công nghiệp toàn cầu vào phân bổ danh mục đầu tư: Công cụ ước tính ma trận hiệp phương sai lớn dựa trên yếu tố đơn giản với dữ liệu tần số cao". Tạp chí Thông kê Kinh tế và Kinh doanh. 34(4): 489-503.

Fan, J., Y. Liao và M. Mincheva. (2013). "Ước tính hiệp phương sai lớn bằng cách xác định ngưỡng bỏ sung trực giao chính". Tạp chí của Hiệp hội Thống kê Hoàng gia, B. 75: 603-680.

Fan, J., Y. Liao và W. Wang. (2016b). "Phân tích thành phần chính dự kiến trong mô hình nhân tố". Biên niên sử thống kê. 44(1): 219.

Fan, J., Y. Liao và J. Yao. (2015). "Nâng cao sức mạnh trong các thử nghiệm cắt ngang có chiều cao". Kinh tế lượng. 83(4): 14977- 1541.

Feng, G., S. Giglio và D. Xiu. (2020). "Thuần hóa vườn thú nhân tố: Thủ nghiệm các nhân tố mới". Tạp chí Tài chính. 75(3): 1327-1370.

- Feng, G., J. He và NG Polson. (2018). "Học sâu để dự đoán lợi nhuận tài sản". bản in trước arXiv arXiv:1804.09314.
- Freyberger, J., A. Neuhierl và M. Weber. (2020). "Phân tích các đặc điểm một cách không tham số". Tạp chí nghiên cứu tài chính. 33(5): 2326-2377.
- Friedman, JH (2001). "Xấp xỉ hàm tham lam: một cỗ máy tăng cường độ dốc". Biên niên sử thống kê: 1189-1232.
- Frost, PA và JE Savarino. (1986). "Phương pháp tiếp cận theo kinh nghiệm của Bayes để lựa chọn danh mục đầu tư hiệu quả". Tạp chí Phân tích Tài chính và Định lượng. 21(3): 293-305. (Truy cập ngày 30/01/2023).
- Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai và A. Walther. (2022). "Có thể dự đoán là không bình đẳng? Tác dụng của học máy đối với thị trường tín dụng". Tạp chí Tài chính. 77(1): 5-47.
- Gabaix, X. (2012). "Những thảm họa hiếm gặp có thể thay đổi: Một khuôn khổ được giải quyết chính xác cho mười câu đố về tài chính vĩ mô". Tạp chí Kinh tế hàng quý. 127: 645-700.
- Gagliardini, P., E. Ossola và O. Scaillet. (2016). "Phi bù rủi ro thay đổi theo thời gian trong các tập dữ liệu vốn chủ sở hữu cắt ngang lớn". Kinh tế lượng. 84(3): 985-1046.
- Gagliardini, P. và D. Ronchetti. (2019). "So sánh các mô hình định giá tài sản theo khoảng cách Hansen-Jagannathan có điều kiện*". Tạp chí Kinh tế lượng Tài chính. 18(2): 333-394.
- Garcia, D., X. Hu và M. Rohrer. (2022). "Màu sắc của ngôn từ tài chính". Công nghệ. trả lời. Đại học Colorado ở Boulder.
- Garleanu, N. và LH Pedersen. (2013). "Giao dịch năng động với lợi nhuận và chi phí giao dịch có thể dự đoán được". Tạp chí Tài chính. 68(6): 2309-2340.
- Gentzkow, M., B. Kelly và M. Taddy. (2019). "Văn bản dưới dạng dữ liệu". tạp chí của Văn học kinh tế. 57(3): 535-74.
- Geweke, J. và G. Chu. (1996). "Đo lường sai số định giá của lý thuyết định giá chênh lệch giá". Việc xem xét các nghiên cứu tài chính. 9(2): 557-587 .
- Gibbons, MR, SA Ross và J. Shanken. (1989). "Một thử nghiệm về hiệu quả của một danh mục đầu tư nhất định". Kinh tế lượng: Tạp chí của Hiệp hội Kinh tế lượng: 1121-1152.

- Giglio, S., B. Kelly và D. Xiu. (2022a). "Mô hình nhân tố, học máy và định giá tài sản".
Đánh giá thường niên về kinh tế tài chính . 14: 1-32.
- Giglio, S., Y. Liao và D. Xiu. (2021a). "Hàng ngàn bài kiểm tra Alpha".
Đánh giá các nghiên cứu tài chính. 34(7): 3456-3496.
- Giglio, S. và D. Xiu. (2021). "Định giá tài sản với các yếu tố bị bỏ qua".
Tạp chí Kinh tế Chính trị. 129(7): 1947-1990.
- Giglio, S., D. Xiu và D. Zhang. (2021b). "Thử nghiệm tài sản và các yếu tố yêu".
Công nghệ. trả lời. Đại học Yale và Đại học Chicago.
- Giglio, S., D. Xiu và D. Zhang. (2022b). "Dự đoán khi các yếu tố yêu". Công nghệ.
trả lời. Đại học Yale và Đại học Chicago.
- Glaeser, EL, MS Kincaid và N. Naik. (2018). "Tầm nhìn máy tính
và Bất động sản: Vẽ ngoài có quan trọng và Quyết định có khuyến khích không
Nhìn". Công nghệ. trả lời. Đại học Harvard.
- Goodfellow, I., Y. Bengio và A. Courville. (2016). Học kĩ càng. MIT
nhấn.
- Goulet Coulombe, P. và M. Göbel. (2023). "Tối đa hóa danh mục đầu tư có thể học
được bằng máy". Có sẵn tại SSRN 4428178.
- Goyal, A. và A. Saretto. (2022). "Lợi nhuận quyền chọn cổ phiếu có bắt thường
không ? IPCA nói không". IPCA nói không (19 tháng 8 năm 2022).
- Gu, S., B. Kelly và D. Xiu. (2020a). "Định giá tài sản mã hóa tự động
Người mẫu". Tạp chí Kinh tế lượng.
- Gu, S., B. Kelly và D. Xiu. (2020b). "Định giá tài sản theo kinh nghiệm thông
qua học máy". Tạp chí nghiên cứu tài chính. 33(5): 2223-2273 .
- Guizarro-Ordonez, J., M. Pelger và G. Zanotti. (2022). "Trọng tài thống kê học
sâu ". Công nghệ. trả lời. Đại học Stanford.
- Hamblay, B., R. Xu và H. Yang. (2022). "Những tiến bộ gần đây trong học tập tăng
cường trong lĩnh vực tài chính". Công nghệ. trả lời. Đại học Oxford.
- Hansen, LP và R. Jagannathan. (1997). "Đánh giá lỗi đặc điểm kỹ thuật trong các mô hình hệ số
chiết khấu ngẫu nhiên". Tạp chí Tài chính. 52: 557-590.
- Hansen, LP và SF Richard. (1987). "Vai trò của thông tin điều hòa trong việc suy
ra các hạn chế có thể kiểm chứng được ngụ ý bởi các mô hình định giá tài sản
động". Kinh tế lượng: Tạp chí của Hiệp hội Kinh tế lượng: 587-613.

- Hansen, LP và KJ Singleton. (1982). "Ước tính các biến công cụ tổng quát của các mô hình kỳ vọng hợp lý phi tuyến tính".
Kinh tế lượng. 50(5): 1269-1286.
- Harvey, CR và Y. Liu. (2020). "Những khám phá sai lầm (và bị bỏ sót) trong kinh tế tài chính". Tạp chí Tài chính. 75(5): 2503-2553.
- Harvey, CR, Y. Liu và H. Zhu. (2016). "... Và mặt cắt ngang của lợi nhuận kỳ vọng". Đánh giá các nghiên cứu tài chính. 29(1): 5-68.
- Harvey, CR (2017). "Diễn văn của Chủ tịch: Quan điểm khoa học về kinh tế tài chính". Tạp chí Tài chính. 72(4): 1399-1440.
- Harvey, CR và CHUNG TÔI Ferson. (1999). "Các biến điều kiện và mặt cắt ngang của lợi nhuận hàng tồn kho". Tạp chí Tài chính. 54: 1325- 1360.
- Hastie, T., A. Montanari, S. Rosset và RJ Tibshirani. (2019).
"Những điều bất ngờ trong phép nội suy bình phương tối thiểu không có đường vân có chiều cao". bản in trước arXiv arXiv:1903.08560.
- Haugen, RA và NL Baker. (1996). "Tính tương đồng trong các yếu tố quyết định lợi nhuận kỳ vọng của cổ phiếu". Tạp chí Kinh tế Tài chính. 41(3): 401-439.
- Hayek, FA (1945). "Việc sử dụng kiến thức trong xã hội". Người Mỹ Tạp chí kinh tế. 35(4): 519-530.
- Anh ấy, A., S. He, D. Rapach và G. Chu. (2022a). "Tỷ suất sinh lợi cổ phiếu kỳ vọng trong mặt cắt ngang: Phương pháp tiếp cận tổng hợp". Giấy làm việc .
- Anh ấy, K., X. Zhang, S. Ren và J. Sun. (2016). "Học tập dư thừa sâu để nhận dạng hình ảnh". Trong: Hội nghị IEEE 2016 về Thị giác máy tính và Nhận dạng mẫu (CVPR). 770-778. doi: 10.1109/ CVPR.2016.90.
- Anh ấy, S., M. Yuan và G. Chu. (2022b). "Danh mục đầu tư chính: Một lưu ý". Giấy làm việc .
- Anh ấy, X., G. Feng, J. Wang và C. Wu. (2021). "Dự đoán lợi nhuận trái phiếu doanh nghiệp cá nhân". Công nghệ. trả lời. Đại học Thành phố Hồng Kông.
- Anh ấy, Z., B. Kelly và A. Manela. (2017). "Định giá tài sản trung gian: Bằng chứng mới từ nhiều loại tài sản". Tạp chí Kinh tế Tài chính. 126(1): 1-35.
- Hochreiter, S. và J. Schmidhuber. (1997). "Trí nhớ dài hạn".
Tính toán thần kinh. 9: 1735-1780.

- Hong, H., T. Lim và JC Stein. (2000). "Tin xấu lan truyền chậm: Quy mô, phạm vi đưa tin của nhà phân tích và khả năng sinh lời của các chiến lược tạo đà". *Tạp chí Tài chính*. 55(1): 265-295.
- Hornik, K., M. Stinchcombe và H. White. (1989). "Mạng chuyển tiếp tiếp liệu đa lớp là các công cụ xấp xỉ phổ quát". *Mạng lưới thần kinh*. 2(5): 359-366.
- Hornik, K., M. Stinchcombe và H. White. (1990). "Xấp xỉ phổ quát của một ánh xạ chưa biết và các dẫn xuất của nó bằng cách sử dụng các mạng tiếp liệu đa lớp". *Mạng lưới thần kinh*. 3(5): 551-560.
- Hou, K., C. Xue và L. Zhang. (2018). "Sao chép sự bất thường". Các Đánh giá các nghiên cứu tài chính. 33(5): 2019-2133.
- Hu, A. và S. Ma. (2020). "Tương tác giữa con người và đầu tư tài chính: Một cách tiếp cận dựa trên video". Có sẵn tại SSRN.
- Huang, D., F. Jiang, K. Li, G. Tong và G. Chu. (2021). "PCA được chia tỷ lệ: Một cách tiếp cận mới để giảm kích thước". *Khoa học quản lý, sắp xuất bản*.
- Huang, D., F. Jiang, J. Tu và G. Chu. (2014). "Tâm lý nhà đầu tư phù hợp: Một công cụ dự đoán mạnh mẽ về lợi nhuận chứng khoán". *Tạp chí nghiên cứu tài chính*. 28(3): 791-837. issn: 0893-9454.
- Huang, J., JL Horowitz và F. Wei. (2010). "Lựa chọn biên trong các mô hình cộng tính không tham số". *Biên niên sử thống kê*. 38(4): 2282-2313.
- Huberman, G. (1982). "Một cách tiếp cận đơn giản đối với lý thuyết định giá chênh lệch giá". *Tạp chí Lý thuyết kinh tế*. 28(1): 183-191.
- Ingersoll, JE (1984). "Một số kết quả trong lý thuyết định giá chênh lệch giá". *Tạp chí Tài chính*. 39(4): 1021-1039.
- Israel, R., B. Kelly và TJ Moskowitz. (2020). "Máy móc có thể học" Tài chính?" *Tạp chí Quản lý đầu tư*. 18(2): 23-36.
- Israelov, R. và BT Kelly. (2017). "Dự báo sự phân bố của quyền chọn trả về". Có sẵn tại SSRN 3033242.
- Jacot, A., F. Gabriel và C. Hongler. (2018). "Hạt nhân tiếp tuyến thần kinh: Sự hội tụ và khái quát hóa trong mạng lưới thần kinh". *bản in trước arXiv arXiv :1806.07572*.
- Jegadeesh, N. và D. Wu. (2013). "Sức mạnh của từ ngữ: Một cách tiếp cận mới để phân tích nội dung". *Tạp chí Kinh tế Tài chính*. 110(3): 712-729.

- Jensen, Tl, B. Kellly, C. Seminario-Amez và LH Pedersen. (2022). "Học máy và Biên giới hiệu quả có thể triển khai". Công nghệ. trả lời. Trường kinh doanh Copenhagen.
- Jensen, TI, B. Kelly và LH Pedersen. (2021). "Có khung hoảng nhẫn rộng trong tài chính không?" Tạp chí Tài chính, sắp xuất bản.
- Jiang, F., G. Tang và G. Chu. (2018). "Đặc điểm công ty và chứng khoán Trung Quốc". Tạp chí Khoa học và Kỹ thuật Quản lý. 3(4): 259-283.
- Jiang, J., B. Kelly và D. Xiu. (2022). "(Tái) tưởng tượng (trong) xu hướng giá". Tạp chí Tài chính, sắp xuất bản.
- Jiang, J., B. Kelly và D. Xiu. (2023). "Lợi nhuận kỳ vọng và mô hình ngôn ngữ lớn". Công nghệ. trả lời. Đại học Chicago và Đại học Yale .
- Jobson, JD và B. Korkie. (1980). "Ước tính cho danh mục đầu tư hiệu quả của Markowitz". Tạp chí của Hiệp hội Thống kê Hoa Kỳ. 75(371): 544-554. (Truy cập ngày 30/01/2023).
- Joror, P. (1986). "Ước tính Bayes-Stein cho phân tích danh mục đầu tư". Tạp chí phân tích tài chính và định lượng. 21(3): 279-292. (Truy cập ngày 30/01/2023).
- Jurado, K., SC Ludvigson và S. Ng. (2015). "Đo lường sự không chắc chắn". Tạp chí kinh tế Mỹ. 105(3): 1177-1216.
- Kan, R., X. Wang và G. Chu. (2022). "Lựa chọn danh mục đầu tư tối ưu với rủi ro ước tính: Không có trường hợp tài sản phi rủi ro". Khoa học quản lý, sắp xuất bản.
- Kan, R. và C. Zhang. (1999). "Thử nghiệm hai lần đối với các mô hình định giá tài sản với các yếu tố vô dụng". Tạp chí Tài chính. 54(1): 203-235.
- Kan, R. và G. Chu. (2007). "Lựa chọn danh mục đầu tư tối ưu với sự không chắc chắn về tham số". Tạp chí phân tích tài chính và định lượng. 42(3): 621-656.
- Ké, T., B. Kelly và D. Xiu. (2019). "Dự đoán kết quả trả về bằng dữ liệu văn bản". Công nghệ. trả lời. Đại học Harvard, Đại học Yale và Đại học Chicago.
- Kelly, B., S. Malamud và LH Pedersen. (2020a). "Danh mục đầu tư chính-sự tử". Giấy làm việc.
- Kelly, B., S. Malamud và K. Chu. (2022a). "Tính phức tạp trong Dự đoán trở lại". Công nghệ. trả lời. Đại học Yale.

- Kelly, B., A. Manela và A. Moreira. (2018). "Lựa chọn văn bản". Đang làm việc giấy.
- Kelly, B., T. Moskowitz và S. Pruitt. (2021). "Hiểu về Động lượng và Sự đảo chiều". Tạp chí Kinh tế Tài chính. 140(3): 726-743.
- Kelly, B., D. Palhares và S. Pruitt. (Sắp có). "Mô hình hóa lợi nhuận trái phiếu doanh nghiệp". Tạp chí Tài chính.
- Kelly, B. và S. Pruitt. (2013). "Kỳ vọng của thị trường trong mặt cắt ngang của giá trị hiện tại". Tạp chí Tài chính. 68(5): 1721-1756.
- Kelly, B. và S. Pruitt. (2015). "Bộ lọc hồi quy ba bước: Một cách tiếp cận mới để dự báo bằng cách sử dụng nhiều yếu tố dự đoán". Tạp chí Kinh tế lượng. 186(2): 294-316. issn: 0304-4076.
- Kelly, B., S. Pruitt và Y. Su. (2020b). "Đặc điểm là hiệp phương sai : Mô hình thống nhất về rủi ro và lợi nhuận". Tạp chí Kinh tế Tài chính.
- Kelly, BT, S. Malamud và K. Chu. (2022b). "Đức tính phức tạp ở mọi nơi". Có sẵn tại SSRN.
- Kim, S., R. Korajchot và A. Neuhierl. (2020). "Danh mục đầu tư chênh lệch giá". Tạp chí nghiên cứu tài chính, sắp xuất bản.
- Koopmans, TC (1947). "Đo lường không cần lý thuyết". Tạp chí Kinh tế và Thống kê. 29(3): 161-172.
- Kosowski, R., A. Timmermann, R. Wermers và H. White. (2006). "Các "ngôi sao" quý tương hỗ có thực sự chọn được cổ phiếu không? Bằng chứng mới từ phân tích bootstrap". Tạp chí Tài chính. 61(6): 2551-2595.
- Kozak, S. (2020). "Thủ thuật hạt nhân cho mặt cắt ngang". Có sẵn tại SSRN 3307895.
- Kozak, S., S. Nagel và S. Santosh. (2018). "Mô hình nhân tố giải thích". Tạp chí Tài chính. 73(3): 1183-1223.
- Kozak, S., S. Nagel và S. Santosh. (2020). "Thu nhỏ mặt cắt ngang". Tạp chí Kinh tế Tài chính. 135(2): 271-292.
- Ledoit, O. và M. Wolf. (2004). "Em yêu, anh đã thu nhỏ ma trận hiệp phương sai mẫu". Tạp chí Quản lý danh mục đầu tư. 30: 110-119.
- Ledoit, O. và M. Wolf. (2012). "Ước tính độ co phi tuyển của ma trận hiệp phương sai chiều lớn". Biên niên sử thống kê. 40: 1024-1060.

- Leippold, M., Q. Wang và W. Chu. (2022). "Học máy trên thị trường chứng khoán Trung Quốc". *Tạp chí Kinh tế Tài chính*. 145(2, Phần A): 64-82.
- Lettau, M. và S. Ludvigson. (2001). "Tiêu dùng, tổng tài sản và lợi nhuận cổ phiếu kỳ vọng". *Tạp chí Tài chính*. 56(3): 815-849 .
- Lettau, M. và M. Pelger. (2020a). "Ước tính giá tài sản tiềm ẩn Các nhân tố". *Tạp chí Kinh tế lượng*. 218: 1-31.
- Lettau, M. và M. Pelger. (2020b). "Các yếu tố phù hợp với chuỗi thời gian và mặt cắt ngang của lợi nhuận chứng khoán". *Đánh giá các nghiên cứu tài chính*. 33(5): 2274-2325.
- Lewellen, J. (2015). "Mặt cắt ngang của lợi nhuận cổ phiếu kỳ vọng". *Đánh giá tài chính quan trọng*. 4(1): 1-44.
- Li, K., F. Mai, R. Shen và X. Yan. (2021). "Đo lường văn hóa doanh nghiệp bằng cách sử dụng máy học". *Tạp chí nghiên cứu tài chính*. 34(7): 3265-3315.
- Li, SZ và Y. Tang. (2022). "Dự báo rủi ro tự động". *Công nghệ. trả lời. Rutgers, Đại học Bang New Jersey*.
- Light, N., D. Maslov và O. Rytchkov. (2017). "Tổng hợp thông tin về mặt cắt ngang của lợi nhuận hàng tồn kho: Phương pháp tiếp cận biến đổi tiềm ẩn". *Tạp chí nghiên cứu tài chính*. 30(4): 1339-1381.
- Lo, AW và AC MacKinlay. (1990). "Thành kiến theo dõi dữ liệu trong các thử nghiệm mô hình định giá tài sản tài chính". *Đánh giá các nghiên cứu tài chính*. 3(3): 431-467.
- Loughran, T. và B. McDonald. (2011). "Khi nào trách nhiệm pháp lý không phải là trách nhiệm pháp lý? Phân tích văn bản, từ điển và 10-Ks". *Tạp chí Tài chính*. 66(1): 35-65.
- Loughran, T. và B. McDonald. (2020). "Phân tích văn bản trong tài chính". *Đánh giá thường niên về kinh tế tài chính*. 12: 357-375.
- Lucas Jr, RE (1976). "Đánh giá chính sách kinh tế lượng: Một phê bình". Trong: Chuỗi hội nghị Carnegie-Rochester về chính sách công. Tập. 1. Bắc-Hà Lan. 19-46.
- Ludvigson, SC và S. Ng. (2010). "Phân tích nhân tố phần bù rủi ro trái phiếu ". Trong: Sổ tay kinh tế học thực nghiệm và tài chính. Ed. của A. Ulah và DEA Giles. Tập. 1. Chapman và Hall, Boca Raton, FL. Chap. 12. 313-372.

Ludvigson, SC (2013). "Chương 12 - Những tiền bộ trong việc định giá tài sản dựa trên tiêu dùng: Thủ nghiệm thực nghiệm". Trong: biên tập. của GM Constantinides, M. Harris và RM Stulz. Tập. 2. Sổ tay Kinh tế Tài chính. Khác. 799-906.

Ludvigson, SC và S. Ng. (2007). "Mối quan hệ rủi ro-lợi nhuận thực nghiệm: Phương pháp phân tích nhân tố". Tạp chí Kinh tế Tài chính. 83(1): 171-222.

Lynch, AW và P. Balduzzi. (2000). "Khả năng dự đoán và chi phí giao dịch: Tác động đến các quy tắc và hành vi tái cân bằng". Tạp chí Tài chính. 55(5): 2285-2309.

Lyonnet, V. và LH Stern. (2022). "Phân bổ vốn mạo hiểm (Mis) trong thời đại AI". Có sẵn tại SSRN 4260882.

Malloy, CJ, TJ Moskowitz và A. Vissing-Jorgensen. (2009). "Rủi ro tiêu dùng dài hạn của cổ đông và lợi nhuận tài sản". Tạp chí Tài chính. 64(6): 2427-2479.

Manela, A. và A. Moreira. (2017). "Tin tức ngụ ý sự biến động và lo ngại về thảm họa". Tạp chí Kinh tế Tài chính. 123(1): 137-162.

Markowitz, H. (1952). "Lựa chọn danh mục đầu tư". Tạp chí Tài chính. 7(1): 77-91.

Martin, IW và S. Nagel. (2021). "Hiệu quả thị trường trong thời đại lớn dữ liệu". Tạp chí Kinh tế Tài chính.

Mehra, R. và EC Prescott. (1985). "Phản bù vốn chủ sở hữu: Một câu đố". Tạp chí Kinh tế tiền tệ. 15(2): 145-161.

Menzly, L., T. Santos và P. Veronesi. (2004). "Hiểu về khả năng dự đoán". Tạp chí Kinh tế Chính trị. 112(1): 1-47. (Truy cập ngày 02/06/2023).

Merton, RC (1973). "Mô hình định giá tài sản vốn liên thời gian". Kinh tế lượng. 41: 867-887.

Michaud, RO (1989). "Bí ẩn tôi ưu hóa Markowitz: 'Tôi ưu hóa' có tôi ưu không? Tạp chí phân tích tài chính. 45(1): 31-42. (Truy cập ngày 30/01/2023).

Mittnik, S., N. Robinzonov và M. Spindler. (2015). "Biến động của thị trường chứng khoán: Xác định các động lực chính và bản chất tác động của chúng". Tạp chí Tài chính Ngân hàng. 58: 1-14.

- Moritz, B. và T. Zimmermann. (2016). "Các loại danh mục đầu tư có điều kiện dựa trên cây : Mối quan hệ giữa lợi nhuận chứng khoán trong quá khứ và tương lai". Công nghệ. trả lời. Đại học Ludwig Maximilian München.
- Nagel, S. và K. Singleton. (2011). "Ước tính và đánh giá các mô hình định giá tài sản có điều kiện". Tạp chí Tài chính. 66(3): 873-909. (Truy cập ngày 20/02/2023).
- Nishii, R. (1984). "Thuộc tính tiệm cận của tiêu chí lựa chọn biến trong hồi quy bội". Biên niên sử thống kê. 12(2): 758-765.
- Novy-Marx, R. (2014). "Dự đoán hiệu suất bất thường về chính trị, thời tiết, hiện tượng nóng lên toàn cầu, vết đen mặt trời và các vì sao". Tạp chí Kinh tế Tài chính. 112(2): 137-146.
- Obaid, K. và K. Pukthuanthong. (2022). "Một bức tranh đáng giá ngàn lời nói: Đô lường tâm lý nhà đầu tư bằng cách kết hợp học máy và ảnh từ tin tức". Tạp chí Kinh tế Tài chính. 144: 273-297 .
- Ohlson, JA (1980). "Các tỷ số tài chính và dự đoán xác suất phá sản". Tạp chí nghiên cứu kế toán. 18(1): 109-131.
- Onatski, A. (2009). "Kiểm định giả thuyết về số lượng nhân tố trong mô hình nhân tố lớn". Kinh tế lượng. 77(5): 1447-1479.
- Onatski, A. (2010). "Xác định số lượng các yếu tố từ phân phối thực nghiệm của giá trị riêng". Tạp chí Kinh tế và Thống kê. 92: 1004-1016.
- Onatski, A. (2012). "Tiệm cận của hàm ước lượng thành phần chính của mô hình nhân tố lớn với nhân tố ảnh hưởng yếu". Tạp chí Kinh tế lượng. 168: 244-258.
- Mục sư, L. (2000). "Các mô hình lựa chọn danh mục đầu tư và định giá tài sản". Các Tạp chí Tài chính. 55(1): 179-223.
- Mục sư, L. và RF Stambaugh. (2000). "So sánh các mô hình định giá tài sản : Quan điểm đầu tư". Tạp chí Kinh tế Tài chính. 56: 335-381.
- Mục sư, L. và RF Stambaugh. (2003). "Rủi ro thanh khoản và lợi nhuận cỗ phiếu kỳ vọng". Tạp chí Kinh tế Chính trị. 111(3): 642-685.
- Pesaran, H. và T. Yamagata. (2017). "Thử nghiệm Alpha trong các mô hình định giá nhân tố tuyến tính với số lượng lớn chứng khoán". Công nghệ. trả lời.

Petersen, MA (2008). "Ước tính lối tiêu chuẩn trong bộ dữ liệu bảng tài chính: So sánh các phương pháp tiếp cận". Tạp chí nghiên cứu tài chính. 22(1): 435-480.

Pukthuanthong, K., R. Roll và A. Subrahmanyam. (2019). "Giao thức xác định yếu tố". Đánh giá các nghiên cứu tài chính. 32(4): 1573-1607 .

Radford, A., J. Wu, R. Child, D. Luân, D. Amodei, I. Sutskever, et al. (2019). "Mô hình ngôn ngữ là những người học đa nhiệm không được giám sát". Blog OpenAI. 1(8): 9.

Rahimi, A. và B. Recht. (2007). "Các tính năng ngẫu nhiên dành cho máy hạt nhân quy mô lớn ". Trong: NIPS. Tập. 3. Số 4. Người trích dẫn. 5.

Rapach, D. và G. Chu. (2013). "Chương 6 - Dự báo sự quay trở lại của hàng tồn kho". Trong: Sổ tay dự báo kinh tế. Ed. của G. Elliott và A. Timmermann. Tập. 2. Sổ tay dự báo kinh tế. Khác. 328-383.

Rapach, D. và G. Chu. (2022). "Định giá tài sản: Khả năng dự đoán theo chuỗi thời gian-nó".

Rapach, DE, JK Strauss và G. Chu. (2010). "Dự đoán phần bù vốn cổ phần ngoài mẫu : Dự báo kết hợp và liên kết với nền kinh tế thực". Tạp chí nghiên cứu tài chính. 23(2): 821-862.

Rapach, DE, JK Strauss và G. Chu. (2013). "Khả năng dự đoán lợi nhuận chứng khoán quốc tế: vai trò của Hoa Kỳ là gì?" Tạp chí Tài chính. 68(4): 1633-1662.

Đúng hơn là AM, A. Agarwal và V. Sastry. (2015). "Mạng lưới thần kinh tái phát và mô hình kết hợp để dự đoán lợi nhuận chứng khoán". Hệ thống chuyên gia với các ứng dụng. 42(6): 3234-3241.

Cuộn, R. (1977). "Phê phán các bài kiểm tra lý thuyết định giá tài sản". Tạp chí Kinh tế Tài chính. 4: 129-176.

Rosenberg, B. (1974). "Các thành phần ngoài thị trường của hiệp thương sai trong lợi nhuận bảo mật". Tạp chí phân tích tài chính và định lượng. 9(2): 263-274.

Rosenberg, liên doanh và RF Engle. (2002). "Hạt nhân định giá theo kinh nghiệm". Tạp chí Kinh tế Tài chính. 64(3): 341-372.

Ross, SA (1976). "Lý thuyết chênh lệch giá về định giá tài sản vốn". Tạp chí Lý thuyết kinh tế. 13(3): 341-360.

- Rossi, AG (2018). "Dự đoán lợi nhuận thị trường chứng khoán bằng máy học hỏi". Đại học Georgetown.
- Rossi, AG và A. Timmermann. (2015). "Mô hình hóa rủi ro hiệp phương sai trong ICAPM của Merton". Tạp chí nghiên cứu tài chính. 28(5): 1428-1461 .
- Samuelson, PA (1965). "Lý thuyết hợp lý về giá chứng quyền". công nghiệp-Đánh giá quản lý thử nghiệm. 6(2): 13-39.
- Schaller, H. và SV Norden. (1997). "Chuyển đổi cơ chế trong lợi nhuận thị trường chứng khoán". Kinh tế tài chính ứng dụng. 7(2): 177-191.
- Schapire, RE (1990). "Sức mạnh của khả năng học hỏi yếu". Máy móc Học hỏi. 5(2): 197-227.
- Sezer, OB, MU Gudelek và AM Ozbayoglu. (2020). " Dự báo chuỗi thời gian tài chính với học sâu: Đánh giá tài liệu có hệ thống: 2005-2019". Máy tính mềm ứng dụng. 90: 106-181.
- Shanken, J. (1992a). "Về việc ước tính các mô hình định giá Beta". Đánh giá các nghiên cứu tài chính. 5: 1-33.
- Shanken, J. (1992b). "Hiện trạng của lý thuyết định giá chênh lệch giá". Tạp chí Tài chính. 47(4): 1569-1574.
- Shiller, RJ (1981). "Giá cổ phiếu có biến động quá nhiều đến mức không thể giải thích được bằng những thay đổi về cổ tức sau đó không?" Tạp chí kinh tế Mỹ . 71(3): 421-436.
- Simon, F., S. Weibels và T. Zimmermann. (2022). "Chính sách danh mục đầu tư tham số sâu ". Công nghệ. trà lời. Đại học Köln.
- Singh, R. và S. Srivastava. (2017). "Dự đoán chứng khoán bằng cách sử dụng deep learning ". Công cụ và ứng dụng đa phương tiện. 76(18): 18569-18584.
- Spigler, S., M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli và M. Wyart. (2019). "Quá trình chuyển đổi gây nhiễu từ quá ít tham số sang quá tham số hóa ảnh hưởng đến việc khai quát hóa trong học sâu". Tạp chí Vật lý A: Toán -lý thuyết. 52(47): 474001.
- Stock, JH và MW Watson. (2002). "Dự báo bằng cách sử dụng các thành phần chính từ một số lượng lớn các yếu tố dự đoán". Tạp chí của Hiệp hội Thống kê Hoa Kỳ. 97(460): 1167-1179.
- Đá, M. (1977). "Sự tương đương tiềm cận của việc lựa chọn mô hình bằng xác thực chéo và tiêu chí của Akaike". Tạp chí của Hiệp hội Thống kê Hoàng gia Series B. 39(1): 44-47.

- Sullivan, R., A. Timmermann và H. White. (1999). "Dữ liệu rình mò, hiệu suất quy tắc giao dịch kỹ thuật và bootstrap". Tạp chí Tài chính. 54(5): 1647-1691.
- Taddy, M. (2013). "Hồi quy nghịch đảo đa thức để phân tích văn bản". Tạp chí của Hiệp hội Thống kê Hoa Kỳ. 108(503): 755-770.
- Tetlock, PC (2007). "Đưa nội dung đến tâm lý nhà đầu tư: Vai trò của truyền thông trên thị trường chứng khoán". Tạp chí Tài chính. 62(3): 1139-1168.
- Tú, J. và G. Chu. (2010). "Kết hợp các mục tiêu kinh tế vào các ưu tiên Bayesian: Lựa chọn danh mục đầu tư theo thông số không chắc chắn". Tạp chí phân tích tài chính và định lượng. 45(4): 959-986.
(Truy cập ngày 30/01/2023).
- Van Binsbergen, JH và R. Koijen. (2010). "Hồi quy dự đoán: Phương pháp tiếp cận giá trị hiện tại". Tạp chí Tài chính. 65(4): 1439-1471 .
- Van Binsbergen, JH và CC Opp. (2019). "Những điều bất thường thực sự". Tạp chí Tài chính. 74(4): 1659-1706.
- Wachter, J. (2013). "Liệu rủi ro thiên tai hiếm gặp thay đổi theo thời gian có thể giải thích được sự biến động tổng hợp của thị trường chứng khoán không?" Tạp chí Tài chính. 68: 987-1035.
- Wachter, JA (2006). "Một mô hình dựa trên tiêu dùng của cấu trúc kỳ hạn lãi suất". Tạp chí Kinh tế Tài chính. 79(2): 365-399 .
- Welch, tôi và A. Goyal. (2008). "Một cái nhìn toàn diện về hiệu suất thực nghiệm của dự đoán phần bù vốn cổ phần". Tạp chí nghiên cứu tài chính. 21(4): 1455-1508.
- Tráng, H. (2000). "Kiểm tra thực tế về việc thu thập dữ liệu". Kinh tế lượng. 68(5): 1097-1126.
- Yogo, M. (2006). "Giải thích dựa trên tiêu dùng về lợi nhuận kỳ vọng của cổ phiếu". Tạp chí Tài chính. 61(2): 539-580.
- Yuan, M. và G. Chu. (2022). "Tại sao sự đa dạng hóa 1/N ngây thơ không quá ngây thơ và làm thế nào để đánh bại nó?" Có sẵn tại SSRN.

Zhang, S., S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, XV Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, PS Koura, A. Sridhar, T. Wang và L. Zettlemoyer. (2022). "OPT: Mô hình ngôn ngữ Transformer được đào tạo trước mở ". arXiv: [2205.01068 \[cs.CL\]](https://arxiv.org/abs/2205.01068).