BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN MINH HẢI

NGHIÊN CỨU PHƯƠNG PHÁP GIẢM CHIỀU BIẾN DỰA TRÊN HÀM NHÂN VÀ ỨNG DỤNG TRONG BÀI TOÁN DỰ BÁO KIM NGẠCH XUẤT KHẨU

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội - Năm 2024

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN MINH HẢI

NGHIÊN CỨU PHƯƠNG PHÁP GIẢM CHIỀU BIẾN DỰA TRÊN HÀM NHÂN VÀ ỨNG DỤNG TRONG BÀI TOÁN DỰ BÁO KIM NGẠCH XUẤT KHẨU

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN
Mã số: 9 48 01 04

Xác nhận của Học viện Khoa học và Công nghệ	**Người hướng dẫn 1** *(Ký, ghi rõ họ tên)*	**Người hướng dẫn 2** *(Ký, ghi rõ họ tên)*
KT. GIÁM ĐỐC PHÓ GIÁM ĐỐC Nguyễn Thị Trung	PGS.TS Đỗ Văn Thành	PGS.TS Nguyễn Đức Dũng

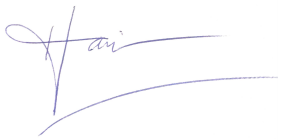
Hà Nội - Năm 2024

LỜI CAM ĐOAN

Tôi xin cam đoan Luận án "*Nghiên cứu phương pháp giảm chiều biến dựa trên hàm nhân và ứng dụng trong bài toán dự báo kim ngạch xuất khẩu*" là Nghiên cứu nghiên cứu của tôi. Các Nghiên cứu được viết chung với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Những kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong các Nghiên cứu nào khác.

Luận án được hoàn thành trong thời gian tôi làm NCS tại phòng Nhận dạng và Công nghệ tri thức, Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Tác giả luận án



NCS. Nguyễn Minh Hải

LỜI CẢM ƠN

Luận án tiến sĩ "*Nghiên cứu phương pháp giảm chiều biến dựa trên hàm nhân và ứng dụng trong bài toán dự báo kim ngạch xuất khẩu*" được thực hiện tại Viện Công nghệ Thông tin, Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, dưới sự hướng dẫn khoa học của PGS.TS. Đỗ Văn Thành và PGS.TS. Nguyễn Đức Dũng.

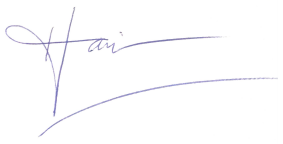
Tôi xin được bày tỏ lòng biết ơn sâu sắc đến hai thầy hướng dẫn là PGS. TS. Đỗ Văn Thành và PGS.TS. Nguyễn Đức Dũng. Trong quá trình học tập, nghiên cứu và thực hiện luận án tôi đã nhận được sự hướng dẫn tận tình, các định hướng khoa học quan trọng và những bài học sâu sắc từ các thầy hướng dẫn. Các thầy cũng đã luôn tận tâm động viên, khuyến khích và chỉ dẫn giúp đỡ tôi hoàn thành được bản luận án này.

Tôi xin chân thành cảm ơn các Ban Lãnh đạo Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Viện Công nghệ thông tin, Học viện Khoa học và Công nghệ, Ban quản lý Tòa nhà Vườn ươm và thầy PGS.TS Ngô Quốc Tạo, NCS Nguyễn Thị Thanh Mai, TS. Nguyễn Thị Phương, Phòng Nhận dạng và Công nghệ Tri thức, Viện Công nghệ thông tin đã luôn giúp đỡ, tạo điều kiện thuận lợi trong việc lưu trú cũng như quá trình học tập, nghiên cứu và thực hiện luận án này.

Tôi xin cảm ơn Ban Giám hiệu, các thầy cô giảng viên Khoa Khoa học Cơ bản, Trường Đại học Công Nghiệp thành phố Hồ Chí Minh đã tạo điều kiện giúp đỡ tôi trong suốt thời gian học tập và nghiên cứu.

Cuối cùng, tôi xin bày tỏ lòng biết ơn sâu sắc tới Bố, Mẹ, Anh, Chị em trong gia đình hai bên Nội, bên Ngoại, Vợ và các con đã cho tôi điểm tựa vững chắc, tạo động lực để tôi hoàn thành luận án này.

Tác giả



NCS. Nguyễn Minh Hải

MỤC LỤC

Danh mục hình

Danh mục bảng

Danh mục các từ viết tắt

STT	Từ viết tắt	Nội dung	Giải thích
1	AIC	Akaike information criteria	Tiêu chuẩn thông tin Akaike
2	ARDL	Autoregressive Distributed Lag	Trễ phân bố tự hồi quy
3	ARIMA model	Autoregressive Intergrated Moving Average Model	Mô hình trung bình trượt tích hợp tự hồi quy
4	BE	Bridge Equation	Phương trình bắc cầu
5	BIC	Bayesian information criteria	Tiêu chuẩn thông tin Bayes
6	BLUE	The Best, Linear, and Unbiased Estimate	Ước lượng không chệch, tuyến tính và tốt nhất.
7	DFM	Dynamic Factor Model	Mô hình nhân tố động (DFM)
8	EAW-MIDAS	Exponential Almon weighting MIDAS	Mô hình MIDAS trọng số Almon hàm mũ
9	KPCA	Kernel Principal Component Analysis	Phân tích thành phân chính hàm nhân
10	LASSO	Least Absolute Shrinkage and Selection Operator	Toán tử lựa chọn và co rút tuyệt đối nhỏ nhất
11	MIDAS	Mixed Data Sampling	Lấy mẫu dữ liệu hỗn hợp
12	PAW-MIDAS	Polynomial Almon weighting MIDAS	Mô hình MIDAS trọng số Almon đa thức
13	PCA	Principal Component Analysis	Phân tích thành phần chính
14	RMSE	Root Mean Squared Error	Sai số bình phương trung bình chuẩn
15	ROBSPCA	Robust Sparse Principal Component Analysis	Phân tích thành phân chính thưa vững

16	RSPCA	Random Sparse Principal Component Analysis	Phân tích thành phân chính thưa ngẫu nhiên
17	SPCA	Sparse Principal Component Analysis	Phân tích thành phân chính thưa
18	STEP-MIDAS	STEP weighting MIDAS	Mô hình MIDAS trọng số STEP
19	U-MIDAS	Unrestricted MIDAS	Mô hình MIDAS không hạn chế

Danh mục các thuật ngữ

STT	Thuật ngữ	Giải thích
1	Biến tần suất cao/ tần suất thấp	Tần suất là nói về kỳ (thời gian) thu thập dữ liệu. Biến có kỳ thu thập dữ liệu ngắn hơn được gọi là biến có (hoặc ở) tần suất cao hơn.
2	Chiết xuất các nhân tố	Là quá trình biến đổi tập các biến giải thích thành tập các biến mới (gọi là các nhân tố) có số lượng ít hơn nhiều nhưng giữ được những thông tin quan trọng của các biến giải thích.
3	Chuỗi thời gian dừng (Y_t)	Chuỗi Y_t được gọi là dừng nếu kỳ vọng và phương sai của nó không đổi; Tự hiệp phương sai của nó chỉ phụ thuộc độ dài trễ, không phụ thuộc vào thời điểm lấy trễ. C*huỗi thời gian dừng không có tính xu thế và tính mùa vụ*.
4	Độ trễ tối ưu/ Số lượng biến trễ	Trễ tối ưu (hay độ dài trễ) của một biến là số lượng tối đa các biến trễ của biến đó có trong mô hình dự báo để độ chính xác dự báo của mô hình là cao nhất.
5	Độ trễ riêng tối ưu	Độ trễ chung tối ưu: là độ dài trễ áp dụng thống nhất cho tất cả các biến có trong mô hình dự báo, theo đó độ chính xác dự báo của mô hình là cao nhất. Độ trễ riêng tối ưu: là độ dài trễ tối ưu cho riêng từng biến trong mô hình để độ chính xác dự báo của mô hình là cao nhất.
6	Sai phân (có/không có mùa vụ)	Là phép toán thường được sử dụng để biến đổi chuỗi thời gian không dừng thành chuỗi dừng. Giả sử chuỗi thời gian $Y_t = \{y_0, y_1, y_2, \dots, y_t\}$, sai phân bậc 1 $D(Y_t)$ của chuỗi này được xác định như sau:

		D(Y_t) = {NA, $y_{1-}\, y_0\, , y_{2-}\, y_1, \dots , y_t\, -\, y_{t-1}$}. Sai phân bậc k $(k > 1)$ của một chuỗi thời gian là sai phân bậc 1 của sai phân bậc $k - 1$.
7	Số quan sát	Là số lượng mẫu (quan sát) trong tập dữ liệu.
8	Chiều biến	Là số lượng các biến trong tập dữ liệu.
9	Cân chỉnh trung bình	Là phép biến đổi dữ liệu của chuỗi thời gian thành chuỗi mới sao tổng giá trị dữ liệu của các quan sát của chuỗi đó bằng 0.
10	Biểu thị (biểu diễn) tuyến tính	Giả sử x_i là một chuỗi thời gian (hay là một véc tơ trong $\mathbb{R}^N$), khi đó nếu $v = \sum_{i=1}^{m} \alpha_i . x_i$, $\alpha_i \in \mathbb{R}$ thì v được gọi là biểu thị (biểu diễn) tuyến tính qua các $x_i, \forall i = 1, \dots , m$
11	Phương pháp OLS	Phương pháp ước lượng bình phương tuyến tính nhỏ nhất.
12	Biến cứng/biến mềm	- Biến cứng là những biến mà dữ liệu của nó được *thu thập theo định kỳ* thời gian thường bởi các cơ quan, tổ chức thống kê. - Biến mềm là những biến mà dữ liệu được thu thập thông qua các hoạt động khảo sát, điều tra hoặc thông qua các phương tiện truyền thông đại chúng, các mạng xã hội và thường không theo định kỳ.
13	Skewness, Kurtosis và Jarque-Bera	- Skewness - là thước đo sự bất đối xứng của phân phối dữ liệu của chuỗi thời gian. - Kurtosis – là thước đo lường đỉnh (peakedness) và độ phẳng (flatness) của phân phối dữ liệu của chuỗi thời gian. - Jarque-Bera là kiểm định thống kê được sử dụng để kiểm tra xem chuỗi dữ liệu có phân phối chuẩn hay không.

MỞ ĐẦU

1. Cơ sở và động lực nghiên cứu

Các tập dữ liệu thế giới thực trong lĩnh vực kinh tế - tài chính thường là dữ liệu chuỗi thời gian ở đó số lượng các biến nói chung là lớn, thậm chí lớn hơn nhiều số quan sát, và người ta không thể xây dựng được mô hình dự báo và thực hiện dự báo trên các tập dữ liệu như vậy bằng các kỹ thuật thống kê. Để vượt qua thách thức này hiện có hai cách tiếp cận chủ yếu nhất là học sâu và giảm chiều dữ liệu.

Cách tiếp cận học sâu được xem là phù hợp nhất trên tập dữ liệu chuỗi thời gian là sử dụng mô hình học sâu mạng nơtron bộ nhớ ngắn dài (LSTM) [1], [2], [3], [4], mô hình mạng các đơn vị định kỳ kiểm soát (GRU) [5], và mô hình Transformer chuỗi thời gian [6], [7]. Các mô hình học sâu LSTM, GRU và Transformer bị hạn chế trong việc xử lý dữ liệu tuần tự đầu vào có sự phụ thuộc lâu dài, trong liên kết các công thức lan truyền ngược theo thời gian, trong xử lý tính mùa vụ và gặp vấn đề về số biến lớn và độ dốc biến mất (vanishing gradient) [8], [9]. Có thể nói rằng đến nay việc ứng dụng các phương pháp học sâu nêu trên trong các bài toán dự báo trên tập dữ liệu chuỗi thời gian lớn (hay tập dữ liệu của một số lớn các biến chuỗi thời gian) trong các lĩnh vực kinh tế - tài chính vẫn ở giai đoạn sơ khai, còn nhiều hạn chế [4], [5], [11].

Nghiên cứu [12] tìm thấy nhiều bằng chứng cho thấy việc kết hợp các kỹ thuật giảm chiều và kỹ thuật học máy để xây dựng mô hình dự báo là cách tiếp cận thống trị trong xây dựng mô hình dự báo trên các tập dữ liệu chuỗi thời gian lớn. Các nghiên cứu [13], [14], [15], [16], [17] cho thấy độ chính xác dự báo của các mô hình được xây dựng dựa vào các mô hình nhân tố, ở đó các nhân tố được chiết xuất từ tập dữ liệu ban đầu bằng các phương pháp giảm chiều PCA hoặc SPCA luôn bằng hoặc cao hơn so với các mô hình dự báo chuẩn khác. Nghiên cứu mới đây [17] cũng đánh giá rằng độ chính xác dự báo của mô hình được xây dựng trên tập dữ liệu chuỗi thời gian lớn theo cách tiếp cận 3 bước là: lựa chọn biến, sử dụng phương pháp giảm chiều PCA, và hồi quy rừng ngẫu nhiên kinh tế (Macroeconomic Random Forest) là cao nhất so với các mô hình được xây dựng theo nhiều cách tiếp cận khác bao gồm cách tiếp cận sử dụng các kỹ thuật học sâu, xích markov, hồi quy lượng tử, ước lượng bình phương tuyến tính nhỏ nhất, …

PCA là phương pháp giảm chiều tuyến tính điển hình. Nghiên cứu [18] chỉ ra rằng PCA là phương pháp giảm chiều tuyến tính tốt nhất do nó bảo toàn cấu trúc hiệp phương sai và phương sai cực đại của tập dữ liệu ban đầu. Bằng thực nghiệm các nghiên cứu [19], [20] cho thấy trên các tập dữ liệu thế giới thực không có phương pháp giảm chiều nào trong 12 phương pháp giảm chiều phi tuyến hàng đầu là tốt hơn phương pháp PCA mặc dù với các tập dữ liệu nhân tạo, cả 12 phương pháp đó đều cho kết quả giảm chiều khá tốt. Nghiên cứu [21] chỉ ra rằng phương pháp giảm chiều PCA là không hiệu quả với các tập dữ liệu không xấp xỉ một siêu phẳng. Như vậy, kết quả nghiên cứu trong [19], [20] tiết lộ rằng các tập dữ liệu thế giới thực được thực nghiệm trong các nghiên cứu đó có vẻ gần xấp xỉ một siêu phẳng. Tuy nhiên thực tế cho thấy các tập dữ liệu chuỗi thời gian thế giới thực không phải lúc nào cũng như vậy.

Những trình bày ở trên là động lực để Luận án nghiên cứu đề xuất một phương pháp giảm chiều biến mới trên tập dữ liệu chuỗi thời gian lớn. Các nghiên cứu [13], [14], [15], [16] và nhất là [17], [19] và [20] đã gợi ý phương pháp này cần phải là mở rộng tự nhiên của phương pháp PCA (tức là trong những trường hợp đặc biệt, phương pháp được đề xuất là phương pháp PCA), khắc phục được hạn chế của phương pháp PCA được chỉ ra trong nghiên cứu [21] là có thể được sử dụng để giảm chiều tập dữ liệu chuỗi thời gian lớn không xấp xỉ một siêu phẳng, và hiệu suất giảm chiều của phương pháp được đề xuất cần bằng hoặc cao hơn hiệu suất giảm chiều của phương pháp PCA. Ở đây hiệu suất của một phương pháp giảm chiều được đo bằng sai số dự báo bình phương trung bình chuẩn (RMSE). Nó đóng vai trò như là hàm mất mát (hàm LOSS).

Mục đích của giảm chiều là tăng tính hiệu quả (tốn ít thời gian và bộ nhớ) và tính dễ giải thích cho các mô hình dự báo được xây dựng trên tập dữ liệu lớn sử dụng phương pháp giảm chiều. Việc đề xuất một quy trình hoặc thuật toán dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất và áp dụng quy trình hoặc thuật toán đó để dự báo các chỉ số kinh tế - tài chính quan trọng cũng cần được nghiên cứu khảo sát. Với mọi quốc gia, dự báo kim ngạch xuất khẩu của toàn nền kinh tế cũng như từng ngành kinh tế luôn là một trong những nội dung dự báo kinh tế vĩ mô quan trọng nhất. Việt Nam có nền kinh tế mở, ở đó kim ngạch xuất,

nhập khẩu chiếm tỷ trọng rất cao trong tổng sản phẩm quốc nội (GDP) vì thế việc dự báo kim ngạch xuất khẩu càng quan trọng và cần thiết hơn. Cùng với tiến trình hội nhập quốc tế ngày càng sâu rộng, các yếu tố tác động đến kim ngạch xuất khẩu của Việt Nam ngày càng lớn. Vấn đề dự báo kim ngạch xuất khẩu trên tập dữ liệu lớn đã được đặt ra. Vì vậy, việc đề xuất quy trình/thuật toán dự báo sử dụng phương pháp giảm chiều được đề xuất và ứng dụng nó trong dự báo kim ngạch xuất khẩu theo tháng của Việt Nam cũng là một trong những động lực nghiên cứu chính để NCS thực hiện Luận án "NGHIÊN CỨU PHƯƠNG PHÁP GIẢM CHIỀU BIẾN DỰA TRÊN HÀM NHÂN VÀ ỨNG DỤNG TRONG BÀI TOÁN DỰ BÁO KIM NGẠCH XUẤT KHẨU".

Cụ thể luận án tập trung nghiên cứu đề xuất phương pháp giảm chiều trên các tập dữ liệu chuỗi thời gian lớn khắc phục được hạn chế và có hiệu suất giảm chiều nổi trội hơn một số phương pháp giảm chiều hiện được sử dụng phổ biến và được xem là hiệu quả nhất trong lĩnh vực kinh tế - tài chính và đề xuất quy trình/thuật toán dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất và ứng dụng của nó trong lĩnh vực kinh tế - tài chính, mà trước hết là lĩnh vực xuất khẩu.

2. Mục tiêu, đối tượng, phạm vi và phương pháp nghiên cứu

2.1 Mục tiêu nghiên cứu của luận án

Mục tiêu tổng quát của luận án này là nghiên cứu đề xuất phương pháp giảm chiều biến hiệu quả trên các tập dữ liệu chuỗi thời gian lớn và ứng dụng của chúng trong dự báo trong lĩnh vực kinh tế - tài chính.

Mục tiêu cụ thể của luận án như sau:

- Đề xuất phương pháp giảm chiều mới khắc phục được nhược điểm của các phương pháp giảm chiều đang được ứng dụng rộng rãi, hiệu quả trong lĩnh vực kinh tế - tài chính. Hiệu suất giảm chiều của phương pháp được đề xuất không nhỏ thua hiệu suất giảm chiều của các phương pháp hiện được ứng dụng phổ biến trong lĩnh vực kinh tế - tài chính.

- Đề xuất quy trình/thuật toán dự báo (có điều kiện cũng như không có điều kiện) trên các tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được

đề xuất và ứng dụng quy trình/thuật toán này để thực hiện dự báo chỉ số kim ngạch xuất khẩu Việt Nam trên tập dữ liệu của một số lớn các chỉ số kinh tế - tài chính.

2.2 Đối tượng nghiên cứu

Các phương pháp giảm chiều trên các tập dữ liệu chuỗi thời gian lớn và ứng dụng của chúng trong việc xây dựng mô hình dự báo cũng như mô hình nowcast trong lĩnh vực kinh tế - tài chính.

2.3 Phạm vi nghiên cứu

Các phương pháp giảm chiều dữ liệu thuộc họ PCA và các phiên bản phát triển của nó trên các tập dữ liệu chuỗi thời gian, ở đó số lượng các biến là rất lớn và ứng dụng của chúng trong lĩnh vực kinh tế - tài chính, trước hết tập trung vào lĩnh vực xuất khẩu.

2.4 Phương pháp nghiên cứu

- Phương pháp phân tích, tổng hợp được sử dụng trong việc phân tích và tổng hợp nguồn tài liệu và nội dung liên quan đến các phương pháp giảm chiều trên các tập dữ liệu chuỗi thời gian lớn bao gồm phương pháp lựa chọn thuộc tính và học thuộc tính, và ứng dụng của các phương pháp giảm chiều trong việc xây dựng mô hình dự báo trên các tập dữ liệu chuỗi thời gian có cùng tần suất lấy mẫu và có tần suất lấy mẫu hỗn hợp. Từ đó phát hiện các khoảng trống nghiên cứu.

- Phương pháp nghiên cứu lý thuyết được sử dụng để đề xuất phương pháp giảm chiều mới đối với các tập dữ liệu chuỗi thời gian lớn, cụ thể là đề xuất phương pháp giảm chiều biến dựa vào kỹ thuật hàm nhân.

- Phương pháp so sánh và thực nghiệm được sử dụng để đánh giá hiệu suất giảm chiều biến của phương pháp được đề xuất so với các phương pháp khác như phương pháp PCA và các phương pháp SPCA bao gồm SPCA, phương pháp SPCA được ngẫu nhiên hóa (RSPCA), và phương pháp SPCA vững (ROBSPCA).

- Phương pháp mô hình hóa được sử dụng để thực hiện dự báo (có điều kiện và không điều kiện) trong lĩnh vực kinh tế - tài chính bằng mô hình định lượng được xây dựng trên tập dữ liệu chuỗi thời gian lớn ứng dụng phương pháp giảm chiều biến được đề xuất.

2.5 Các tập dữ liệu

Các tập dữ liệu chuỗi thời gian thế giới thực trong một số lĩnh vực kinh tế - tài chính được sử dụng trong Luận án bao gồm:

- 07 tập dữ liệu được thu thập từ cơ sở dữ liệu UCI có tên là Residential Building [22], S&P 500, DJI, và Nasdaq [23], Air Quality [24], Appliances Energy [25], và SuperConductivity [26].

- Các tập dữ liệu thực của nền kinh tế Việt Nam được ký hiệu EXP, VN30, CPI, VIP, IIP được thu thập từ các nguồn: Tổng cục thống kê Việt Nam (GSO); công ty Fiinpro chuyên cung cấp dịch vụ dữ liệu tài chính và kinh doanh; các chỉ số chứng khoán trong nước chẳng hạn rổ VN30 được thu thập trên trang web; các số liệu tài chính như giá cả thế giới của một số loại hàng hóa, một số chỉ số chứng khoán quốc tế như NASDAD, S&P 500, NIKKEI,…, được thu thập từ Quỹ tiền tệ quốc tế IMF[1], cục dự trữ liên bang Mỹ FED[2], liên minh Châu Âu EUROSTAT[3]. Một số số liệu điều tra được thu thập từ một số cuộc khảo sát được tổ chức thường xuyên như chỉ số người quản trị mua hàng PMI[4].

Các tập dữ liệu này được sử dụng để thực nghiệm đánh giá hiệu suất giảm chiều biến do Luận án đề xuất. Tập dữ liệu EXP còn được sử dụng để xây dựng mô hình dự báo kim ngạch xuất khẩu theo tháng của Việt Nam. Đặc trưng thống kê của các tập dữ liệu đó sẽ được trình bày chi tiết trong một chương nội dung của Luận án.

3. Ý nghĩa lý luận và thực tiễn của luận án

Nội dung nghiên cứu của luận án có ý nghĩa quan trọng về khía cạnh:

- Ý nghĩa khoa học: Cung cấp một giải pháp giảm chiều biến trên các tập dữ liệu chuỗi thời gian lớn có thể xấp xỉ một siêu phẳng hoặc không và ứng dụng của nó trong các bài toán dự báo trên các tập dữ liệu chuỗi thời gian lớn có tần suất lấy mẫu giống nhau hoặc khác nhau (còn được gọi là tần suất hỗn hợp).

- Ý nghĩa thực tiễn: Các kết quả nghiên cứu của luận án có thể ứng dụng được ngay vào thực tế của cuộc sống. Độ chính xác dự báo của các mô hình được xây dựng

[1] www.imf.org
[2] www.fred.stlouisfed.org
[3] https:/lec.europa.eu/eurostat
[4] https://www.pmi.spglobal.com

trên các tập dữ liệu lớn bằng sử dụng phương pháp giảm chiều được đề xuất là rất cao.

4. Những đóng góp chính của luận án

- Đề xuất phương pháp giảm chiều biến trên các tập dữ liệu chuỗi thời gian lớn dựa vào kỹ thuật hàm nhân (gọi tắt KTPCA). Nó là mở rộng tự nhiên của phương pháp PCA, có thể được sử dụng để giảm chiều biến trên các tập dữ liệu xấp xỉ hoặc không xấp xỉ một siêu phẳng. Hiệu suất giảm chiều của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất (gọi tắt là KTPCA lặp) là bằng hoặc cao hơn các phương pháp giảm chiều PCA, SPCA, RSPCA, và ROBSPCA trên các tập dữ liệu lấy mẫu tần suất giống nhau cũng như hỗn hợp, trong đó các mô hình nowcast/dự báo được xây dựng dựa trên các nhân tố được chiết xuất bằng các phương pháp KTPCA, PCA, SPCA, RSPCA và ROBSPCA. Liên quan đến đóng góp này là các bài báo [CT3], [CT6] thuộc danh mục các Nghiên cứu của Luận án.

- Đề xuất thuật toán dự báo có và không có điều kiện trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều biến KTPCA lặp và ứng dụng của nó để thực hiện dự báo có và không có điều kiện kim ngạch xuất khẩu. Độ phức tạp tính toán của thuật toán được đề xuất là đa thức bậc 3 của số lượng biến và số quan sát cùng với độ chính xác dự báo bằng ứng dụng thuật toán đó là khá cao. Cụ thể, với mô hình dự báo kim ngạch xuất khẩu sử dụng quy trình/thuật toán dự báo không điều kiện thì % sai số dự báo trung bình của mô hình này là cao hơn % sai số dự báo của mô hình ARIMA(2,1,2) đến 2.38 điểm %, làm tăng độ chính xác dự báo lên 63,6%, trong khi với mô hình dự báo kim ngạch xuất khẩu sử dụng quy trình/thuật toán dự báo có điều kiện thì % sai số dự báo trung bình của mô hình này là cao hơn % sai số dự báo của mô hình cầu xuất khẩu đến 1.62 điểm %, làm tăng độ chính xác dự báo lên 52.9%. Điều này cho thấy triển vọng ứng dụng của phương pháp giảm chiều cùng thuật toán dự báo sử dụng phương pháp giảm chiều đó để không chỉ dự báo kim ngạch xuất khẩu mà còn có thể dự báo các chỉ tiêu kinh tế - tài chính khác trên các tập dữ liệu chuỗi thời gian lớn.

Liên quan đến đóng góp này là các bài báo [CT1], [CT2], [CT4] [CT5] thuộc danh mục các Nghiên cứu của luận án.

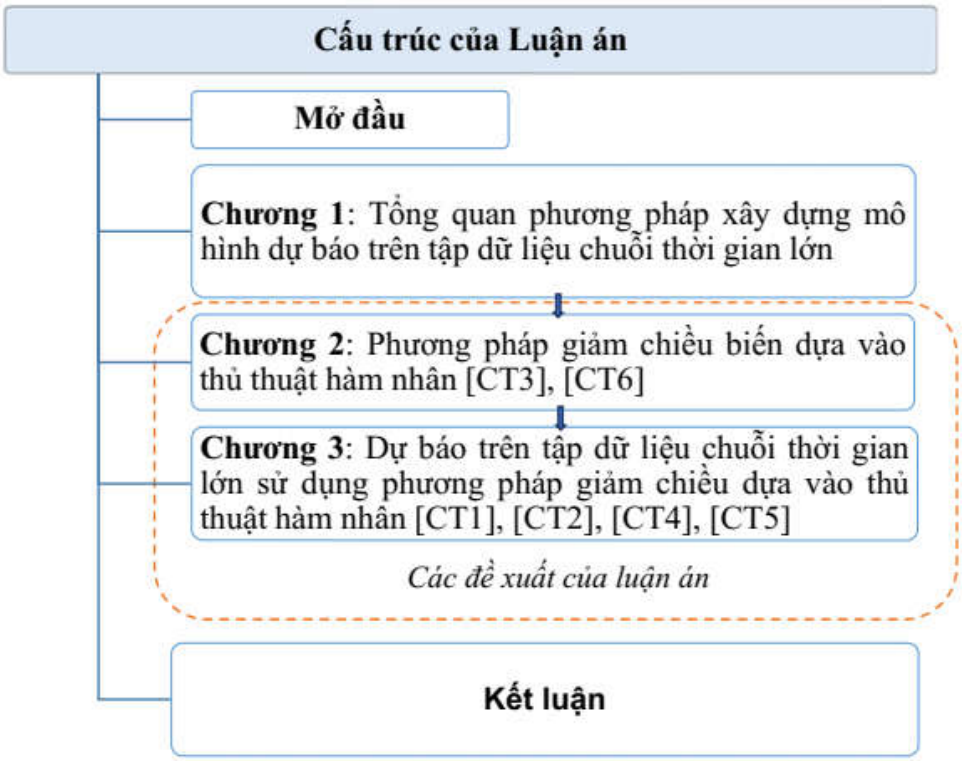
5. Cấu trúc của luận án

Cấu trúc của luận án gồm:

- **Phần mở đầu**: Trình bày cơ sở lý thuyết và động lực nghiên cứu của luận án; mục tiêu, đối tượng, phạm vi nghiên cứu; phương pháp nghiên cứu; những đóng góp chính và cấu trúc của luận án.

- **Chương 1**: Tổng quan về phương pháp xây dựng mô hình dự báo và mô hình nowcast trên tập dữ liệu chuỗi thời gian lớn; xác định vấn đề và phạm vi nghiên cứu, một số kiến thức liên quan và cuối cùng là một số kết luận.

- **Chương 2**: Đề xuất phương pháp giảm chiều biến cho các tập dữ liệu chuỗi thời gian lớn dựa vào kỹ thuật hàm nhân, gọi là KTPCA, và so sánh hiệu suất giảm chiều biến của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất với hiệu suất giảm chiều biến của các phương pháp PCA và họ SPCA trên các tập dữ liệu có cùng hoặc không cùng tần suất lấy mẫu, và cuối cùng là một số kết luận.



Hình 0.1: Cấu trúc của luận án

- **Chương 3**: Đề xuất thuật toán dự báo có và không có điều kiện trên các tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất, và ứng

dụng thuật toán này để dự báo có và không có điều kiện kim ngạch xuất khẩu theo tháng của Việt Nam.

Phần kết luận trình bày những đóng góp nghiên cứu chính của luận án và hạn chế của Luận án.

CHƯƠNG 1. TỔNG QUAN PHƯƠNG PHÁP XÂY DỰNG MÔ HÌNH DỰ BÁO TRÊN TẬP DỮ LIỆU CHUỖI THỜI GIAN LỚN

1.1 Tổng quan các nghiên cứu trong và ngoài nước

Những thuật ngữ như: Prediction, Forecast, Nowcast và Foresight đều nói về dự báo nhưng chúng có một số điểm khác biệt. Theo từ điển tiếng Anh của Đại học Cambridge:

- "Prediction" là một nhận định về những gì mà ta nghĩ sẽ xảy ra trong tương lai và thường được gọi là dự đoán. Như vậy để dự đoán tương lai người làm dự báo có thể sử dụng dữ liệu lịch sử hoặc không.

- "Forecast" là một nhận định về tương lai được tính toán từ dữ liệu lịch sử. Nói cách khác "Forecast" là "Prediction" được thực hiện, tính toán từ dữ liệu lịch sử. Thuật ngữ này thường được gọi là dự báo.

- "Foresight" là nói về khả năng phán đoán chính xác điều gì sắp xảy ra. Giống như "Prediction", những phán đoán ấy có thể được tính toán, rút ra từ dữ liệu lịch sử hoặc không. Khác với "Prediction", "Foresight" – thường là phán đoán cho tương lai dài, thậm chí rất dài và thường được thực hiện bằng phương pháp định tính (phân tích định tính).

- Theo nghiên cứu [27], "Nowcast" là "Prediction" về hiện tại, tương lai gần và quá khứ mới đây. Trong trường hợp các tập dữ liệu lịch sử là tập dữ liệu chuỗi thời gian thì "nowcast" là dự báo biến phụ thuộc ở tần suất lấy mẫu thấp theo các biến giải thích ở một số tần suất lấy mẫu khác cao hơn. Chẳng hạn việc dự báo chỉ số GDP ở tần suất lấy mẫu theo quý (hay GDP quý) theo các biến kinh tế được lấy mẫu *theo tháng* như kim ngạch xuất nhập khẩu, chỉ số phát triển công nghiệp, chỉ số giá tiêu dùng, đầu tư từ ngân sách nhà nước,…; theo các biến được lấy mẫu *theo tuần* như các loại lãi suất tiền gửi theo tháng, quý, năm được các ngân hàng công bố hàng tuần; và *theo ngày* như chỉ số chứng khoán của 30 công ty có giá trị vốn hóa lớn nhất của Việt Nam trên thị trường chứng khoán, … chính là nowcasting GDP. Các mô hình nowcast cho phép cập nhật dự báo theo luồng dữ liệu thời gian thực (theo dữ liệu ở tần suất cao hơn ngay khi chúng được công bố ở những thời điểm có thể rất khác

nhau). Khi các tần suất lấy mẫu trong tập dữ liệu là giống nhau thì bài toán nowcast sẽ trở thành bài toán dự báo.

Thuật ngữ "Nowcast" được nói đến lần đầu năm 1981 [28] và được định nghĩa một cách chính xác năm 2006 [27]. Theo đó nó là một sự kết hợp của "now" (hiện tại) và "cast" (dự báo) với mong muốn dự báo kinh tế-xã hội cũng có thể được thực hiện theo cách của dự báo thời tiết.

Các mô hình nowcast sử dụng các thông tin có sẵn, kịp thời và đáng tin cậy để hình thành các dự báo cho các biến quan tâm [9], [29], [30]. Việc sử dụng kịp thời các thông tin, dữ liệu tin cậy có thể có đã nói rằng thông tin dữ liệu được sử dụng trong các mô hình nowcast là rất lớn. Nó không chỉ gồm dữ liệu thống kê (được gọi là dữ liệu cứng) mà còn gồm những thông tin dữ liệu khác không phải là dữ liệu thống kê (được gọi là dữ liệu mềm). Những dữ liệu được tạo ra bởi các cuộc điều tra hay được thu thập từ các phương tiện thông tin đại chúng, các mạng xã hội là thuộc loại dữ liệu mềm. Các mô hình nowcast cho phép cập nhật dự báo theo luồng dữ liệu thời gian thực và bằng việc sử dụng mô hình nowcast, người ta có phản ứng kịp thời và chính xác trước các biến động ngày càng nhanh và khó lường của các hiện tượng tự nhiên, kinh tế, xã hội như hiện nay.

Việc xây dựng các mô hình nowcast là rất cần thiết để hỗ trợ công tác chỉ đạo điều hành và hoạch định chính sách của các cơ quan chính phủ, để hỗ trợ các hoạt động sản xuất kinh doanh của các doanh nghiệp nhất là những doanh nghiệp quy mô lớn, có quan hệ trao đổi thương mại cao với các doanh nghiệp bên ngoài.

1.1.1 Các nghiên cứu ngoài nước

1.1.1.1 Phương pháp xây dựng mô hình dự báo trên tập dữ liệu tần suất lấy mẫu giống nhau

Giả sử $Y_t = (y_1, \dots, y_t) \in \mathbb{R}^t$ và $\boldsymbol{X_t} = [x_{1,t}, x_{2,t}, \dots, x_{m,t}] \in \mathbb{R}^{t \times m}$ tương ứng là biến phụ thuộc (hay biến cần quan tâm) và tập các biến giải thích; *m* và *t* tương ứng là số lượng các biến và các quan sát. Mô hình dự báo biến Y_t theo các biến giải thích $\boldsymbol{X}_t$ có dạng**:**

$$Y_t = F(Y_{t-k}, \boldsymbol{X}_{t-p}) + u_t \; với \; 1 \le k \le t-1; \; 0 \le p \le t-1 \qquad (1.1)$$

ở đây u_t là phần dư với giả định là nhiễu trắng, Y_{t-k} là trễ bậc k của biến Y_t $(k \geq 1)$; $F(.)$ là hàm tuyến tính hoặc phi tuyến, ở dạng ẩn hoặc ở dạng tường minh. Trong thực hành ứng dụng hàm $F(.)$ được ước lượng từ *t* quan sát đã cho của biến phụ thuộc và biến giải thích. Hàm *F*(.) thường được xác định bằng phương pháp hồi quy hoặc các bộ phân lớp tùy thuộc biến phụ thuộc nhận giá trị số hay giá trị phân loại. Khi số lượng biến *m* là lớn hơn số quan sát *t* hoặc khi số lượng biến *m* là rất lớn thì các kỹ thuật hồi quy theo mô hình phương trình (1.1) là không thể thực hiện được. Các nhà mô hình hóa kinh tế gọi đó là "lời nguyền về chiều" (the curse of dimensionality).

Các nghiên cứu [9], [30], [31], [32], [33], [34], [35] đã tổng quan các phương pháp và kỹ thuật dự báo được sử dụng trên các tập dữ liệu chuỗi thời gian lớn. Có thể nói học sâu và giảm chiều là 2 cách tiếp cận chủ yếu nhất được sử dụng khi thực hiện dự báo hoặc phân lớp trên các tập dữ liệu lớn.

a. Dự báo sử dụng phương pháp học sâu

Học sâu là mô hình mạng nơtron nhiều lớp, đã được chứng minh là có đặc tính nhận dạng mẫu tốt. Về bản chất học sâu mạng nơtron [36] là một mô hình hồi quy phi tuyến dựa vào độ dốc giảm dần, ở đó hàm $F(.)$ trong mô hình (1.1) ở trên không được xác định một cách tường minh. Về nguyên tắc có thể tìm được hàm $F(.)$ tối ưu nhưng để tìm được hàm như vậy thì chi phí thời gian và tính toán tốn kém và có thể là không phù hợp với yêu cầu nhanh chóng và kịp thời.

Cách tiếp cận học sâu được xem là phù hợp nhất trên tập dữ liệu chuỗi thời gian là sử dụng mô hình mạng nơtron bộ nhớ ngắn dài (LSTM) [1], [2], [3], [4], [11]; mô hình mạng các đơn vị định kỳ được kiểm soát (GRU) [5], và mô hình Transformer chuỗi thời gian [6], [7]. Theo [9] các mô hình học sâu LSTM, GRU chỉ phù hợp với tập dữ liệu chuỗi thời gian ở đó số quan sát là lớn trong khi số biến (hay chiều biến) là không quá lớn. Theo [8] các mô hình học sâu nói trên vẫn bị hạn chế trong việc xử lý dữ liệu đầu vào có tính tuần tự, trong liên kết các công thức lan truyền ngược theo thời gian, và nhất là khi xử lý dữ liệu lớn có sự phụ thuộc lâu dài. Do đó việc xử lý tính mùa vụ cũng như xác định chính xác độ trễ tối ưu của các biến dữ liệu trong các mô hình này là bị hạn chế. Quá trình đào tạo các mô hình LSTM và GRU cũng gặp phải vấn đề về biến số và độ dốc (gradient) [8], đòi hỏi chi phí thời gian và tính toán lớn.

Mô hình học sâu Transformer đã đạt được hiệu suất vượt trội về xử lý ngôn ngữ tự nhiên và thị giác máy tính. Trong số nhiều ưu điểm của Transformer, khả năng nắm bắt sự phụ thuộc và tương tác ở phạm vi dài của mô hình này đã thu hút nhiều nhà nghiên cứu xây dựng mô hình dự báo chuỗi thời gian dựa vào mô hình Transformer. Trung tâm của Transformer là khả năng tự chú ý. Nó cho phép một lớp được kết nối đầy đủ với các trọng số được sinh ra dựa trên sự giống nhau theo cặp của các mẫu đầu vào. Kết quả là nó chia sẻ cùng một đường dẫn tối đa như các lớp được kết nối đầy đủ nhưng với số lượng tham số ít hơn nhiều, khiến nó phù hợp để lập mô hình sự phụ thuộc lâu dài. Tuy nhiên các kết quả đạt được của mô hình Transformer mới sơ khai ban đầu. Vấn đề xử lý tính mùa vụ và tính chu kỳ của dữ liệu chuỗi thời gian bằng sử dụng mô hình Transformer vẫn còn nhiều hạn chế. Thông qua nghiên cứu thực nghiệm, nghiên cứu [10] cho thấy mô hình dựa trên mạng nơtron đa lớp đơn giản có thể đạt được kết quả dự báo tốt hơn so với mô hình Transformer chuỗi thời gian. Cho đến thời điểm này việc ứng dụng phương pháp học sâu LSTM, GRU, hay Transformer trong các bài toán dự báo trên tập dữ liệu lớn của các biến giải thích chuỗi thời gian trong lĩnh vực kinh tế - tài chính vẫn còn nhiều hạn chế [5], [6], [7], [11].

b. Dự báo sử dụng phương pháp giảm chiều

Nghiên cứu [37] có thể được xem là nghiên cứu đầu tiên về việc xây dựng mô hình dự báo trên tập dữ liệu chuỗi thời gian lớn bằng sử dụng phương pháp giảm chiều PCA. Nghiên cứu này cho rằng có thể thay thế một số lớn các biến giải thích bằng một số ít các nhân tố ẩn (hidden factor), đó là các thành phần chính được chiết xuất bằng phương pháp PCA. Nghiên cứu này cũng cho biết dấu hiện để nhận biết một tập dữ liệu có xấp xỉ một siêu phẳng hay không. Theo đó, một tập dữ liệu sẽ không xấp xỉ một siêu phẳng nếu khi tăng tỷ lệ tích lũy phương sai thì số lượng các nhân tố thành phần chính sẽ tăng rất nhanh.

Quy trình dự báo sử dụng phương pháp giảm chiều nói chung gồm 2 Giai đoạn chính như được thể hiện trong Hình 1.1 ở dưới [38], [39], [40]. Nội dung chính của Giai đoạn 1 là thực hiện giảm chiều dữ liệu. Giai đoạn này nói chung gồm 2 bước [38]. Bước 1 thực hiện phương pháp giảm chiều lựa chọn thuộc tính nhằm chọn ra các biến có tác động thực sự đến sự biến đổi của biến cần được dự báo. Bước 2 sử

dụng phương pháp học thuộc tính nhằm chuyển đổi tập dữ liệu của một số lượng lớn các biến được lựa chọn ở Bước 1 thành tập dữ liệu của một số nhỏ các biến mới nhưng vẫn nắm bắt được những thông tin quan trọng trong tập dữ liệu ban đầu. Tập các biến mới sẽ được dùng để thay thế cho tập các biến giải thích trong các bài toán dự báo trên tập dữ liệu lớn. Giai đoạn 2 sử dụng kỹ thuật hồi quy hoặc kỹ thuật phân lớp tùy thuộc giá trị của biến phụ thuộc nhận giá trị số hay giá trị phân loại để xây dựng tương ứng mô hình dự báo hay bộ phân lớp. Việc thực hiện dự báo hoặc phân lớp trên tập dữ liệu kiểm thử nhằm kiểm định và đánh giá chất lượng của mô hình dự báo hoặc bộ phân lớp, nếu mô hình hoặc bộ phân lớp đó được chấp nhận thì nó sẽ được sử dụng để thực hiện dự báo biến phụ thuộc hoặc phân lớp các tập dữ liệu đầu vào mới.

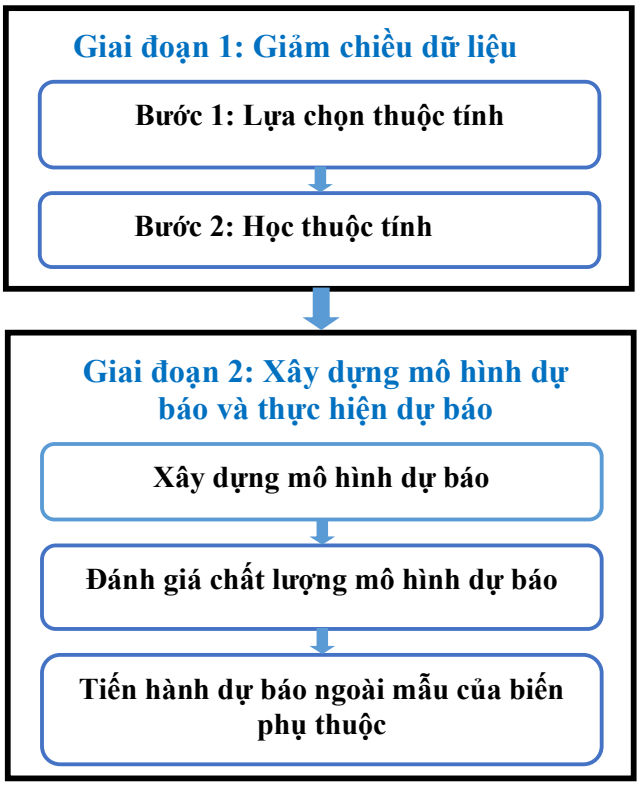
Bước 1: Lựa chọn thuộc tính (hay lựa chọn biến) là kỹ thuật nhằm lựa chọn tập con các biến có ảnh hưởng đến sự thay đổi của biến phụ thuộc bằng cách loại bỏ các biến không hoặc ít liên quan đến biến phụ thuộc (các biến gây nhiễu) hoặc các biến dư thừa với mục đích dự báo hoặc phân lớp trong tập các biến giải thích.

Các kỹ thuật lựa chọn biến được chia thành 03 loại theo 03 cách tiếp cận khác nhau bao gồm [41]:

- Phương pháp tiếp cận bộ lọc (Filter): Các biến được sắp xếp theo một số tiêu chí nào đó và sau đó lựa chọn các biến có tiêu chí đạt trên một ngưỡng xác định. Những phương pháp lọc thuộc tính điển hình như: Lọc Chi - Bình phương, Lọc Tương quan, Lọc dựa vào Entropy, Lọc Rừng ngẫu nhiên [42]. Trong ứng dụng thực tế, việc lựa chọn thuộc tính theo cách tiếp cận lọc trong các bài toán dự báo trong lĩnh vực kinh tế - tài chính là việc kết hợp sử dụng lý thuyết kinh tế và độ đo hệ số tương quan Pearson đối với các biến liên tục (nhận giá trị số) hoặc độ đo thông tin tương hỗ dựa vào entropy đối với các biến phân loại.

- Phương pháp tiếp cận bọc (Wrapper): Sử dụng thuật toán để tìm kiếm tập con các biến đắt giá (biến có trên toàn bộ tập dữ liệu ban đầu bằng cách đánh giá chất lượng của các tập con các biến. Chất lượng của các tập con các biến được chọn thường được đánh giá thông qua độ chính xác dự báo hoặc độ chính xác phân lớp tương ứng của thuật toán dự báo hoặc bộ phân lớp. Các kỹ thuật học máy có thể được sử dụng trong các cách tiếp cận này là: học Máy véc tơ hỗ trợ, Cây quyết định, Mạng Bayes,

Thuật toán k người láng giềng gần nhất, Thuật toán vét cạn, Thuật toán leo đồi, Thuật toán tham lam, Mạng Nơtron, Luật kết hợp, Giải thuật di truyền, Phân cụm dữ liệu,…



Hình 1.1: Hai giai đoạn chính trong quy trình xây dựng mô hình dự báo trên tập dữ liệu có số chiều cao [38]

- Phương pháp tiếp cận nhúng (Embeded): Các biến được xếp hạng ngay trong quá trình thực thi việc học chứ không phải sau khi kết thúc quá trình học như phương pháp tiếp cận bọc. Trong lĩnh vực kinh tế - tài chính, một số phương pháp lựa chọn thuộc tính theo cách tiếp cận nhúng được xem là hiệu quả và được ứng dụng rộng rãi cho đến thời điểm này là hồi quy RIDGE, hồi quy Bayes, hồi quy LASSO, hồi quy LASSO thích nghi (A-LASSO), và hồi quy lưới đàn hồi (Elastic Net) [9]. Các mô hình này là những kỹ thuật lựa chọn tập con các biến trong các bài toán dự báo khi tập các biến giải thích là lớn. Tuy nhiên khi tập các biến giải thích là rất lớn, rõ ràng việc sử dụng các kỹ thuật hồi quy theo các phương pháp nêu trên là khó khả thi vì bản chất việc hồi quy vẫn phải được thực hiện trên tất cả các biến để chọn ra tập con biến phù hợp. Do đó cần phải thực hiện thêm các phương pháp hoặc kỹ thuật giảm chiều biến khác.

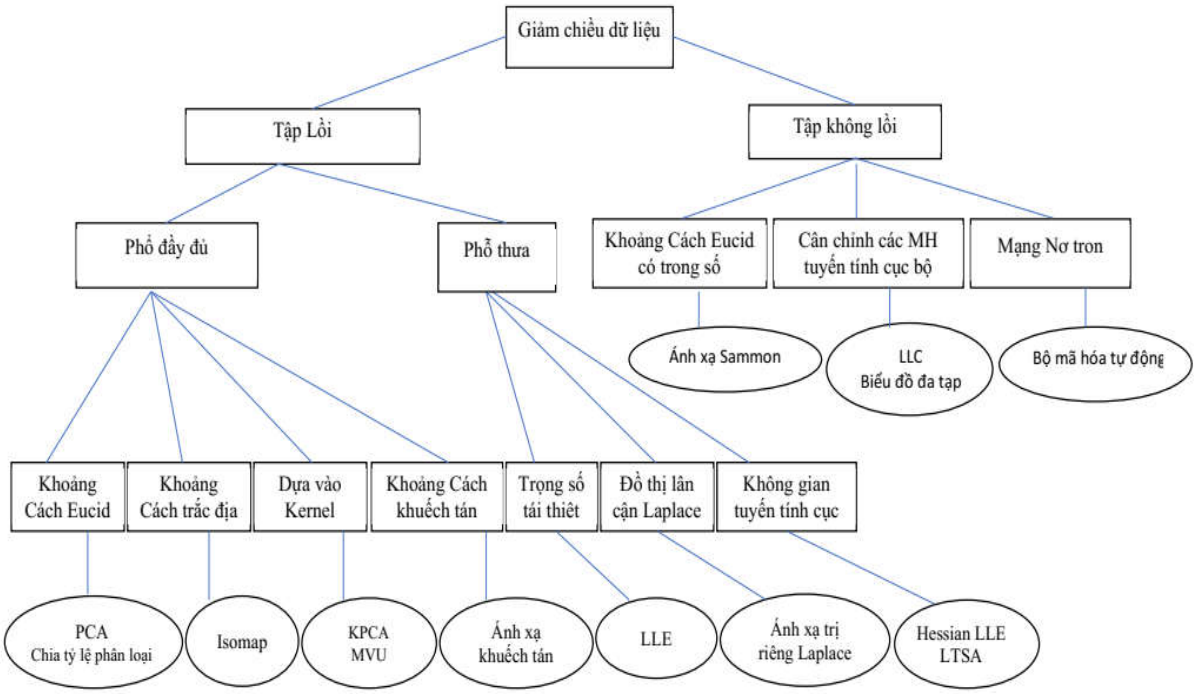
Trong 03 cách tiếp cận trên, mỗi cách tiếp cận đều có những lợi thế và bất lợi riêng của nó. Tiêu chí để phân biệt là tốc độ tính toán và nguy cơ xảy ra hiện tượng "Overfit". Theo tiêu chí về tốc độ tính toán thì phương pháp lọc là nhanh hơn so với phương pháp tiếp cận nhúng và phương pháp tiếp cận bọc là chậm nhất. Ngược lại, theo tiêu chí "Overfit" thì phương pháp bọc là xử lý tốt hơn phương pháp tiếp cận nhúng và phương pháp tiếp cận lọc nói chung là thấp nhất [43].

Các kỹ thuật lựa chọn biến cũng còn được phân theo kỹ thuật học có giám sát, bán giám sát và không giám sát. Khi đó mỗi kỹ thuật lựa chọn biến như vậy lại được phân theo 03 cách tiếp cận nêu trên. Kỹ thuật lựa chọn biến không giám sát bao gồm lựa chọn thuộc tính không giám sát theo cách tiếp cận lọc, bọc và nhúng đang được quan tâm hiện nay bởi người ta nhận thấy rằng các kỹ thuật như vậy còn cho phép phát hiện mối quan hệ giữa các biến giải thích tốt hơn so với kỹ thuật lựa chọn biến theo cùng cách tiếp cận của kỹ thuật học có giám sát [44]. Điều đó có nghĩa là việc sử dụng kỹ thuật lựa chọn biến không giám sát có thể chọn được những biến có liên quan đắt giá và loại bỏ các biến dư thừa trong mô hình dự báo biến phụ thuộc tốt hơn so với sử dụng kỹ thuật lựa chọn biến có giám sát.

Bước 2: ***Học thuộc tính (chuyển đổi biến)*****:** nhằm xây dựng một tập các biến mới nhỏ hơn rất nhiều từ tập các biến giải thích ban đầu nhưng vẫn nắm giữ được những thông tin quan trọng nhiều nhất có thể trong tập các biến này.

Suy cho cùng các kỹ thuật giảm chiều học thuộc tính đều có thể được quy về giải quyết bài toán tối ưu. Hình 1.2 trình bày một cách phân loại các kỹ thuật giảm chiều học thuộc tính theo cách tiếp cận giải quyết bài toán tối ưu lồi hoặc không [45]. Trong các kỹ thuật học này, PCA là phương pháp tối ưu lồi. Nó là phương pháp học không giám sát và là phương pháp học siêu phẳng điển hình nhằm chuyển đổi tập dữ liệu từ không gian chiều cao về không gian chiều thấp hơn nhiều mà vẫn giữ được cấu trúc hiệp phương sai và cực đại hóa phương sai của tập dữ liệu ban đầu [18]. Tập dữ liệu trong không gian chiều thấp hơn là các thành phần chính được chọn, ở đó mỗi thành phần chính là kết quả của một phép chiếu tuyến tính của tập dữ liệu ban đầu được cân chỉnh trung bình lên một véc tơ riêng của ma trận hiệp phương sai của tập dữ liệu đầu vào. Tỷ lệ phần trăm của phương sai tích lũy của k thành phần chính ứng với các giá trị riêng lớn nhất cũng chính là tỷ lệ phần trăm thông tin của tập dữ liệu

ban đầu mà k thành phần chính này (cũng được gọi là nhân tố thành phần chính) nắm giữ được. Trong thực hành người ta thường chỉ lấy ra k nhân tố thành phần chính đầu tiên tương ứng với k giá trị riêng lớn nhất sao cho tỷ lệ phần trăm của phương sai tích lũy của k thành phần chính đó từ 70% trở lên làm tập các nhân tố mới thay thế tập các biến giải thích ban đầu.



Hình 1.2: Phân loại các kỹ thuật giảm chiều học thuộc tính điển hình [19]

Nghiên cứu [46] mới đây đã tiến hành so sánh thực nghiệm hiệu suất giảm chiều của các phương pháp học thuộc tính PCA, KPCA, LDA, MDS, SVD, LLE, Isomap, LE, ICA, và T-SNE với việc sử dụng bộ phân lớp SVM với hàm nhân Gauss trên 3 tập dữ liệu chéo thế giới thực. Ở đây PCA, LDA, SVD, và ICA là phương pháp học tuyến tính sử dụng phép chiếu ngẫu nhiên, trong khi các phương pháp còn lại đều là phương pháp học đa tạp (manifold). Nghiên cứu này đã chỉ ra rằng trong hầu hết các trường hợp, các phương pháp học phi tuyến hoạt động tốt hơn phương pháp học tuyến tính và các phương pháp học đa tạp hoạt động tốt hơn phương pháp dựa vào phép chiếu ngẫu nhiên. Tuy nhiên kết luận của nghiên cứu [46] được rút ra từ kết quả thực nghiệm trên các tập dữ liệu chéo chứ không phải dữ liệu chuỗi thời gian và số lượng tập dữ liệu được thực nghiệm chỉ là 3, còn khá nhỏ.

Với tập dữ liệu thực nghiệm lớn hơn rất nhiều, nghiên cứu [19] cũng so sánh đánh giá hiệu suất giảm chiều của 12 kỹ thuật giảm chiều phi tuyến hàng đầu, bao gồm Phân tích thành phần chính với hàm nhân (KPCA) [47], [48], Isomap, Maximum Variance Unfolding, Locally Linear Embedding (LLE), Laplacian Eigenmaps (LE), Hessian LLE, Multilayer Autoencoders, Diffusion Maps, Multidimensional Scaling, Local Tangent Space Analysis, Locally Linear Coordination, và Manifold Charting. Kết quả cho thấy mặc dù 12 kỹ thuật trên có thể giảm chiều tốt với các tập dữ liệu nhân tạo, tuy nhiên không có kỹ thuật nào trong số 12 kỹ thuật này giảm chiều tốt hơn phương pháp PCA trên các tập dữ liệu thế giới thực bao gồm cả các tập dữ liệu chuỗi thời gian [19].

Nghiên cứu [20] đã thực nghiệm dự báo lợi tức chỉ số S&P500 ETF (SPY) theo tần suất ngày bằng cách sử dụng kết hợp ba phương pháp giảm chiều gồm PCA, PCA vững mờ (FRPCA) và KPCA, sau đó mạng nơron nhân tạo (ANN) được sử dụng để phân loại trên tập dữ liệu của 60 biến kinh tế - tài chính. Kết quả thực nghiệm cho thấy, tương tự như nghiên cứu [19], PCA + ANN cho độ chính xác phân loại cao hơn một chút so với KPCA + ANN và FRPCA + ANN.

Các nghiên cứu [19], [20] đã tiết lộ rằng trong thế giới thực các tập dữ liệu lớn đa phần có thể gần xấp xỉ một siêu phẳng chứ không khẳng định rằng nó luôn là như vậy và trong thực tế có thể tìm thấy rất nhiều bằng chứng cho thấy các tập dữ liệu thế giới thực không phải luôn xấp xỉ một siêu phẳng và khi đó phương pháp PCA là không hiệu quả, thậm chí có nhiều trường hợp là không thể thực hiện được [21].

Phân tích 13 phương pháp giảm chiều nêu trên có thể nhận thấy rằng trừ PCA, các phương pháp giảm chiều còn lại đều là kỹ thuật học đa tạp (hay manifold) tức là chúng phù hợp với các tập dữ liệu ở đó các điểm dữ liệu của nó là xấp xỉ một đa tạp, nhưng làm thế nào để biết các điểm dữ liệu của tập dữ liệu lớn là xấp xỉ một đa tạp lại là một thách thức lớn khác. Trong số 12 phương pháp giảm chiều phi tuyến nêu trên có nhiều phương pháp về bản chất là được phát triển từ ý tưởng bảo toàn cấu trúc hiệp phương sai và cực đại hóa phương sai của phương pháp PCA, chẳng hạn như các phương pháp KPCA, Isomap, Maximum Variance Unfolding, Diffusion Maps là như vậy. Cùng với phương pháp PCA, các phương pháp này đều bảo toàn khoảng cách [45] và được ứng dụng rất thành công để giảm chiều trong các bài toán nhận

dạng ảnh và chữ viết tay, nhưng trừ phương pháp PCA và các phát triển của nó, chưa có những ứng dụng ấn tượng của 12 phương pháp giảm chiều phi tuyến nêu trên trong các bài toán dự báo trên tập dữ liệu chuỗi thời gian lớn [5], [11].

Ngoài ra có thể thấy ý tưởng của phương pháp KPCA là được phát triển từ phương pháp PCA [47], [48] và phương pháp học máy véc tơ hỗ trợ (SVM) [49], [50]. Cụ thể, ý tưởng chính của phương pháp KPCA là sử dụng ánh xạ $\boldsymbol{\Phi}$ (có thể là tuyến tính hoặc không) để chuyển các điểm dữ liệu trong không gian đầu vào thành các điểm dữ liệu mới trong không gian có số chiều cao hơn (thậm chí có số chiều vô hạn) được gọi là không gian đặc trưng. Không gian đặc trưng có cấu trúc của một không gian véc tơ tái tạo (reproducing kernel Hilbert space). Ánh xạ $\boldsymbol{\Phi}$ được chọn sao cho trong không gian đặc trưng các điểm dữ liệu của tập $\boldsymbol{\Phi}(\mathbf{X})$ ($\mathbf{X}$ là tập các véc tơ đầu vào) là xấp xỉ một siêu phẳng và khi đó ta có thể thực hiện phương pháp PCA trên tập dữ liệu $\boldsymbol{\Phi}(\mathbf{X})$ trong không gian đặc trưng. Tuy nhiên, việc xác định được một cách tường minh ánh xạ $\boldsymbol{\Phi}$ cũng như không gian đặc trưng tương ứng là rất khó. Giải pháp khắc phục là sử dụng kỹ thuật hàm nhân, đó là thay vì phải tìm tập dữ liệu $\boldsymbol{\Phi}(\mathbf{X})$ và ma trận hiệp phương sai của nó trong không gian đặc trưng, ta chỉ cần tính ma trận $\mathbf{K} = [k_{ij}]$ với $k_{ij} = \boldsymbol{\kappa}(x_i, x_j)$, ở đây $\boldsymbol{\kappa}$ là hàm đối xứng xác định dương hoặc bán xác định dương được gọi là hàm nhân và x_i, x_j là các điểm dữ liệu của tập dữ liệu ban đầu [47], [48].

Trong lĩnh vực kinh tế - tài chính, đối với các bài toán dự báo/nowcast trên tập dữ liệu chuỗi thời gian lớn, cho đến nay người ta chủ yếu dừng ở việc sử dụng các phương pháp PCA và SPCA để chiết xuất các nhân tố thành phần chính. Kỹ thuật hồi quy điển hình được sử dụng trong xây dựng mô hình dự báo và mô hình nowcast trên tập dữ liệu chuỗi thời gian lớn tương ứng là mô hình trễ phân bố tự hồi quy ARDL [51] và mô hình DFM hoặc lọc Kalman [13], [52] tùy thuộc tập dữ liệu đó có tần suất lấy mẫu giống nhau hoặc lấy mẫu hỗn hợp. Thông qua thực nghiệm, nhiều bài báo đã chỉ ra rằng độ chính xác dự báo của các mô hình được xây dựng dựa vào mô hình ARDL nhân tố đối với bài toán dự báo và mô hình DFM đối với bài toán nowcast là cao hơn độ chính xác dự báo của các mô hình chuẩn (benchmark) khác, ở đây các nhân tố được chiết xuất từ tập dữ liệu lớn bằng các phương pháp PCA hoặc SPCA [12], [14], [15].

Phương pháp SPCA được đề xuất dựa vào lập luận rằng những thành phần chính được xác định bởi phương pháp PCA là tổ hợp tuyến tính của tất cả các biến giải thích đầu vào, điều này có vẻ khiên cưỡng vì có thể có những thành phần chính chỉ là tổ hợp tuyến tính của một vài biến giải thích như vậy [53], [54]. Khi đó mô hình dự báo được xây dựng dựa vào các nhân tố thành phần chính được chiết xuất bằng phương pháp SPCA không chỉ giải thích tốt hơn mà còn có thể cho độ chính xác dự báo cao hơn. Cũng như phương pháp PCA, phương pháp SPCA được phát triển thành nhiều phiên bản khác nhau trong đó đáng lưu ý là các phương pháp RSPCA và ROBSPCA. Về bản chất các phương pháp thuộc họ SPCA là được phát triển dựa vào sự kết hợp của phương pháp PCA và các mô hình hồi quy thưa trong đó nhất là hồi quy LASSO và hồi quy mạng đàn hồi. Phương pháp SPCA là phương pháp giảm chiều tuyến tính và tương tự như phương pháp PCA, nó cũng không phù hợp để giảm chiều các tập dữ liệu không xấp xỉ một siêu phẳng.

1.1.1.2 Phương pháp xây dựng mô hình nowcast trên tập dữ liệu lớn tần suất hỗn hợp

Các mô hình dự báo được xây dựng trên các tập dữ liệu có tần suất lấy mẫu như nhau. Khi đó để dự báo một biến phụ thuộc theo một tần suất nào đó thì các biến giải thích cũng phải ở tần suất như vậy. Những mô hình đó chưa thực sự phù hợp để dự báo các biến kinh tế vĩ mô. Trong nền kinh tế có rất nhiều hoạt động khác nhau, dẫn đến việc thống kê và ban hành số liệu của các biến kinh tế vĩ mô khác nhau cũng được thực hiện theo các tần suất khác nhau. Chẳng hạn, trong hầu hết các nền kinh tế, chỉ số GDP chỉ có thể thống kê được ở tần suất quý, trong khi nhiều chỉ số khác như kim ngạch xuất khẩu, chỉ số phát triển công nghiệp, chỉ số giá tiêu dùng, lãi suất, cung tiền M2, đầu tư xây dựng cơ bản từ ngân sách nhà nước (hay đầu tư công), …, có thể được thống kê hàng tháng. Các chỉ số này đều là những biến giải thích quan trọng trong mô hình dự báo GDP. Do đó vấn đề xây dựng các mô hình dự báo trên các tập dữ liệu tần suất hỗn hợp đã được đặt ra.

Các nghiên cứu [55], [56], [57] nhấn mạnh vai trò của thông tin, dữ liệu thời gian thực trên các phương tiện thông tin đại chúng, các mạng xã hội trong việc nowcasting kịp thời các hoạt động kinh tế - tài chính. Nói cách khác nowcast liên quan chặt chẽ với dữ liệu lớn và để xây dựng mô hình nowcast được sử dụng để cập nhật dự báo theo các luồng dữ liệu thời gian thực như vậy cần phải sử dụng phương

pháp, kỹ thuật mới có sự kết hợp với các phương pháp, kỹ thuật của ngành công nghệ thông tin.

Các nghiên cứu [9], [15], [30], [31], [32], [33], [58] cho thấy phương pháp mô hình hóa dự báo hiệu quả trên tập dữ liệu lớn tần suất hỗn hợp kinh tế vĩ mô là sử dụng mô hình DFM và bộ lọc Kalman, trong đó mô hình DFM được ứng dụng nhiều hơn. Mô hình DFM gồm 02 loại là mô hình phương trình bắc cầu (BE) nhân tố và mô hình lấy mẫu dữ liệu hỗn hợp nhân tố (MIDAS) [31], [34], [59], ở đây các nhân tố được chiết xuất từ tập dữ liệu của các biến giải thích đầu vào.

a. Bộ lọc Kalman: Bộ lọc này được đề xuất năm 1960 bởi Kalman [60], [61]. Bộ lọc Kalman là một hệ các phương trình toán học. Nó cung cấp một giải pháp tính toán đệ quy của phương pháp ước lượng bình phương tuyến tính nhỏ nhất (OLS). Bộ lọc này rất mạnh ở một số khía cạnh: hỗ trợ các ước tính về quá khứ, hiện tại và thậm chí cả trong tương lai và nó có thể làm như vậy ngay cả khi tính chính xác của hệ thống được mô hình hóa là không được xác định rõ. Bộ lọc Kalman đã được ứng dụng trong điều khiển chuyển động của tàu vũ trụ Apollo, điều khiển tự động các phương tiện giao thông trên bộ và trên biển. Bộ lọc Kalman đã và đang là chủ đề nghiên cứu mở rộng và ứng dụng, đặc biệt trong lĩnh vực điều khiển tự động và xe tự lái. Bộ lọc này đang được liên kết chặt chẽ với lĩnh vực thị giác máy tính.

Bộ lọc Kalman ước tính trạng thái $z \in \mathbb{R}^m$ của một quá trình điều khiển thời gian rời rạc. Ở dạng tổng quát, bộ lọc Kalman [61] có thể được biểu diễn bởi phương trình vi phân ngẫu nhiên tuyến tính có dạng:

$$x_{k+1} = A_k x_k + B u_k + w_k \tag{1.7}$$

với bộ giá trị đo đạc $z \in \mathbb{R}^N$ được xác định bởi:

$$z_k = H_k x_k + v_k \tag{1.8}$$

ở đây w_k và v_k tương ứng biểu diễn nhiễu của quá trình và của phép đo đạc. Chúng được thừa nhận là biến ngẫu nhiên độc lập, có phân phối chuẩn với kỳ vọng bằng 0 và phương sai không đổi (dĩ nhiên phương sai của chúng nói chung là khác nhau). Ma trận *A* cấp $m \times m$ trong phương trình (1.7) biểu diễn quan hệ của trạng thái x ở thời điểm *k* với trạng thái này ở thời điểm *k+1* có sự vắng mặt của hoặc là hàm dẫn

xuất hoặc là nhiễu quá trình. Ma trận B cấp $m \times p$ biểu diễn quan hệ giữa đầu vào điều khiển $u \in \mathbb{R}^p$ và trạng thái của x. Ma trận H cấp $m \times N$ trong phương trình (1.8) thể hiện quan hệ của trạng thái của x với giá trị đo đạc z_k .

Mô hình được biểu diễn bởi các phương trình (1.7), (1.8) cũng được gọi là mô hình không gian trạng thái (nghĩa là mô hình ước lượng trạng thái ẩn của hệ thống theo cách tối ưu về mặt thống kê). Bộ lọc Kalman đã được ứng dụng trong dự báo kinh tế và cho độ chính xác dự báo khá cao nhưng đòi hỏi chi phí tính toán rất lớn nên trong lĩnh vực kinh tế - tài chính mô hình DFM được sử dụng phổ biến hơn [9].

b. Mô hình DFM: Mô hình DFM được đề xuất bởi Geweke (1977). Mô hình này giả thiết rằng p nhân tố ẩn, động không được quan sát có thể nắm bắt được thông tin của tập dữ liệu gồm m biến giải thích đầu vào $\boldsymbol{X_t}$ và p nhỏ hơn rất nhiều so với m. Trong trường hợp tổng quát, nó có dạng như sau [62]:

$$\boldsymbol{X_t} = \boldsymbol{\Lambda f_t} + \varepsilon_t \tag{1.9}$$

$$\boldsymbol{f_t} = \psi(L)\boldsymbol{f_{t-1}} + \eta_t \tag{1.10}$$

ở đây, L là toán tử trễ lùi, $\boldsymbol{X_t} = [X_{1,t}, X_{2,t}, \dots, X_{m,t}]$, trong đó $X_{i,t} = (x_{i,1}, x_{i,2}, \dots, x_{i,N}) \epsilon \mathbb{R}^N$; $\boldsymbol{f_t}$ là p nhân tố ẩn; $\boldsymbol{\Lambda}$ là ma trận trọng số của các nhân tố cấp $m \times p$; ε_t là véc tơ của các lỗi có đặc điểm riêng, chúng có thể có tương quan yếu [63].

Nghiên cứu [37] chỉ ra rằng p thành phần chính đầu tiên của tập dữ liệu có thể ước lượng nhất quán p nhân tố ẩn không được quan sát theo các giả thiết của mô hình DFM. Nếu $\boldsymbol{W}$ là ma trận cấp $N \times p$ của p véc tơ riêng đầu tiên của ma trận hiệp phương sai $\boldsymbol{S_X}$ của $\boldsymbol{X_t}$, tức $\boldsymbol{S_X} = \frac{1}{N}\boldsymbol{X_t^T}.\boldsymbol{X_t}$ thì các nhân tố tại thời điểm t được ước lượng bởi:

$$\hat{\boldsymbol{f}}_{\boldsymbol{t}} = \boldsymbol{W^T X_t} \tag{1.11}$$

Khi đó dự báo trước h bước ngoài mẫu của biến phụ thuộc y_t được xác định bằng cách hồi quy biến y_{t+h} theo $\hat{\boldsymbol{f}}_{\boldsymbol{t}}, \hat{\boldsymbol{f}}_{\boldsymbol{t-1}}, \dots, \hat{\boldsymbol{f}}_{\boldsymbol{t-q+1}}$. Nói cách khác:

$$\hat{\boldsymbol{y}}_{\boldsymbol{t+h}} = \hat{\boldsymbol{f}}_{\boldsymbol{t}}^{\boldsymbol{T}}\delta_{\boldsymbol{1}} + \hat{\boldsymbol{f}}_{\boldsymbol{t-1}}^{\boldsymbol{T}}\delta_2 + \cdots + \hat{\boldsymbol{f}}_{\boldsymbol{t-q-1}}^{\boldsymbol{T}}\delta_q \tag{1.12}$$

ở đây $\delta_i \epsilon R^p$ là véc tơ của các tham số được ước lượng bằng phương pháp OLS, nó tương ứng với trễ thứ *i* trong phép hồi quy phụ. Khi các biến giải thích ở tần suất khác với tần suất của biến phụ thuộc và số nhân tố là nhỏ, để thực hiện được việc hồi quy biến y_{t+h} trên $\hat{\boldsymbol{f}}_{\boldsymbol{t}}, \hat{\boldsymbol{f}}_{\boldsymbol{t-1}}, \ldots, \hat{\boldsymbol{f}}_{\boldsymbol{t-q+1}}$ người ta phải biểu diễn mô hình DFM dưới dạng mô hình không gian trạng thái nhân tố [31], nghĩa là mô hình hóa các mối quan hệ của biến phụ thuộc với các nhân tố.

Việc dự báo biến phụ thuộc bằng sử dụng mô hình DFM được thực hiện theo thủ tục hai bước. Biến phụ thuộc cần được dự báo là hàm tuyến tính của các biến giải thích X_t. Bằng cách thay thế (1.10) vào (1.11), và đặt $\hat{\theta}_t = \widehat{W}.\hat{\delta}_t$, thì phương trình (1.11) có thể được viết dưới dạng:

$$\hat{y}_{t+h} = X_t^T \hat{\theta}_1 + X_t^T \hat{\theta}_2 + \cdots + X_t^T \hat{\theta}_q \qquad (1.13)$$

và như vậy trong trường hợp số nhân tố được chiết xuất từ $\boldsymbol{X_t}$ không lớn, người ta còn có thể ước lượng các hệ số $\hat{\theta}_i$ bằng cách khác đó là sử dụng mô hình hồi quy RIDGE, LASSO, hoặc lưới đàn hồi [9].

Như đã đề cập ở trên, mô hình DFM bao gồm mô hình hồi quy BE nhân tố và mô hình hồi quy MIDAS nhân tố [9], [13], trong đó các nhân tố được chiết xuất từ tập các biến giải thích ban đầu bằng một số phương pháp giảm chiều. Phương pháp tiếp cận mô hình hồi quy BE [64] đưa ra một giải pháp thuận tiện để lọc và tổng hợp các biến được đặc trưng bởi các tần suất khác nhau. Tuy nhiên, việc tổng hợp có thể dẫn đến mất thông tin hữu ích. Vấn đề này đã dẫn đến sự phát triển của phương pháp mô hình hóa dự báo trên các tập dữ liệu tần suất hỗn hợp được gọi là hồi quy MIDAS [65]. Việc so sánh các ý tưởng chính trong cách tiếp cận của các mô hình hồi quy BE và MIDAS đã được đề cập trong nghiên cứu [66].

Mô hình hồi quy MIDAS bao gồm các mô hình hồi quy MIDAS không bị hạn chế (U-MIDAS) và mô hình hồi quy MIDAS bị hạn chế. Trong loại mô hình MIDAS thứ nhất, các tham số của các thành phần tần suất cao trong mô hình hồi quy ở tần suất thấp là không bị hạn chế, trong khi đó trong loại mô hình MIDAS thứ hai chúng là bị hạn chế bởi những điều kiện ràng buộc như yêu cầu phải tuân theo những quy luật nào đó. Mô hình hồi quy MIDAS bị hạn chế là rất phong phú vì có vô vàn cách để đưa ra các điều kiện hạn chế hoặc ràng buộc về các tham số của thành phần tần

suất cao. Trong thực tế ứng dụng người ta thường tập trung vào các mô hình hồi quy MIDAS ở đó các tham số của biến giải thích tần suất cao thay đổi theo từng bước (STEP-MIDAS), tuân theo quy luật đa thức (PAW-MIDAS), tuân theo quy luật hàm mũ bậc 2 (EAW-MIDAS), tuân theo quy luật hàm mũ bêta (B-MIDAS),… [67]. Trong các mô hình đã nêu, các mô hình hồi quy BE, U-MIDAS, PAW-MIDAS, và STEP-MIDAS được ước lượng bằng phương pháp bình phương tuyến tính nhỏ nhất trong khi mô hình EAW-MIDAS được ước lượng bằng phương pháp tối ưu phi tuyến.

Nghiên cứu [52] đã nghiên cứu quan hệ giữa hồi quy MIDAS và bộ lọc Kalman trên các tập dữ liệu tần suất hỗn hợp. Do bộ lọc Kalman liên quan đến một hệ phương trình, trong khi hồi quy MIDAS liên quan đến một phương trình duy nhất nên hiệu suất của hồi quy MIDAS có thể kém hơn, nhưng nó có thể ít bị lỗi ước lượng tham số và/hoặc lỗi kỹ thuật hơn. Các tác giả xem xét hồi quy MIDAS và bộ lọc Kalman khớp nhau như thế nào trong các trường hợp lý tưởng, ở đó các thành phần của quá trình ngẫu nhiên, độ trễ của các biến tần suất thấp và tần suất cao đều được xác định một cách chính xác. Kết quả thực nghiệm cho thấy độ chính xác dự báo của các mô hình được xây dựng dựa vào bộ lọc Kalman và mô hình MIDAS là tương tự như nhau. Trong hầu hết các trường hợp, bộ lọc Kalman cho độ chính xác dự báo cao hơn một chút, nhưng độ phức tạp tính toán của nó lớn hơn rất nhiều [52].

Nghiên cứu [68] đã thực nghiệm so sánh và kết luận rằng mô hình hồi quy MIDAS và mô hình hồi quy BE có sai số dự báo (RMSE) thấp hơn so với mô hình không gian trạng thái. So sánh 3 phương pháp dự báo này, bài báo cũng cho thấy mô hình hồi quy BE sử dụng tập biến nhỏ (≤ 6 biến) hoạt động tốt hơn so với sử dụng tập biến trung bình (14 biến) hoặc tập biến lớn (34 biến). Hiệu suất tốt nhất thuộc về mô hình hồi quy MIDAS khi sử dụng tập biến trung bình. Ngược lại, mô hình DFM cho thấy hiệu suất khả quan hơn trên tập biến lớn.

Nghiên cứu [17] mới đây đã đề xuất quy trình 3 bước bao gồm lựa chọn thuộc tính, chiết xuất nhân tố và hồi quy rừng ngẫu nhiên kinh tế để thực hiện nowcast tốc độ tăng trưởng thương mại thế giới hàng năm trên tập dữ liệu của 536 biến kinh tế - tài chính tần suất lấy mẫu hỗn hợp. Kết quả cho thấy độ chính xác dự báo theo quy trình được đề xuất là tốt hơn so với các cách tiếp cận khác, bao gồm cả cách tiếp cận

sử dụng các kỹ thuật học mạng nơtron, xích markov, ước lượng bình phương tuyến tính nhỏ nhất, hồi quy lượng tử,…

Bản chất quy trình 3 bước trong nghiên cứu [17] là như sau: xuất phát từ thực tế rằng các yếu tố có tác động đến tăng trưởng thương mại là rất lớn, nghiên cứu này xem tập các yếu tố như là một rừng. Trước hết ở Bước 1, nghiên cứu này sử dụng phương pháp hồi quy góc nhỏ để loại bỏ những biến không hoặc ít liên quan đến sự biến động của thương mại thế giới. Hồi quy góc nhỏ là một phương pháp hồi quy biến phụ thuộc trên tập lớn của các biến giải thích được thực hiện theo cách mở rộng dần dần. Ở mỗi vòng lặp người ta bổ sung vào một biến vào phương trình hồi quy và quan sát xem tốc độ thay đổi của phần dư (biểu hiện qua gradient của nó) có giảm dần không, nếu không biến này bị loại bỏ, nếu có biến này được giữ lại và bổ sung biến mới, nếu gradient vẫn giảm dần thì giữ lại biến có trị tuyệt đối hệ số góc cao và loại bỏ biến có hệ số góc thấp. Và quá trình cứ như vậy, kết quả cuối cùng nhận được một tập các biến được xem là tương quan cao với biến phụ thuộc. Tập các yếu tố (hay biến giải thích) còn lại vẫn rất lớn, Bước 2 sẽ phân tập các yếu tố này (rừng) thành các cụm (hay cây) và thực hiện việc chiết xuất các nhân tố trên từng cây. Bước 3 sẽ xây dựng các mô hình dự báo thương mại thế giới trên từng cây bằng sử dụng mô hình nhân tố động sau đó kết hợp kết quả dự báo tăng trưởng thương mại từ các kết quả dự báo của biến này trên các cây thành phần. Cách tiếp cận 3 bước là khá tương tự như hồi quy rừng ngẫu nhiên. Ở hồi quy rừng ngẫu nhiên, kết quả dự báo là trung bình số học các kết quả dự báo ở các cây thành phần, trong cách tiếp cận 3 bước, kết quả dự báo nhận được bằng việc thực hiện phương pháp hồi quy đa biến của biến phụ thuộc theo các biến dự báo ở các cây con. Cách tiếp cận hồi quy như vậy được gọi là phương pháp kết hợp dự báo và hiện tại có nhiều phương pháp hồi quy khác nhau để kết hợp dự báo [69]. Kết hợp dự báo là một phương pháp dự báo. Độ chính xác dự báo sử dụng phương pháp kết hợp kết quả dự báo của nhiều mô hình khác được chứng minh là cao hơn độ chính xác dự báo theo mỗi mô hình thành phần [69]. Trong cách tiếp cận 3 bước, nếu xét riêng từng bước thì cách lựa chọn có vẻ hợp lý nhưng khi các bước được kết hợp thực hiện cùng nhau thì giải pháp đề xuất như vậy chưa thực sự thuyết phục. Chẳng hạn vì nội dung cơ bản của Bước 2 là thực hiện phương pháp giảm chiều học thuộc tính nên ở Bước 1, các biến cần được lựa chọn sao cho không

gây mâu thuẫn (nhiễu) hoặc dư thừa là đủ, nếu chọn theo cách tối ưu hơn sẽ làm mất nhiều thông tin có giá trị trong xây dựng mô hình dự báo.

Tương tự như trường hợp dự báo trên tập dữ liệu lớn có tần suất lấy mẫu giống nhau, cho đến nay khi dự báo trên tập dữ liệu có tần suất lấy mẫu hỗn hợp, theo cách tiếp cận 3 bước, các nhân tố đều được chiết xuất bằng sử dụng phương pháp giảm chiều PCA. Như đã trình bày ở trên phương pháp PCA là không hiệu quả khi áp dụng cho các tập dữ liệu (các cây) không xấp xỉ một siêu phẳng. Khi đó kết quả dự báo theo quy trình dự báo 3 bước trong nghiên cứu [17] cũng bị hạn chế.

Để thực hiện nowcasting trên tập dữ liệu chuỗi thời gian tần suất hỗn hợp, cần phải giải quyết 03 thách thức sau [4]:

Một là: Xử lý việc học trên các tập dữ liệu tần suất lấy mẫu hỗn hợp, trong đó nhất là đề xuất giải pháp hoặc kỹ thuật để có thể phân lớp/hồi quy biến phụ thuộc ở tần suất thấp theo các biến giải thích ở một vài tần suất khác cao hơn.

Hai là : Xử lý những vấn đề liên quan đến dữ liệu lớn, trong đó nhất làm cách nào để có thể thực hiện kỹ thuật phân lớp hoặc hồi quy trên tập dữ liệu hỗn hợp của một số rất lớn các biến.

Ba là: Xử lý dữ liệu rách (ragged-edge data), ở đây dữ liệu rách liên quan đến tình trạng dữ liệu của các biến khác nhau được phổ biến ở nhiều thời điểm rất khác nhau và tập dữ liệu của các biến như vậy bị xộc xệch, có nhiều quan sát ở đó có biến có dữ liệu, có biến không có dữ liệu.

Trong 03 thách thức nêu trên, thách thức thứ hai là lớn nhất và được cộng đồng quan tâm nghiên cứu nhiều nhất. Đây cũng là thách thức mà Luận án tập trung nghiên cứu và đề xuất phương pháp giải quyết.

1.1.2 Các nghiên cứu trong nước

Khác với tình hình nghiên cứu sôi động ở ngoài nước, tình hình nghiên cứu trong nước về xây dựng mô hình dự báo/mô hình nowcast trên tập dữ liệu chuỗi thời gian lớn trong lĩnh vực kinh tế - xã hội nói chung và kinh tế - tài chính nói riêng vẫn còn hạn chế. Nhóm nghiên cứu [70] đã dự báo tăng trưởng xuất khẩu của Việt Nam bằng sử dụng mô hình véc tơ tự hồi quy tần suất hỗn hợp (MF_VAR) và mô hình MIDAS trên tập dữ liệu kinh tế - tài chính. Kết quả dự báo cho thấy mô hình MIDAS

cho kết quả dự báo tốt hơn mô hình MF_VAR và mang lại hiệu quả cao trong ngắn hạn trên tập dữ liệu thực nghiệm. Tuy nhiên các biến giải thích tần suất cao trong các nghiên cứu này là nhỏ và nghiên cứu đã không phải thực hiện bất kỳ một phương pháp giảm chiều nào đối với tập dữ liệu của các biến đầu vào.

Đối với bài toán phân lớp trên tập dữ liệu lớn: hiện đã có nhiều nhóm nghiên cứu sử dụng các phương pháp giảm chiều trong các bài toán phân lớp và nhận dạng mẫu. Tuy nhiên, các phương pháp giảm chiều trong các bài toán này thường thuộc vào nhóm lựa chọn thuộc tính. Một trong những nhóm nghiên cứu điển hình theo hướng tiếp cận này là nhóm nghiên cứu của PGS.TS. Nguyễn Long Giang và cộng sự, Viện CNTT, Viện Hàn lâm Khoa học Việt Nam. Nhóm nghiên cứu các phương pháp giảm chiều lựa chọn thuộc tính (hay trích chọn thuộc tính) chủ yếu dựa vào lý thuyết tập thô [71], [72], [73], [74]. Các thuộc tính được trích chọn được sử dụng chủ yếu cho các bài toán phân lớp hay dự báo xu thế.

Luận án tiến sĩ [43] đã tổng quan, so sánh hiệu suất giảm chiều lựa chọn thuộc tính theo 3 cách tiếp cận lọc, bọc, và nhúng, đồng thời ứng dụng của các tiếp cận ấy trong bài toán dự báo và phân lớp. Và chưa được như tên gọi, luận án chưa cải tiến một cách có ý nghĩa hoặc đề xuất phương pháp lựa chọn thuộc tính mới theo một trong 3 cách tiếp cận đã nêu. Các bài toán ứng dụng phương pháp giảm chiều lựa chọn thuộc tính trong luận án còn giản đơn.

Đề tài nghiên cứu khoa học cấp bộ - Bộ Tài chính [75] có thể được xem là nghiên cứu trong nước đầu tiên về sử dụng phương pháp giảm chiều *học thuộc tính* trong các bài toán trên các tập dữ liệu tần suất hỗn hợp. Tuy nhiên các mô hình nowcast được xây dựng trong nghiên cứu này chỉ được xây dựng dựa trên mô hình phương trình bắc cầu (BE), phương pháp giảm chiều học thuộc tính cũng như phương pháp xác định độ trễ của các thành phần tần suất cao trong mô hình nowcast chưa được làm rõ. Hiệu suất của phương pháp giảm chiều cũng chưa được so sánh và đánh giá.

1.2 Các vấn đề còn tồn tại

Từ tổng quan, đánh giá các nghiên cứu liên quan ở trong và ngoài nước về việc xây dựng mô hình dự báo và mô hình nowcast trên các tập dữ liệu chuỗi thời gian lớn

của các biến giải thích tương ứng lấy mẫu tần suất giống nhau và hỗn hợp cho thấy cách tiếp cận học sâu đang được quan tâm nghiên cứu nhưng cách tiếp cận này vẫn còn ở giai đoạn đầu và còn có nhiều hạn chế. Hiện tại các kỹ thuật này chưa thể học được trên các tập dữ liệu có hàng chục nghìn thậm chí hàng trăm nghìn biến giải thích chuỗi thời gian nếu chúng không được sử dụng kết hợp với những kỹ thuật giảm chiều dữ liệu.

Quy trình xây dựng các mô hình dự báo hoặc bộ phân lớp trên tập dữ liệu lớn thường gồm 2 giai đoạn, trước hết là thực hiện một số kỹ thuật giảm chiều để tìm và/hoặc sinh ra tập dữ liệu mới có số chiều nhỏ hơn rất nhiều so với tập dữ liệu ban đầu nhưng vẫn nắm giữ được các thông tin quan trọng trong tập dữ liệu ban đầu, tiếp theo sử dụng kỹ thuật học hồi quy hoặc học phân lớp trên tập dữ liệu mới ấy. Dù kỹ thuật học hồi quy hoặc học phân lớp có tiên tiến và có thể học được trên tập dữ liệu lớn, khi xây dựng mô hình dự báo hoặc bộ phân lớp trên các tập dữ liệu lớn cũng cần sử dụng kết hợp kỹ thuật này với phương pháp giảm chiều, điều đó không chỉ làm cho mô hình dự báo/bộ phân lớp hoạt động hiệu quả hơn, nhanh hơn mà việc diễn giải các kết quả dự báo hoặc phân lớp cũng thuận lợi, dễ dàng hơn.

Việc giảm chiều trên tập dữ liệu lớn thường kết hợp các kỹ thuật lựa chọn thuộc tính và kỹ thuật học thuộc tính. Trong lĩnh vực kinh tế - tài chính kỹ thuật lựa chọn thuộc tính được sử dụng phổ biến trong các ứng dụng thực tế là độ đo hệ số tương quan Pearson đối với tập dữ liệu giá trị số và độ đo thông tin tương hỗ dựa vào entropy đối với tập dữ liệu phân loại, trong khi đó kỹ thuật học thuộc tính hiện được sử dụng phổ biến và hiệu quả nhất là PCA và SPCA. Không may, các kỹ thuật PCA và SPCA chỉ giảm chiều hiệu quả đối với các tập dữ liệu ở đó các điểm dữ liệu xấp xỉ một siêu phẳng, trong khi các tập dữ liệu thế giới thực không phải luôn như vậy.

Luận án tập trung nghiên cứu giải pháp để khắc phục tồn tại này của các phương pháp PCA và SPCA. Cụ thể, luận án tập trung nghiên cứu:

1) Đề xuất phương pháp giảm chiều mới được xem là mở rộng tự nhiên của phương pháp PCA đồng thời khắc phục được nhược điểm của phương pháp PCA trên các tập dữ liệu không xấp xỉ một siêu phẳng, và có hiệu suất giảm chiều cao hơn hoặc bằng hiệu suất giảm chiều của các phương pháp PCA và SPCA trong các bài toán dự

báo và nowcast tương ứng trên các tập dữ liệu lấy mẫu tần suất giống nhau và hỗn hợp.

2) Đề xuất quy trình hoặc thuật toán dự báo sử dụng phương pháp giảm chiều được đề xuất và ứng dụng của nó trong việc dự báo một chỉ số kinh tế vĩ mô quan trọng trên tập dữ liệu lớn.

1.3 Một số kiến thức cơ sở

1.3.1 Các loại dữ liệu kinh tế - tài chính

Trong lĩnh vực kinh tế - tài chính có 3 loại dữ liệu [76]: dữ liệu chéo, dữ liệu chuỗi thời gian, và dữ liệu mảng (Panel data), trong đó dữ liệu chuỗi thời gian chiếm phần lớn. Dữ liệu chéo là dữ liệu được thu thập tại cùng một thời điểm cho nhiều đối tượng thuộc cùng một loại. Thể hiện rõ nhất của dữ liệu chéo là các dữ liệu điều tra ở đó thứ tự lấy mẫu (hay các quan sát) là không quan trọng. Dữ liệu chuỗi thời gian là dữ liệu được thu thập theo định kỳ thời gian. Với loại dữ liệu này, thứ tự lấy mẫu (hay quan sát) là rất quan trọng và không thể thay đổi hoặc loại bỏ bất kỳ một quan sát nào, bởi vì trong mỗi chuỗi thời gian thường ẩn chứa các quy luật về sự thay đổi của nó cũng như về mối quan hệ của nó với một số chuỗi thời gian khác. Dữ liệu mảng là dữ liệu vừa có tính chất của dữ liệu chéo vừa có tính chất của dữ liệu chuỗi thời gian [76], [77].

Dữ liệu chuỗi thời gian xuất hiện phổ biến trong các hệ thống thống kê nhà nước, bộ ngành, địa phương và các doanh nghiệp. Đặc điểm của dữ liệu chuỗi thời gian là có tính xu hướng, tính chu kỳ, tính mùa vụ, tính bất thường và biến đổi ngẫu nhiên. Với sự phát triển của khoa học công nghệ, nhất là công nghệ thông tin, việc thu thập dữ liệu như vậy ngày càng dễ dàng, nhanh chóng và đầy đủ hơn.

1.3.2 Phân loại dự báo

Hiện có rất nhiều cách tiếp cận khác nhau để phân loại dự báo. Chẳng hạn nếu phân loại theo thời gian xa nhất của dự báo (còn được gọi là đường chân trời của dự báo) thì có dự báo ngắn hạn, trung hạn và dài hạn. Thông thường, nếu thời gian xa nhất của dự báo không lớn hơn 3 lần kỳ dữ liệu thì dự báo đó được gọi là dự báo ngắn hạn, nếu thời gian xa nhất của dự báo bằng 4-5 lần kỳ dữ liệu thì đó là dự báo trung hạn, và dự báo dài hạn là dự báo có độ xa nhất của dự báo từ 6 lần kỳ dữ liệu trở lên.

Chẳng hạn, nếu kỳ thu thập dữ liệu là tháng thì dự báo cho từ 1-3 tháng tiếp theo là dự báo ngắn hạn, dự báo cho 4-5 tháng tiếp theo là dự báo trung hạn, dự báo từ 6 tháng tiếp theo trở lên là dự báo dài hạn. Các phương pháp dự báo ngắn hạn, trung hạn và dài hạn nói chung là khác nhau. Dự báo có điều kiện và không điều kiện là một cách phân loại khác. Khác nhau cơ bản giữa dự báo có điều kiện và dự báo không điều kiện là ở chỗ để dự báo có điều kiện biến phụ thuộc, ta cần phải thực hiện dự báo các biến giải thích ngoại sinh ở trong mô hình, trong khi với dự báo không điều kiện thì không cần phải thực hiện như vậy [77].

1.3.2.1 Mô hình dự báo có điều kiện

Phương pháp dự báo có điều kiện thường được sử dụng khi người làm dự báo cảm nhận rằng tương lai có thể diễn ra không gần giống như hiện tại và quá khứ. Khi đó người ta thường kết hợp dự báo bằng mô hình định lượng được xây dựng với phương pháp phán xử [78] hoặc là xây dựng các kịch bản dự báo. Kịch bản được xây dựng dựa vào phân tích, dự báo các biến ngoại sinh. Kịch bản dự báo biến phụ thuộc được hình thành khi các biến giải thích ngoại sinh được dự báo bằng sử dụng mô hình phụ của nó được gọi là kịch bản cơ sở. Ngoài kịch bản cơ sở cần phải xây dựng một số kịch bản dự báo khác theo các bộ giả định khác nhau về giá trị của các biến ngoại sinh [79], [80]. Khi người dự báo cảm nhận thấy tương lai không có những biến động bất thường thì có thể lấy kết quả dự báo của kịch bản cơ sở làm kết quả dự báo cuối cùng. Khi đó độ chính xác dự báo biến phụ thuộc không chỉ phụ thuộc vào chất lượng của mô hình dự báo của nó mà còn phụ thuộc vào độ chính xác dự báo của các biến ngoại sinh. Mô hình dự báo biến phụ thuộc theo các biến giải thích có dạng phương trình (1.1) ở trên [77].

1.3.2.2 Mô hình dự báo không điều kiện

Các mô hình hồi quy đơn biến và đa biến đều có thể được sử dụng để xây dựng mô hình dự báo không điều kiện.

a. Các mô hình dự báo đơn biến

Mô hình dự báo không điều kiện đơn biến là mô hình được xây dựng chỉ dựa vào chính biến cần được dự báo và nó không chứa bất kỳ biến giải thích ngoại sinh nào khác nên không cần phải dự báo các biến ngoại sinh. Các mô hình đơn biến điển

hình được sử dụng phổ biến trong các ứng dụng thực tiễn là mô hình tự hồi quy có xu thế AR(p), mô hình trung bình trượt tích hợp tự hồi quy ARIMA [81], và mô hình làm trơn hàm mũ Holt-Winter [82].

Hai nhược điểm chính của các mô hình đơn biến là: (1) độ chính xác dự báo của các mô hình đơn biến thường không cao bằng mô hình dự báo không điều kiện đa biến bởi vì thực chất các mô hình đơn biến chỉ là trường hợp riêng của mô hình dự báo không điều kiện đa biến bằng cách loại bỏ các biến giải thích ngoại sinh ra khỏi mô hình dự báo biến phụ thuộc mặc dù các biến đó cũng có ảnh hưởng đến sự thay đổi của biến phụ thuộc và (2) không cho biết các yếu tố nào là nguyên nhân chính tác động đến sự thay đổi của biến phụ thuộc.

b. Mô hình đa biến

Trong lĩnh vực kinh tế - tài chính thường tồn tại nhiều chỉ số mà sự biến động của chúng có quan hệ ổn định với những biến động của một số chỉ số khác [83]. Mối quan hệ đó thường được biểu diễn thông quan các mô hình dự báo. Do vậy các thông tin về một số chỉ số có thể được sử dụng để giám sát và dự báo một số chỉ số khác được gọi là các chỉ số dẫn báo (hay chỉ số báo trước). Chỉ số dẫn báo thường được sử dụng trong xây dựng mô hình dự báo không điều kiện.

Để xây dựng mô hình dự báo không điều kiện đa biến, trước hết cần phải xác định xem trong tập các biến giải thích ban đầu có các biến nào là chỉ số dẫn báo của biến phụ thuộc [84]. Việc xác định các chỉ số dẫn báo thường được dựa vào lý thuyết đồ thị hoặc mô hình toán học. Trong lĩnh vực kinh tế - tài chính, kiểm định nhân quả Granger [84] thường được sử dụng để phát hiện các chỉ số dẫn báo của biến phụ thuộc.

Cụ thể, quan hệ nhân quả Granger được xác định như sau: Giả sử Y_t và X_t lần lượt là các biến chuỗi thời gian dừng, khi đó biến X_t là nguyên nhân Granger của biến Y_t nếu:

$$Y_t = \sum_{j=1}^{q} \beta_j Y_{t-j} + \sum_{i=0}^{p} \alpha_i X_{t-i} + e_t \qquad (1.14)$$

và $\sum_{i=1}^{p} \alpha_i^2 \neq 0.$

ở đây e_t là phần dư được giả định là nhiễu trắng, α_i, β_j là các tham số được ước lượng; p và q lần lượt là độ trễ tối ưu của X_t và Y_t [84].

Có thể thấy về bản chất quan hệ nhân quả Granger là quan hệ tuyến tính giữa biến X_t và biến Y_t. Trong thực tế ứng dụng, quan hệ này thường được xác định dựa vào việc ước lượng mô hình (1.14) ở dạng sau [84]:

$$Y_t = \sum_{j=1}^{p} \beta_j Y_{t-j} + \sum_{i=0}^{p} \alpha_i X_{t-i} + e_t \quad (1.15)$$

Khi đó, biến X_t được gọi là nguyên nhân Granger (hay chỉ số dẫn báo) của biến Y_t với độ trễ p. Độ trễ này được xác định chủ yếu dựa vào tri thức miền ứng dụng. Chỉ số dẫn báo X_t cũng được nói là được phát hiện bằng sử dụng kiểm định nhân quả Granger.

Giả sử, $Y_t, \mathbf{X}$ lần lượt là biến phụ thuộc và tập các biến giải thích. Tập $\{X_{1,t}, X_{2,t}, \ldots, X_{m,t}\}$ là các chỉ số dẫn báo của Y_t trong **X,** ở đó các chỉ số dẫn báo được xác định bằng kiểm định nhân quả Granger với độ trễ p. Mô hình dự báo không điều kiện của Y_t theo tập các chỉ số dẫn báo $X_1, X_2, \ldots, X_m$ có dạng:

$$Y_t = \sum_{j=1}^{p} \beta_j Y_{t-j} + \sum_{k=1}^{m} \sum_{i=1}^{p} \alpha_{k,i} X_{k,t-i} + e_t \quad (1.16)$$

Trong mô hình (1.16), khi loại bỏ thành phần các biến giải thích ra khỏi mô hình thì mô hình phương trình (1.16) trở thành mô hình AR(p). Đó là mô hình đơn biến [81] để dự báo không điều kiện biến Y_t.

1.3.3 Dữ liệu lớn

1.3.3.1 Khái niệm về dữ liệu lớn

Dữ liệu lớn được định nghĩa khá khác nhau trong các tài liệu. Nghiên cứu [85] đã phân tích một danh sách khá toàn diện các định nghĩa hiện có về dữ liệu lớn và đề xuất một định nghĩa mới nhằm khắc phục những hạn chế của các định nghĩa trước đó, theo đó "Dữ liệu lớn (Big data) là tài sản *Thông tin* được đặc trưng bởi *Khối lượng lớn, Tốc độ nhanh* và *Đa dạng*, và đòi hỏi phải có các *Công nghệ* và *phương pháp phân tích* đặc thù để chuyển đổi nó thành *Giá trị*".

1.3.3.2 Nhận diện một tập dữ liệu lớn

Nghiên cứu [9] đã giới thiệu 3 cách để nhận diện dữ liệu lớn bao gồm:

- Thứ nhất, dữ liệu lớn được nhận diện dựa vào một số đặc trưng, trong đó những đặc trưng quan trọng nhất là 5 chữ "V" [9], [86]. Đó là Volume (lượng dữ liệu được tạo ra và lưu trữ là lớn), Variety (kiểu và bản chất của dữ liệu là đa dạng), Velocity (tốc độ dữ liệu được tạo ra và được xử lý để đáp ứng các nhu cầu và thách thức là cao), Value và Veracity (chất lượng và giá trị của dữ liệu được xác thực).

- Thứ hai, dữ liệu lớn có thể được nhận diện thông qua số lượng các biến và số lượng các quan sát theo ba dạng cơ bản [9], [58], đó là "TALL" có số lượng biến (m) vừa phải trong khi số quan sát (N) rất lớn ($N \gg m$) , "FAT" có số lượng biến rất lớn nhưng số quan sát vừa phải ($m \gg N$) và "HUGE" có số lượng biến và số quan sát là rất lớn.

- Cuối cùng, dữ liệu lớn được nhận diện thông qua nội dung và nguồn dữ liệu như dữ liệu từ mạng xã hội (social networks) chẳng hạn Facebook, Twitter, Blog, hình ảnh, video…, hay từ các hệ thống kinh doanh truyền thống như dữ liệu do cơ quan nhà nước cung cấp (hồ sơ y tế, bảo hiểm xã hội), dữ liệu do doanh nghiệp tạo ra (giao dịch thương mại, ngân hàng) hoặc từ kết nối vạn vật (dữ liệu do máy tạo ra) chẳng hạn dữ liệu từ cảm biến, dữ liệu từ hệ thống máy tính (Log, Web Log) [9].

Mặc dù chưa có định nghĩa thống nhất về dữ liệu lớn nhưng có thể nói đặc trưng cơ bản nhất của dữ liệu lớn được phát biểu trong định lý HACE [87]: " Dữ liệu lớn bắt đầu với khối lượng lớn, không đồng nhất, các nguồn tự trị với việc tổ chức phân tán và phi tập trung, và cách khám phá các mối quan hệ trong tập dữ liệu là phức tạp và luôn tiến hóa giữa các dữ liệu".

1.3.3.3 Thách thức của dữ liệu lớn

Trong nghiên cứu [87] cũng chỉ rõ những thách thức chủ yếu của khai phá dữ liệu với dữ liệu lớn ở 03 cấp độ. Thứ nhất, cấp độ khai thác dữ liệu, tập trung vào truy cập dữ liệu và quy trình tính toán thực tế bởi vì dữ liệu lớn thường có khối lượng tính toán gia tăng rất nhanh và lưu trữ ở nhiều vị trí khác nhau nên rất cần nền tảng điện toán phù hợp, hiệu quả nhằm lưu trữ dữ liệu quy mô lớn, phân tán để tính toán. Thứ hai, cấp độ ngữ nghĩa, dữ liệu lớn và tri thức miền ứng dụng đề cập đến nhiều khía

cạnh liên quan đến các quy định, chính sách, tri thức người dùng và thông tin miền. Hai vấn đề quan trọng nhất ở cấp độ này là chia sẻ dữ liệu và quyền riêng tư, và tri thức về lĩnh vực và ứng dụng. Cuối cùng, cấp độ các thuật toán khai thác dữ liệu lớn bao gồm học cục bộ và học kết hợp cho nhiều nguồn thông tin, khai thác từ dữ liệu thưa, không chắc chắn và không đầy đủ cũng như khai thác dữ liệu động và phức hợp.

Trong luận án này tập trung giải quyết các bài toán dữ liệu lớn theo cách nhận diện thứ 2 trên các tập dữ liệu chuỗi thời gian lớn nhận giá trị số, ở đó số biến thường là lớn hoặc là số biến lớn hơn nhiều so với số quan sát. Phương pháp được luận án sử dụng theo cách tiếp cận giảm chiều dữ liệu, cụ thể khi đó Bước 1 sử dụng phương pháp/kỹ thuật lựa chọn thuộc tích để loại bỏ các biến nhiễu, biến không hoặc ít liên quan trong khi Bước 2 sử dụng phương pháp/ kỹ thuật học thuộc tính để chiết xuất các nhân tố.

1.3.4 Giảm chiều dữ liệu

Giả sử $\mathbf{X} = [X_1, X_2, \dots, X_m]_{N\times m}$ là một tập dữ liệu đầu vào, ở đây $X_i^T = (x_{i_1}, x_{i_2}, \dots, x_{i_N}) \in \mathbb{R}^{\mathbf{N}}$; *N, m*, và X_i lần lượt là số quan sát, số biến giải thích trong **X,** và véc tơ dữ liệu (hay biến giải thích hoặc biến giải thích trong **X**).

Ký hiệu $\chi_j = (x_{1_j}, x_{2_j}, \dots, x_{m_j})$ là điểm dữ liệu của tập **X,** $\mathrm{Y_h} = (y_{\mathrm{h}_1}, y_{\mathrm{h}_2}, \dots, y_{h_N})^{\mathrm{T}} \in \mathbb{R}^{\mathbf{N}}$, trong đó *h =1, 2, ..., p,* và $\mathbf{Y} = [\mathrm{Y_1}, \mathrm{Y_2}, \dots, \mathrm{Y_p}]$.

Giảm chiều biến là một ánh xạ $\boldsymbol{\mathcal{R}}$: $\mathbb{R}^{\boldsymbol{m}} \longrightarrow \mathbb{R}^{\boldsymbol{p}}$

$$(x_{1_j}, x_{2_j}, \dots, x_{m_j}) \mapsto (y_{1_j}, y_{2_j}, \dots, y_{p_j}),$$

sao cho p ≪ m và tập dữ liệu $[\mathrm{Y_1}, \mathrm{Y_2}, \dots, \mathrm{Y_p}] = \boldsymbol{\mathcal{R}}(\mathbf{X})$ nắm bắt thông tin quan trọng của **X** nhiều như có thể.

Như vậy giảm chiều thực chất là sử dụng một số kỹ thuật để biến đổi tập dữ liệu ban đầu thành tập dữ liệu của một số biến mới ít hơn rất nhiều nhưng vẫn nắm giữ được những thông tin quan trọng từ tập dữ liệu ban đầu nhiều như có thể.

1.3.4.1 Độ đo hệ số tương quan Pearson:

Giả sử $X = (x_i)$, $Y = (y_i)$, ở đây *i = 1,.., m*, là các biến chuỗi thời gian vô hướng. Hệ số tương quan Pearson giữa biến *X* và biến *Y* được xác định theo công thức (1.17) dưới đây [88]:

$$Corr(X,Y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} \quad (1.17)$$

trong đó, $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{m}$, $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{m}$ và khi đó, đặt $P(X,Y) = |Corr(X,Y)|$ được gọi là độ đo hệ số tương quan Pearson. Độ đo này thường được sử dụng để lựa chọn các biến đắt giá bằng cách loại bỏ các biến dư thừa và các biến không hoặc ít liên quan (biến nhiễu) trong các bài toán dự báo trên các tập dữ liệu chuỗi thời gian lớn vô hướng.

Giả sử *Y,* X_1, X_2 là các biến chuỗi thời gian; α, β tương ứng là các ngưỡng có liên quan (relevant threshold) và ngưỡng dư thừa (redundancy threshold) do người dùng xác định. Biến X_1 được gọi là có liên quan đối với biến Y nếu $P(Y, X_1) \geq \alpha$; Biến X_2 được gọi là dư thừa đối với Y nếu tồn tại biến X_1 sao cho $P(Y, X_1) \geq P(Y, X_2)$ và $P(X_1, X_2) \geq \beta$.

Trong trường hợp tập các biến có liên quan và không dư thừa vẫn còn rất lớn thì ta không thể thực hiện hồi quy biến phụ thuộc theo tập các biến đó. Khi đó, cần phải thực hiện những phương pháp giảm chiều khác trước khi thực hiện các thuật toán hồi quy hoặc phân lớp.

1.3.4.2 Phương pháp PCA

Phương pháp PCA có thể được sử dụng để giảm chiều. Phương pháp này được trình bày tóm tắt như sau:

Giả sử, tập dữ liệu **X** là ma trận được cân chỉnh trung bình, tức là $\sum_{j=1}^{N} x_{i_j} = 0, \forall i = 1, \ldots, m$. Kí hiệu $m\mathbf{R} = \mathbf{X}^T\mathbf{X}$, ở đây **R** là ma trận hiệp phương sai của tập **X** và là ma trận vuông cấp $m \times m$, đối xứng và xác định dương. Do đó **R** được chéo hóa bởi ma trận gồm các véc tơ riêng trực giao và các giá trị riêng nhận giá trị dương của **R** [18].

Cụ thể, $\mathbf{R} = \mathbf{E}.\mathbf{D}.\mathbf{E}^{\mathbf{T}}$, trong đó **R**, **E, D** là các ma trận vuông cấp $m \times m$, $\mathbf{D} = diag(\lambda_1, \lambda_2, \ldots, \lambda_m)$ là ma trận đường chéo, ở đó $\lambda_1, \lambda_2, \ldots, \lambda_m$ là các giá trị riêng của **R,** mỗi cột e_i của ma trận trực giao **E** là một véc tơ riêng tương ứng với trị riêng λ_i của **R**, và $\frac{\lambda_i}{m}$ là phương sai của e_i.

Thành phần chính thứ i, ký hiệu PC_i, là chiếu của $\mathbf{X}$ lên véc tơ riêng thứ i. Véc tơ cột PC_i được xác định như sau [18]:

$$PC_i = \boldsymbol{X}.e_i \tag{1.18}$$

Giả sử các giá trị riêng λ_i được sắp xếp theo thứ tự giảm dần. Phần trăm phương sai tích lũy của p thành phần chính đầu tiên được xác định theo công thức sau:

$$PCV(p) = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{m} \lambda_i} * 100;\ \ p = 1, \dots m \tag{1.19}$$

Phần trăm giá trị riêng trên tổng thể chỉ ra rằng p thành phần chính đầu tiên nắm giữ $PCV(p)$(%) thông tin của tập dữ liệu ban đầu [18].

Giảm chiều bằng phương pháp PCA là chọn p véc tơ thành phần chính đầu tiên sao cho phần trăm phương sai tích lũy ứng với các thành phần chính được chọn lớn hơn một ngưỡng được xác định trước.

Có nhiều cách để xác định số lượng thành phần chính (PC) theo phương pháp PCA bao gồm: Xác thực chéo (Cross-validation), Phần trăm phương sai tích lũy (Cumulative variance percentage), Kiểm tra sàng lọc (Scree test), Giá trị riêng trung bình (Average eigenvalues), Hàm sai số nhúng (Embedded error function), Tiêu chí thông tin Akaike (Akaike information criterion), Tiêu chí độ dài mô tả tối thiểu (Minimum description length criterion), Phương sai của sai số tái thiết (Variance of reconstruction error) [89]. Luận án này sử dụng phần trăm phương sai tích lũy để chọn số lượng những PC đầu tiên sao cho chúng có thể giữ lại ít nhất q (%) phương sai tổng thể của tập dữ liệu, ở đây q là ngưỡng do người dùng xác định và thường lớn hơn 70%.

Giả sử $PCV(p) \geq q$, thì p thành phần chính đầu tiên được chọn để thay thế tập m biến giải thích được xác định bởi công thức sau [18]:

$$\mathbf{PC}_{N\times \mathrm{p}} = \mathbf{X}_{\mathrm{N}\times \mathrm{m}}.\mathbf{E}_{\mathrm{m}\times \mathrm{p}} \tag{1.20}$$

ở đây, $\mathbf{E}_{\mathrm{m}\times \mathrm{p}}$ là ma trận tương ứng với p véc tơ cột đầu tiên được chọn trong ma trận **E.** Như vậy phương pháp PCA có thể được viết dưới dạng giả code như sau:

Thuật toán PCA [46]

Input: $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{m}}$

Output: $\mathbf{Y} \in \mathbb{R}^{\mathrm{N}\times\mathrm{p}}$

Begin

1. Xây dựng ma trận hiệp phương sai ($\mathbf{X}^{\mathbf{T}}\boldsymbol{X}$)
2. Tìm giá trị riêng và véc tơ riêng của ma trận $\mathbf{X}^{\mathbf{T}}\boldsymbol{X}$
3. Sắp xếp các véc tơ riêng theo các giá trị riêng theo thứ tự giảm dần
4. Xây dựng ma trận **E** (m × p) với p véc tơ riêng đầu tiên
5. Cân chỉnh trung bình **X** và sử dụng **E** để thu được không gian con mới **Y** = **X.E**

End.

Tương tự như cách tiếp cận giảm chiều biến, ký hiệu $N\widetilde{\mathbf{R}} = \mathbf{X}.\mathbf{X}^{\boldsymbol{T}}$, $\widetilde{\mathbf{R}}$ là ma trận hiệp phương sai của tập $\mathbf{X}^{\boldsymbol{T}}$ và là ma trận vuông đối xứng xác định dương cấp *N,* và **V** là ma trận vuông của các véc tơ riêng của $\widetilde{\mathbf{R}}$ dưới dạng cột. Khi đó, *p* thành phần chính đầu tiên là phép chiếu của **X** lên *p* véc tơ riêng đầu tiên của ma trận **V** được xác định như sau:

$$\widetilde{\boldsymbol{PC}}_{m\times\mathrm{p}} = \mathbf{X}^{\mathbf{T}}{}_{\mathrm{m}\times\mathrm{N}}.\mathbf{V}_{\mathrm{N}\times\mathrm{p}} \tag{1.21}$$

Công thức (1.21) được sử dụng để giảm số lượng quan sát của tập dữ liệu **X**. Nghiên cứu [90] đã chỉ ra mối quan hệ kép giữa các cặp véc tơ riêng và giá trị riêng của ma trận **R** với các cặp véc tơ riêng và giá trị riêng của ma trận $\widetilde{\mathbf{R}}$.

Hiện tại, các thành phần chính được chiết xuất bằng phương pháp PCA có thể được tìm theo 2 cách tiếp cận. Cách tiếp cận thứ nhất là sử dụng công cụ đại số tuyến tính và cách tiếp cận thứ hai là giải quyết bài toán tối ưu. Thuật toán PCA được trình bày ở trên là theo cách tiếp cận thứ nhất. Cụ thể giả sử $\mathbf{PC} = [PC_1, PC_2, \dots, PC_m] \in \mathbf{R}^{\mathbf{N}\times\mathbf{m}}$ là tập tất cả các thành phần chính và theo (1.20) nó có thể được biển diễn theo công thức sau:

$$\mathbf{PC} = \mathbf{X}.\mathbf{E} \tag{1.22}$$

ở đây, $\mathbf{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_m}] \in \mathbf{R}^{\boldsymbol{m}\times\boldsymbol{m}}$ được gọi là ma trận trọng số, $\boldsymbol{e_i}, i = 1, \dots, m$ là các cột của $\mathbf{E}$ và được biểu thị như là các hướng chính hoặc véc tơ các trọng số.

Do ma trận trọng số trực chuẩn $\mathbf{E}$ được tính toán dựa vào công cụ đại số tuyến tính, nên các thành phần chính có phương sai cực đại trong tập dữ liệu đầu vào [18]. Vì vậy có thể chuyển bài toán tìm ma trận $\mathbf{E}$ thành bài toán cực đại phương sai như cách tiếp cận thứ hai. Theo cách tiếp cận này có thể xác định các thành phần chính bằng cách giải bài toán tối ưu sau [91]:

$$\min_{\mathbf{E}} f(\mathbf{E}) = \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{X} - \mathbf{X}\mathbf{E}\mathbf{E}^{\mathbf{T}}\|_{\mathbf{F}}^{\mathbf{2}} \quad (1.23)$$

$$\text{sao cho } \mathbf{E}\mathbf{E}^{\mathbf{T}} = \mathbf{I}$$

ở đây $\|.\|_{\mathbf{F}}$ là chuẩn Frobenius của ma trận.

1.3.4.3 Họ phương pháp SPCA

- *Phương pháp SPCA*: là sự kết hợp giữa hồi quy thưa và các thành phần chính để tìm ra một tập hợp các véc tơ có trọng số thưa, tức là các véc tơ chỉ có một vài giá trị khác không [53]. Với lập luận tương tự như trên, nghiên cứu [91] đề xuất phương pháp tìm các thành phần chính thưa trong [53] bằng sử dụng phép chiếu biến như là một chiến lược tối ưu hóa như sau:

$$\min_{\mathbf{E},\mathbf{G}} f(\mathbf{E}, \mathbf{G}) = \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{X} - \mathbf{X}\mathbf{G}\mathbf{E}^{\mathbf{T}}\|_{\mathbf{F}}^{\mathbf{2}} + \psi(\mathbf{G}) \quad (1.24)$$

$$\text{sao cho } \mathbf{E}\mathbf{E}^{\mathbf{T}} = \mathbf{I}$$

trong đó $\mathbf{E}$ và $\mathbf{G}$ là các ma trận vuông cấp m, $\mathbf{G}$ là ma trận trọng số thưa và $\mathbf{E}$ là ma trận trực chuẩn, ψ biểu thị một bộ điều chỉnh thưa LASSO (chuẩn L_1) hoặc là lưới đàn hồi. Khi đó, tập các thành phần chính được xác định bởi công thức $\mathbf{PC} = \mathbf{X} \,.\, \mathbf{G}$.

- *Phương pháp RSPCA*: Để tăng tốc độ tính toán, người ta đã sử dụng phép xấp xỉ ma trận đã cho thành tích của một số ma trận cấp thấp hơn (hạng thấp hơn). Nghiên cứu [91] đã giới thiệu cách phân tách ma trận dữ liệu theo chiều thấp nhằm nắm bắt thông tin thiết yếu của tập dữ liệu ban đầu, sau đó ứng dụng phương pháp RSPCA sử dụng phép chiếu biến như một chiến lược tối ưu hóa để xác định các véc tơ thành phần chính. Phương pháp RSPCA sử dụng phép chiếu biến như một chiến lược tối ưu hóa thực chất là giải bài toán tối ưu sau:

$$\min_{\mathbf{E},\mathbf{G}} f(\mathbf{E},\mathbf{G}) = \frac{\mathbf{1}}{\mathbf{2}} \left\| \widehat{\mathbf{X}} - \widehat{\mathbf{X}}\mathbf{G}\mathbf{E}^{\mathbf{T}} \right\|_{\mathbf{F}}^{\mathbf{2}} + \psi(\mathbf{G}) \quad (1.25)$$

$$\text{sao cho } \mathbf{E}\mathbf{E}^{\mathbf{T}} = \mathbf{I}$$

ở đây, $\widehat{\mathbf{X}} \in \mathbf{R}^{\mathbf{h}\times\mathbf{m}}$ được ký hiệu là bản phác thảo của $\mathbf{X} \in \mathbf{R}^{\mathbf{N}\times\mathbf{m}}$ chiều *h* được chọn lớn hơn một chút so với hạng của ma trận đích *k* [91].

- *Phương pháp SPCA vững (ROBSPCA)*: Để vượt qua thách thức rằng trong nhiều tình huống thực tế, dữ liệu bị hỏng do lỗi đo lường hoặc do các tác động khác, người ta thường sử dụng việc phân rã ma trận dữ liệu thành các thành phần thưa và có hạng thấp. Nghiên cứu [91] đã đề xuất phiên bản robust SPCA (gọi tắt là ROBSPCA) bằng cách sử dụng phép chiếu biến như một chiến lược tối ưu hóa. Nó cũng được xây dựng dựa vào ý tưởng tách một ma trận dữ liệu thành một ma trận hạng thấp và một ma trận thưa và sử dụng những thông tin trước đó như là bộ điều chỉnh tính thưa. Cụ thể hơn, phương pháp ROBSPCA sử dụng phép chiếu biến như là một chiến lược tối ưu được xác định như sau:

$$\min_{\mathbf{E},\mathbf{G}} f(\mathbf{E},\mathbf{G}) = \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{X} - \mathbf{X}\mathbf{G}\mathbf{E}^{\mathbf{T}}\|_{\mathbf{F}}^{\mathbf{2}} + \psi(\mathbf{G}) + \boldsymbol{\gamma}\|\boldsymbol{S}\|_{\mathbf{1}} \quad (1.26)$$

$$\text{sao cho } \mathbf{E}\mathbf{E}^{\mathbf{T}} = \mathbf{I}$$

ở đây, ma trận $\boldsymbol{S}$ nắm bắt các giá trị ngoại lai bị hỏng nặng trong tập dữ liệu ban đầu.

Nghiên cứu [91] cũng chỉ ra rằng phương pháp tìm các thành phần chính, thành phần chính thưa bằng sử dụng phương pháp PCA và SPCA thông qua giải bài toán tối ưu cho kết quả giống như tìm các thành phần chính bằng phương pháp PCA và SPCA truyền thống [53], [54]. Tuy nhiên các phương pháp PCA, SPCA, RSPCA, và ROBSPCA được thực hiện theo cách tiếp cận thứ hai hiệu quả hơn nhiều về mặt tốc độ tính toán.

1.3.4.4 Thủ thuật hàm nhân

Giả sử **X** là tập dữ liệu không xấp xỉ một siêu phẳng. Ý tưởng của thủ thuật hàm nhân là ánh xạ tập dữ liệu **X** vào không gian có số chiều cao hơn, có thể là vô hạn chiều được gọi là không gian đặc trưng $\boldsymbol{\mathcal{H}}$ thông qua một hàm phi tuyến **Φ:**

$$\boldsymbol{\Phi}: \mathbf{X} \longrightarrow \boldsymbol{\mathcal{H}}$$

$$\mathbf{X} \longrightarrow \boldsymbol{\Phi}(\mathbf{X}) \quad (1.27)$$

sao cho trong không gian đặc trưng $\boldsymbol{\mathcal{H}}$, tập dữ liệu ban đầu **X** được biến đổi thành tập $\boldsymbol{\Phi}(\mathbf{X})$ xấp xỉ một siêu phẳng và ta có thể thực hiện phương pháp PCA để giảm chiều tập $\boldsymbol{\Phi}(\mathbf{X})$ trong không gian đặc trưng. Theo cách tiếp cận này việc xác định được hàm phi tuyến $\boldsymbol{\Phi}$ và không gian đặc trưng $\boldsymbol{\mathcal{H}}$ là rất khó.

Một cách tiếp cận khác được xem là tương đương như cách tiếp cận trên nhưng tránh được những khó khăn đã nêu cũng như giảm chi phí tính toán tốn kém là sử dụng thủ thuật hàm nhân bằng cách chọn hàm $\boldsymbol{\kappa}$ đối xứng, xác định dương (hoặc bán xác định dương). Khi đó:

$$\boldsymbol{\kappa}(X_1, X_2) = \boldsymbol{\Phi}(X_1).\boldsymbol{\Phi}(X_2) \tag{1.28}$$

$\boldsymbol{\kappa}$ được gọi là hàm nhân và thay vì phải tìm không gian đặc trưng $\boldsymbol{\mathcal{H}}$ và $\boldsymbol{\Phi}(\mathbf{X})$ để từ đó xác định được ma trận hiệp phương sai của nó ta chỉ cần tìm ma trận hàm nhân $\mathbf{K} = [\boldsymbol{\kappa}(X, X')]$ với $X, X' \in \mathbf{X}$.

1.3.4.5 Phương pháp KPCA

KPCA là phương pháp giảm chiều tập dữ liệu **X** không xấp xỉ một siêu phẳng [92]. Ý tưởng chính của phương pháp KPCA là thực hiện một phép biến đổi $\boldsymbol{\Phi}$ là hàm có thể là phi tuyến tính từ không gian véc tơ đầu vào $\mathbb{R}^{\boldsymbol{m}}$ vào không gian chiều cao hơn $\mathbb{R}^D$, có thể vô hạn chiều, được gọi là không gian đặc trưng $\boldsymbol{\mathcal{H}}$ sao cho tập $\mathbf{X} \in \mathbb{R}^{\mathbf{m}}$ (không xấp xỉ một siêu phẳng) biến đổi thành tập $\boldsymbol{\Phi}(\mathbf{X}) \in \mathbb{R}^D$ (xấp xỉ một siêu phẳng), ở đây $D \gg m$. Phương pháp KPCA có thể được trình bày dưới dạng giả code như sau:

Thuật toán KPCA [46]

Input: $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{m}}$.

Output: $\mathbf{Y} \in \mathbb{R}^{\mathrm{N}\times\mathrm{p}}$.

Begin

1: Xác định tập dữ liệu tuyến tính $\boldsymbol{\Phi}(\mathbf{X})$ là ảnh của tập dữ liệu **X** qua ánh xạ phi tuyến với hàm nhân $\boldsymbol{\Phi}$.

2: Ứng dụng phương pháp PCA cho $\boldsymbol{\Phi}(\mathbf{X})$ để thu được tập dữ liệu chiều biến được giảm **Y**.

End.

Có thể thấy rằng ma trận $\boldsymbol{C}^* = \boldsymbol{\Phi}(\mathbf{X}).\boldsymbol{\Phi}(\mathbf{X})^{\mathbf{T}}$ là ma trận vuông cấp $D \times D$ của $\boldsymbol{\Phi}(\mathbf{X})$ trong không gian $\mathbb{R}^D$ là rất khó xác định vì hai lý do: thứ nhất $\boldsymbol{\Phi}$ rất khó xác định và thứ hai là số chiều $D \times D$ là rất lớn và có thể vô hạn chiều nếu như hàm nhân là hàm Gauss, do đó nỗ lực chéo hóa ma trận $\boldsymbol{C}^*$ là không thể thực hiện được. Tuy nhiên, có thể sử dụng ma trận Gramm $\boldsymbol{G} = \boldsymbol{\Phi}(\mathbf{X})^{\mathbf{T}}.\boldsymbol{\Phi}(\mathbf{X})$ là ma trận vuông cấp $N \times N$ được cân chỉnh trung bình. Việc chéo hóa ma trận $\boldsymbol{G}$ được thực hiện theo cách giống như ma trận $\boldsymbol{C}^*$. Vì vậy trong thực tế phép biến đổi $\boldsymbol{\Phi}$ chủ yếu là nhằm xác định được ma trận $\boldsymbol{G}$ [92]. Ma trận $\boldsymbol{G}$ được xác định thông qua ma trận hàm nhân **K,** ở đây:

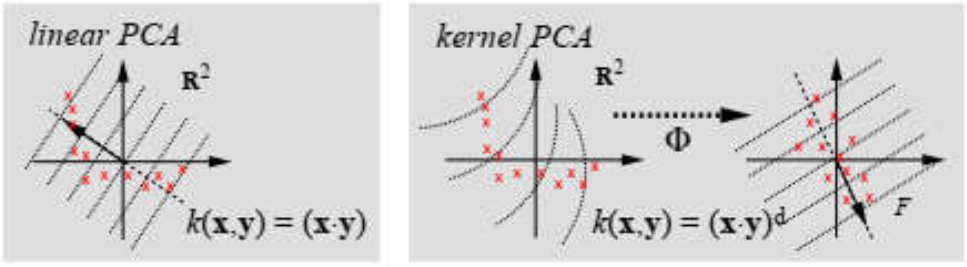
$$[\boldsymbol{K}]_{i,j} = \boldsymbol{\Phi}(\chi_i)^{\mathbf{T}}.\boldsymbol{\Phi}(\chi_j) = \boldsymbol{\kappa}(\chi_i, \chi_j), \mathrm{i}, \mathrm{j} = 1, \dots, N \qquad (1.29)$$

trong đó các χ_i là các *điểm dữ liệu* của tập dữ liệu ban đầu **X, κ** là hàm đối xứng, xác định dương (hoặc nửa xác định dương). Và ma trận $\boldsymbol{G}$ được cân chỉnh trung bình sẽ được xác định bởi công thức:

$$\boldsymbol{G} = \mathbf{K} - \mathbf{K}.\mathbf{1}_N - \mathbf{1}_N.\mathbf{K} + \mathbf{1}_N.\mathbf{K}.\mathbf{1}_N \qquad (1.30)$$

ở đây $\mathbf{1}_N(\mathrm{i}, \mathrm{j}) = 1/\boldsymbol{N}$ [47], [48], [92].

Phương pháp KPCA xác định các thành phần chính phi tuyến (PC) bằng cách chiếu $\boldsymbol{\Phi}(\mathbf{X})$ lên các véc tơ riêng của ma trận Gramm $\boldsymbol{G}$. Trong nhiều nghiên cứu người ta thường ký hiệu ma trận Gramm của ma trận hàm nhân **K** là $\boldsymbol{K_c}$ để nhắc rằng nó là ma trận hàm nhân được cân chỉnh trung bình**.** Việc thực hiện theo cách như vậy được gọi là thủ thuật hàm nhân.



Hình 1.3: Phương pháp giảm chiều PCA và KPCA [47]

Trong thực tế có rất nhiều các hàm nhân thỏa mãn là hàm đối xứng và xác định dương (hoặc bán xác định dương), chẳng hạn như hàm nhân đa thức, hàm nhân Gauss, hàm nhân Sigmoid,… [93].

Nhận xét 1.1: Trong trường hợp hàm nhân đa thức có dạng $\boldsymbol{\kappa}(\chi_i, \chi_j) = < \chi_i, \chi_j >$ khi đó ma trận hàm nhân của tập dữ liệu $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{m}}$ theo phương pháp KPCA là ma trận cấp $N \times N$ trong khi ma trận hiệp phương sai của tập dữ liệu $\mathbf{X}$ có cấp $m \times m$. Như vậy, phương pháp KPCA không trở thành phương pháp PCA khi hàm hạt nhân trở thành tích vô hướng của hai véc tơ. Tuy nhiên nghiên cứu [94] đã chỉ ra rằng các giá trị riêng và véc tơ riêng của ma trận hiệp phương sai của $\mathbf{X}$ là $\mathbf{X}^{\boldsymbol{T}}_{\boldsymbol{m}\times\boldsymbol{N}}.\mathbf{X}_{\boldsymbol{N}\times\boldsymbol{m}}$ có liên quan chặt chẽ theo nghĩa có thể tính toán được từ các giá trị riêng, véc tơ riêng của ma trận $\mathbf{X}_{\boldsymbol{N}\times\boldsymbol{m}}.\mathbf{X}^{\boldsymbol{T}}_{\boldsymbol{m}\times\boldsymbol{N}}$ tương ứng với chúng**.** Kết quả này cho thấy phương pháp KPCA được xem là sự mở rộng tự nhiên của phương pháp PCA, nghĩa là các thành phần chính phi tuyến được xác định bằng sử dụng phương pháp KPCA cũng được tìm thấy bằng phương pháp PCA.

1.3.5 Mô hình DFM

1.3.5.1 Mô hình BE nhân tố

Mô hình BE là mô hình hồi quy tuyến tính liên kết các biến ở tần suất cao với các biến ở tần suất thấp hơn. Phương pháp này cho phép ước tính sớm các biến tần suất thấp bằng cách sử dụng các biến ở tần suất cao hơn [13], [34].

Mô hình phương trình BE nhân tố được đề xuất như sau [13]:

$$y_t = \alpha + \sum_{i=1}^{m} \beta_i x_{i,t} + \sum_{j=1}^{k} \gamma_i F_{j,t} + u_t \qquad (1.31)$$

trong đó y_t, $x_{\mathrm{i,t}}$ lần lượt là biến phụ thuộc tần suất thấp tại thời điểm t và các biến giải thích ở cùng tần suất với tần suất biến phụ thuộc $\mathrm{y_t}$; $F_{j,t}$ là các nhân tố cùng tần suất với $\mathrm{y_t}$ và được tích hợp tương ứng từ các nhân tố $F^H_{j,t/s}$ ở tần suất cao hơn. Ở đây $F^H_{j,t/s}$ được chiết xuất từ một tập lớn các biến giải thích ban đầu $z^H_{j,t/s}$ được lấy mẫu ở tần suất cao hơn với *S* giá trị ứng với mỗi giá trị tần suất thấp bằng cách sử dụng phương pháp chuyển đổi thuộc tính để giảm chiều biến. $F^H_{j,t/s}$ cũng như $z^H_{j,t/s}$ được gọi là các thành phần tần suất cao trong mô hình tần suất hỗn hợp; u_t là phần dư (sai số) của mô hình.

Do các hệ thống kinh tế - tài chính đều có quán tính và các biến kinh tế vĩ mô thường tồn tại tự tương quan, nên cần thêm các biến trễ của các biến vào trong mô hình dự báo các chỉ tiêu kinh tế - tài chính, vì vậy mô hình BE nhân tố có thể được mở rộng thêm để bao gồm các biến trễ của biến phụ thuộc cũng như các biến giải thích. Khi đó, mô hình (1.30) có thể được viết dưới dạng:

$$y_t = \sum_{k=1}^{P} b_k y_{t-k} + \sum_{i=1}^{m}\sum_{j=0}^{r_i} \beta_{i_j} x_{i,t-j} + \sum_{j=1}^{k}\sum_{h=0}^{p_j} \gamma_{j_h} F_{j,t-h} + c + u_t \qquad (1.32)$$

ở đây, r_i *(i =1, ..., m)*, p_j *(j =1, ..., k)* và *p* tương ứng là độ trễ tối ưu của các biến $x_{i,t}$, $F_{j,t}$ và y_t. Các độ trễ tối ưu có thể được xác định bằng cách sử dụng tiêu chuẩn thông tin Akaike (AIC) hoặc tiêu chuẩn thông tin Bayes (BIC) [74], [76].

Mô hình (1.36) có thể viết lại như sau:

$$\psi(L) y_t = \sum_{i=1}^{m} \beta_i(L)\, x_{i,t} + \sum_{j=1}^{k} \gamma_j(L)\, F_{j,t} + c + u_t \qquad (1.33)$$

ở đây, L là biểu thị toán tử trễ lùi, $\psi(L) = 1 - \sum_{j=1}^{p} b_j L^j$, $\beta_i(L) = \sum_{j=0}^{r_i} \beta_{i_j} L^j$, và $\gamma_j(L) = \sum_{h=0}^{p_j} \gamma_{j_h} L^h$.

Nhận xét 1.2: Khi S = 1, các mô hình BE nhân tố (1.31), (1.32), và (1.33) là mô hình trễ phân bố tự hồi quy (ARDL) [81] có dạng:

$$y_t = \boldsymbol{\mathcal{F}_1}(y_t, F_{j,t}) = \sum_{k=1}^{q} b_k y_{t-k} + \sum_{j=1}^{p}\sum_{h=0}^{p_j} \gamma_{j_h} F_{j,t-h} + c + u_t \qquad (1.34)$$

trong đó, u_t là sai số của mô hình với giả định là nhiễu trắng. c, b_k, và γ_{j_h} là các tham số ước lượng; p_j, (*j =1, ..., p*) và *q* tương ứng là các độ dài trễ tối ưu của các biến $F_{j,t}$ và Y_t; $F_{j,t}$ (*j =1, ..., p*) là các nhân tố được chiết xuất từ tập **X** bằng sử dụng phương pháp giảm chiều biến.

Cách tiếp cận mô hình BE nhân tố cung cấp một giải pháp thuận tiện để lọc và tổng hợp các biến ở các tần suất khác nhau [64]. Tuy nhiên, việc tổng hợp có thể dẫn

đến mất thông tin hữu ích [12], [65]. Vấn đề này đã dẫn đến sự phát triển mô hình tần suất hỗn hợp được gọi là MIDAS [65]. Mô hình này cho phép thực hiện hồi quy trực tiếp biến phụ thuộc ở tần suất thấp theo các biến giải thích được lấy mẫu ở các tần suất cao hơn khác nhau.

1.3.5.2 Mô hình MIDAS nhân tố

Mô hình MIDAS nhân tố có dạng [67]:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + f\left(\{F_{t/S}^H\}, \theta, \lambda\right) + u_t \qquad (1.35)$$

ở đây y_t, $x_{\mathrm{i,t}}$ lần lượt là biến phụ thuộc tần suất thấp tại thời điểm t và các biến giải thích ở cùng tần suất với tần suất biến $\mathrm{y_t}$; $\{F_{t/s}^H\}$ là tập các nhân tố được chiết xuất từ tập lớn các biến giải thích được lấy mẫu ở một tần suất cao hơn nào đó với số lượng giá trị tần suất cao tương ứng với một giá trị tần suất thấp là S; $\psi(L) = 1 - \sum_{j=1}^{P} b_j L^j$; $\beta_i(L) = \sum_{j=0}^{r_i} \beta_{i_j} L^j$; f là hàm mô tả tác động của dữ liệu tần suất cao trong hồi quy tần suất thấp; b = (b_k), $\beta_i = \left(\beta_{i_j}\right)$, θ, và λ là các véc tơ tham số cần được ước lượng.

a. Mô hình MIDAS không bị hạn chế (U-MIDAS)

Nếu ta chỉ muốn bao gồm các thành phần ở tần suất cao hơn làm biến giải thích trong hồi quy tần suất thấp thì mô hình (1.35) có thể được đưa về dạng sau [67]:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{\tau=0}^{k-1} {F_{(t-\tau)/S}^H}^T \theta_\tau + u_t \qquad (1.36)$$

ở đây, ${F_{(t-\tau)/S}^H}^T$ là τ giai đoạn tần suất cao trước t (được gọi là trễ tần suất cao thứ τ tại thời điểm t). Mỗi θ_τ riêng biệt được liên kết với mỗi nhân tố trễ tần suất cao của *S*. Số lượng hệ số θ_τ có thể lớn hơn *S*. Trong trường hợp các hệ số này không bị ràng buộc, thì mô hình (1.36) được gọi là mô hình U-MIDAS.

b. Mô hình MIDAS trọng số STEP (STEP-MIDAS)

Trong mô hình STEP-MIDAS, các hệ số trên dữ liệu tần suất cao bị hạn chế bằng cách sử dụng hàm STEP. Cụ thể, mô hình STEP-MIDAS được tạo ra từ mô hình (1.36) và có dạng [67], [95]:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{\tau=0}^{k-1} (F^H_{(t-\tau)/S})^T \varphi_\tau + u_t \qquad (1.37)$$

ở đây, δ là độ dài STEP; k là số trễ tần suất cao ($k \leq S$) và $\varphi_\tau = \theta_i$, $i = int(\tau/\delta)$.

c. Mô hình MIDAS trọng số Almon đa thức (PAW-MIDAS)

Đối với tần suất cao đến k, hệ số hồi quy của các thành phần tần suất cao trong mô hình (1.37) được mô hình hóa dưới dạng đa thức bậc p trong tham số MIDAS là θ và mô hình hồi quy MIDAS có dạng như sau [67], [95]:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{\tau=0}^{k-1} (F^H_{(t-\tau)/S})^T (\sum_{j=0}^{p} \tau^j \theta_j) + u_t \qquad (1.38)$$

trong đó k là độ trễ lớn nhất được chọn ($k < S$ hoặc $k > S$); và p là bậc đa thức Almon. Do đó, số lượng hệ số cần được ước lượng phụ thuộc vào bậc đa thức chứ không phải số lượng trễ tần suất cao. Mô hình (1.38) được gọi là mô hình *PAW-MIDAS*. Bằng cách sắp xếp lại, mô hình (1.38) được viết lại như sau:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{i=0}^{p} {Z_{i,t}}^T \theta_i + u_t \qquad (1.39)$$

trong đó, $Z_{i,t} = \sum_{\tau=0}^{k-1} \tau^i F^H_{(t-\tau)/S}$

d. Mô hình MIDAS trọng số Almon mũ (EAW-MIDAS)

Mô hình EAW-MIDAS là mô hình MIDAS trễ phân bố không phải là đa thức. Mô hình này sử dụng trọng số mũ và đa thức trễ bậc 2. mô hình EAW-MIDAS có dạng [67], [95]:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{\tau=0}^{k-1} (F^H_{(t-\tau)/S})^T (\frac{\exp(\tau\theta_1 + \tau^2\theta_2)}{\sum_{j=0}^{k} \exp(j\theta_1 + j^2\theta_2)}) + u_t \qquad (1.40)$$

Mô hình (1.40) được viết lại như sau:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \sum_{\tau=0}^{k} Z_{\tau,t}{}^{T} + u_t \quad (1.41)$$

$$Z_{\tau,t} = \left(\frac{\exp(\tau\theta_1+\tau^2\theta_2)}{\sum_{j=0}^{k}\exp(j\theta_1+j^2\theta_2)}\right) F^{H}_{(t-\tau)/S}$$

trong đó, *k* là số lượng trễ được chọn; hàm trọng số mũ và đa thức trễ phụ thuộc vào hai hệ số MIDAS θ_1 vả θ_2.

Có thể thấy, mô hình (1.38) nếu ta chỉ cần thêm tổng trọng số bằng nhau (hoặc trung bình các trọng số) của biến giải thích tần suất cao trong hồi quy tần suất thấp, thì mô hình này có dạng:

$$\psi(L)y_t = \sum_{i=1}^{m} \beta_i(L)x_{i,t} + \left(\sum_{\tau=0}^{S-1} F^{H}_{(t-\tau)/S}\right)^{T}\lambda + u_t \quad (1.42)$$

Véc tơ tham số λ được liên kết với một biến giải thích và mô hình (1.42) gần giống như mô hình được xác định bởi công thức (1.39). Mô hình này được gọi là mô hình trọng số tổng hợp bằng nhau.

Nhận xét 1.3: Các mô hình (1.36) và (1.42) có thể được coi là hai thái cực của mô hình MIDAS. Mô hình (1.36) cung cấp tính linh hoạt cao nhất nhưng yêu cầu một số lượng lớn các hệ số. Các mô hình từ (1.37) đến (1.40) được coi là nằm giữa hai mô hình này. Bằng cách đưa ra các hạn chế khác nhau về ảnh hưởng của các biến tần suất cao ở các độ trễ khác nhau, người ta có thể tạo ra các mô hình MIDAS trung gian giữa mô hình U-MIDAS (1.36) và mô hình trọng số tổng hợp bằng nhau (1.42). Những hạn chế như vậy đối với ảnh hưởng của các biến tần suất cao có thể được thực hiện thông qua các hàm trọng số MIDAS. Khi *S = 1* thì mô hình MIDAS (1.35) cũng là mô hình ARDL (1.34) [81].

Nhận xét trên cho thấy rằng trên tập dữ liệu của các biến giải thích có cùng tần suất lấy mẫu với biến phụ thuộc, các bài toán nowcast là các bài toán dự báo trên tập dữ liệu chuỗi thời gian lớn.

1.3.6 Quy trình mô hình hóa dự báo kinh tế - tài chính

Mô hình hóa dự báo kinh tế - tài chính là một quá trình đã được thừa nhận là chuẩn công nghiệp CRISP-DM (CRoss Industry Standard Process for Data Mining) [96]. Quá trình này khá tương tự như quá trình khai phá tri thức từ các cơ sở dữ liệu (hay KDD) và có thể được chia thành 6 pha theo trình tự như sau [96]:

1. *Xác định và định nghĩa vấn đề Kinh doanh/Nghiên cứu* (Understanding Business/Research Objective): Mục đích của pha này là hiểu lĩnh vực cần được dự báo, hiểu nhu cầu tri thức của người sử dụng, xác định rõ vấn đề và tạo tập dữ liệu phục vụ dự báo.

2. *Hiểu dữ liệu* (Data Understanding): Làm quen với dữ liệu, làm sạch dữ liệu chẳng hạn như loại bỏ các dữ liệu tạp, dữ liệu bất thường và lỗi, và bổ sung dữ liệu bị mất.

3. *Chuẩn bị dữ liệu* (Data Preparation): Chuyển đổi dữ liệu sang dạng phù hợp, rút gọn kích thước dữ liệu thông qua việc tìm các thuộc tính hữu ích, giảm bớt số chiều và biến đổi dữ liệu để nhận được các bất biến, xây dựng tập dữ liệu để chạy mô hình.

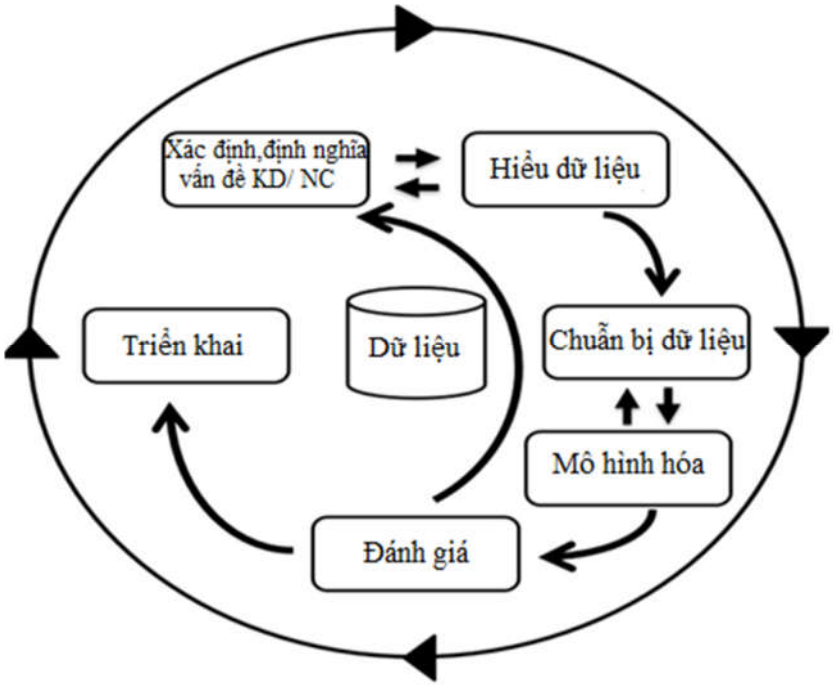
4. *Xây dựng mô hình dự báo* (modelling): Trên cơ sở nhiệm vụ phân tích và dự báo kinh tế - tài chính cần lựa chọn phương pháp dự báo (hay kỹ thuật học máy) phù hợp, sau đó thực hiện xây dựng mô hình. Mô hình dự báo được xây dựng trên tập dữ liệu huấn luyện (Training dataset).

5. *Đánh giá* (Evaluation): Thực hiện dự báo kiểm định, chấp nhận mô hình nhận được ở Pha 4. Nội dung chính của Pha này là sử dụng mô hình để dự báo ngoài mẫu biến phụ thuộc với độ xa nhất của dự báo ngoài mẫu bằng số quan sát trong tập dữ liệu kiểm thử. Sau đó, so sánh kết quả dự báo ngoài mẫu với số liệu thống kê thực tế (được rút ra từ tập dữ liệu kiểm thử (Testing dataset) để quyết định có chấp nhận mô hình dự báo này hay không?

6. *Triển khai* (Deployment): Trong mô hình hóa kinh tế - tài chính, triển khai ở đây là sử dụng mô hình dự báo được xây dựng để dự báo tương lai. Nếu sai số dự báo của mô hình (được đo bằng RMSE hoặc phần trăm sai số tuyệt đối) là nhỏ thì mô hình dự báo là được chấp nhận, ta có thể sử dụng mô hình này để dự báo tương lai.

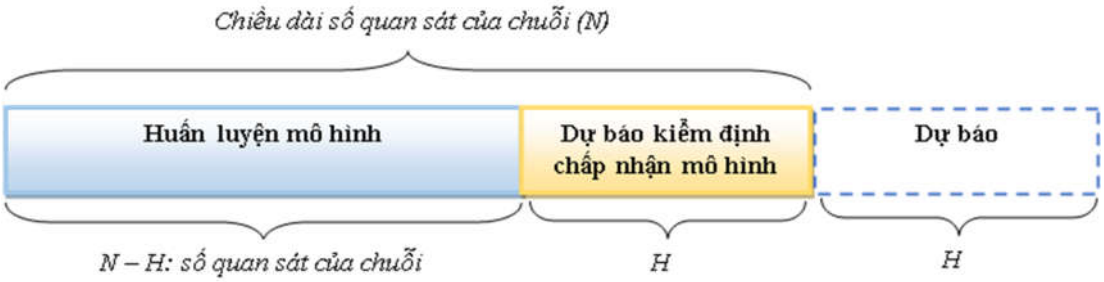
Để dự báo tương lai, ta cần ước lượng lại các tham số của mô hình trên toàn bộ tập dữ liệu (bao gồm các tập dữ liệu huấn luyện và kiểm thử) và sử dụng mô hình mới được cập nhật để thực hiện dự báo ngoài mẫu.

Quá trình mô hình hóa dự báo kinh tế - tài chính nêu trên được mô tả tóm tắt trong Hình 1.4 ở dưới [96], ở đó vòng tròn bên ngoài thể hiện trình tự thực hiện các pha của quá trình này, trong khi các mũi tên bên trong cho thấy trong quá trình xây dựng mô hình và thực hiện dự báo, trong nhiều trường hợp ta phải quay lại các pha trước đó để phân tích và đưa ra các điều chỉnh cần thiết.



Hình 1.4: Quá trình mô hình hóa dự báo kinh tế - tài chính [96]

Trong lĩnh vực kinh tế, 03 pha cuối cùng của mô hình chuẩn công nghiệp CRISP-DM là quan trọng nhất trong quy trình mô hình hóa dự báo kinh tế - tài chính được trình bày trong Hình 1.4. Hình 1.5 là thể hiện mối quan hệ của 3 pha cuối cùng dưới góc nhìn của mô hình hóa kinh tế. Hình này cho thấy số lượng các quan sát trong tập dữ liệu kiểm thử bằng độ xa nhất của dự báo.



Hình 1.5: Ba pha cuối của quá trình mô hình hóa kinh tế - tài chính

1.3.7 Các tiêu chuẩn đánh giá độ chính xác của mô hình

Để đánh giá độ chính xác dự báo của mô hình được xây dựng bằng phương pháp hồi quy, người ta thường sử dụng một số trong các tiêu chuẩn sau:

- Sai số tuyệt đối trung bình *MAE* (Mean Absolute Error)

$$MAE = \frac{\sum_{t=1}^{N} \left| Y_t - \hat{Y}_t \right|}{N} \tag{1.43}$$

- Sai số bình phương trung bình MSE (Mean Squared Error)

$$MSE = \frac{\sum_{t=1}^{N} \left(Y_t - \hat{Y}_t \right)^2}{N} \tag{1.44}$$

- Sai số phần trăm tuyệt đối trung bình *MAPE* (Mean Absolute Percent Error)

$$MAPE = \frac{\sum_{t=1}^{N} \frac{\left| Y_t - \hat{Y}_t \right|}{Y_t}}{N} \tag{1.45}$$

- Sai số bình phương trung bình chuẩn *RMSE* (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N} \left(Y_t - \hat{Y}_t \right)^2}{N}} \tag{1.46}$$

- Phần trăm sai số dự báo so với giá trị thực tế

$$\%\ Sai\ số\ dự\ báo = \left| \frac{\left(Y_t - \hat{Y}_t \right) * 100}{Y_t} \right| \tag{1.47}$$

trong đó, N, Y_j, $\hat{Y}_j$ lần lượt là số quan sát, giá trị thực tế và giá trị dự báo của mô hình.

Tiêu chuẩn RMSE và tiêu chuẩn MSE được xem là như nhau. Chúng được sử dụng như hàm mất mát (loss) trong các bài toán học mạng nơron và các bài toán học

khác ở đó biến cần được dự báo nhận giá trị liên tục (hay giá trị số). Nếu tiêu chuẩn RMSE được dành cho những người phát triển các mô hình dự báo thì tiêu chuẩn % Sai số dự báo dành cho những người sử dụng kết quả dự báo của mô hình do tính dễ hiểu, trực quan, gần gũi với đời sống thực.

Luận án này sử dụng các tiêu chuẩn RMSE theo công thức (1.46) và phần trăm sai số dự báo so với giá trị thực tế theo công thức (1.47) để đánh giá độ chính xác dự báo của mô hình.

1.4 Kết luận Chương 1

Trong chương này, luận án đã trình bày một số thuật ngữ tiếng Anh mà khi dịch sang tiếng Việt đều có nghĩa gần với thuật ngữ dự báo. Chương này đã tổng quan những nghiên cứu liên quan ở trong và ngoài nước để xác định khoảng trống nghiên cứu, từ đó xác định vấn đề và phạm vi nghiên cứu của luận án.

Chương này cũng trình bày một số kiến thức cơ bản cần thiết phục vụ cho các chương nghiên cứu tiếp theo. Chương 2 tiếp theo sẽ trình bày đề xuất phương pháp giảm chiều biến dựa vào kỹ thuật hàm nhân, được gọi tắt là phương pháp KTPCA và đánh giá hiệu suất giảm chiều của phương pháp đó.

CHƯƠNG 2. PHƯƠNG PHÁP GIẢM CHIỀU BIẾN DỰA VÀO KỸ THUẬT HÀM NHÂN

Chương này sẽ đề xuất phương pháp giảm chiều mới dựa vào kỹ thuật hàm nhân như là sự mở rộng tự nhiên khác của phương pháp PCA. Nó được gọi là phương pháp KTPCA. Việc thực nghiệm đánh giá hiệu suất giảm chiều của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất (gọi tắt là KTPCA lặp) trên các tập dữ liệu tần suất lấy mẫu giống nhau cũng như tần suất lấy mẫu hỗn hợp so với hiệu suất giảm chiều biến của các phương pháp PCA, SPCA, RSPCA, và ROBSPCA cũng được trình bày trong Chương này.

2.1 Phương pháp giảm chiều biến dựa vào kỹ thuật hàm nhân

Giả sử $\mathbf{X} = [X_1, X_2, \dots, X_m]$ là tập dữ liệu của các biến giải thích chuỗi thời gian, $X_i \in \mathbb{R}^{\mathbf{N}}, i = 1, \dots, m$; m là rất lớn. Không mất tính tổng quát, **X** là ma trận đã được cân chỉnh trung bình, tức là $\sum_{j=1}^{N} x_{i_j} = 0\,, \forall \mathrm{i} = 1, \dots, \mathrm{m}$.

2.1.1 Phương pháp giảm chiều dựa vào kỹ thuật hàm nhân

Chương 1 đã chỉ rõ mặc dù phương pháp giảm chiều KPCA là sự mở rộng tự nhiên của phương pháp PCA. Với các tập dữ liệu tuyến tính thì PCA là phương pháp giảm chiều tốt nhất và với tập dữ liệu chỉ xấp xỉ tuyến tính thì hiệu suất giảm chiều của phương pháp KPCA không tốt bằng phương pháp PCA. Vấn đề xác định mức độ xấp xỉ tuyến tính của tập dữ liệu để hiệu suất giảm chiều của phương pháp PCA còn tốt hơn phương pháp KPCA vẫn là vấn đề mở. Luận án chưa nghiên cứu giải quyết vấn đề này.

Tuy nhiên ý tưởng của phương pháp KPCA gợi ý để luận án đề xuất phương pháp giảm chiều mới dựa trên hàm nhân và được gọi là KTPCA để phân biệt nó với phương pháp KPCA. Phương pháp này khác với phương pháp KPCA ở chỗ:

- Ma trận hàm nhân xác định bởi $\boldsymbol{K} = [\boldsymbol{\kappa}(X_i, X_j)]$, được tính toán trên các véc tơ dữ liệu đầu vào $\mathrm{X}_i, \mathrm{X}_j$. Như vậy ma trận hàm nhân trong phương pháp KTPCA là khác với ma trận hàm nhân trong phương pháp KPCA là được tính trên các điểm dữ liệu χ_i, χ_j như theo công thức (1.29).

- Thay vì *chiếu tập dữ liệu* $\boldsymbol{\Phi}(\boldsymbol{X})$ *được cân chỉnh trung bình* lên các véc tơ riêng của ma trận hàm nhân trong không gian đặc trưng $\boldsymbol{\mathcal{H}}$, nhưng vì không gian này không được chỉ ra cụ thể, nên việc tìm các thành phần chính hàm nhân được tính toán bằng sử dụng hàm điểm. Trong khi đó, phương pháp KTPCA *chiếu tập dữ liệu đầu vào* $\boldsymbol{X}$ *được cân chỉnh trung bình* lên tập các véc tơ riêng của ma trận hàm nhân $\boldsymbol{K}$ trong không gian đầu vào. Vì không gian đầu vào đã được xác định tường minh nên có thể thực hiện phép chiếu này một cách dễ dàng.

- Phương pháp KTPCA khác với phương pháp KPCA ở việc sử dụng hàm nhân. Phương pháp KPCA sử dụng thủ thuật hàm nhân để tạo ra không gian đặc trưng, ở đó tập dữ liệu đầu vào là xấp xỉ một siêu phẳng. Phương pháp KTPCA không tiếp cận theo cách như vậy và nó được xem là phương pháp sử dụng kỹ thuật hàm nhân.

Giả sử các giá trị riêng của ma trận hàm nhân được sắp xếp theo thứ tự giảm dần và q(%) là ngưỡng phần trăm giá trị riêng tích lũy do người dùng xác định, q(%) thường lớn hơn 70%. Giả sử $PCV(k) \geq q$, thế thì p nhân tố thành phần chính được chọn để thay thế cho tập m biến giải thích đầu vào bằng sử dụng phương pháp KTPCA được xác định như sau:

$$\mathbf{PC}_{N\times \mathrm{p}} = \mathbf{X}_{N\times m}.\tilde{\mathbf{E}}_{m\times p} \tag{2.1}$$

ở đây, $\tilde{\mathbf{E}}_{\mathrm{m}\times \mathrm{p}}$ là ma trận của p véc tơ riêng đầu tiên tương ứng với các trị riêng lớn nhất của ma trận hàm nhân $\boldsymbol{K}$. Nói cách khác thuật toán giảm chiều bằng sử dụng phương pháp KTPCA có thể được viết dưới dạng giả code như sau:

Thuật toán KTPCA

Input: $\mathbf{X} \in \mathbb{R}^{N\times m}$, hàm nhân $\boldsymbol{\kappa}$ và q(%)

Output: $\mathbf{Y} \in \mathbb{R}^{N\times p},\ p \ll m$

Begin

1. Xây dựng ma trận hàm nhân $\boldsymbol{K} = [\boldsymbol{\kappa}(X_i, X_j)]$
2. Tìm giá trị riêng và véc tơ riêng của ma trận hàm nhân
3. Sắp xếp các véc tơ riêng tương ứng với các giá trị riêng theo thứ tự giảm dần

4. Xây dựng ma trận $\widetilde{\boldsymbol{E}}_{m\times p}$ với p véc tơ riêng đầu tiên
5. Cân chỉnh trung bình $\mathbf{X}$ và sử dụng $\widetilde{\boldsymbol{E}}_{m\times p}$ để thu được không gian con mới

$\mathbf{Y} = \mathbf{X}.\widetilde{\boldsymbol{E}}_{m\times p}$

End

Như vậy có thể thấy rằng phương pháp KTPCA là một sự kết hợp ý tưởng giảm chiều của hai phương pháp KPCA và PCA. Khi hàm nhân $\boldsymbol{\kappa}$ là tích vô hướng của hai véc tơ đầu vào, tức là $\boldsymbol{\kappa}(X_i, X_j) = < X_i, X_j >$ thì ma trận hàm nhân $\boldsymbol{K}$ trở thành ma trận hiệp phương sai, và phương pháp KTPCA trở thành phương pháp PCA. Đó là điều mà luận án mong muốn.

Trong khi sử dụng phương pháp KTPCA để giảm chiều biến, điều cốt yếu là phải chọn hàm nhân phù hợp sao cho RMSE của mô hình dự báo biến phụ thuộc theo các nhân tố được chiết xuất tương ứng với hàm nhân này là nhỏ nhất. Cũng như phương pháp KPCA, cho đến thời điểm này chưa có tiêu chuẩn nào để lựa chọn được hàm nhân tối ưu như vậy cho phương pháp KTPCA. Do đó, hàm nhân phù hợp nhất để giảm chiều dữ liệu bằng phương pháp KTPCA chỉ có thể được xác định bằng quá trình thử và sai dựa vào mô hình có RMSE tốt nhất. Phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất được gọi là KTPCA lặp.

Do có rất nhiều hàm thỏa mãn điều kiện đối xứng xác định dương nên các hàm nhân cũng rất đa dạng và phong phú. Nhiều nghiên cứu gợi ý rằng trong ứng dụng thực tiễn thì hàm nhân đa thức $\boldsymbol{\kappa}(X_i, X_j) = (c_1 < X_i, X_j > + c_2)^d$ và hàm nhân Gauss $\boldsymbol{\kappa}(X_i, X_j) = \exp\left(-\frac{\|\mathrm{X_i} - \mathrm{X_j}\|^2}{2\rho^2}\right)$, trong đó $c_1 > 0, c_2 \geq 0,\ d \in \mathbb{N}$, và $\rho > 0$ là các tham số do người dùng chọn là được sử dụng phổ biến nhất [34], [93]. Đối với hàm nhân Gauss, tham số ρ^2 được khuyến nghị là nên được chọn xung quanh giá trị là khoảng cách trung bình tối thiểu giữa hai véc tơ dữ liệu (ρ_0^2) của các biến giải thích [97]. Cụ thể giá trị này được xác định bởi công thức sau:

$$\rho_0^2 = c.\frac{1}{m}\sum_{i=1}^{m} min_{i\neq j} \left\|X_{\mathrm{i}} - X_{\mathrm{j}}\right\|^2 \qquad (2.2)$$

ở đây, c là tham số do người dùng chọn và được gọi là tham số điều khiển.

Trong khi đó đối với hàm nhân đa thức, khi d = 1, $c_1 = 1$, và $c_2 = 0$, $\boldsymbol{\kappa}(X_i, X_j) = < X_i, X_j >$, tức là $\boldsymbol{\Phi}(X_i) = X_i$, khi đó $\boldsymbol{\kappa}(X_i, X_j)$ là tích vô hướng của 02 véc tơ X_i, X_j và $\mathbf{K} = [\boldsymbol{\kappa}(X_i, X_j)]$ là ma trận hiệp phương sai của **X**, và phương pháp KTPCA và phương pháp PCA là trùng nhau.

Bảng 2.1 ở dưới tóm tắt các phương pháp PCA, KPCA và KTPCA. Qua đó cho thấy điểm khác nhau chủ yếu của các phương pháp này.

Các phương pháp giảm chiều học thuộc tính có thể được phân loại theo các tiêu chí khác nhau như Học phi tuyến tính hoặc học tuyến tính; Học có giám sát hoặc không giám sát; Học dựa trên phép chiếu ngẫu nhiên hoặc dựa trên đa tạp, Phương pháp giảm chiều dựa trên tối ưu lồi hoặc không lồi [19] [46] và Bảo toàn khoảng cách giữa các điểm dữ liệu hay không [98]. Xem xét tất cả các cách phân loại trên, PCA là phương pháp học tập dựa trên phép chiếu ngẫu nhiên, tuyến tính, không giám sát, phương pháp lồi và bảo toàn khoảng cách giữa các điểm dữ liệu.

Bảng 2.1: Sự khác nhau của các phương pháp PCA, KPCA, và KTPCA

PCA [18]	**KPCA** [47]	**KTPCA**
- Tập dữ liệu $\boldsymbol{X} \in \mathbb{R}^{N\times m}$ được cân chỉnh trung bình - Tìm trị riêng và véc tơ riêng của ma trận hiệp phương sai của $\boldsymbol{X}$ - Sắp xếp véc tơ riêng theo giá trị riêng - p nhân tố đầu tiên được xác định bởi: $\boldsymbol{PC}_{N\times \mathrm{p}} = \boldsymbol{X}_{\mathrm{N}\times \mathrm{m}}.\boldsymbol{E}_{\mathrm{m}\times \mathrm{p}}$	- Tập dữ liệu $\boldsymbol{X} \in \mathbb{R}^{N\times m}$ - Xác định ma trận hàm nhân $\boldsymbol{K} = [\boldsymbol{\kappa}(\chi_i, \chi_j)]$, χ_i là véc tơ điểm dữ liệu của $\boldsymbol{X}$ và ma trận Gramm *cấp* $N \times N$*:* - $\boldsymbol{K_c} = \boldsymbol{K} - \boldsymbol{K}.\mathbf{1_N} - \mathbf{1_N}.\boldsymbol{K} + \mathbf{1_N}.\boldsymbol{K}.\mathbf{1_N}$ - Tìm trị riêng, véc tơ riêng của $\boldsymbol{K_c}$ - Thành phần chính hàm nhân được xác định thông qua hàm điểm: $f_v(\boldsymbol{\Phi}(Z)) = v.\boldsymbol{\Phi}(Z) = \sum_{i=1}^{m} \alpha_i \ \boldsymbol{\Phi}(\chi_i).\boldsymbol{\Phi}(Z) = \sum_{i=1}^{m} \alpha_i \boldsymbol{\kappa}(\chi_i, Z)$, ở đây Z là điểm dữ liệu của **X**.	- Tập dữ liệu $\boldsymbol{X} \in \mathbb{R}^{N\times m}$ được cân chỉnh trung bình - Xác định ma trận hàm nhân $\boldsymbol{K_{m\times m}} = [\boldsymbol{\kappa}(X_i, X_j)]$, X_i là véc tơ dữ liệu của $\boldsymbol{X}$**.** - Tìm trị riêng và véc tơ của ma trận $\boldsymbol{K}$ ứng với hàm nhân k; - p nhân tố được xác định bởi: $\boldsymbol{PC}_{N\times p} = \boldsymbol{X}_{\mathrm{N}\times \mathrm{m}}.\widetilde{\boldsymbol{E}}_{\mathrm{m}\times \mathrm{p}}$

So với phương pháp PCA, KTPCA được phân loại tương tự như phương pháp PCA. Sự khác biệt quan trọng nhất giữa hai phương pháp này là phương pháp KTPCA có thể làm giảm chiều của các tập dữ liệu xấp xỉ một siêu phẳng hoặc không. Khi đó, KTPCA là một phương pháp học dựa trên đồ thị (học trên các tập dữ liệu xấp xỉ đa tạp và bảo toàn khoảng cách giữa các điểm dữ liệu) trong khi PCA thì không như vậy.

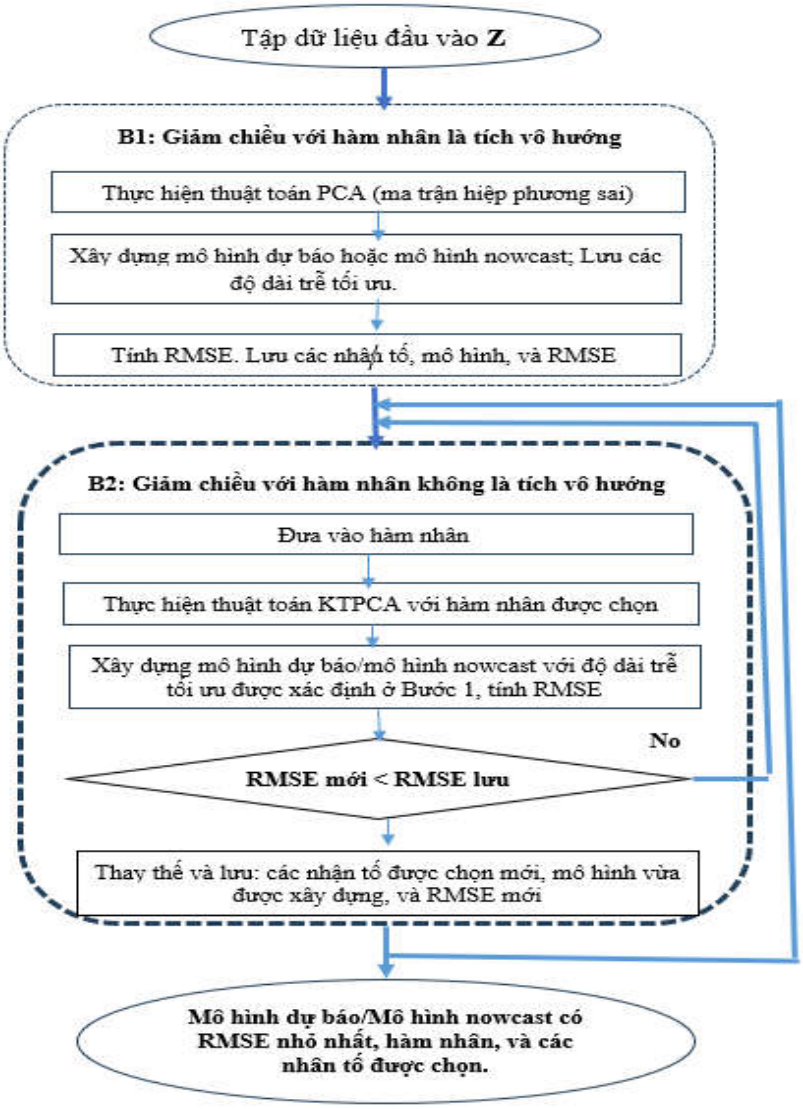
So với phương pháp KPCA, phương pháp KTPCA và KPCA đều là phương pháp học không giám sát, học dựa trên đồ thị và là phương pháp giảm chiều dựa trên tối ưu lồi [19]. Khác với phương pháp KPCA, phương pháp KTPCA được triển khai trên ma trận hàm nhân không được xây dựng trên các điểm dữ liệu mà trên các véc tơ dữ liệu của tập dữ liệu đầu vào. Nó không chiếu ảnh của tập dữ liệu gốc trong không gian đặc trưng lên các véc tơ riêng của ma trận hàm nhân mà chiếu tập dữ liệu gốc được cân chỉnh trung bình trong không gian đầu vào lên các véc tơ riêng của ma trận hàm nhân.

Cả ba thuật toán giảm chiều PCA, KPCA và KTPCA đều là các kỹ thuật tối ưu toàn cục. Hơn nữa, các nhân tố thành phần chính được chiết xuất bằng một trong các phương pháp KTPCA và PCA được thể hiện tường minh hơn nên chúng giải thích tốt hơn so với các nhân tố thành phần chính được chiết xuất bằng phương pháp KPCA.

2.1.2 Giảm chiều bằng sử dụng phương pháp KTPCA lặp

Việc giảm chiều biến bằng sử dụng phương pháp KTPCA lặp được trình bày trong Hình 2.1 bên dưới. Hình 2.1 cho thấy đó cũng là quy trình xây dựng mô hình dự báo và mô hình nowcast với hàm nhân là phù hợp nhất tùy thuộc vào tập dữ liệu đầu vào được lấy mẫu tần suất giống nhau hay tần suất lấy mẫu hỗn hợp. Trong trường hợp xây dựng mô hình dự báo thì các mô hình này được xây dựng dựa vào mô hình ARDL theo phương trình (1.34), còn trong trường hợp xây dựng mô hình nowcast thì các mô hình đó được xây dựng dựa vào mô hình BE theo phương trình (1.32) hoặc MIDAS bao gồm mô hình U-MIDAS theo phương trình (1.36) hoặc một trong số các mô hình MIDAS bị hạn chế chẳng hạn như các phương trình (1.37), (1.39) và (1.41).

Phương pháp KTPCA lặp được mô tả trong Hình 2.1 gồm 2 bước chính. Nội dung chi tiết từng bước được trình bày cụ thể như sau:



Hình 2.1: Phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất

a. Bước 1: Sử dụng phương pháp PCA

Giả sử, tập dữ liệu X đã được căn chỉnh trung bình, tức là $\sum_{j=1}^{N} x_{i_j} = 0, \forall i = 1, \ldots, m$.. Theo Hình 2.1, việc giảm chiều biến bằng phương pháp KTPCA lặp luôn được bắt đầu bằng sử dụng phương pháp PCA để chiết xuất các nhân tố thành phần chính, cụ thể như sau:

- Tính các giá trị riêng và véc tơ riêng của ma trận hiệp phương sai của **X**;

- Sắp xếp các véc tơ riêng theo các giá trị riêng giảm dần của các giá trị riêng tương ứng;

- Xác định số lượng các nhân tố được chọn dựa trên phần trăm giá trị riêng tích lũy của tổng số giá trị riêng theo công thức (1.19). Giả sử p là số lượng các nhân tố nhỏ nhất sao cho phần trăm phương sai tích lũy tương ứng với các nhân tố này lớn hơn ngưỡng phần trăm giá trị riêng tích lũy được xác định trước q(%), tức là $PCV(p) \geq q$. Ký hiệu **SPC** = $\{FAC_i$, i = 1,..., p$\}$;

- Xây dựng mô hình nowcast/dự báo của biến phụ thuộc Y theo các nhân tố được chọn FAC_i, ký hiệu là $\hat{Y}_t$.

- Tính giá trị RMSE của mô hình theo công thức (1.46) và được ký hiệu là ERR. Lưu ERR, các nhân tố FAC_i và mô hình $\hat{Y}_t$.

b. Bước 2: Sử dụng phương pháp KTPCA lặp

Bước này là một quy trình lặp theo các hàm nhân bằng phương pháp KTPCA, trong đó hàm nhân κ không phải là tích trong của hai véc tơ. Cụ thể, Bước này thực hiện các nội dung sau:

- Chọn một hàm nhân là đối xứng, xác định dương (hoặc bán xác định dương). Trong ứng dụng thực tế các hàm nhân này thường là hàm nhân đa thức hoặc hàm nhân Gauss với các tham số được chọn như trình bày ở mục trên.

- Tính ma trận hàm nhân **K.**

- Thực hiện phương pháp PCA trên ma trận hàm nhân **K**, xác định tập các nhân tố được chọn theo công thức (3.5), ký hiệu $\boldsymbol{FC} = (FAC_i, i \geq 1)$.

- Xây dựng mô hình nowcast/mô hình dự báo dựa trên mô hình hồi quy đã chọn ở Bước 1, ký hiệu YF.

- Tính giá trị $RMSE$ của mô hình theo công thức (1.46), ký hiệu $ERRF$.

- So sánh giá trị $RMSE$ của mô hình dự báo mới được xây dựng với $RMSE$ đang được lưu, nếu $ERRF \leq ERR$ thì thực hiện thay thế $ERR \leftarrow ERRF, \boldsymbol{SPC} \leftarrow \boldsymbol{FC}$ và $\hat{Y}_t \leftarrow \boldsymbol{YF}$.

Việc tiếp tục hay kết thúc quá trình lặp là tùy thuộc vào người dùng. Quá trình lặp càng nhiều với các hàm nhân được đưa vào thử nghiệm càng phù hợp thì mô hình được xây dựng cho độ chính xác dự báo càng cao. Vào cuối quá trình lặp này, tương ứng với hàm nhân phù hợp nhất trong số các hàm nhân được thử nghiệm, ta nhận được các nhân tố được dùng để thay thế tập các biến giải thích ban đầu, mô hình dự

báo hoặc mô hình nowcast của biến phụ thuộc, và RMSE của mô hình này. Do đó, việc giảm chiều biến bằng sử dụng phương pháp KTPCA lặp trong các bài toán dự báo và nowcast trên tập dữ liệu lớn các biến giải thích chuỗi thời gian và quá trình xây dựng các mô hình dự báo hoặc mô hình nowcast đã được tích hợp trong một.

Theo Hình 2.1 có thể thấy rằng mô hình dự báo hoặc mô hình nowcast được xây dựng sử dụng phương pháp giảm chiều KTPCA lặp luôn cho độ chính xác dự báo bằng hoặc cao hơn độ chính xác dự báo của mô hình được xây dựng sử dụng phương pháp giảm chiều PCA.

Việc xác định chính xác độ dài trễ tối ưu chung cho tất cả các biến trong mô hình là rất quan trọng. Trong quy trình xây dựng mô hình dự báo và mô hình nowcast sử dụng phương pháp KTPCA lặp, độ dài trễ tối ưu chung của các biến trong mô hình được xác định ngay trong Bước 1. Ở các vòng lặp trong Bước 2, luôn sử dụng độ trễ tối ưu chung như vậy. Việc xác định độ trễ tối ưu chung cho các biến trong mô hình dự báo và mô hình nowcast có một số khác biệt, và cũng tiêu tốn nhiều thời gian. Phần thực nghiệm so sánh đánh giá hiệu suất giảm chiều của phương pháp KTPCA lặp sẽ cho thấy điều đó.

Hình 2.1 là cơ sở để phát triển một chương trình máy tính cho phép xây dựng tự động các mô hình nowcast trên tập dữ liệu chuỗi thời gian lớn tần suất hỗn hợp bằng sử dụng phương pháp giảm chiều biến KTPCA lặp. Hình 2.1 có thể được mã hóa trong môi trường của các ngôn ngữ như Python, R hoặc Matlab dựa vào việc sử dụng một số công cụ có sẵn.

2.2 Hiệu suất giảm chiều biến của phương pháp KTPCA lặp

Hiệu suất giảm chiều biến của một phương pháp giảm chiều nào đó được đo bằng RMSE của mô hình nowcast hoặc mô hình dự báo được xây dựng tương ứng dựa vào mô hình DFM hoặc mô hình ARDL nhân tố, trong đó các nhân tố được chiết xuất từ tập dữ liệu lớn của các biến giải thích ở tần suất cao hơn cũng như các biến giải thích có cùng tần suất với biến phụ thuộc bằng sử dụng phương pháp KTPCA lặp. RMSE càng nhỏ, hiệu suất của phương pháp giảm chiều càng cao.

Trong phần này, luận án tập trung thực nghiệm so sánh hiệu suất giảm chiều của phương pháp KTPCA lặp với các phương pháp PCA, SPCA, RSPCA và

ROBSPCA và hiệu suất giảm chiều của phương pháp PCA với họ phương pháp SPCA. Đó là những phương pháp được sử dụng hiệu quả và phổ biến nhất hiện nay khi xây dựng mô hình nowcast/mô hình dự báo trên tập dữ liệu chuỗi thời gian lớn trong lĩnh vực kinh tế - tài chính. Các thực nghiệm so sánh được thực hiện trên các tập dữ liệu thế giới thực, các mô hình dự báo và mô hình nowcast được xây dựng tương ứng dựa vào mô hình ARDL nhân tố và 05 mô hình DFM bao gồm các mô hình BE nhân tố, U-MIDAS nhân tố, STEP-MIDAS nhân tố, PAW-MIDAS nhân tố, và EAW-MIDAS nhân tố. Việc chiết xuất các nhân tố bởi các phương pháp PCA, SPCA, RSPCA, và ROBSPCA được thực hiện bằng sử dụng gói "Sparsepca" [99], trong khi việc chiết xuất chúng bởi phương pháp KTPCA được thực hiện bằng công cụ tự phát triển dựa vào các gói "Kernlab" [100], "Caret" [101] và "Midas-r" [95].

Như đã đề cập ở trên, trên tập dữ liệu của các biến giải thích có cùng tần suất lấy mẫu của biến phụ thuộc, các mô hình BE, U-MIDAS và MIDAS hạn chế trở thành mô hình ARDL theo phương trình (1.33) và các mô hình nowcast trở thành mô hình dự báo được xây dựng trên tập dữ liệu tần suất lấy mẫu giống nhau. Tuy nhiên, do việc lựa chọn trễ tối ưu của các thành phần tần suất cao trong mô hình nowcast và các biến trong mô hình dự báo là khá khác nhau. Vì vậy, việc so sánh hiệu suất giảm chiều của phương pháp KTPCA lặp đã được tiến hành trên các tập dữ liệu của các biến giải thích có tần suất lấy mẫu giống nhau cũng như hỗn hợp.

2.2.1 Đối với các tập dữ liệu tần suất lấy mẫu giống nhau

2.2.1.1 Tập dữ liệu thực nghiệm

Các tập dữ liệu được sử dụng cho thực nghiệm bao gồm 04 tập dữ liệu thực của nền kinh tế Việt Nam và 07 tập dữ liệu trong cơ sở dữ liệu UCI[5]. Chúng được đặt tên là EXP, VN30, CPI, VIP, Residential Building [22], S&P 500, DJI & Nasdaq [23], Air Quality [24], Appliances Energy [25], và SuperConductivity [26]. Các tập dữ liệu EXP, VN30, CPI và S&P 500 đều không chứa thông tin dư thừa hoặc nhiễu.

Ngoài ra, dữ liệu trong các tập EXP và CPI đã được chuyển đổi thành giá trị số tương đối (%) so với tháng cùng kỳ năm trước, trong khi dữ liệu trong tập VN30 và S&P 500 được giữ nguyên ở dạng ban đầu. Tập dữ liệu Residential Building được

[5] Datasets - UCI Machine Learning Repository

giữ nguyên sau khi xóa thuộc tính Zip codes. Tập dữ liệu S&P500, DJI, NASDAQ và Air Quality đều được bổ sung dữ liệu còn thiếu bằng phương pháp trung bình trượt có trọng số. Trọng số phụ thuộc vào từng tập dữ liệu. Tập dữ liệu S&P 500, DJI và NASDAQ bao gồm các quan sát từ ngày 01 tháng 11 năm 2010 đến ngày 26 tháng 10 năm 2017 trong tập dữ liệu ban đầu tương ứng của chúng, trong khi tập dữ liệu Air Quality được thu thập theo mỗi giờ bao gồm các quan sát từ trưa ngày 11 tháng 03 năm 2004, đến trưa ngày 04 tháng 04 năm 2005 trong tập dữ liệu ban đầu.

Bảng 2.2: Các đặc tính thống kê của các tập dữ liệu thực nghiệm

Các tập dữ liệu	Loại tập dữ liệu	Loại thuộc tính	Số quan sát	Số biến	Dữ liệu khuyết thiếu	Biến phụ thuộc	Tần suất
EXP	Chuỗi thời gian	Thực	60	63	No	Kim ngạch xuất khẩu	Tháng
VN30	Chuỗi thời gian	Thực	366	34	No	Chỉ số VN30	Ngày
CPI	Chuỗi thời gian	Thực	72	102	No	Chỉ số CPI	Tháng
VIP	Chuỗi thời gian	Thực	60	265	No	Giá trị sản xuất các ngành	Tháng
Residential Building	Đa biến	Thực	371	27[6]	No	Giá bán	
S&P500	Chuỗi thời gian	Thực	1760	52	Yes	Chỉ số S&P500	Ngày
DJI	Chuỗi thời gian	Thực	1760	81	Yes	Chỉ số Dow Jones	Ngày
NASDAQ	Chuỗi thời gian	Thực	1760	81	Yes	Chỉ số Nasdaq	Ngày
Air Quality	Chuỗi thời gian	Thực	9348	12	Yes	Khí CO	Giờ
Appliances Energy	Chuỗi thời gian	Thực	19704	23	No	Sử dụng năng lượng của thiết bị (wh)	10 phút
SuperConduct.	Đa biến	Thực	21263	81	No	Nhiệt độ tới hạn	

[6] : Loại bỏ cột V1: zip codes

Tập dữ liệu Appliances Energy bao gồm các quan sát từ 17:50 ngày 11 tháng 01 năm 2016 đến 11:50 sáng ngày 27 tháng 05 năm 2016. Tập dữ liệu này được cập nhật 10 phút một lần. Tập dữ liệu cuối cùng, SuperConductivity, được lấy cùng tên với tập dữ liệu huấn luyện.

Bảng 2.2 ở trên cho thấy một số đặc tính thống kê của các tập dữ liệu này. Trong bảng này, số lượng của các thuộc tính (gọi tắt là số thuộc tính) là số lượng các biến giải thích không bao gồm biến phụ thuộc.

Sự biến động của hiện tượng tự nhiên, kinh tế, xã hội đều có quán tính nên trong các mô hình dự báo cũng như mô hình nowcast thường phải bao gồm các biến trễ của biến phụ thuộc và của các biến giải thích. Trong lĩnh vực kinh tế - tài chính, khi số biến giải thích trong các mô hình dự báo hoặc mô hình nowcast có từ 7-16 thì mô hình ấy được gọi là có số biến trung bình, cao hơn thế được gọi là mô hình có số biến lớn [68]. Bảng 2.2 cũng cho thấy trong 11 tập dữ liệu thế giới thực bao gồm các tập dữ liệu có số biến lớn hơn số quan sát (tập dữ liệu EXP, CPI, VIP), hoặc có số biến lớn ngoại trừ tập dữ liệu Air Quality.

2.2.1.2 Phương pháp thực nghiệm

Để so sánh hiệu suất giảm chiều biến của phương pháp KTPCA lặp với các phương pháp PCA, SPCA, RSPCA và ROBSPCA, trên 11 tập dữ liệu thực nghiệm, luận án thống nhất chỉ chọn 06 hàm nhân khác nhau để thực nghiệm với phương pháp KTPCA, trong đó 03 hàm nhân đa thức và 03 hàm nhân Gauss. Cụ thể, các hàm nhân thực nghiệm được chọn như sau: trong 03 hàm nhân đa thức luôn có hàm nhân đa thức đặc biệt $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,1,0)$, khi đó phương pháp KTPCA và PCA là như nhau; đối với tập dữ liệu EXP, VN30, CPI, Air Quality và Appliances Energy, 02 hàm nhân đa thức còn lại có dạng $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,2,0.5)$ và $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,3,0.5)$ trong khi đối với các tập dữ liệu khác, 02 hàm nhân đa thức là $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(0.5,2,0.5)$ và $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(0.5,3,0.5)$. Đối với hàm nhân Gauss có tham số ρ^2, giá trị tham số này của 03 hàm nhân được chọn bằng, nhỏ hơn, và lớn hơn giá trị ρ_0^2, và chúng được ký hiệu là $\boldsymbol{GA_4}, \boldsymbol{GA_5}$, và $\boldsymbol{GA_6}$, tương ứng. Mức độ nhỏ hơn hoặc lớn hơn giá trị ρ_0^2 phụ thuộc vào từng tập dữ liệu thực nghiệm và dựa vào phân tích số lượng các nhân tố được chiết xuất bằng phương pháp KTPCA với tham số

hàm nhân Gauss ρ^2 được chọn xung quanh giá trị ρ_0^2. Mô hình ARDL theo phương trình (1.34) được sử dụng để xây dựng mô hình dự báo trên tập dữ liệu của các biến giải thích có cùng tần suất lấy mẫu.

Các mô hình dự báo được xây dựng dựa vào công thức (1.34) sử dụng phương pháp ước lượng bình phương tuyến tính nhỏ nhất. Tiêu chuẩn lựa chọn số lượng các nhân tố được chiết xuất là tỷ lệ phần trăm giá trị riêng tích lũy của chúng [89]. Ngoại trừ tập dữ liệu EXP có độ trễ tối ưu của tất cả các biến trong mô hình dự báo được xác định dựa vào lý thuyết kinh tế [102] và là 6, trong khi đó, đối với 10 tập dữ liệu còn lại, độ trễ tối ưu của tất cả các biến được xác định chính xác bằng cách sử dụng kết hợp tiêu chuẩn thông tin Akaike (AIC) và tính mùa vụ của các tập dữ liệu chuỗi thời gian [81]. Do đó, độ trễ tối ưu của các nhân tố được chiết xuất bằng cách sử dụng các phương pháp giảm chiều biến khác nhau cho mỗi tập dữ liệu nói chung là khác nhau.

Tất cả các nhân tố đều được kiểm tra tính dừng và được chuyển thành chuỗi thời gian dừng trước khi thực hiện ước lượng mô hình dự báo và trong tất cả các mô hình ước lượng, tất cả các biến đều có ý nghĩa thống kê cao, ít nhất ở mức dưới 10%, Các điều kiện để mô hình ước lượng là tốt nhất, tuyến tính và không chệch (gọi tắt là BLUE) đều được thỏa mãn [51].

2.2.1.3 Kết quả

Khoảng cách trung bình tối thiểu giữa hai véc tơ cột của 11 tập dữ liệu được sử dụng cho việc thực nghiệm được tính theo công thức (2.2) và được trình bày trong Bảng 2.3 cho các tập dữ liệu tương ứng. Giá trị này là gợi ý quan trọng để chọn các hàm nhân Gauss phù hợp $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{GA}(\rho^2)$ khi thực hiện phương pháp KTPCA trên một tập dữ liệu tương ứng nhất định.

Bảng 2.3: Khoảng cách trung bình tối thiểu giữa hai véc tơ cột của các tập dữ liệu

Các tập dữ liệu	*EXP*	*VN30*	*CPI*	*VIP*	*Res. Buil.*	*S&P500*
Khoảng cách trung bình tối thiểu giữa hai véc tơ dữ liệu của các biến giải thích (=ρ_0^2)	$e^{-0.5639}$	$e^{7.046}$	$e^{1.461}$	$e^{34.906}$	$e^{26.919}$	$e^{15.426}$
	DJI	NASDAQ	AirQuality	App. Energy	SuperCond.	
	$e^{15.171}$	$e^{12.971}$	$e^{18.977}$	$e^{13.595}$	$e^{22.353}$	

Với ngưỡng phần trăm giá trị riêng tích lũy là 75% cho tất cả các phương pháp giảm chiều biến nói trên và tất cả các tập dữ liệu thực nghiệm, kết quả của việc giảm chiều biến, RMSE của các mô hình dự báo được xây dựng theo các nhân tố được chiết xuất bởi các phương pháp PCA, SPCA, RSPCA, ROBSPCA như cũng như phương pháp KTPCA với các hàm nhân $\boldsymbol{PL_1}, \boldsymbol{PL_2}, \boldsymbol{PL_3}, \boldsymbol{GA_4}, \boldsymbol{GA_5}$, và $\boldsymbol{GA_6}$ được trình bày trong Bảng A1 trong Phụ lục. Ở đây, $\boldsymbol{PL_1}, \boldsymbol{PL_2}$, và $\boldsymbol{PL_3}$ lần lượt là ký hiệu của các hàm nhân đa thức bậc nhất, bậc hai và bậc ba. Các giá trị tham số của hàm nhân đa thức $\boldsymbol{PL_2}$ và $\boldsymbol{PL_3}$ hơi khác nhau tùy thuộc vào tập dữ liệu thực nghiệm cụ thể như được giới thiệu trong Phần 2.2.1.2. Các giá trị của tham số ρ^2 trong hàm nhân Gauss $\boldsymbol{GA_5}$ và $\boldsymbol{GA_6}$ cho mỗi tập dữ liệu thực nghiệm được trình bày trong Bảng A2 trong Phụ lục.

a. *Hiệu suất giảm chiều của KTPCA lặp so với PCA, SPCA, RSPCA và ROBSPCA*

Được chiết xuất từ Bảng A1 trong Phụ lục, Bảng 2.4 tóm tắt các kết quả giảm chiều biến của các phương pháp KTPCA lặp, PCA, SPCA, RSPCA và ROBSPCA trên 11 tập dữ liệu thực nghiệm của các biến giải thích có cùng tần suất lấy mẫu.

Đối với tập dữ liệu EXP, các biến giải thích trong tập dữ liệu này bao gồm một số chỉ số kinh tế theo tần suất hàng tháng, một số biến tài chính trên thị trường chứng khoán và thị trường tiền tệ trên thế giới và trong nước, giá thế giới của một số sản phẩm đầu vào và đầu ra của nền kinh tế ở tần suất hàng ngày. Tuy nhiên, chúng được tổng hợp với tần suất hàng tháng. Theo nghiên cứu [102], khi xây dựng mô hình dự báo trên tập dữ liệu của các biến kinh tế - tài chính theo tần suất hàng tháng bằng phương pháp hồi quy, độ trễ tối ưu của tất cả các biến trong mô hình nói chung là 6, 12, thậm chí là 24. Bảng 2.4 cho thấy nếu độ trễ tối ưu được xác định theo cách như vậy và bằng 6, thì không thể thực hiện được việc ước lượng mô hình dự báo trên các nhân tố được chiết xuất bởi các phương pháp PCA, SPCA, RSPCA và ROBSPCA. Ví dụ nếu phương pháp giảm chiều biến là PCA thì số lượng nhân tố thành phần chính được chọn là 10. Khi đó, chúng ta không thể hồi quy biến phụ thuộc trên tập dữ liệu gồm 60 quan sát và 76 biến giải thích bao gồm 10 nhân tố được chọn + (10 nhân tố + 01 biến phụ thuộc) được trễ từ 1 đến 6. Tuy nhiên, nếu phương pháp giảm chiều biến là KTPCA thì thách thức trên có thể được giải quyết dễ dàng.

Cũng cần lưu ý rằng khi thực hiện giảm chiều biến của tập dữ liệu bằng phương pháp KTPCA với hàm nhân Gauss, nếu giá trị của tham số ρ^2 nhỏ hơn giá trị ρ_0^2 của tập dữ liệu này, thì số lượng các nhân tố được chọn có xu hướng tăng lên. Ngược lại, số lượng này có xu hướng giảm nếu giá trị của tham số ρ^2 lớn hơn giá trị ρ_0^2 (để biết thêm chi tiết, xem Bảng A1 - Phụ lục). Điều đó tương tự với hàm nhân đa thức, cụ thể, khi bậc hàm nhân đa thức tăng lên thì số nhân tố được chọn theo phương pháp KTPCA cũng có xu hướng giảm. Như vậy, có thể nói rằng phương pháp KTPCA đã khắc phục được những hạn chế của phương pháp PCA và các phương pháp SPCA trong việc giảm chiều biến của các tập dữ liệu lớn, trong đó số lượng quan sát trong tập dữ liệu nhỏ hơn số lượng các biến giải thích hoặc số lượng nhân tố tăng rất nhanh khi tăng tỷ lệ phần trăm giá trị riêng tích lũy.

Bảng 2.4: Hiệu suất giảm chiều của phương pháp KTPCA lặp

Tập dữ liệu	Phương pháp	KTPCA lặp	PCA	SPCA	RSPCA	ROBSPCA
EXP	SL nhân tố	6	14	10	10	10
	RMSE	0.0104	NA	NA	NA	NA
VN30	SL nhân tố	14	14	14	14	15
	RMSE	**0.1819**	0.1895	0.1968	0.1968	0.2054
CPI	SL nhân tố	6	4	4	4	4
	RMSE	**0.4452**	1.4836	1.0659	1.0673	1.0659
VIP	SL nhân tố	4	4	4	4	4
	RMSE	**672.66**	715.96	826.28	1373.57	2642.83
Residential Building	SL nhân tố	2	1	1	1	1
	RMSE	**919.9**	1152.4	1152.5	1152.5	1151.2
S&P500	SL nhân tố	2	1	1	1	1
	RMSE	**61.60**	161.415	161.441	161.441	161.441
DJI	SL nhân tố	1	1	1	1	1
	RMSE	**91.82**	**91.82**	309.24	309.24	309.23
NASDAQ	SL nhân tố	1	1	1	1	1
	RMSE	**81.05**	365.97	85.47	85.47	85.46
Air Quality	SL nhân tố	5	1	1	1	1
	RMSE	**50.297**	71.459	71.499	71.499	71.427
App. Energy	SL nhân tố	6	3	3	3	3
	RMSE	**98.81**	101.74	101.76	101.76	101.75
SuperConductivity	SL nhân tố	2	2	2	2	2
	RMSE	**26.094**	27.314	27.332	27.332	27.319

Trong đó, ký hiệu NA là "No Available" nghĩa là dữ liệu không xác định.

Bảng 2.4 cho thấy rằng trong 8/11 trường hợp của tập dữ liệu thực nghiệm, hàm nhân thích hợp nhất trong số 06 hàm nhân được chọn là hàm nhân Gauss. Trong 03/11 trường hợp còn lại (tương ứng với tập dữ liệu VIP, DJI và NASDAQ), hàm nhân phù hợp nhất là hàm nhân đa thức, trong đó chỉ 1/3 trường hợp (tương ứng với tập dữ liệu DJI), hàm nhân phù hợp nhất là tích vô hướng của hai véc tơ và khi đó phương pháp KTPCA và PCA là như nhau. Trong 8/11 trường hợp hàm nhân phù hợp nhất là hàm nhân Gauss, không có trường hợp nào giá trị của tham số ρ^2 của hàm nhân phù hợp nhất lớn hơn giá trị ρ_0^2 của tập dữ liệu tương ứng, chỉ 1/8 trường hợp trong đó giá trị của tham số ρ^2 bằng giá trị ρ_0^2 (đối với tập dữ liệu VN30). Và như vậy, trong 7/8 trường hợp còn lại, giá trị của tham số ρ^2 nhỏ hơn giá trị ρ_0^2 tương ứng của nó.

Bảng 2.5: Hiệu suất giảm chiều của các phương pháp (RMSE)

Các phương pháp	*DS2*	*DS3*	*DS4*	*DS5*	*DS6*
KTPCA lặp	**0.1819**	**0.4452**	**672.6600**	**919.9000**	**61.6000**
PCA	0.1895	1.4836	715.9608	1152.3950	161.4154
SPCA	0.1968	1.0660	826.2757	1152.5310	161.4407
RSPCA	0.1968	1.0673	1373.5670	1152.5310	161.4407
ROBSPCA	0.2054	1.0659	2642.8340	1151.2470	161.4410
Các phương pháp	*DS7*	*DS8*	*DS9*	*DS10*	*DS11*
KTPCA lặp	**91.8236**	**81.0500**	**50.2970**	**98.8100**	**26.0940**
PCA	**91.8236**	365.9698	71.45873	101.7423	27.3143
SPCA	309.2405	85.4666	71.4989	101.7635	27.3318
RSPCA	309.2405	85.4666	71.4989	101.7635	27.3318
ROBSPCA	309.2349	85.4621	71.4266	101.7468	27.3193

Lưu ý: Ký hiệu DS1 đến DS11 trong Bảng 2.5 tương ứng được gán cho 11 tập dữ liệu thực nghiệm trong Bảng 2.2.

Được trích từ Bảng 2.4, Bảng 2.5 ở trên so sánh hiệu suất giảm chiều biến của phương pháp KTPCA lặp với các phương pháp PCA, SPCA, RSPCA và ROBCPCA. Chúng được thực hiện trong những trường hợp độ trễ của các biến trong mô hình được chọn tối ưu. Tập dữ liệu EXP không được đề cập trong Bảng 2.5 và các mã từ DS1 đến DS11 được gán cho 11 tập dữ liệu thực nghiệm trong Bảng 2.2 kể từ trên

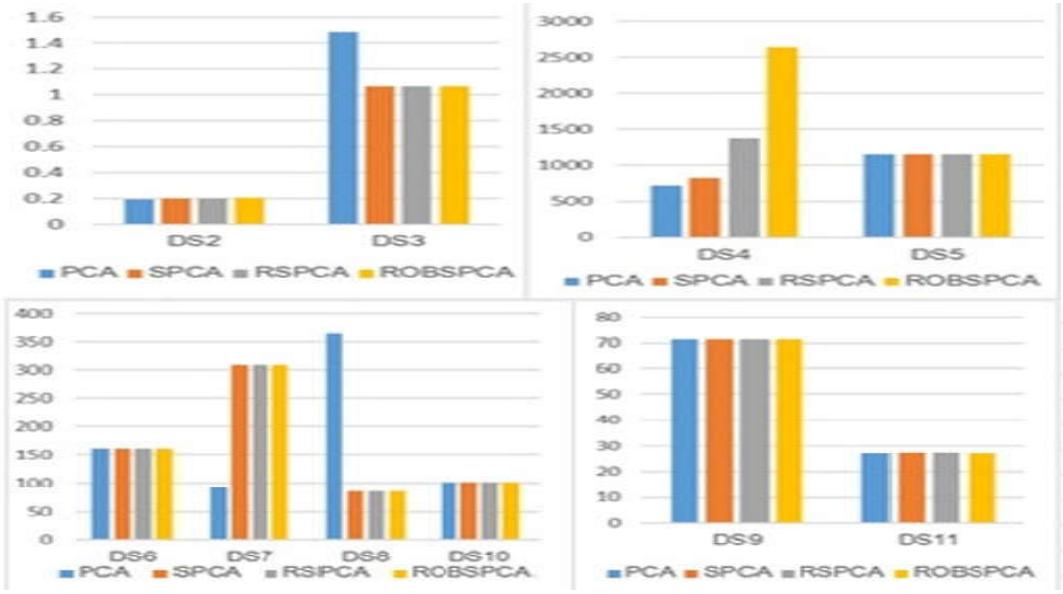
xuống dưới. Bảng 2.4 và Bảng 2.5 cho thấy đối với mỗi tập trong số 10 tập dữ liệu thực nghiệm cuối cùng, luôn có thể tìm được hàm nhân thích hợp để RMSE của mô hình dự báo của biến phụ thuộc theo các nhân tố được chiết xuất bằng phương pháp KTPCA là bằng hoặc nhỏ hơn RMSE của các mô hình dự báo được xây dựng dựa vào các nhân tố được chiết xuất bởi các phương pháp PCA, SPCA, RSPCA và ROBSPCA. Cụ thể, chỉ 1/10 trường hợp hiệu suất của phương pháp KTPCA lặp bằng với hiệu suất của phương pháp PCA, SPCA, RSPCA và ROBSPCA (đối với tập dữ liệu DJI hoặc DS7), 9/10 trường hợp (tức là đối với 9 tập dữ liệu còn lại), hiệu suất của phương pháp KTPCA lặp cao hơn so với các phương pháp nói trên.

Từ phân tích trên, có thể kết luận rằng hiệu suất giảm chiều biến của phương pháp KTPCA lặp là bằng hoặc cao hơn so với các phương pháp PCA và họ SPCA.

b. *Hiệu suất giảm chiều của PCA so với họ SPCA*

Hiệu suất giảm chiều biến của phương pháp PCA được coi là cao hơn so với phương pháp SPCA nếu nó cao hơn hiệu suất giảm chiều của tất cả các phương pháp SPCA, RSPCA và ROBSPCA. Ngược lại, nếu hiệu suất giảm chiều của một trong các phương pháp SPCA, RSPCA, và ROBSPCA cao hơn hiệu suất giảm chiều của phương pháp PCA thì ta nói hiệu suất giảm chiều của họ SPCA là cao hơn phương pháp PCA.

Được tạo ra từ dữ liệu trong Bảng 2.5 (ngoại trừ dữ liệu liên quan đến phương pháp KTPCA lặp), Hình 2.2 cho thấy có 6/10 trường hợp tương ứng với các tập dữ liệu DS2, DS5, DS6, DS9, DS10 và DS11, ở đó hiệu suất giảm chiều biến của các phương pháp PCA, SPCA, RSPCA và ROBSPCA được coi là xấp xỉ bằng nhau; 2/10 trường hợp tương ứng với tập dữ liệu DS4 và DS7, ở đó hiệu suất của phương pháp PCA cao hơn so với họ phương pháp SPCA, và 2/10 trường hợp còn lại tương ứng với tập dữ liệu DS3 và DS8, ở đó hiệu suất của phương pháp PCA kém hơn của họ SPCA. Như vậy, hiệu suất giảm chiều biến của các phương pháp PCA và họ SPCA là cạnh tranh. Kết quả này trái ngược với niềm tin lâu nay rằng hiệu suất giảm chiều của phương pháp SPCA dường như là cao hơn phương pháp PCA.



Hình 2.2: So sánh hiệu suất giảm chiều của PCA và họ SPCA

2.2.2 Đối với các tập dữ liệu tần suất hỗn hợp

Trong phần này, mô hình hồi quy được sử dụng để xây dựng các mô hình nowcast là mô hình BE nhân tố, U-MIDAS nhân tố và một số mô hình MIDAS bị hạn chế khác nhân tố bao gồm các mô hình STEP-MIDAS nhân tố, PAW-MIDAS nhân tố, và EAW-MIDAS nhân tố.

2.2.2.1 Tập dữ liệu thực nghiệm

Các tập dữ liệu được sử dụng để thực nghiệm được thể hiện trong Bảng 2.6. Cụ thể, gồm 07 tập dữ liệu trong cơ sở dữ liệu UCI[7] được giới thiệu trong Bảng 2.2 và 03 tập dữ liệu thực về nền kinh tế Việt Nam, trong đó tập CPI trong Bảng 2.2, tập dữ liệu RGDP và IIP là mới. Tuy nhiên khác với các tập dữ liệu đã có trong Bảng 2.2, trong Bảng 2.6, các biến phụ thuộc trong các tập dữ liệu trong Bảng 2.2 được tổng hợp ở tần suất thấp hơn để chúng là tập dữ liệu lấy mẫu tần suất hỗn hợp. Giá trị của biến phụ thuộc tổng hợp được xác định theo một trong hai cách: cách thứ nhất nó là giá trị trung bình cộng của S giá trị của biến này và cách thứ hai nó là tổng của S giá trị, trong đó S là số lượng giá trị tần suất cao cho mỗi giá trị tần suất thấp. Giá trị S là khác nhau đối với các tập dữ liệu khác nhau. Cụ thể, giá trị của biến phụ thuộc tổng hợp của các tập dữ liệu S&P 500, DJI, Nasdaq, Air Quality, Super Conductivity,

[7] Datasets - UCI Machine Learning Repository

Residential Building và CPI, được tính theo cách thứ nhất, trong khi đối với tập dữ liệu Appliances Energy được tính theo cách thứ hai.

Luận án giả định rằng số ngày làm việc trong các tháng là như nhau và bằng 20 ngày mỗi tháng. Giả định này gần với số ngày làm việc thực tế trong các tháng. Vì vậy, trong Bảng 2.6, khi S bằng 20, chúng ta hiểu rằng biến phụ thuộc ở tần suất hàng tháng, trong khi các biến giải thích ở tần suất hàng ngày.

Bảng 2.6: Các đặc tính thống kê của các tập dữ liệu thực nghiệm

Các đặc điểm thống kê	RGDP	CPI	IIP	Air Quality	App. Energy
Đặc điểm của tập dữ liệu	Chuỗi thời gian	Chuỗi thời gian	Chuỗi thời gian	Chuỗi thời gian	Chuỗi thời gian
Loại thuộc tính	Thực	Thực	Thực	Thực	Thực
Số biến tần suất thấp	3	3	1	1	1
Số biến tần suất cao	87	102	42	12	27
Tổng số quan sát	72	72	1840	9348	19704
Số quan sát tần suất thấp	24	24	92	779	3284
S - số lượng giá trị tần suất cao cho một giá trị tần suất thấp [8]	3	3	20	12	6
Dữ liệu khuyết thiếu	No	No	Yes	Yes	No
Biến phụ thuộc	Tốc độ tăng trưởng GDP	Lạm phát giá tiêu dùng	Chỉ số sản xuất công nghiệp	Khí CO	Sử dụng năng lượng của thiết bị
Các đặc điểm thống kê	Res. Build.	S&P 500	DJI	NASDAQ	SuperCond.
Đặc điểm của tập dữ liệu	Dữ liệu chéo	Chuỗi thời gian	Chuỗi thời gian	Chuỗi thời gian	Dữ liệu chéo
Loại thuộc tính	Thực	Thực	Thực	Thực	Thực
Số biến tần suất thấp	1	1	1	1	1
Số biến tần suất cao	27	52	81	81	81
Tổng số quan sát	366	1760	1760	1760	21260
Số quan sát tần suất thấp	122	88	88	88	1063
S - số lượng giá trị tần suất cao cho một giá trị tần suất thấp	3	20	20	20	20
Dữ liệu khuyết thiếu	No	Yes	Yes	Yes	No
Biến phụ thuộc	Giá bán	Chỉ số S&P500	Chỉ số DJI	Chỉ số NASDAQ	Nhiệt độ tới hạn

[8] : Tổng số quan sát (hay số quan sát tần suất cao) = S * số quan sát tần suất thấp.

Trong tập dữ liệu IIP, biến phụ thuộc với tần suất hàng tháng là chỉ số sản xuất công nghiệp (viết tắt là IIP), trong khi tất cả các biến giải thích khác được lấy theo tần suất ngày làm việc (5 ngày mỗi tuần). Tập dữ liệu này bao gồm nhiều dữ liệu về tài chính quốc tế (đặc biệt là thị trường chứng khoán), giá thế giới của một số sản phẩm thiết yếu của nền kinh tế Việt Nam. Dữ liệu quốc tế có thể bị thiếu ở một số ngày làm việc. Dữ liệu bị khuyết thiếu được xử lý theo cách tương tự như đã được đề cập ở mục 2.2.1.1.

Ngoài ra, trong tập dữ liệu RGDP và CPI, có một số biến giải thích khác có cùng tần suất quý với biến phụ thuộc. Các biến còn lại đều ở tần suất hàng tháng. Dữ liệu trong các tập này đã được chuyển đổi và chúng đều là số tương đối (%) thể hiện sự thay đổi của các biến so với cùng kỳ năm trước. Do đó, trong 10 tập dữ liệu thực nghiệm, chỉ có 02 tập dữ liệu đầu tiên có một số biến giải thích ở cùng tần suất thấp như biến phụ thuộc, trong khi 08 tập dữ liệu còn lại, tất cả các biến giải thích đều ở tần suất cao hơn tần suất của biến phụ thuộc.

2.2.2.2 Phương pháp thực nghiệm

Để xây dựng các mô hình nowcast, trước tiên, biến phụ thuộc ở tần suất thấp, các biến giải thích ở cùng tần suất với biến phụ thuộc và các nhân tố được chiết xuất từ các biến giải thích tần suất cao hơn được chuyển thành chuỗi thời gian dừng. Tiêu chuẩn để lựa chọn số lượng các nhân tố ở tần suất cao cũng là tỷ lệ phần trăm giá trị riêng tích lũy của chúng [89]. Các mô hình nowcast đều được ước lượng trong điều kiện lý tưởng, đó là độ trễ của các biến giải thích tần suất cao được xác định chính xác. Cụ thể là:

- Đối với các mô hình nowcast được xây dựng dựa vào mô hình BE nhân tố (1.37), tất cả các nhân tố tần suất cao được tổng hợp thành các nhân tố tần suất thấp giống như biến phụ thuộc. Trong 09 tập dữ liệu đầu tiên, giá trị của các nhân tố ở tần suất thấp là giá trị trung bình của S giá trị của các nhân tố ở tần suất cao hơn, trong khi ở tập dữ liệu cuối cùng, giá trị của các nhân tố tần suất thấp là tổng S giá trị của các nhân tố ở tần suất cao hơn. Độ trễ đơn tối ưu của biến phụ thuộc và các biến giải thích được xác định bằng tiêu chuẩn AIC, ở đây, độ trễ đơn là độ trễ được áp dụng cho tất cả các biến ở cùng tần suất. Ngoài ra, tất cả các biến trong các mô hình nowcast đều có ý nghĩa thống kê, ít nhất ở mức < 10%.

- Đối với các mô hình nowcast được xây dựng dựa vào mô hình MIDAS nhân tố, độ trễ đơn tối ưu của biến phụ thuộc và các biến giải thích ở cùng tần suất với biến phụ thuộc được xác định như trong mô hình dự báo được xây dựng dựa vào mô hình BE nhân tố. Ngược lại, độ trễ tối ưu của mỗi nhân tố ở tần suất cao hơn thường khác nhau và được xác định theo một trong hai cách tiếp cận sau:

(i) Cách tiếp cận thứ nhất: Xác định độ trễ chung (được gọi độ trễ đơn) tối ưu cho tất cả các nhân tố và độ trễ riêng tối ưu của từng nhân tố tần suất cao dựa vào RMSE. Ở đây, độ trễ riêng tối ưu của mỗi nhân tố tần suất cao không vượt quá độ trễ đơn tối ưu.

(ii) Cách tiếp cận thứ hai: Độ trễ đơn tối đa do người sử dụng xác định (được gọi là độ trễ cố định), độ trễ riêng tối ưu của từng nhân tố tần suất cao được xác định chính xác dựa vào RMSE và không được vượt quá độ trễ cố định này.

Cách tiếp cận đầu tiên phù hợp với các mô hình STEP-MIDAS (1.37) và PAW-MIDAS (1.39) và số lượng các nhân tố tần suất cao trong các mô hình nowcast không quá lớn, thường không vượt quá 10. Cách tiếp cận này chỉ được ứng dụng cho các mô hình EAW-MIDAS (1.41) và U-MIDAS (1.36) khi số lượng các nhân tố trong các mô hình nowcast là khá nhỏ. Nói cách khác đối với các mô hình EAW-MIDAS (1.41) và U-MIDAS (1.36), độ trễ riêng tối ưu của mỗi nhân tố tần suất cao nói chung được xác định theo cách tiếp cận thứ hai.

Việc so sánh hiệu suất giảm chiều biến của phương pháp KTPCA LẶP và các phương pháp PCA, SPCA, RSPCA, và ROBSPCA cũng được thực hiện trên 06 hàm nhân đã được đề cập trong Phần 2.2.1.2.

2.2.2.3 Kết quả

Ngoại trừ 02 tập dữ liệu RGDP và IIP, 8 tập dữ liệu còn lại trong Bảng 2.6 được lấy từ các tập dữ liệu tương ứng cùng tên trong Bảng 2.2. Hơn nữa, số lượng biến giải thích tần suất cao và số quan sát trên 8 tập dữ liệu này là không thay đổi so với các tập dữ liệu tương ứng được trình bày trong Bảng 2.2. Do đó, khoảng cách trung bình tối thiểu giữa 2 véc tơ cột trên 8 tập này được xác định như trong Bảng 2.3. Khoảng cách này trong hai tập dữ liệu RGDP và IIP tương ứng là $\rho_0^2 = \exp(1.464)$ và $\rho_0^2 = \exp(8.978)$.

Với cùng ngưỡng tỷ lệ phần trăm giá trị riêng tích lũy là 75% cho tất cả các phương pháp giảm chiều biến được đề cập ở trên, cho tất cả các tập dữ liệu thực nghiệm và 05 mô hình hồi quy: BE, PAW-MIDAS, STEP-MIDAS, U-MIDAS và EAW-MIDAS, kết quả giảm chiều biến, RMSE của các mô hình dự báo theo các nhân tố được chiết xuất bởi các phương pháp giảm chiều biến và các hàm nhân thích hợp nhất trong số 06 hàm nhân được thực nghiệm được trình bày trong Bảng B (phần Phụ lục), ở đó những ký hiệu SET1 đến SET10 tương ứng với mười tập dữ liệu thực nghiệm trong Bảng 2.6 kể từ trên xuống dưới. Cột cuối cùng trong Bảng B chỉ ra hàm nhân thích hợp nhất trong số 06 hàm nhân được thực nghiệm cho phương pháp KTPCA theo từng mô hình hồi quy nhân tố, số lượng các nhân tố được chọn, và phần trăm giá trị riêng tích lũy của chúng.

Bảng B cũng chỉ ra rằng đối với tập dữ liệu SET3, độ trễ riêng tối ưu của các nhân tố được chiết xuất từ tập dữ liệu này trong các mô hình nowcast được xác định theo cách tiếp cận thứ hai vì số lượng các nhân tố được chọn khi đó là khá lớn. Khi các mô hình nowcast được xây dựng dựa vào mô hình U-MIDAS trên các tập dữ liệu từ SET6 đến SET10, để đảm tỷ lệ phần trăm giá trị riêng tích lũy của các nhân tố đã chọn lớn hơn ngưỡng được xác định trước thì chỉ cần chọn một hoặc hai nhân tố. Vì vậy, độ trễ riêng tối ưu của các nhân tố tần suất cao được chọn theo cách tiếp cận đầu. Với các mô hình nowcast được xây dựng dựa vào mô hình EAW-MIDAS trên các tập dữ liệu từ SET6 đến SET10, độ trễ cố định của các nhân tố tần suất cao được xác định bằng độ trễ đơn tối ưu trong các mô hình nowcast được xây dựng dựa vào mô hình U-MIDAS.

Cột cuối cùng trong Bảng B cho thấy các hàm nhân cụ thể thích hợp nhất trong 06 hàm nhân được thực nghiệm để giảm số lượng biến cho mỗi tập dữ liệu cụ thể và theo một phương pháp hồi quy cụ thể.

a. Hiệu suất giảm chiều của KTPCA lặp so với PCA, SPCA, RSPCA, và ROBSPCA

Bảng 2.7 dưới đây được rút ra từ Bảng B trong phần Phụ lục. Bảng này bao gồm năm bảng phụ 3a, 3b, 3c, 3d và 3e chứa RMSE của các mô hình nowcast được xây dựng dựa vào các mô hình BE nhân tố, các mô hình U-MIDAS, STEP-MIDAS,

PAW-MIDAS, và EAW-MIDAS nhân tố. Ở đây, các nhân tố được chiết xuất từ các tập dữ liệu thực nghiệm nói trên bằng phương pháp PCA, SPCA, RSPCA, ROBSPCA, và KTPCA lặp.

Bảng 2.7 cũng cho thấy đối với tất cả 10 tập dữ liệu thực nghiệm và 05 loại mô hình hồi quy nhân tố động vừa nêu, hiệu suất giảm chiều biến bằng sử dụng phương pháp KTPCA lặp luôn cao nhất. Cụ thể, đối với tất cả 05 mô hình hồi quy, luôn có thể chọn được một hàm nhân sao cho RMSE của mô hình nowcast được xây dựng trên các nhân tố được chiết xuất bằng phương pháp KTPCA tương ứng với hàm nhân này nhỏ hơn hoặc bằng RMSE của các mô hình nowcast được xây dựng trên các nhân tố được chiết xuất bằng một trong các phương pháp PCA, SPCA, RSPCA, và ROBSPCA. Ví dụ: trong trường hợp mô hình nowcast được xây dựng dựa vào mô hình hồi quy STEP-MIDAS nhân tố (xem Bảng 3b), hàm nhân phù hợp nhất trong 06 hàm nhân được thực nghiệm sử dụng phương pháp giảm chiều KTPCA trên các tập dữ liệu SET1, SET4 và SET5 là tích vô hướng của hai véc tơ (khi đó phương pháp KTPCA và PCA là như nhau). Đồng thời, các hàm nhân phù hợp nhất cho phương pháp giảm chiều này trên tập dữ liệu SET6 và 06 tập dữ liệu còn lại lần lượt là hàm nhân đa thức bậc 2 và hàm nhân Gauss (xem phần Phụ lục). Kết quả tương tự khi các mô hình nowcast được xây dựng dựa vào 04 loại hồi quy khác ở trên.

Bảng 2.7: Hiệu suất giảm chiều biến của các phương pháp được đề xuất

3a.BE	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA lặp**
SET1	**0.00049**	0.00079	0.00079	0.00079	**0.00049**
SET2	**0.00018**	0.00049	0.00051	0.00049	**0.00018**
SET3	1.34898	1.20384	1.04437	1.54530	**0.56932**
SET4	0.61523	0.61105	0.61040	0.61106	**0.59286**
SET5	377.62520	377.26180	377.26200	377.06180	**360.13100**
SET6	565.51470	565.52300	565.52300	565.51600	**513.61890**
SET7	**4.30740**	4.30760	4.30760	4.30760	**4.30740**
SET8	57.10330	56.43210	56.43210	56.43210	**56.29750**
SET9	18.59450	18.59410	18.59410	18.54890	**18.34790**
SET10	13.53810	13.53970	13.54250	13.54290	**13.36620**

3b.STEP	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA lặp**
SET1	**0.00744**	0.00973	0.00972	0.00973	**0.00744**
SET2	0.00824	0.00439	0.00439	0.00439	**0.00395**
SET3	26.52232	21.39361	28.86856	28.13315	**8.75305**
SET4	**0.63003**	0.63004	0.63004	**0.63004**	**0.63004**
SET5	**385.19720**	385.68000	385.68000	385.34540	**385.19720**
SET6	430.84120	430.83730	430.83730	430.83970	**421.70900**
SET7	259.88440	259.80830	257.66440	259.80650	**72.78710**
SET8	4101.59300	4101.95800	4101.95800	4102.27500	**1024.70800**
SET9	1419.76700	1419.80700	1419.80700	1419.75600	**687.29870**
SET10	14.34250	14.34620	14.34620	14.34310	**13.96490**
3c.PAW	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA lặp**
SET1	**0.00003**	0.00021	0.00020	0.00021	**0.00003**
SET2	**0.00147**	0.00183	0.00182	0.00183	**0.00147**
SET3	1.12680	0.73420	0.75080	0.62080	**0.04330**
SET4	0.62980	0.62930	0.64020	0.62930	**0.61740**
SET5	384.40070	384.41150	384.32180	384.32700	**384.01710**
SET6	404.33890	399.47980	399.47980	399.48000	**399.34980**
SET7	40.70190	42.84440	42.84440	42.84440	**33.61590**
SET8	337.80480	337.80250	337.80250	337.80260	**311.39130**
SET9	107.96670	107.96660	107.96660	107.90060	**107.03020**
SET10	13.95800	13.95800	13.95800	13.95800	**13.94850**
3d.EAW	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA lặp**
SET1	0.00523	0.00527	0.00528	0.00527	**0.00454**
SET2	0.00691	0.00547	0.00742	0.00547	**0.00509**
SET3	4.49830	4.71740	4.35610	4.31460	**4.18100**
SET4	0.47620	0.47650	0.47650	0.47610	**0.43920**
SET5	385.45490	385.45150	385.45150	385.45970	**385.00000**
SET6	504.90740	504.90760	504.90760	504.90690	**379.01570**
SET7	**2.80600**	2933.00000	2.95300	2.95300	**2.50600**
SET8	240.00000	239.70000	239.70000	239.50000	**118.90000**

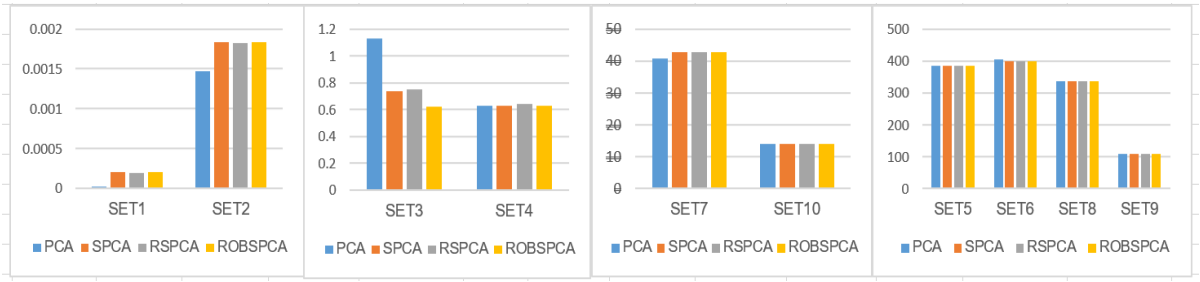
SET9	82.22790	82.12540	82.12540	82.03570	**36.36560**
SET10	13.93220	13.93100	13.93100	13.93220	**13.93020**
3e.U	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA lặp**
SET1	0.00204	0.00095	0.00092	0.00095	**0.00070**
SET2	**0.00011**	0.00252	0.00296	0.00251	**0.00011**
SET3	**0.02830**	0.98600	0.31090	0.66320	**0.02830**
SET4	0.40540	0.40580	0.40580	0.40550	**0.33300**
SET5	376.98510	377.40160	377.40160	376.80080	**351.20000**
SET6	430.11820	430.17320	430.17320	430.12860	**359.12290**
SET7	0.00070	0.00006	0.00006	0.00006	**0.00005**
SET8	2.93200	2.93100	2.93100	2.93100	**2.93000**
SET9	0.89930	0.89920	0.89920	0.89200	**0.80410**
SET10	14.02310	14.02190	14.02190	14.02310	**13.91150**

Lưu ý: Ký hiệu SET1 đến SET10 ở Bảng 2.7 tương ứng với mười tập dữ liệu thực nghiệm trong Bảng 2.6

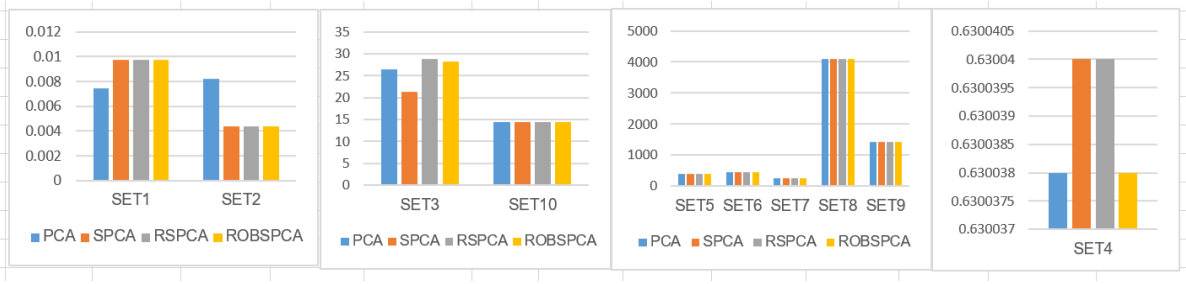
Đối với tập dữ liệu SET3, nếu sử dụng các phương pháp giảm chiều PCA và họ SPCA thì cần phải chọn ít nhất 12 nhân tố để có tỷ lệ phần trăm giá trị riêng tích lũy tối thiểu là 75%, trong khi đối với 09 tập dữ liệu còn lại, số nhân tố được chọn ít hơn nhiều. Điều này tiết lộ rằng mức độ xấp xỉ một siêu phẳng của tập dữ liệu SET3 thấp hơn so với 9 tập dữ liệu còn lại. Trong bối cảnh đó, cần sử dụng phương pháp KTPCA để giảm chiều biến. Khi đó có thể giảm đáng kể số lượng các nhân tố nhưng vẫn đảm bảo tỷ lệ phần trăm giá trị riêng tích lũy vượt quá ngưỡng cho trước bằng cách chọn hàm nhân phù hợp nhất cho mục đích này.

b. Hiệu suất giảm chiều của PCA và họ SPCA

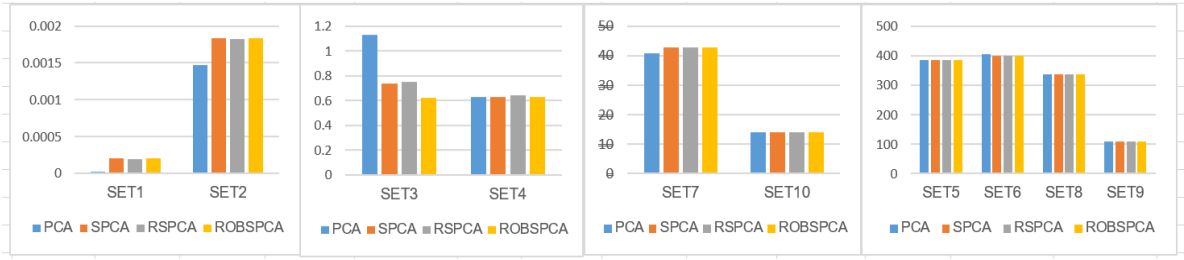
Các hình 2.3, 2.4, 2.5, 2.6, và 2.7 dưới đây được vẽ từ các bảng con 3a, 3b, 3c, 3d và 3e tương ứng trong Bảng 2.7 ở trên. Các hình này thể hiện kết quả so sánh hiệu suất giảm chiều biến của phương pháp PCA và họ SPCA trên 10 tập dữ liệu thực nghiệm nói trên và theo các mô hình hồi quy nhân tố là mô hình BE, mô hình MIDAS với hàm trọng số STEP bậc ba (STEP3-MIDAS), mô hình MIDAS với hàm trọng số ALMON đa thức bậc hai (PAW2-MIDAS), mô hình EAW-MIDAS và U-MIDAS.



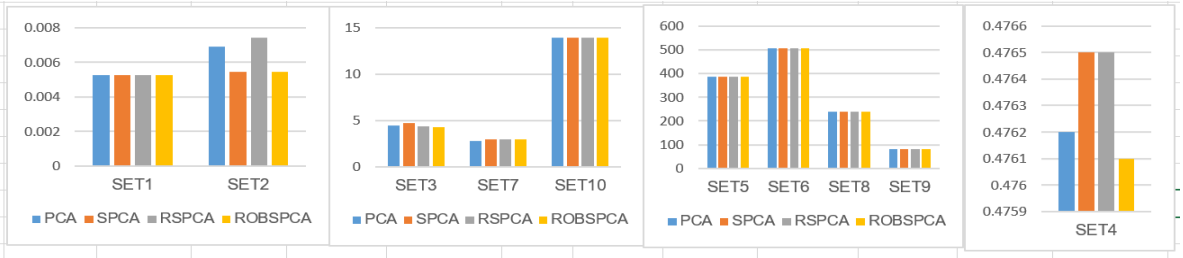
Hình 2.3: Hiệu suất giảm chiều dựa vào mô hình BE



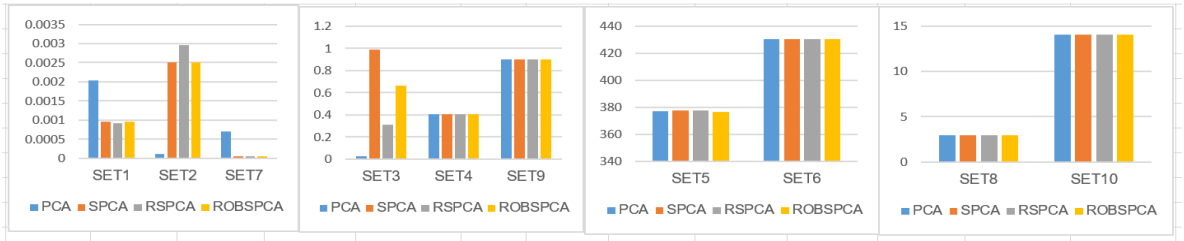
Hình 2.4: Hiệu suất giảm chiều dựa vào mô hình STEP3-MIDAS



Hình 2.5: Hiệu suất giảm chiều dựa vào mô hình PAW2-MIDAS



Hình 2.6: Hiệu suất giảm chiều dựa vào mô hình EAW-MIDAS



Hình 2.7: Hiệu suất giảm chiều dựa vào mô hình U-MIDAS

Hình 2.7 cho thấy hiệu suất giảm chiều biến của phương pháp PCA xấp xỉ bằng, cao hơn và thấp hơn hiệu suất giảm chiều của họ SPCA tương ứng trên 6/10 tập dữ liệu (bao gồm SET4, SET5, SET6, SET8, SET9, và SET10), 2/10 tập dữ liệu (SET2 và SET3), và 2/10 tập dữ liệu (SET1 và SET7).

Bảng 2.8: Hiệu suất giảm chiều của PCA so với họ SPCA

Mô hình DFM	*Bằng*	*Cao hơn*	*Thấp hơn*
BE	SET4, SET5, SET6, SET8, SET9, SET10	SET1, SET2, SET3	SET7
STEP3-MIDAS	SET5, SET6, SET7, SET8, SET9, SET10	SET1, SET4	SET2, SET3
PAW2-MIDAS	SET4, SET5, SET6, SET8, SET9, SET10	SET1, SET2, SET7	SET3
EAW-MIDAS	SET1, SET5, SET6, SET8, SET9, SET10	SET3, SET4, SET7	SET2
U-MIDAS	SET4, SET5, SET6, SET8, SET9, SET10	SET2, SET3	SET1, SET7

Bảng 2.8 trình bày tóm tắt kết quả so sánh hiệu suất giảm chiều biến của phương pháp PCA và họ SPCA theo các mô hình hồi quy nhân tố đã đề xuất. Ba cột cuối cùng của bảng này hiển thị tập dữ liệu thực nghiệm trong đó hiệu suất của phương pháp PCA xấp xỉ bằng, cao hơn và thấp hơn hiệu suất giảm chiều của họ phương pháp SPCA. Như vậy hiệu suất giảm chiều biến của các phương pháp SPCA không cao hơn phương pháp PCA. Hiệu suất giảm chiều của các phương pháp này là cạnh tranh.

2.3 Kết Luận Chương 2

- Chương này trình bày phương pháp giảm chiều dựa vào kỹ thuật hàm nhân (gọi tắt KTPCA). Sự khác biệt của phương pháp này so với các phương pháp KPCA và PCA đã được chú ý làm rõ. Phương pháp KTPCA trở thành phương pháp PCA khi hàm nhân là tích vô hướng của hai véc tơ nên nó là mở rộng tự nhiên của phương PCA. Phương pháp KTPCA đã khắc phục được hạn chế của phương pháp PCA là có thể giảm chiều các tập dữ liệu không xấp xỉ một siêu phẳng. Hiệu suất giảm chiều của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất là bằng hoặc cao hơn

so với các phương pháp PCA, SPCA, RSPCA, và ROBSPCA trên các tập dữ liệu tần suất lấy mẫu giống nhau cũng như hỗn hợp, trong đó các nhân tố được sử dụng để xây dựng các mô hình nowcast/dự báo được chiết xuất bằng các phương pháp KTPCA, PCA và họ SPCA.

- Chương này đã không so sánh hiệu suất giảm chiều của phương pháp KTPCA với phương pháp KPCA vì các nghiên cứu [19] và [20] đã thực hiện so sánh hiệu suất giảm chiều của phương pháp PCA với phương pháp KPCA và đều cho kết luận trên các tập dữ liệu thế giới thực, hiệu suất giảm chiều của phương pháp PCA là cao hơn phương pháp KPCA.

- Chương này cũng cho thấy hiệu suất giảm chiều đối với cả hai loại tập dữ liệu có tần suất lấy mẫu giống nhau và hỗn hợp của phương pháp PCA và họ SPCA là cạnh tranh. Điều này là khác với niềm tin đã tồn tại lâu nay là họ phương pháp SPCA có hiệu suất giảm chiều nổi trội hơn phương pháp PCA.

Kết quả nghiên cứu của chương này được công bố trên Nghiên cứu [CT3], [CT6] phần danh mục Nghiên cứu của tác giả.

Chương 3 tiếp theo sẽ trình bày chi tiết đề xuất thuật toán dự báo không và có điều kiện sử dụng phương pháp giảm chiều được đề xuất trong Chương này.

CHƯƠNG 3. DỰ BÁO TRÊN TẬP DỮ LIỆU CHUỖI THỜI GIAN LỚN SỬ DỤNG PHƯƠNG PHÁP GIẢM CHIỀU DỰA VÀO KỸ THUẬT HÀM NHÂN

Chương 3 trình bày thuật toán dự báo không và có điều kiện trên tập dữ liệu lớn sử dụng phương pháp giảm chiều KTPCA lặp được đề xuất ở Chương 2. Các mô hình dự báo được xây dựng dựa vào mô hình ARDL nhân tố theo phương trình (1.34) đối với mô hình dự báo có điều kiện và theo phương trình (1.16) đối với mô hình dự báo không điều kiện, trong đó các nhân tố được chiết xuất bằng phương pháp KTPCA lặp. Việc mô hình hóa dự báo kim ngạch xuất khẩu của Việt Nam theo tần suất tháng sử dụng thuật toán được đề xuất cũng được trình bày trong Chương này.

3.1 Quy trình dự báo không và có điều kiện sử dụng phương pháp KTPCA lặp

Quy trình dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều KTPCA lặp được phát triển dựa vào quy trình mô hình hóa dự báo kinh tế - tài chính được trình bày trong mục 1.3.6 Chương 1 có tính đến phương pháp giảm chiều này.

Hình 2.1 ở Chương 2 cho thấy việc thực hiện giảm chiều bằng phương pháp KTPCA lặp và xây dựng mô hình dự báo được kết hợp trong một. Cụ thể sau khi kết thúc việc giảm chiều bằng phương pháp KTPCA lặp thì cũng nhận được mô hình dự báo có RMSE tốt nhất. Điều này gợi ý rằng quá trình dự báo chuỗi thời gian sử dụng phương pháp KTPCA lặp có một số khác biệt so với quá trình mô hình hóa dự báo chuỗi thời gian. Sự khác biệt ấy chủ yếu thuộc về 03 pha cuối cùng của quy trình mô hình hóa dự báo chuỗi thời gian.

Mục này sẽ trình bày quy trình dự báo trên tập dữ liệu lớn sử dụng phương pháp giảm chiều KTPCA lặp. Quy trình này được đề xuất sao cho nó có thể là cơ sở để phát triển thành chương trình tin học cho phép dự báo (không và có điều kiện) tự động cho biến phụ thuộc trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất.

Cách tiếp cận dự báo có điều kiện thường được sử dụng khi người làm dự báo nhận thấy có thể có những yếu tố bất thường như thiên tai, dịch họa, biến động địa

chính trị trong nước và thế giới, và nhất là chính sách điều hành nền kinh tế của chính phủ thay đổi có tác động mạnh đến sự thay đổi của các biến trong mô hình. Phương pháp dự báo có điều kiện đã được trình bày tóm tắt trong mục 1.3.2.1 Chương 1. Khác với dự báo có điều kiện, cách tiếp cận dự báo không điều kiện thường được sử dụng để dự báo ngắn hạn biến phụ thuộc khi các nhà dự báo cảm nhận rằng các chỉ số dẫn báo của biến phụ thuộc không có những thay đổi bất thường trong ngắn hạn.

Quy trình thực hiện dự báo không và có điều kiện của biến phụ thuộc Y trên tập dữ liệu lớn của các biến giải thích $\boldsymbol{X}$ sử dụng phương pháp giảm chiều biến KTPCA lặp được mô tả trong Hình 3.1 bên dưới, ở đó $Y, \boldsymbol{X}$ có cùng tần suất lấy mẫu.

Hình 3.1 bao gồm hai hình 3.1a và 3.1b, tương ứng mô tả quy trình dự báo có điều kiện và không điều kiện trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều KTPCA lặp. Cả hai quy trình này có thể được chia thành bốn giai đoạn. Nội dung chính cần thực hiện ở các giai đoạn cơ bản là giống nhau, song vẫn có một số khác biệt. Cụ thể, nội dung chính của các giai đoạn trong hai quy trình dự báo đó được trình bày tóm tắt như sau:

Giai đoạn 1*: Xử lý dữ liệu*

Trong Giai đoạn này, cả hai quy trình đều thực hiện loại bỏ các giá trị ngoại lai, bổ sung giá trị bị thiếu, sau đó chuyển đổi dữ liệu về cùng một dạng có thể so sánh được. Việc khắc phục dữ liệu bị thiếu (missing data) phụ thuộc vào vị trí bị thiếu:

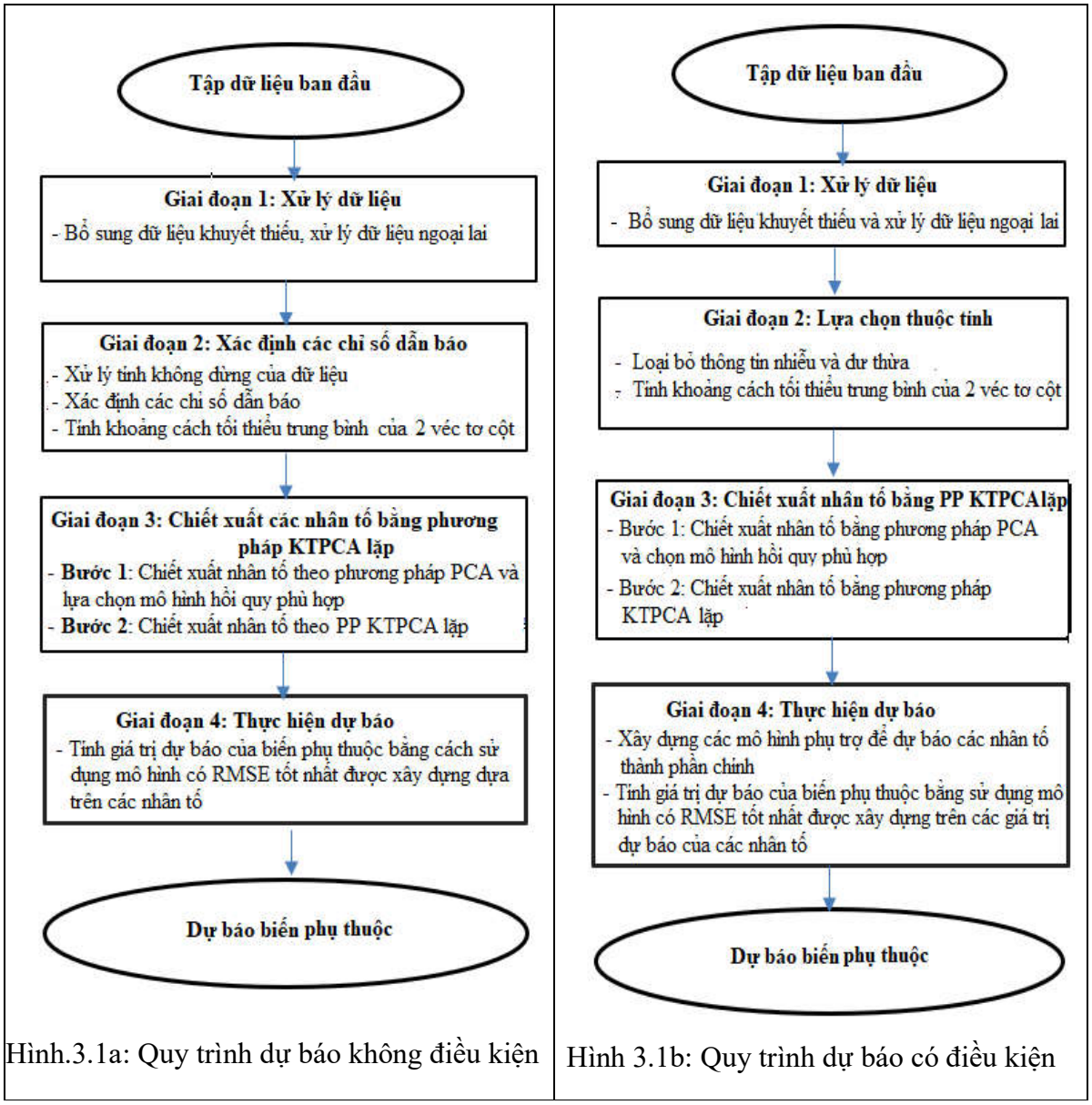
- Trường hợp các dữ liệu bị thiếu xảy ra phía đầu của các quan sát và/hoặc ở phía cuối các quan sát, ta sử dụng phương pháp ngoại suy bằng sử dụng mô hình AR(p) có xu thế hoặc xây dựng mô hình dự báo biến đó theo một số biến khác mà theo lý thuyết kinh tế chúng có quan hệ chặt chẽ với nhau.

- Ngược lại, nếu giá trị khuyết thiếu không nằm ở giữa quan sát đầu tiên và cuối cùng, ta có thể sử dụng phương pháp nội suy chẳng hạn như làm trơn hàm mũ [103] hoặc mô hình trung bình trượt phụ thuộc vào tỷ lệ các giá trị bị thiếu so với giá trị hiện có là ít hay nhiều.

Giai đoạn 2*: Lựa chọn thuộc tính*

Nội dung chính của Giai đoạn này bao gồm:

- Đối với quy trình dự báo không điều kiện: Đó là xác định các chỉ số dẫn báo của biến phụ thuộc theo mô hình (1.15) hoặc (1.16) trên tập các biến giải thích. Để tránh hồi quy giả mạo theo mô hình (1.15) hoặc (1.16), tất cả các biến đều phải được đưa về *chuỗi dừng* trước khi thực hiện việc xác định các chỉ số dẫn báo của biến phụ thuộc.



Hình.3.1a: Quy trình dự báo không điều kiện

Hình 3.1b: Quy trình dự báo có điều kiện

Hình 3.1: Quy trình dự báo không và có điều kiện

- Đối với quy trình dự báo có điều kiện: Đó là lựa chọn những biến có giá trị đối với mục đích dự báo của biến phụ thuộc bằng cách loại bỏ những biến không hoặc

ít liên quan hoặc dư thừa với mục đích dự báo của biến phụ thuộc bằng sử dụng độ đo hệ số tương quan Pearson theo công thức (1.17). Nội dung này cũng có thể được xem là xử lý thông tin nhiễu (dữ liệu ngoại lai) và dư thừa trong tập dữ liệu ban đầu của các biến giải thích. Cần lưu ý là, ở Giai đoạn 2 trong quy trình dự báo có điều kiện, các biến giải thích không được đưa về chuỗi dừng vì nếu làm như vậy có thể dẫn đến làm mất thông tin và ảnh hưởng đến việc xác định chính xác các biến không dư thừa và có tương quan cao với biến phụ thuộc bằng sử dụng hệ số tương quan Pearson.

Như vậy về bản chất mục đích của Giai đoạn 2 ở cả hai quy trình dự báo không và có điều kiện là như nhau đó là lựa chọn những biến "đắt giá" để đưa vào các mô hình dự báo, nhưng tính chất đòi hỏi các biến đó là khác nhau trong mỗi loại dự báo. Trong dự báo không điều kiện thì các biến được chọn phải có tính chất dẫn báo trong khi trong dự báo có điều kiện thì chỉ cần các biến có tương quan cao và không dư thừa với mục đích dự báo biến phụ thuộc. Khác với tính tương quan, tính dẫn báo không có tính chất bắc cầu nên tính dư thừa không được đặt ra đối với dự báo không điều kiện.

- Kết thúc Giai đoạn này, nếu tập các chỉ số dẫn báo hoặc biến giải thích ban đầu được chọn còn lớn thì cần tính khoảng cách trung bình tối thiểu của 2 véc tơ dữ liệu trong tập dữ liệu của các biến được chọn theo công thức (2.2) và chuyển sang Giai đoạn 3 tiếp theo. Không mất tính tổng quát, ta ký hiệu tập dữ liệu mới là $\boldsymbol{Z}$.

***Giai đoạn 3**: Chiết xuất nhân tố bằng phương pháp KTPCA lặp*

Nội dung chính của Giai đoạn 3 trong cả hai quy trình dự báo đều giống nhau, đó là thực hiện phương pháp giảm chiều biến bằng phương pháp KTPCA lặp. Giai đoạn này đã được trình bày chi tiết trong mục 2.2.3 Chương 2. Đây là quy trình lặp theo các hàm nhân của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất. Kết quả thực nghiệm ở Chương 2 cho thấy, quá trình này cũng có thể được xem là quá trình heuristic, vì từ kết quả giảm chiều bằng phương pháp KTPCA với hàm nhân đa thức là tích vô hướng của hai véc tơ hay hàm nhân Gauss với tham số ρ_0^2 = khoảng cách trung bình tối thiểu giữa hai véc tơ dữ liệu đã gợi ý nên chọn hàm nhân đa thức

cũng như hàm nhân Gauss với tham số ρ^2 như thế nào thì hiệu suất giảm chiều sẽ cao hơn.

Trong Giai đoạn 3, việc chiết xuất các nhân tố và xây dựng mô hình dự báo của biến phụ thuộc theo các nhân tố được thực hiện kết hợp vào nhau. Do đó, khi Giai đoạn này kết thúc, ta sẽ nhận được mô hình dự báo không hoặc có điều kiện với RMSE nhỏ nhất cùng với các nhân tố trong mô hình tương ứng với hàm nhân phù hợp nhất trong số các hàm nhân được thực nghiệm.

***Giai đoạn 4**: Thực hiện dự báo*

Nội dung chính của Giai đoạn này trong cả hai quy trình là tính các giá trị dự báo của biến phụ thuộc theo giá trị dự báo của các biến ngoại sinh trong mô hình bằng sử dụng mô hình được xây dựng ở Giai đoạn 3. Sự khác nhau của hai quy trình dự báo có hoặc không có điều kiện trong Giai đoạn này là:

- Trong dự báo có điều kiện, để nhận được các giá trị dự báo trong tương lai của biến phụ thuộc, trước hết ta cần xây dựng các mô hình phụ để dự báo các nhân tố ngoại sinh trong mô hình và thực hiện dự báo các nhân tố này bằng sử dụng mô hình dự báo phụ của nó. Luận án này sử dụng mô hình AR(p) có xu thế để dự báo các nhân tố ngoại sinh. Cụ thể, giả sử Y_t là biến chuỗi thời gian dừng. Mô hình hồi quy AR(p) có xu thế bậc hai của biến Y_t có dạng:

$$Y_t = a.tr + b.tr^2 + \sum_{j=1}^{p} c_j.Y_{t-j} + d + u_t \qquad (3.1)$$

trong đó phần dư u_t là nhiễu trắng; a, b, c_j, d là các tham số ước lượng của mô hình; tr, tr^2 là các biến xu thế [81]. Các biến tr, tr^2 được đưa vào mô hình (3.1) để phản ánh các yếu tố có liên quan đến sự biến đổi của biến Y_t nhưng chúng có thể chưa được phản ánh trong các trễ của biến này. Thành phần d + $a.tr + b.tr^2$ phản ánh xu thế của biến Y_t.

- Trong khi đó, ở quy trình dự báo không điều kiện thì không cần làm như vậy, vì có thể tính được giá trị tương lai của biến phụ thuộc theo các giá trị hiện tại và quá khứ của các nhân tố ngoại sinh bằng sử dụng mô hình dự báo của biến phụ thuộc đã được xây dựng trong Giai đoạn 3.

Quy trình dự báo có điều kiện và không điều kiện trên các tập dữ liệu chuỗi thời gian lớn ở trên cho thấy giảm chiều học thuộc tính sử dụng phương pháp KTPCA lặp là giai đoạn rất quan trọng trong quy trình này. Các Giai đoạn 2, 3, và 4 nhằm xây dựng các mô hình dự báo không và có điều kiện trên tập dữ liệu lớn đã được xử lý, vì thế nó cũng được xem là quy trình xây dựng mô hình dự báo theo cách tiếp cận 3 bước được đề xuất trong nghiên cứu mới đây [17]. Trong nghiên cứu này, các tác giả đã thực nghiệm và khẳng định độ chính xác dự báo của các mô hình được xây dựng trên các tập dữ liệu chuỗi thời gian lớn theo cách tiếp cận 3 bước là: lựa chọn biến bằng sử dụng phương pháp hồi quy góc nhỏ, học thuộc tính bằng sử dụng phương pháp PCA, và hồi quy rừng ngẫu nhiên là cao nhất so với độ chính xác dự báo của các mô hình được xây dựng theo tất cả các cách tiếp cận khác bao gồm cả cách tiếp cận sử dụng các kỹ thuật học sâu mạng nơtron, xích markov, hồi quy lượng tử, ước lượng bình phương tuyến tính nhỏ nhất, …

Quy trình dự báo không và có điều kiện đều thể hiện rõ cách tiếp cận 3 bước trong xây dựng mô hình dự báo trong nghiên cứu [17]. Dưới đây sẽ phân tích kỹ hơn sự giống nhau và khác biệt giữa luận án và nghiên cứu [17] theo từng bước (hay giai đoạn) của cách tiếp cận đó.

Giai đoạn 2 trong Quy trình dự báo và Bước 1 trong nghiên cứu [17] đều nhằm mục đích lựa chọn biến. Trong nghiên cứu [17], việc lựa chọn biến được thực hiện bằng phương pháp hồi quy góc nhỏ mà về bản chất là sử dụng độ đo hệ số tương quan Pearson như luận án này đề xuất. Khác với nghiên cứu [17], ở Giai đoạn này (hay Bước 1), luận án còn loại bỏ những biến dư thừa điều đó sẽ làm tăng hiệu quả cho việc xây dựng mô hình dự báo trong khi nghiên cứu [17] không thực hiện nội dung này. Để xây dựng mô hình dự báo không điều kiện, ở Bước này, luận án đề xuất phương pháp lựa chọn biến là các chỉ số dẫn báo bằng sử dụng quan hệ nhân quả Granger [76].

Giai đoạn 3 hay Bước 2: Nghiên cứu [17] sử dụng phương pháp PCA để chiết xuất nhân tố, nên độ chính xác dự báo sẽ bị hạn chế nếu tập dữ liệu được sử dụng để xây dựng mô hình không xấp xỉ một siêu phẳng. Luận án sử dụng phương pháp KTPCA lặp sẽ khắc phục được nhược điểm này;

Giai đoạn 4 hay Bước 3: Nghiên cứu [17] sử dụng phương pháp hồi quy rừng ngẫu nhiên, trong khi luận án sử dụng phương pháp hồi quy có các trễ của biến phụ thuộc và các biến giải thích (hay mô hình trễ phân bố tự hồi quy ARDL). Bản chất của hồi quy rừng ngẫu nhiên *kinh tế* là sự phát triển của hồi quy rừng ngẫu nhiên theo cách như sau: tập dữ liệu lớn được phân thành nhiều nhóm con (hay cụm hoặc cây con), sau đó xây dựng mô hình dự báo biến phụ thuộc trên mỗi nhóm con để nhận được chuỗi dự báo trong mẫu (hay chuỗi được làm phù hợp – fitted) của biến phụ thuộc trên mỗi nhóm con. Khác với học rừng ngẫu nhiên, chuỗi dự báo của biến phụ thuộc không phải là giá trị trung bình (đối với biến dự báo nhận giá trị rời rạc) hay lớp xuất hiện nhiều nhất (nếu biến phụ thuộc nhận giá trị phân loại) mà được tìm bằng cách hồi quy biến phụ thuộc theo các chuỗi dự báo của các nhóm con. Hồi quy rừng ngẫu nhiên kinh tế được xem là một phương pháp xây dựng mô hình dự báo trong lĩnh vực kinh tế-tài chính. Người ta đã chứng minh được rằng độ chính xác dự báo được thực hiện trên nhiều nhóm con theo cách như vậy là cao hơn so với thực hiện theo cách tương tự chỉ trên một nhóm. Nói cách khác sử dụng phương pháp hồi quy rừng ngẫu nhiên kinh tế dựa vào mô hình trễ phân bố tự hồi quy sẽ cho độ chính xác dự báo cao hơn nếu chỉ dựa vào mô hình này trên một tập dữ liệu cho.

Bảng 3.1 trình bày tóm tắt kết quả so sánh cách tiếp cận xây dựng mô hình dự báo có điều kiện trong luận án này với cách tiếp cận 3 bước trong xây dựng mô hình dự báo trong nghiên cứu [17].

Bảng 3.1: So sánh hai cách tiếp cận xây dựng mô hình dự báo có điều kiện

Luận án so với nghiên cứu [17]	Giai đoạn 2- Bước 1: Lựa chọn biến	Giai đoạn 3- Bước 2: Học thuộc tính	Giai đoạn 4- Bước 3: Phương pháp hồi quy
Luận án	Sử dụng phương pháp hồi quy góc nhỏ, nhưng xử lý dữ liệu dư thừa. Đánh giá: tốt hơn	Sử dụng phương pháp giảm chiều thực hiện cho cả tập dữ liệu xấp xỉ hoặc không xấp xỉ một siêu phẳng. Đánh giá: tốt hơn	Mô hình ARDL trên các nhân tố được chiết xuất từ tập dữ liệu của tất cả các biến đầu vào. Đánh giá kém hơn.
Nghiên cứu [17]	Sử dụng phương pháp hồi quy góc nhỏ, nhưng không xử	Sử dụng phương pháp giảm chiều PCA (là trường hợp riêng của	Hồi quy rừng ngẫu nhiên kinh tế. Bản chất của nó là phân các biến

	lý dữ liệu dư thừa. Đánh giá: kém hơn	phương pháp giảm chiều trong luận án) cho cả các tập dữ liệu không xấp xỉ siêu phẳng. Đánh giá: kém hơn	giải thích thành các nhóm con, xây dựng mô hình dự báo biến phụ thuộc trên các nhóm con bằng sử dụng mô hình ARDL, sau đó kết hợp các kết quả dự báo biến phụ thuộc của các mô hình thành phần. Đánh giá tốt hơn

Bảng 3.1 cho thấy quy trình dự báo được đề xuất ở trên khá tương tự với cách tiếp cận dự báo 3 bước trong [17] và có thể điều chỉnh để sử dụng phương pháp hồi quy rừng ngẫu nhiên kinh tế ở trong quy trình này. Khi đó độ chính xác dự báo theo quy trình sẽ được cải thiện hơn nữa. Hiện tại luận án chưa thực hiện theo cách như vậy.

Các thuật toán dự báo (có điều kiện và không điều kiện) trên các tập dữ liệu chuỗi thời gian lớn được đề xuất trong phần tiếp theo là được xây dựng dựa vào các quy trình dự báo được giới thiệu trong phần này.

3.2 Thuật toán dự báo trên tập dữ liệu chuỗi thời gian lớn

Các thuật toán này được xây dựng theo quy trình được đề xuất trong Hình 3.1. Giả sử $\mathbf{X_t} = [X_{1,t}, X_{2,t}, \dots, X_{m,t}] \in \mathbb{R}^{\mathbf{t}\times\mathbf{m}}$ là tập dữ liệu của các biến chuỗi thời gian, $X_{i,t} \in \mathbb{R}^{\mathbf{t}}, i = 1, \dots, m$; $Y_t \in \mathbb{R}^{\mathbf{t}}$ là biến phụ thuộc, trong đó *m* và *t* lần lượt là số lượng biến và số lượng quan sát; *m* là rất lớn.

Vấn đề là xây dựng một thuật toán cho phép tự động thực hiện dự báo có không hoặc có điều kiện của biến phụ thuộc Y_t theo tập các biến giải thích $\mathbf{X_t}$.

Các thuật toán dự báo trên tập dữ liệu chuỗi thời gian lớn được đề xuất trong phần tiếp theo được xây dựng dựa vào các quy trình dự báo ở trên.

3.2.1 Thuật toán dự báo có điều kiện

Không mất tính tổng quát, giả sử tập dữ liệu của các biến giải thích $\mathbf{X_t}$ được cân chỉnh trung bình. Tập dữ liệu này được sử dụng để chiết xuất các nhân tố bằng sử dụng phương pháp KTPCA ứng với mỗi hàm nhân được đưa vào thử nghiệm.

Thuật toán dự báo có điều kiện trên tập dữ liệu chuỗi thời gian lớn được trình bày dưới dạng giả mã như sau:

THUẬT TOÁN 1a: **CONF algorithm**

Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t \times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, α và β: các ngưỡng liên quan (relevant threshold) và ngưỡng dư thừa (redundancy threshold), q(%): ngưỡng giá trị riêng tích lũy.

Output: $\hat{Y}_{t+h}$: dự báo trước h bước tại thời điểm t của Y_t trên $\mathbf{X_t}$.

Begin

1. Cho h - thời điểm xa nhất của dự báo;
2. *Repetition* ← "Yes";
3. *FeatureSelection* ($\mathbf{X_t}$, Y_t); // *α và β: các ngưỡng liên quan và dư thừa*
4. Center $\mathbf{X_t}$;
5. Tính khoảng cách tối thiểu trung bình của 2 véc tơ dữ liệu của các biến giải thích $\mathbf{X_t}$;
6. Tính ma trận hiệp phương sai $\boldsymbol{K}$ của $\mathbf{X_t}$;
7. *FeatureLearning*($\boldsymbol{K}$, *q*);// *q(%): ngưỡng giá trị riêng tích lũy*
8. Lưu các nhân tố được giữ lại, mô hình dự báo trên tập các nhân tố được giữa lại, và RMSE của mô hình này;
9. **Repeat**
10. Nhập một hàm nhân $\boldsymbol{\kappa}$: $\mathbb{R}^{\mathbf{t}}$ x $\mathbb{R}^{\mathbf{t}} \longrightarrow \mathbb{R}$;

    // chọn hàm nhân theo các nghiên cứu [34],[93],[97]
11. Tính ma trận hàm nhân $\boldsymbol{K}$;
12. *FeatureLearning* ($\boldsymbol{K}, q$);
13. **if** *RMSE* của mô hình vừa được xây dựng < *RMSE* đang được lưu **then**

    Thay tập các nhân tố đang lưu, mô hình dự báo đang lưu, RMSE đang lưu tương ứng bằng tập các nhân tố mới được giữ lại, mô hình dự báo mới được xây dựng, và RMSE của mô hình này;
14. **end**
15. **Until (***Repetition* = "No")

16. *Forecast (*Mô hình dự báo biến Y_t, Tập các nhân tố được giữ lại*);*
17. Return $\hat{Y}_{t+h}$;

End.

Trong thuật toán này, hàm *FeatureSelection* trong dòng lệnh 3 thực hiện lựa chọn các biến có giá trị bằng cách sử dụng độ đo hệ số tương quan Pearson để loại bỏ các biến giải thích không hoặc ít liên quan hoặc các biến dư thừa đối với biến phụ thuộc. Thủ tục *FeatureLearning* trong dòng lệnh 7 và 12 thực hiện chiết xuất các nhân tố từ tập dữ liệu đầu vào được cân chỉnh trung bình bằng phương pháp PCA. Số lượng các nhân tố được chọn là số nguyên nhỏ nhất sao cho tỷ lệ phần trăm giá trị riêng tích lũy của các nhân tố được chọn là lớn hơn ngưỡng *q(%)* do người dùng xác định. Thủ tục này cũng ứng dụng phương pháp ước lượng bình phương tuyến tính nhỏ nhất trên các nhân tố được chọn để loại bỏ các nhân tố không có ý nghĩa thống kê và sau đó tính toán RMSE của mô hình ước lượng cuối cùng. Các dòng lệnh từ 9 đến 15 thể hiện quy trình lặp thực hiện phương pháp KTPCA trên tập dữ liệu đầu vào được cân chỉnh trung bình và kiểm tra xem RMSE của mô hình dự báo mới được xây dựng có nhỏ hơn RMSE đang được lưu hay không. Nếu đúng, thay thế tập các nhân tố đang được lưu, mô hình dự báo đang được lưu, và RMSE đang được lưu tương ứng bằng các nhân tố mới được chọn, mô hình dự báo mới được xây dựng, và RMSE của mô hình này. Thủ tục *Forecast* trong dòng lệnh 16 thực hiện dự báo trước *h* bước của các nhân tố $FAC_{i,t}$, (i ≥ 1) tại thời điểm *t* và sử dụng mô hình dự báo đã được xây dựng để tính toán giá trị trước *h* bước của biến phụ thuộc Y_t tại thời điểm đó.

Hàm *FeatureSelection*, thủ tục *FeatureLearning* và thủ tục *Forecast* được giới thiệu chi tiết hơn bên dưới.

THUẬT TOÁN 2a: *FeatureSelection* Algorithm

Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, *α và β: các ngưỡng liên quan và dư thừa*, $Y_t \in \mathbb{R}^{\mathbf{t}}$.

Output: Tập các biến có liên quan và không dư thừa trong $\mathbf{X_t}$.

begin

1. Loại bỏ các biến ít hoặc không liên quan đến Y_t;
2. Order **($\mathbf{X_t}$)** // Sắp xếp các biến theo thứ tự giảm dần của độ đo Pearson;

3. Loại bỏ các biến dư thừa trong $\mathbf{X_t}$;
4. return $\mathbf{X_t}$;

end;

THUẬT TOÁN 3a: *FeatureLearning* Procedure

Input: Ma trận $\mathbf{K_{m\times m}}$; q(%): tỷ lệ phần trăm giá trị riêng tích lũy;

Output: Tập các nhân tố được giữ lại; mô hình dự báo Y_t theo các nhân tố được giữ lại, và RMSE của mô hình này;

begin

1. Tính giá trị riêng và véc tơ riêng của ma trận **K;**
2. Sắp xếp các véc tơ riêng theo thứ tự giảm dần của các giá trị riêng tương ứng;
3. Chiết xuất các nhân tố bằng cách chiếu tập dữ liệu $\mathbf{X_t}$, đã được cân chỉnh trung bình, lên các véc tơ riêng;
4. Tạo dựng tập hợp gồm p nhân tố đầu tiên sao cho % giá trị riêng tích lũy của chúng là số không nhỏ hơn q(%) đã cho;
5. Xây dựng mô hình dự báo Y_t trên các nhân tố được giữ lại dựa trên mô hình trễ phân bố tự hồi quy ARDL;
6. Tính RMSE của mô hình dự báo vừa được xây dựng.

end;

THUẬT TOÁN 4a: *Forecast* Algorithm

Input: Tập nhân tố được giữa lại cuối cùng; mô hình dự báo Y_t theo các nhân tố được giữ lại;

Output: $\hat{Y}_{t+h}$: dự báo trước h bước của biến Y_t tại thời điểm t.

begin

1. Xây dựng mô hình dự báo phụ cho các nhân tố trong mô hình dự báo biến Y_t dựa trên mô hình tự hồi quy có xu thế bậc 2 AR(p);
2. Thực hiện dự báo h-bước ngoài mẫu cho các nhân tố bằng sử dụng các mô hình dự báo phụ tương ứng;

3. Tính $\hat{Y}_{t+h}$ bằng sử dụng mô hình dự báo của biến Y_t;
4. return $\hat{Y}_{t+h}$;

end;

Một điều cần lưu ý là trong thuật toán này, độ trễ tối ưu chung của biến phụ thuộc và các nhân tố được chiết xuất bằng phương pháp KTPCA trong các mô hình dự báo ở các vòng lặp khác nhau là như nhau và được xác định chính xác khi thực hiện giảm chiều biến bằng phương pháp PCA. Điều này là để tạo thuận lợi cho việc xây dựng chương trình máy tính cho thuật toán được đề xuất.

3.2.2 Thuật toán dự báo không điều kiện

Trong các mô hình dự báo không điều kiện, khoảng thời gian dự báo phụ thuộc vào độ trễ nhỏ nhất của tất cả các chỉ số dẫn báo nằm trong mô hình và luôn có thể dự báo không điều kiện cho biến phụ thuộc trước ít nhất 1 bước. Các chỉ số dẫn báo có thể được xác định theo mô hình (1.15) hoặc (1.16) nhưng trong ứng dụng thực tế, người ta thường sử dụng mô hình (1.16) với cùng độ trễ p cho tất cả các biến giải thích. Luận án này xác định các chỉ số dẫn báo theo cách như vậy. Khi đó, độ dài trễ p được xác định bằng cách kết hợp kiến thức miền ứng dụng với tiêu chuẩn thông tin Akaike (AIC) hoặc Bayesian (BIC) [81]. Thuật toán dự báo không điều kiện của biến phụ thuộc trên một tập dữ liệu lớn các biến giải thích được trình bày dưới dạng giả mã như sau:

THUẬT TOÁN 1b: **UNCONF algorithm**

Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, q(%): ngưỡng giá trị riêng tích lũy;

Output: $\hat{Y}_{t+h}$: dự báo trước h bước ngoài mẫu được thực hiện tại thời điểm t của biến Y_t // h ít nhất là 1 nhưng không được xác định trước;

Begin

1. Xác định độ trễ chung p cho tất cả các biến;
2. *Repetition* ← "Yes";
3. *LeadingIndicatorSelection* ($\mathbf{X_t}$, Y_t);
4. Center $\mathbf{X_t}$;

5. Tính khoảng cách tối thiểu trung bình của 2 véc tơ dữ liệu của các biến giải thích
6. Tính ma trận hiệp phương sai **K** của $\mathbf{X_t}$
7. *FeatureLearning*(***K***, q)**;**
8. Lưu các nhân tố được giữ lại, mô hình dự báo trên tập các nhân tố được giữa lại, và RMSE của mô hình này.
9. **Repeat**
10. Nhập một hàm nhân $\boldsymbol{\kappa}$**:** $\mathbb{R}^{\mathbf{t}}$ x $\mathbb{R}^{\mathbf{t}} \longrightarrow \mathbb{R}$
11. Tính ma trận hàm nhân ***K***;
12. *FeatureLearning* (***K***, q);
13. **if** *RMSE* của mô hình vừa được xây dựng **<** *RMSE* đang được lưu **then**

    Thay tập các nhân tố đang lưu, mô hình dự báo đang lưu, RMSE đang lưu tương ứng bằng tập các nhân tố mới được giữ lại, mô hình dự báo mới được xây dựng, và RMSE của mô hình này
14. **end**
15. **Until (***Repetition* = "No")
16. *Calculate*($\widehat{\mathrm{Y}}_{\mathrm{t+h}}$, Mô hình dự báo biến $\mathrm{Y_t}$)
17. **Reture** ($\hat{Y}_{t+h}$);

End.

Cấu trúc của thuật toán dự báo không điều kiện (**UNCONF Algorithm)** tương tự như cấu trúc của thuật toán CONF và chỉ khác thuật toán CONF ở 3 dòng lệnh 1, 3 và 16. Cụ thể, dòng lệnh 1 xác định độ trễ chung cho tất cả các biến trong việc tìm kiếm các chỉ số dẫn báo của biến phụ thuộc. Dòng lệnh 3 gọi hàm *LeadingIndicatorSelection* để xác định tất cả các chỉ số dẫn báo có ý nghĩa thống kê của biến Y_t và dòng lệnh 16 gọi hàm *Calculate* để tính các giá trị dự báo ngoài mẫu của biến phụ thuộc bằng sử dụng mô hình dự báo biến Y_t. Thuật toán UNCONF cũng được giải thích khá giống với thuật toán CONF dựa vào các dòng lệnh thay đổi trên.

Hàm *LeadingIndicatorSelection* trong dòng lệnh 3 của thuật toán UNCONF là như sau.

THUẬT TOÁN 2b: *LeadingIndicatorSelection* Algorithm

Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t}\times\mathbf{m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, p là độ trễ chung;

Output: Tập các chỉ số dẫn báo của Y_t với trễ p trong $\mathbf{X_t}$, ký hiệu $\boldsymbol{X}^*$;

begin

1. Chuyển biến Y_t và các biến trong $\mathbf{X_t}$ thành các chuỗi thời gian dừng;
2. Cho mức ý nghĩa thống kê: α;
3. $\boldsymbol{X}^* \leftarrow \{\emptyset\}$;
4. **for** mỗi biến trong $\mathbf{X_t}$ thực hiện
5. Xây dựng mô hình dự báo biến Y_t theo biến này dựa vào mô hình (2.2);
6. Tính xác suất của thống kê F trong mô hình dự báo;
7. **if** xác suất đó < α **then** biến giải thích đó là chỉ số dẫn báo và được thêm vào tập $\boldsymbol{X}^*$
8. **end for**
9. **reture** ($\boldsymbol{X}^*$);

end;

Đầu vào, Đầu ra và cấu trúc của thuật toán 2b (*LeadingIndicatorSelection*) là khác với thuật toán 2a. Trong thuật toán 2b, dòng lệnh 1 kiểm tra xem biến Y_t và các biến trong $\mathbf{X_t}$ có phải là chuỗi thời gian dừng hay không, nếu không, chuyển đổi chúng thành chuỗi dừng. Kiểm định kiểm tra tính dừng được sử dụng là kiểm định Augmented Dickey-Fuller và phương pháp biến đổi một biến không dừng thành một biến dừng là sử dụng kết hợp log và sai phân hoặc chỉ sai phân tùy thuộc vào giá trị của các biến có dương hay không [76], [77]. Các dòng lệnh từ 4 đến 6 được thực hiện như nhau cho từng biến giải thích. Trong các công cụ thống kê, α thường được chọn là 0.001, 0.01, 0.05 và 0.1. Trong thực tế, nói chung, α được chọn là 0.05.

THUẬT TOÁN 3b: *FeatureLearning* Procedure

Input: Ma trận $\mathbf{K}_{\boldsymbol{N}\times\boldsymbol{g}}$ là ma trận hàm nhân của tập gồm g chỉ số dẫn báo; q(%): tỷ lệ phần trăm giá trị riêng tích lũy.

Output: Tập các nhân tố được giữ lại; mô hình dự báo biến Y_t trên các nhân tố được giữ lại, và RMSE của mô hình này.

begin

1. Tính giá trị riêng và véc tơ riêng của ma trận **K;**
2. Sắp xếp các véc tơ riêng theo thứ tự giảm dần của các giá trị riêng tương ứng;
3. Chiết xuất các nhân tố bằng cách chiếu tập dữ liệu $\mathbf{X_t}$, đã được cân chỉnh trung bình, lên các véc tơ riêng;
4. Tạo dựng tập hợp gồm p nhân tố đầu tiên sao cho % giá trị riêng tích lũy của chúng là số không nhỏ hơn q(%) đã cho;
5. Xây dựng mô hình dự báo Y_t trên các nhân tố được giữ lại của các chỉ số dẫn báo dựa trên mô hình ARDL, ở đó độ trễ của biến phụ thuộc và biến giải thích đã được xác định trước.
6. Tính RMSE của mô hình dự báo vừa được xây dựng;

end;

Thuật toán 3b (*FeatureLearning*) trong thuật toán 3b có đầu vào, đầu ra và cấu trúc tương tự như thuật toán 3a. Nó khác thuật toán 3a chỉ ở dòng lệnh 5. Trong thuật toán UNCONF, độ trễ của biến phụ thuộc và biến giải thích là như nhau. Độ trễ chung này đã được xác định trong thuật toán 1b.

THUẬT TOÁN 4b: *Calculate* Algorithm

Input: Tập các nhân tố được giữ lại cuối cùng; mô hình dự báo biến Y_t theo các nhân tố được giữ lại.

Output: $\hat{Y}_{t+h}$: các dự báo trước h – bước được thực hiện tại thời điểm t cho biến Y_t, $(1 \leq h \leq \text{p})$;

begin

1. Tính $\hat{Y}_{t+h}$ bằng sử dụng mô hình dự báo biến Y_t tại thời điểm t.
2. return $\hat{Y}_{t+h}$;

end;

Mục đích của các thuật toán 4a và 4b là như nhau. Đầu vào – Đầu ra của các thuật toán này là tương tự như nhau, chúng khác nhau chỉ ở giá trị h – bước. Trong thuật

toán 4a, giá trị h do người sử dụng xác định, trong khi trong thuật toán 4b, giá trị h chì được xác định dựa vào độ trễ thấp nhất của tất cả các nhân tố dẫn báo trong mô hình dự báo Y_t. Cụ thể nếu độ trễ thấp nhất đó là k (k=1, 2, …) thì tương ứng ta có thể dự báo không điều kiện k- bước ngoài mẫu biến Y_t. Trong thuật toán 4b, ta không cần dự báo ngoài mẫu các nhân tố dẫn báo mà chỉ cần dựa vào giá trị của các biến này để tính giá trị dự báo h-bước ngoài mẫu của biến Y_t.

Các thuật toán CONF và UNCONF có thể được mã hóa dễ dàng trong môi trường của các ngôn ngữ như Python, R hoặc Matlab dựa vào tham chiếu đến các gói có sẵn. Cụ thể, luận án tham khảo gói Kernlab [104] và gói Caret [105] để mã hóa các thuật toán được đề xuất trong R.

Việc ước lượng độ phức tạp tính toán của thuật toán dự báo không và có điều kiện sẽ được trình bày trong phần tiếp theo dưới đây.

3.2.3 Độ phức tạp tính toán

3.2.3.1 Độ phức tạp tính toán của thuật toán CONF

Gọi m, N tương ứng là số biến và số quan sát của tập dữ liệu đầu vào $\mathbf{X_t}$, q là số lần lặp của phương pháp giảm chiều KTPCA và của việc xây dựng mô hình dự báo trên các nhân tố được chiết xuất bởi phương pháp này.

Độ phức tạp tính toán của thuật toán dự báo có điều kiện phụ thuộc vào độ phức tạp tính toán của: (1) thuật toán *FeatureSelection* (dòng lệnh 3) trong thuật toán CONF, (2) việc tính ma trận hàm nhân (với hàm nhân là tích vô hướng hoặc không phải là tích vô hướng) (dòng lệnh 6 hoặc dòng lệnh 11), (3) thủ tục *FeatureLearning* (dòng lệnh 7 hoặc 12), và (4) thuật toán *Forecast* ở dòng lệnh 16.

Ký hiệu c là chi phí tính toán hệ số tương quan Pearson của hai véc tơ đầu vào, khi đó chi phí để loại bỏ những biến giải thích ít liên quan đến biến phụ thuộc và sắp xếp những biến này theo thứ tự (giảm dần hoặc tăng dần) là $O(c.(m-1) + m.log_2 m)$. Trong khi đó chi phí để loại bỏ những biến giải thích dư thừa không vượt quá $O\big((m-1)+(m-2)+\cdots+2\big) = O\left(\frac{m(m-1)}{2}\right) = O(m^2)$.

Vậy độ phức tạp tính toán của thuật toán *FeatureSelection* là: $O(m^2)$ (3.2)

Các dòng lệnh 6, 7 thực chất là thực hiện phương pháp PCA và xây dựng mô hình dự báo trên các nhân tố được chiết xuất dựa trên mô hình ARDL sử dụng phương pháp ước lượng bình phương tuyến tính nhỏ nhất. Theo [46], độ phức tạp tính toán của thuật toán PCA là $O(N.m^2+ N^3)$, trong khi theo [106], độ phức tạp tính toán của phương pháp ước lượng bình phương nhỏ nhất của biến phụ thuộc trên tập dữ liệu của p nhân tố thành phần chính với N quan sát là $O(p^2.N + p^3) = O(N)$ do p khá nhỏ khi thực hiện phương pháp giảm chiều học thuộc tính.

Vậy độ phức tạp tính toán của các dòng lệnh 6 và 7 là:

$$O(N.m^2+ N^3) \quad (3.3)$$

Cũng theo [46], độ phức tạp tính toán của thuật toán KPCA là $O(N^3)$. Như vậy chi phí tính toán của ma trận hàm nhân trên các véc tơ điểm dữ liệu là không vượt quá $O(N^3)$. Trong phương pháp KTPCA, ma trận hàm nhân ở dòng lệnh 12 được xác định trên các véc tơ đầu vào (hay m biến giải thích), nên chi phí tính toán của ma trận hàm nhân trong phương pháp này không vượt quá $O(m^3)$.

Do đó độ phức tạp tính toán của dòng lệnh 11 và 12 là: $O(N.m^2+ N^3 + m^3)$. Vì có q vòng lặp như vậy nên độ phức tạp tính toán của các dòng lệnh từ 10 đến 16 là:

$$q.O(N.m^2+ N^3 + m^3). \quad (3.4)$$

Độ phức tạp tính toán của thuật toán *Forecast* ở dòng lệnh 16 của thuật toán CONF chủ yếu phụ thuộc vào chi phí tính toán để xây dựng các mô hình phụ để dự báo p nhân tố trong mô hình dự báo có điều kiện. Theo [106] chi phí tính toán để xây dựng một mô hình như vậy là $O((s + 2)^2.N + (s + 2)^3) = O(N)$, ở đây s là độ dài trễ tối ưu của các biến ngoại sinh và có 2 biến xu thế là *tr* và tr^2. Và độ phức tạp tính toán của thuật toán *Forecast* là $p.O(N) = O(N)$ (do p rất nhỏ) (3.5)

Từ (3.2), (3.3), (3.4) và (3.5) ta nhận được độ phức tạp tính toán của thuật toán dự báo có điều kiện CONF là: $q.O(N.m^2+ N^3 + m^3)$. (3.6)

3.2.3.2 Độ phức tạp tính toán của thuật toán UNCONF

Thuật toán dự báo không điều kiện khác thuật toán có điều kiện chủ yếu ở các thuật toán *LeadingIndicatorSelection* và *Calculate.* Do chi phí tính toán của *Calculate* là rất nhỏ so với các thuật toán *FeatureLearning* nên có thể bỏ qua.

Với mỗi biến giải thích, chi phí tính toán để biết biến này có phải là nguyên nhân Granger với s trễ của biến phụ thuộc là $O((2s+1)^2.N + (2s+1)^3 = O(N)$ do s cố định và nhỏ [106]. Do vậy độ phức tạp tính toán của thuật toán *LeadingIndicatorSelection* là:

$$O\big(m.O(N)\big) = O(m.N) \tag{3.7}$$

Lập luận tương tự như thuật toán CONF, ta nhận được độ phức tạp của thuật toán UNCONF là $q.O(N.m^2 + N^3 + m^3)$. Vậy độ phức tạp của thuật toán dự báo, bao gồm dự báo không và có điều kiện là:

$$q.O(N.m^2 + N^3 + m^3). \tag{3.8}$$

Phần tiếp theo dưới đây sẽ trình bày việc mô hình hóa dự báo không và có điều kiện kim ngạch xuất khẩu Việt Nam trên tập dữ liệu chuỗi thời gian lớn sử dụng thuật toán được đề xuất.

3.3 Dự báo kim ngạch xuất khẩu sử dụng thuật toán dự báo

3.3.1 Xác định vấn đề dự báo

Với mọi nền kinh tế, dự báo kim ngạch xuất khẩu luôn là một trong các nội dung quan trọng nhất của dự báo kinh tế vĩ mô. Việc dự báo chính xác kim ngạch xuất khẩu có liên quan chặt chẽ đến nhiều hoạt động của nền kinh tế trong đó nhất là các hoạt động sản xuất, nhập khẩu nguyên, vật liệu cho sản xuất hàng xuất khẩu; hoạt động duy trì hay thay đổi tỷ giá hối đoái, cung tín dụng cho nền kinh tế, …

Việt Nam là nền kinh tế mở rất cao. Tổng kim ngạch xuất, nhập khẩu cao hơn 200% của Tổng sản phẩm trong nước (GDP) trong đó tổng kim ngạch xuất khẩu lớn hơn 100% GDP. Bởi vậy ở Việt Nam hoạt động dự báo kim ngạch xuất khẩu càng quan trọng và cần thiết hơn.

Với sự hội nhập quốc tế ngày càng sâu rộng, các yếu tố tác động đến kim ngạch xuất khẩu của Việt Nam ngày càng nhiều và đa dạng. Việc thu thập dữ liệu như vậy ngày càng dễ dàng và đầy đủ hơn nhờ sự tiến bộ của ngành công nghệ thông tin. Làm thế nào để có thể dự báo được kim ngạch xuất khẩu Việt Nam khi có quá nhiều các yếu tố tác động như vậy là động lực để Luận án nghiên cứu ứng dụng mô hình dự báo không và có điều kiện sử dụng phương pháp giảm chiều dựa vào kỹ thuật

hàm nhân được đề xuất trong Chương 2 vào dự báo kim ngạch xuất khẩu hàng tháng của Việt Nam.

Vấn đề cần được giải quyết ở mục này là: dự báo kim ngạch xuất khẩu tháng của Việt Nam theo tất cả các yếu tố (biến số) trong và ngoài nước tiềm năng có ảnh hưởng đến hoạt động xuất khẩu của Việt Nam.

3.3.2 Các yếu tố tác động đến kim ngạch xuất khẩu và thu thập dữ liệu

3.3.2.1 Các yếu tố tác động đến kim ngạch xuất khẩu

Nghiên cứu [107] về dự báo ngắn hạn kim ngạch xuất khẩu của Ấn Độ đã giả định rằng kim ngạch xuất khẩu có liên quan đến nhu cầu nhập khẩu các sản phẩm sản xuất trong nước của các nước đối tác, đến chỉ số giá xuất khẩu, đến tỷ giá hối đoái và cùng với các biến trễ của chúng. Nghiên cứu [108] đã phân tích tổng thương mại của các nước Trung và Đông Âu (CEE) với các nước thuộc khu vực đồng Euro bằng cách sử dụng mô hình trọng lực tăng cường, trong đó tổng thương mại là tổng kim ngạch xuất, nhập khẩu. Kết quả chỉ ra rằng tổng thương mại của các nước CEE có mối quan hệ tích cực có ý nghĩa thống kê cao với tỷ giá hối đoái thực; với quy mô kinh tế (GDP) của các nước láng giềng, của các nước có ngôn ngữ tương đồng, của các nước thành viên trong liên minh thương mại trong khi khoảng cách địa lý không liên quan đến tổng thương mại.

Một trong những mô hình thường được sử dụng để dự báo kim ngạch xuất khẩu là mô hình cầu xuất khẩu. Mô hình này giả định rằng cung co giãn vô hạn, tức là khi có cầu thì bất kỳ nguồn cung nào cũng có thể được sử dụng để sản xuất phục vụ xuất khẩu. Trong mô hình cầu xuất khẩu, hầu hết các biến số như tỷ giá hối đoái, chỉ số giá và giá tương đối của hàng xuất khẩu đều được sử dụng, trong đó giá tương đối là một trong những yếu tố rất quan trọng quyết định năng lực cạnh tranh của hoạt động xuất khẩu và lợi thế so sánh [109]. Cụ thể, nghiên cứu [109] đề xuất mô hình dự báo tổng kim ngạch xuất khẩu dựa vào mô hình cầu xuất khẩu có dạng tổng quát như sau:

$$X_t = \boldsymbol{f}\left(X_{t-i,} ED_{t-i+1,} ER_{t-i+1,} \boldsymbol{P_{t-i+1}}\right), i \geq 1\} \qquad (3.9)$$

trong đó X_t là kim ngạch xuất khẩu hàng hóa và dịch vụ (thể hiện bằng giá danh nghĩa hoặc thực tế), ED_t là một thước đo tổng hợp của cầu bên ngoài, ER_t là tỷ giá hối đoái

(giá danh nghĩa hoặc giá thực tế) và $\boldsymbol{P_t}$ là véc tơ giá cả, tạo ra động lực giá cho nhóm hàng hóa trên thị trường quốc tế. Trong thực hành dự báo, hiện tại người ta ứng dụng mô hình (3.9) để dự báo kim ngạch xuất khẩu với chỉ một số lượng hạn chế các biến giải thích. Số lượng biến giải thích có trong các mô hình cầu xuất khẩu thường trung bình là 4 nếu không tính đến các biến trễ khác của chúng. Các biến như vậy được gọi là biến cứng.

Nghiên cứu [110] đã sử dụng mô hình cầu xuất khẩu để dự báo tổng kim ngạch xuất khẩu, trong đó các biến giải thích trong mô hình bao gồm chỉ số giá xuất khẩu hàng hóa và dịch vụ (giá danh nghĩa hoặc giá thực tế), tỷ giá hối đoái (giá danh nghĩa hoặc giá thực tế), chỉ số tổng hợp đo lường nhu cầu bên ngoài đối với các sản phẩm được sản xuất trong nước, một véc tơ giá của các mặt hàng xuất khẩu chính của nền kinh tế và các biến trễ của chúng. Hiện mô hình dự báo cầu xuất khẩu là mô hình được sử dụng hiệu quả và phổ biến nhất trong dự báo kim ngạch xuất khẩu của toàn nền kinh tế cũng như từng ngành kinh tế ở các quốc gia.

Nghiên cứu [111] đã đưa ra dự báo chỉ số kim ngạch xuất nhập khẩu của hầu hết các nước châu Âu (gồm 18 nước Đông và Trung Âu, 28 nước EU) để nghiên cứu nền kinh tế của các nước này. Các biến dựa vào khảo sát (gọi là biến mềm) như chỉ số niềm tin kinh doanh, chỉ số niềm tin người tiêu dùng, chỉ số kỳ vọng sản xuất được đưa vào mô hình cầu xuất khẩu để xem liệu sai số dự báo có thấp hơn nếu so với mô hình gồm chỉ các biến cứng như biến giá và chỉ số giá, biến tổng cầu, tỷ giá hối đoái hay không. Ở hầu hết các quốc gia, kết quả cho thấy việc bổ sung các biến mềm vào mô hình dự báo xuất khẩu theo cách tiếp cận mô hình cầu sẽ cho sai số dự báo thấp hơn. Tuy nhiên, việc đưa cả biến cứng và biến mềm vào mô hình cầu xuất khẩu khiến số lượng biến giải thích tăng lên, và trong nhiều trường hợp, việc ước lượng mô hình bằng hồi quy đa biến là không thể.

Trong môi trường giàu dữ liệu hay dữ liệu lớn, nhiều biến số khác cũng ảnh hưởng đến hoạt động xuất khẩu của một nền kinh tế, chẳng hạn như sản xuất công nghiệp, dư nợ tín dụng của các ngành sản xuất, hàng tồn kho, tình hình kinh tế - chính trị trong nước và quốc tế.

Có thể thấy mô hình cầu xuất khẩu mới chỉ đưa vào mô hình một số ít biến đại diện cho mỗi nhóm biến có tác động đến xuất khẩu nhìn từ phía cầu (phía tiêu dùng), điều này có thể bỏ sót nhiều thông tin dẫn đến độ chính xác dự báo của mô hình bị hạn chế. Về nguyên tắc, việc đưa được càng nhiều thông tin không gây nhiễu hoặc dư thừa vào mô hình dự báo thì độ chính xác dự báo của mô hình càng cao.

3.3.2.2 Tập dữ liệu phục vụ dự báo

Các nghiên cứu [109], [110], [111] đã gợi ý những loại dữ liệu cần được thu thập để phục vụ dự báo kim ngạch xuất khẩu theo tháng của Việt Nam.

Tập dữ liệu được thu thập và sử dụng để xây dựng mô hình dự báo kim ngạch xuất khẩu theo tháng của Việt Nam được gọi là **EXP** bao gồm các yếu tố tác động đến kim ngạch xuất khẩu trong mô hình cầu xuất khẩu [109], [110].

Đối với các nền kinh tế thị trường, ngoài phía cầu (hay phía tiêu dùng) người ta còn xem xét nghiên cứu phía cung (hay phía sản xuất) và sức mạnh của thị trường trong việc cân đối cung - cầu. Nghiên cứu của [111], [112] gợi ý rằng tập dữ liệu của các yếu tố phản ánh ba khía cạnh chính đó của nền kinh tế cũng cần được thu thập để phục vụ dự báo kim ngạch xuất khẩu. Tập dữ liệu này cần gồm cả dữ liệu của các biến mềm được thu thập qua các hoạt động điều tra, khảo sát.

Tập dữ liệu thực tế được sử dụng để dự báo kim ngạch xuất khẩu của Việt Nam theo tháng trong luận án này là tập dữ liệu của 161 biến giải thích trong đó có các biến cứng và biến mềm. Các biến cứng bao gồm các biến có tác động đến kim ngạch xuất khẩu trong mô hình cầu xuất khẩu (3.9) và gồm nhiều biến giải thích khác có liên quan hoặc tiềm năng có liên quan đến kim ngạch xuất khẩu nhìn từ phía cung và phản ánh sức mạnh thị trường được thu thập thông qua hệ thống báo cáo thống kê nhà nước. Giá trị của các biến giải thích trong tập này có thể là số tuyệt đối hoặc là số tương đối (%). Trong đó, dữ liệu giá trị số tuyệt đối được thu thập từ tháng 1 năm 2013 đến tháng 6 năm 2019, trong khi dữ liệu số tương đối được thu thập từ tháng 2 năm 2014 đến tháng 6 năm 2019. Vì vậy, số lượng quan sát của những biến giải thích nhận giá trị số tương đối là 65.

Tập hợp tất cả những biến giải thích được giới thiệu trong Bảng 3.2 ở dưới bao gồm tên của biến giải thích, tần suất, ý nghĩa sử dụng của biến và nguồn dữ liệu

của nó. Tập này đã bao gồm tất cả các biến giải thích được sử dụng trong mô hình cầu xuất khẩu. Cụ thể:

Bảng 3.2: Tập dữ liệu phục vụ dự báo kim ngạch xuất khẩu

Các biến giải thích	**Tên Biến**	**Freq.**	**Tính chất**	**Nguồn**
10 biến là các chỉ số sản xuất công nghiệp	$IIPi, i = 1, \dots, 10$	M	Phản ánh phía cung của thị trường	GSO
02 biến là đầu tư và thu ngân sách nhà nước	INV, REV	M		GSO
04 biến là đầu tư trực tiếp từ nước ngoài	$FDIi, i = 1, \dots, 4$	M		Fiinpro
19 biến là các chỉ số tồn kho một số ngành chế biến, chế tạo	$TKi, i = 1, \dots, 19$	M		Fiinpro
07 biến là dư nợ tín dụng một số ngành kinh tế	$DNi, i = 1, \dots, 7$	M		Fiinpro
26 biến là kim ngạch xuất khẩu của 25 ngành sản xuất và tổng kim ngạch xuất khẩu	$EX, EXi,$ $i = 1, \dots, 25$	M		GSO
01 biến là chỉ số giá hàng hóa xuất khẩu Việt Nam;	$PEX,$	M		GSO
01 biến là chỉ số nhà quản trị mua hàng.	PMI	M		Markitecon omics
19 biến là kim ngạch nhập khẩu của toàn nền kinh tế và 18 ngành sản xuất	$NK, NKi,$ $i = 1, \dots, 18$	M	Phản ánh phía cầu của thị trường	Fiinpro
19 biến là các chỉ số tiêu thụ trong lĩnh vực sản suất và chế biến;	$TTi, i = 1, \dots, 19$	M		GSO
05 biến là dịch vụ tiêu dùng và bán lẻ hàng hóa một số ngành kinh tế;	$BLi, i = 1, \dots, 5$	M		Fiinpro
10 biến là các chỉ số lạm phát giá tiêu dùng chung của toàn nền kinh tế và một số mặt hàng	$KPCI, CPI, CPIi,$ $i = 1, \dots, 8$	M		GSO
02 biến là chỉ số giá USD và giá vàng;	$USDI, GOLDI$	M		FRED

05 biến là giá biến đổi thế giới của gạo từ Việt Nam, Thái Lan, cà phê Rubusta, cao su và giá đồng thế giới	$PRICE\_TL$, $PCOFFE$, $PRUPP, PCOPP$	M		FRED
01 biến là chỉ số giá xuất khẩu của thế giới	$PWEX$	M		Europa
17 biến là các chỉ số kim ngạch nhập khẩu của 16 nước phát triển và phần còn lại của toàn thế giới	IMi[9], $i = 1, \dots, 17$	M		Europa
04 biến là chỉ số chứng khoán thế giới và trong nước;	$SP500, DJINDEX$ $VNINDEX, HNX,$	M	Phản ánh sức mạnh của thị trường	Cophieu68
02 biến là tỷ giá hối đoái giữa VND và nhân dân tệ sang USD;	$CHUS, ER$	M		Fred
05 biến là cho vay ngắn, trung và dài hạn và lãi suất huy động;	$LSHD_{NH}, LSHD_{TH},$ $LSHD_{DH}, LSCV_{NH},$ $LSCV\_DH$	M		Fiinpro
03 biến là tiền gửi từ tổ chức và tư nhân và tổng phương tiện thanh toán.	$M2KT, M2DC, M2$	M		Fiinpro

- Các biến ảnh hưởng đến tỷ giá bao gồm tỷ giá hối đoái danh nghĩa giữa VND và USD (*ER*), nhân dân tệ và USD.

- Các biến giá cả bao gồm: chỉ số giá vàng (*GOLDI*) và chỉ số giá USD (02) biến); giá thế giới của gạo Việt Nam, Thái Lan; giá cà phê Robusta, cao su, và giá đồng (05); chỉ số giá xuất khẩu thế giới (01);

- Các biến số được sử dụng xây dựng chỉ số tổng cầu nước ngoài (ED) của nền kinh tế Việt Nam bao gồm 17 biến số kim ngạch nhập khẩu theo giá thực tế của 16 nước đối tác có tỷ trọng nhập khẩu lớn nhất các sản phẩm hàng hóa và dịch vụ của Việt Nam, và phần còn lại của thế giới.

Xét ở khía cạnh cung, cầu, sức mạnh thị trường của nền kinh tế:

- Các biến giải thích phản ánh phía cung bao gồm các biến về chỉ số sản xuất công nghiệp trong một số ngành kinh tế, đầu tư từ ngân sách nhà nước và đầu tư trực

[9] : Bao gồm các quốc gia USA, UK, EU, France, Germany, Italy, Russia, Australia, India, China, Japan, Korea, Malaysia, Singapore, Thailand, Taiwan, và phần còn lại của thế giới.

tiếp nước ngoài, chỉ số tồn kho trong một số ngành sản xuất, dư nợ tín dụng của một số ngành kinh tế, kim ngạch xuất khẩu của một số ngành công nghiệp chế biến chế tạo quan trọng, chỉ số giá xuất khẩu hàng hóa và dịch vụ của Việt Nam, chỉ số người quản trị mua hàng (PMI).

- Các biến giải thích phản ánh phía cầu bao gồm kim ngạch nhập khẩu của một số ngành kinh tế và toàn nền kinh tế, chỉ số tiêu dùng trong một số ngành công nghiệp chế biến chế tạo, tổng mức bán lẻ hàng hóa và doanh thu dịch vụ của toàn nền kinh tế và một số ngành kinh tế, chỉ số giá vàng và đô la, lạm phát giá tiêu dùng của toàn nền kinh tế và của một số rổ hàng hóa và dịch vụ, giá thế giới của gạo Việt Nam và gạo Thái Lan, giá cà phê Robusta thế giới, cao su và đồng, kim ngạch nhập khẩu của các nước đối tác thương mại chính của nền kinh tế và tổng kim ngạch nhập khẩu của thế giới.

- Các biến giải thích phản ánh sức mạnh thị trường bao gồm một số chỉ số chứng khoán trong nước và quốc tế; tỷ giá hối đoái giữa VND/USD và CH/USD; lãi suất tiền gửi và cho vay dài hạn, trung hạn, ngắn hạn; tổng phương tiện thanh toán và tiền gửi từ các tổ chức và khu vực tư nhân. Dữ liệu của các biến giải thích tiềm năng trong Bảng 3.2 bao gồm cả dữ liệu cứng (dữ liệu thống kê) và dữ liệu mềm (thu thập thông qua khảo sát) như chỉ số PMI của người quản trị mua hàng.

Bảng 3.2 cũng cho thấy rằng dữ liệu của tất cả các biến giải thích được thu thập theo tần suất tháng từ 6 nguồn khác nhau, trong đó hai nhà cung cấp dữ liệu chính là Tổng cục Thống kê Việt Nam (GSO) và công ty FiinPro chuyên cung cấp dịch vụ dữ liệu tài chính và kinh doanh.

3.3.3 Dự báo không điều kiện kim ngạch xuất khẩu

Tập dữ liệu EXP là lớn. Để thực hiện dự báo trên tập như vậy, các nhà dự báo kinh tế thường chỉ chọn lựa một vài chỉ số dẫn báo có ý nghĩa thống kê cao để đưa vào mô hình dự báo không điều kiện kim ngạch xuất khẩu. Rõ ràng độ chính xác dự báo theo cách tiếp cận như vậy sẽ bị hạn chế do còn nhiều biến có tác động đến sự thay đổi của kim ngạch xuất khẩu nhưng chưa được đưa vào mô hình dự báo của nó. Hạn chế này dễ dàng được khắc phục bằng ứng dụng thuật toán dự báo không điều kiện trên tập dữ liệu lớn sử dụng phương pháp giảm chiều dựa vào hàm nhân được

đề xuất. Dưới đây sẽ trình bày các kết quả trung gian của việc ứng dụng thuật toán dự báo đó dự báo không điều kiện kim ngạch xuất khẩu hàng tháng của Việt Nam.

3.3.3.1 Giai đoạn 1: Xử lý dữ liệu

Dữ liệu được thu thập từ các nguồn là các cơ quan thống kê chính thức và từ một số công ty chuyên về dịch vụ cung cấp dữ liệu nên các dữ liệu đó cơ bản đã được xử lý, làm sạch. Với tập dữ liệu đã cho, nội dung xử lý dữ liệu ở giai đoạn này là khắc phục tình trạng dữ liệu khuyết thiếu (do dữ liệu được thu thập từ nhiều nguồn khác nhau, nên độ dài dữ liệu không đồng nhất) và chuyển đổi dữ liệu về cùng một dạng có thể so sánh được so với tháng cùng kỳ năm trước.

- *Khắc phục giá trị khuyết thiếu*: Giá trị khuyết thiếu của các biến nếu xảy ra ở những quan sát đầu và/hoặc ở những quan sát cuối sẽ được khắc phục bằng sử dụng phương pháp ngoại suy AR(p) có xu thế hoặc bằng xây dựng mô hình dự báo biến này theo một số biến khác mà theo lý thuyết kinh tế chúng có liên quan với nhau. Ví dụ, biến lạm phát cơ bản của nền kinh tế (KCPI) và lạm phát giá tiêu dùng (CPI) thường có tương quan rất cao. Lạm phát giá tiêu dùng CPI có số liệu đầy đủ từ tháng 1/2013 đến tháng 6/2019, trong khi KCPI chỉ có số liệu từ tháng 4/2015 đến tháng 6/2019. Để có dữ liệu cho KCPI từ tháng 1/2013 đến 3/2015, luận án đã xây dựng một mô hình hồi quy giữa KCPI và CPI chung của toàn nền kinh tế để khắc phục hiện tượng khuyết thiếu dữ liệu này. Nếu giá trị khuyết thiếu không nằm ở các quan sát đầu tiên hoặc cuối cùng, việc khắc phục dữ liệu khuyết thiếu được thực hiện bằng sử dụng phương pháp nội suy chẳng hạn như phương pháp trung bình trượt hoặc phương pháp làm trơn hàm mũ [103] tùy thuộc vào tỷ lệ giá trị khuyết thiếu so với giá trị hiện có là nhiều hay ít. Ví dụ, các chỉ số về tiêu thụ sản phẩm theo tháng trong các ngành sản xuất là khuyết thiếu khá nhiều, bài báo đã sử dụng phương pháp trung bình trượt với số mùa vụ là 12 để xử lý giá trị khuyết thiếu.

- *Chuyển đổi số liệu - xử lý tính không dừng*: Phân tích các biến trong Bảng 3.2 cho thấy phần lớn các biến giải thích là nhận giá trị số tương đối mang tính so sánh với tháng cùng kỳ năm trước như các chỉ số sản xuất công nghiệp, chỉ số tiêu thụ và tồn kho sản phẩm, chỉ số lạm phát giá tiêu dùng,... nhưng cũng có nhiều biến khác nhận giá trị là số tuyệt đối theo giá hiện hành (hay danh nghĩa) không mang ý nghĩa có thể so sánh được như tổng mức bán lẻ, kim ngạch xuất nhập khẩu, giá thế

giới về một số sản phẩm hàng hóa, tỷ giá hối đoái, … và cũng như có một số biến nhận giá trị số tuyệt đối nhưng đã hàm ý có thể so sánh được như các chỉ số chứng khoán, chỉ số PMI.

Để đảm bảo tính lôgic và ý nghĩa nhất quán của dữ liệu, luận án đã chuyển đổi các biến nhận giá trị số tuyệt đối không có ý nghĩa so sánh thành các biến tương ứng của nó nhận giá trị số tương đối mang tính có thể so sánh với tháng cùng năm trước theo công thức sau [112]:

$$DLOG(X_t^*) = LOG\left(\frac{X_t}{X_{t-12}}\right) - LOG\left(\frac{X_{t-1}}{X_{t-13}}\right) \quad (3.10)$$

ở đây, $X_t^* = \frac{X_t}{X_{t-12}}$; X_t là biến số nhận giá trị số tuyệt đối tại thời điểm t; LOG: là logarit của cơ số e.

Việc sử dụng công thức (3.10) để chuyển đổi dữ liệu cũng góp phần xử lý tính không dừng của các biến chuỗi thời gian.

Để xây dựng mô hình và thực hiện kiểm định chấp nhận mô hình được xây dựng, luận án thực hiện chia tập dữ liệu EXP gồm 161 biến với 65 tháng quan sát thành hai tập:

- Tập huấn luyện: gồm 62 tháng quan sát từ tháng 02/2014 đến tháng 03/2019 được sử dụng để xây dựng mô hình dự báo không điều kiện kim ngạch xuất khẩu của Việt Nam theo tần suất tháng.

- Tập kiểm thử: gồm 03 tháng quan sát từ tháng 04/2019 đến tháng 06/2019 được sử dụng để kiểm định chấp nhận mô hình được xây dựng ờ trên.

3.3.3.2 Giai đoạn 2: Xác định các chỉ số dẫn báo

Kiểm định tính dừng của biến kim ngạch xuất khẩu (ký hiệu là *EX*) và 161 biến giải thích trên tập dữ liệu huấn luyện cho thấy hầu hết các biến giải thích và biến phụ thuộc (*EX*) đều dừng (còn được gọi là dừng sai phân bậc 0) và các biến còn lại không dừng nhưng dừng sai phân bậc 1 (tức sai phân bậc 1 của chúng là chuỗi dừng). Khi các biến là chuỗi thời gian dừng thì tính mùa vụ, tính chu kỳ, và tính xu thế của các biến là được xử lý.

Giả sử X là biến dừng sai phân bậc 1, khi đó D(X) là dừng sai phân bậc 0, ở đây D(.) là phép tính sai phân của một biến chuỗi thời gian. Tập dữ liệu mới của tất cả các biến giải thích ban đầu đã được đưa về dừng sai phân bậc 0 được sử dụng để xác định chỉ số dẫn báo của biến phụ thuộc bằng kiểm định nhân quả Granger.

Thực hiện dòng lệnh 3 đến dòng lệnh 5 trong thuật toán *LeadingIndicatorSelection* để kiểm định nhân quả Granger của biến phụ thuộc *EX* với mỗi biến giải thích với độ trễ tối ưu chung được xác định theo lý thuyết kinh tế là 6 như theo gợi ý trong [102]. Với ngưỡng $\alpha < 0.1$, nghĩa là xác suất bác bỏ là $< 10\%$, luận án chọn được 37 biến là các chỉ số dẫn báo của biến phụ thuộc. Bảng 3.3 ở dưới là danh sách các chỉ số dẫn báo có ý nghĩa thống kê đối với biến kim ngạch xuất khẩu *EX*.

Có thể thấy rằng tập các chỉ số dẫn báo vẫn bao gồm khá đầy đủ các yếu tố tác động đến kim ngạch xuất khẩu trong mô hình dự báo cầu xuất khẩu, chẳng hạn các biến liên quan đến giá cả như giá thế giới về đồng PCOPP, cao su PRUBB, giá dầu POIL, chỉ số giá xuất khẩu thể giới PWEX. Các biến số liên quan đến xây dựng chỉ số tổng cầu nước ngoài (ED) của nền kinh tế Việt Nam như IM_TAI, IM_ITA, IM_MAL, IM_IND, …. Tuy nhiên tỷ giá hối đoái giữa VND và USD, giữa nhân dân tệ và USD lại không phải là chỉ số dẫn báo của *EX* của Việt Nam (chi tiết có thể tham khảo Bảng C ở phần Phụ lục). Điều này có thể giải thích là do VND chưa phải là đồng tiền chuyển đổi và tỷ giá hối đoái ở Việt Nam chưa được điều hành theo đúng cơ chế thị trường nên tỷ giá hối đoái VND/USD không có ý nghĩa báo trước cho kim ngạch xuất khẩu của Việt Nam, trong khi đó dù độ mở của nền kinh tế Việt Nam là cao so với GDP nhưng so với tổng thương mại thế giới thì còn rất nhỏ bé (chiếm khoảng 0,2%) nên tác động của tỷ giá hối đoái giữa đồng Nhân dân tệ/Đô la Mỹ (CH/USD) đến kim ngạch xuất khẩu Việt Nam là không rõ rệt.

Một cách tương tự, có thể thấy các chỉ số dẫn báo của kim ngạch xuất khẩu Việt Nam đều phản ánh đầy đủ cả 3 khía cạnh cung, cầu và sức mạnh thị trường của nền kinh tế.

Bảng 3.3: Các chỉ số dẫn báo được chọn của biến EX

No	Biến	Xác suất bác bỏ	No	Biến	Xác suất bác bỏ
1	M2	0.00	20	D(CPI4)	0.05
2	M2KT	0.00	21	NK17	0.06
3	NK9	0.00	22	NK16	0.06
4	M2DC	0.00	23	TT17	0.06
5	IM_TAI	0.01	24	PRUBB	0.07
6	PCOPP	0.01	25	TT3	0.08
7	D(CPI6)	0.01	26	POIL	0.08
8	XK20	0.01	27	NK2	0.08
9	CPI3	0.01	28	PWEX	0.08
10	XK9	0.01	29	IM_INDO	0.08
11	IM_AUS	0.02	30	IM_ITA	0.08
12	IIP	0.02	31	TK13	0.08
13	TK17	0.02	32	XK19	0.09
14	XK6	0.04	33	XK23	0.09
15	NK1	0.04	34	D(CPI8)	0.10
16	XK12	0.04	35	IM_MAL	0.10
17	TK5	0.04	36	IM_IND	0.10
18	D(DN4)	0.04	37	BL4	0.10
19	D(CPI5)	0.05			

Kết thúc Giai đoạn 2, từ tập dữ liệu gồm 161 biến ứng cử viên được thu thập, với xác suất bác bỏ là $< 10\%$ đã loại bỏ 124 biến ứng viên và xác định được 37 biến là các chỉ số dẫn báo của kim ngạch xuất khẩu. Chúng được sử dụng để thay thế cho tập 161 biến các biến ứng viên làm đầu vào cho giai đoạn 3.

3.3.3.3 Giai đoạn 3: Chiết xuất nhân tố và xây dựng mô hình dự báo

Đầu tiên, tập dữ liệu gồm 37 chỉ số dẫn báo được cân chỉnh trung bình và tính khoảng cách trung bình tối thiểu giữa 02 véc tơ cột trên tập dữ liệu này là $\rho_0^2 = 0.3273 \approx e^{-1,12}$. Thực hiện chiết xuất các nhân tố bằng phương pháp KTPCA lặp theo 06 hàm nhân tương ứng được nêu ở cột thứ nhất trong Bảng 3.4 ở dưới và độ trễ

tối ưu được chọn là 6 như theo gợi ý trong [102] với các tập dữ liệu kinh tế - tài chính ở tần suất tháng. Với ngưỡng phần trăm giá trị riêng tích lũy được chọn là 75%, Bảng 3.4 bên dưới trình bày số các nhân tố được chọn, tỷ lệ phần trăm tích lũy giá trị riêng, và $RMSE$ của mô hình dự báo không điều kiện của biến kim ngạch xuất khẩu *EX*.

Dòng đầu tiên trong Bảng 3.4 là kết quả chiết xuất nhân tố bằng phương pháp KTPCA với hàm nhân đa thức là tích vô hướng của hai véc tơ $\boldsymbol{\kappa_1}(x_i, x_j) = \langle x_i, x_j \rangle$. Khi đó nó cũng là kết quả chiết xuất nhân tố bằng sử dụng phương pháp PCA. Kết quả cho thấy với ngưỡng phương sai tích lũy là 75%, thì cần phải chọn 12 nhân tố để đạt tỷ lệ phần trăm phương sai tích lũy trên ngưỡng này và là 77.01%. Ta không thể xây dựng được mô hình dự báo kim ngạch xuất khẩu *EX* theo mô hình (1.16) trên 12 nhân tố được chọn bởi vì với độ trễ tối ưu chung là 6 thì số lượng biến trong mô hình dự báo *EX* = 12*6 (số biến trễ) + 6 (biến trễ của EX) = 78 biến, trong khi số quan sát của tập dữ liệu EXP chỉ 63 quan sát. Tương tự, các hàm nhân Gauss GA_5 và GA_6 với tỷ lệ phần trăm giá trị riêng tích lũy lần lượt là 75.20% và 76.41% thì số nhân tố được chọn tương ứng là 9 và 10 nhân tố cũng cho kết quả tương tự.

Bảng 3.4: Kết quả giảm chiều bằng phương pháp KTPCA LẶP

Hàm nhân	*Dạng hàm nhân*	*Số các nhân tố*	*Tỉ lệ tích lũy (%)*	*RMSE*
Đa thức	$\boldsymbol{\kappa_1}(x_i, x_j) = \langle x_i, x_j \rangle$ hay PCA	12	77.01	Không tiếp tục
	$\boldsymbol{\kappa_2}(x_i, x_j) = \langle x_i, x_j \rangle^{\mathbf{2}}$	2	83.34	0.0228
	$\boldsymbol{\kappa_3}(x_i, x_j) = \langle x_i, x_j \rangle^{\mathbf{3}}$	1	74.83	0.0270
Gauss	$\boldsymbol{GA_4}$: $\rho^2 = e^{-0.4}$	5	76.03	0.0202
	$\boldsymbol{GA_5}$: $\rho^2 = e^{-1,12}$	9	75.20	Không tiếp tục
	$\boldsymbol{GA_6}$: $\rho^2 = e^{-1,2}$	10	76.41	Không tiếp tục

Sự kiện cần phải có một số khá lớn (trên 6) nhân tố mới đạt được ngưỡng phần trăm phương sai tích lũy là 75% thể hiện rằng tập dữ liệu của 37 chỉ số dẫn báo đầu vào là không xấp xỉ một siêu phẳng [37].

Bảng 3.4 cho thấy hàm nhân phù hợp nhất trong số 06 hàm nhân thực nghiệm là hàm nhân Gauss GA_4 với tham số $\rho^2 = e^{-0,4}$ với *RMSE* của mô hình bằng 0.0202

là thấp nhất và 05 nhân tố dẫn báo được chọn để thay thế cho tập dữ liệu gồm 37 chỉ số dẫn báo. Tập dữ liệu của 05 nhân tố này nắm giữ 76.03% thông tin của tập dữ liệu của 37 chỉ số dẫn báo được chọn.

Kiểm tra tính dừng của 5 nhân tố cho thấy tất cả các nhân tố đều dừng. Mô hình dự báo không điều kiện kim ngạch xuất khẩu theo tháng của Việt Nam theo 05 nhân tố có dạng như sau:

$$EX = \underset{0.113}{-0.807EX(-1)***} + \underset{0.099}{0.221EX(-3))**} \underset{0.094}{-0.223\,EX(-4)**} \underset{0.034}{-0.130\,FAC_1(-1)***}$$
$$\underset{0.017}{-0.037FAC_2(-1)**} \underset{0.017}{-0.077FAC_2(-3)***} \underset{0.012}{-0.041FAC_3(-1)***} \underset{0.018}{+0.030FAC_3(-2)*}$$
$$\underset{0.017}{+0.055FAC_3(-3)***} \underset{0.018}{-0.056FAC_3(-5)***} \underset{0.015}{+0.038FAC_3(-6)**} \underset{0.007}{+0.023FAC_4(-1)***}$$
$$\underset{0.007}{-0.013FAC_4(-2)*} \underset{0.007}{-0.026FAC_4(-3)***} \underset{0.008}{+0.030FAC_5(-5)***} \quad (3.11)$$

R^2= 0.8481 D-W stat*:* 1.9420 SMPL*:* 56 sau khi hiệu chỉnh độ trễ các biến.
ở đây, dấu hoa thị cho biết ý nghĩa thống kê của Thống kê t: ***, **, và * biểu thị ý nghĩa của thống kê T tương ứng ở các mức 1%, 5%, và 10%.

3.3.3.4 Giai đoạn 4: Thực hiện dự báo

Giai đoạn này thực hiện dự báo kiểm định chấp nhận mô hình dự báo được xây dựng ở Giai đoạn 3 cũng như để dự báo ngoài mẫu của biến phụ thuộc khi mô hình dự báo được xây dựng ở Giai đoạn đó được cập nhật lại trên toàn bộ các quan sát của tập dữ liệu của các chỉ số dẫn báo.

Dự báo kiểm định chấp nhận mô hình được tiến hành trên tập dữ liệu kiểm thử (testing data set) bao gồm 03 quan sát là các tháng 4/2019, 5/2019, và 6/2019 bằng sử dụng thuật toán *Calculate*(YF, SPC) với mô hình (3.11). Có thể thấy rằng theo mô hình (3.11) và chỉ dựa vào các chỉ số dẫn báo ở tháng hiện tại thì ta chỉ có thể dự báo được kim ngạch xuất khẩu *EX* ở 01 tháng tiếp theo, tức là tháng 4/2019.

Khi biết giá trị của 37 chỉ số dẫn báo ở các tháng 4/2019 và tháng 5/2019, sử dụng phương pháp KTPCA với hàm nhân Gauss với tham số $\rho^2 = e^{-0.4}$ để chiết xuất 5 nhân tố tương ứng có các quan sát gồm cả các tháng 4/2019 và 5/2019, và sử

dụng mô hình (3.11) để dự báo kim ngạch xuất khẩu Việt Nam tương ứng đến tháng 5/2019 và tháng 6/2019.

Các dự báo không điều kiện của kim ngạch xuất khẩu Việt Nam ở các tháng 4/2019, 5/2019 và 6/2019 được thực hiện theo cách như vậy được so sánh với giá trị thống kê thực tế của kim ngạch xuất khẩu ở các tháng này và so với các kết quả dự báo bởi một số mô hình đơn biến điển hình khác bao gồm mô hình AR(p) có xu thế bậc hai với p = 6, mô hình ARIMA, và mô hình Holt -Winter. Kết quả được trình bày trong Bảng 3.5 bên dưới, trong đó ký hiệu *EXF* là giá trị dự báo của *EX* bởi mô hình (3.11), và các mô hình đơn biến AR, ARIMA, và Holt – Winter.

Bảng 3.5: So sánh kết quả dự báo kim ngạch xuất khẩu của các mô hình với thực tế

Mô hình		**Mô hình đề xuất**		**AR(6)**	
Tháng	*EX*	*EXF*	% sai số dự báo	*EXF*	% sai số dự báo
04/2019	20439.83	20299.12	0.69	18891.92	7.57
05/2019	21904.59	21173.66	3.34	20724.46	5.39
06/2019	21427.77	21418.12	0.05	20211.47	5.68
		$RMSE_{OUT}$ = **429.79**	abs(% sai số db) TB = **1.36**	$RMSE_{OUT}$ = 1325.16	abs(% sai số db) TB =6.21
Mô hình		**ARIMA(2, 1, 2)**		**Holt -Winter Add**	
Tháng	*EX*	*EXF*	% sai số dự báo	*EXF*	% sai số dự báo
04/2019	20439.83	19238.68	5.88	19389.46	5.14
05/2019	21904.59	21213.68	3.15	20644.72	5.75
06/2019	21427.77	20958.26	2.19	20349.69	5.03
		$RMSE_{OUT}$= ***844.70***	abs(% sai số db) TB = ***3.74***	$RMSE_{OUT}$= 1133.26	abs(% sai số db) TB = 5.31

ở đây, % sai số dự báo bằng trị tuyệt đối của (giá trị thống kê thực tế - giá trị dự báo)* 100/giá trị thống kê thực tế, $RMSE_{OUT}$ được tính toán theo công thức sau:

$$RMSE_{OUT} = \sqrt{\frac{1}{h}.\sum_{j=1}^{h}\left(Y_{t+j} - \hat{Y}_{t+j}\right)^2} \qquad (3.12)$$

ở đây, $\hat{Y}_{t+1}, \hat{Y}_{t+2}, \dots,$ và $\hat{Y}_{t+h}$ là các giá trị dự báo ngoài mẫu của Y_t. Abs(% sai số db)TB là trung bình của trị tuyệt đối của phần trăm sai số dự báo ở 3 tháng 4, 5, và 6 năm 2019.

Giá trị $RMSE_{OUT}$ càng nhỏ hoặc abs(% sai số db) TB càng nhỏ, độ chính xác của dự báo ngoài mẫu của mô hình càng cao, trong khi đó RMSE càng nhỏ, độ chính xác dự báo trong mẫu của mô hình càng cao.

Bảng 3.5 cho thấy độ chính xác dự báo của mô hình dự báo không điều kiện kim ngạch xuất khẩu được xây dựng dựa vào mô hình ARDL nhân tố theo thuật toán không điều kiện sử dụng phương pháp giảm chiều được đề xuất là cao hơn nhiều độ chính xác dự báo của các mô hình dự báo không điều kiện đơn biến AR(p), ARIMA, và Holt -Winter. % sai số dự báo trung bình 3 tháng 4/2019, 5/2019, và 6/2019 của mô hình dự báo kim ngạch xuất khẩu sử dụng thuật toán không điều kiện cao hơn % sai số dự báo của mô hình dự báo không điều kiện được xây dựng dựa vào mô hình dự báo đơn biến tốt nhất là ARIMA(2,1,2) đến 2.38 điểm %, làm tăng độ chính xác dự báo đến 63.6%.

Vì vậy, ta có thể chấp nhận mô hình dự báo được xây dựng và sử dụng mô hình này để dự báo kim ngạch xuất khẩu cho các tháng ngoài mẫu tiếp theo như tháng 7/2019.

3.3.3.5 Dự báo ngoài mẫu kim ngạch xuất khẩu

Dữ liệu phục vụ dự báo kim ngạch xuất khẩu bao gồm những quan sát đến tháng 6/2019. Để dự báo kim ngạch xuất khẩu ở tháng tiếp theo cần thực hiện những nội dung sau:

- Cập nhật bổ sung các quan sát đến tháng 6/2019 cho 5 nhân tố được chiết xuất từ 37 chỉ số dẫn báo bằng phương pháp KTPCA với hàm nhân Gauss có tham số $\rho^2 = e^{-0.4}$.

- Ước lượng lại mô hình (3.11) với các nhân tố dẫn báo có số quan sát đến tháng 6/2019;

- Sử dụng mô hình (3.11) vừa được ước lượng lại để dự báo kim ngạch xuất khẩu Việt Nam ở tháng 7/2019.

Để thực hiện các nội dung này bằng quy trình dự báo không điều kiện thì chỉ cần thực hiện lại các Giai đoạn 3 và Giai đoạn 4 với hàm nhân Gauss có tham số $\rho^2 = e^{-0.4}$ theo cách tương tự như đã được trình bày ở trên.

Phần tiếp theo dưới đây sẽ trình bày việc dự báo kim ngạch xuất khẩu của Việt Nam hàng tháng bằng sử dụng thuật toán có điều kiện **CONF** trên tập dữ liệu lớn các biến kinh tế - tài chính.

3.3.4 Dự báo có điều kiện kim ngạch xuất khẩu

Tập dữ liệu **EXP** ở trên cũng được sử dụng để thực hiện dự báo kim ngạch xuất khẩu ở nhiều tháng tiếp theo, chẳng hạn là 03 tháng.

3.3.4.1 Giai đoạn 1: Xử lý dữ liệu

Nội dung xử lý dữ liệu ở giai đoạn này giống như nội dung của giai đoạn cùng tên trong quy trình/thuật toán dự báo không điều kiện nên được bỏ qua không trình bày lại.

3.3.4.2 Giai đoạn 2: Lựa chọn biến

Mô hình dự báo *EX* trước hết được xây dựng trên tập dữ liệu huấn luyện từ tháng 2/2014 đến tháng 3/2019 gồm 62 quan sát và việc dự báo kiểm định chấp nhận mô hình này được thực hiện trên tập dữ liệu kiểm thử gồm 3 tháng tiếp theo.

Nội dung cần thực hiện ở Giai đoạn này là loại bỏ những biến không hoặc ít liên quan đến sự biến động của biến phụ thuộc (các biến như vậy được gọi là nhiễu) và loại bỏ những biến dư thừa đối với mục đích dự báo. Trên tập dữ liệu huấn luyện, với ngưỡng liên quan và dư thừa lần lượt là 0.2 và 0.9, bằng sử dụng thuật toán *FeatureSelection*, 98 biến ứng cử viên không liên quan hoặc dư thừa đã bị loại bỏ và chỉ có 63 biến là có liên quan và không dư thừa đối với mục đích dự báo kim ngạch xuất khẩu *EX*. Các biến này được trình bày trong Bảng 3.6 ở dưới. Các biến này là các biến giải thích đầu vào được sử dụng để xây dựng mô hình dự báo có điều kiện của biến *EX*.

Khác với các chỉ số dẫn báo trong Bảng 3.4, các biến giải thích trong Bảng 3.6 được gọi là chỉ số báo đồng thời của kim ngạch xuất khẩu *EX*, nghĩa là sự thay đổi của nó có ảnh hưởng đến sự thay đổi đồng thời của biến *EX*. Phân tích Bảng 3.6 có thể thấy yếu tố phản ánh phía cung (chẳng hạn XK1, XK11, XK2, …) và phía cầu

của nền kinh tế (NK, NK17, NK18, …) có ảnh hưởng rất mạnh đến sự thay đổi của biến *EX*. Chi tiết có thể tham khảo Bảng C ở phần Phụ lục. Tập dữ liệu của 63 biến này được chọn thay thế cho tập dữ liệu 161 biến ứng cử viên ban đầu, tuy nhiên 63 biến này vẫn còn khá lớn nên cần thực hiện giảm chiều trước khi xây dựng mô hình dự báo.

Bảng 3.6: Các biến liên quan, không dư thừa với chỉ số kim ngạch xuất khẩu

No	*Biến*	*Hệ số TQ Pearson*	*No*	*Biến*	*Hệ số TQ Pearson*	*No*	*Biến*	*Hệ số TQ Pearson*
1	XK1	0.931	22	XK6	0.496	43	XK8	0.260
2	XK11	0.846	23	IM_CHI	0.493	44	IIP9	0.252
3	XK2	0.806	24	XK18	0.485	45	XK15	0.239
4	NK	0.780	25	NK11	0.470	46	IIP8	0.239
5	XK16	0.750	26	XK7	0.457	47	IM_MAL	0.207
6	XK17	0.723	27	NK7	0.401	48	IM_GER	-0.230
7	XK20	0.694	28	NK2	0.399	49	TT5	-0.234
8	XK14	0.693	29	IIP2	0.372	50	IM_AUS	-0.241
9	NK10	0.690	30	NK9	0.370	51	TT6	-0.257
10	NK18	0.685	31	NK15	0.368	52	TBL	-0.261
11	XK5	0.653	32	NK14	0.359	53	TT2	-0.272
12	XK19	0.647	33	NK3	0.354	54	BL1	-0.274
13	XK22	0.629	34	IIP5	0.347	55	IM_SIN	-0.284
14	NK13	0.608	35	IIP	0.335	56	TT14	-0.289
15	XK21	0.603	36	IIP4	0.334	57	BL4	-0.303
16	XK23	0.580	37	NK4	0.318	58	TT8	-0.314
17	INVES	0.563	38	XK3	0.302	59	TT19	-0.336
18	XK4	0.545	39	IIP6	0.299	60	TT13	-0.358
19	NK12	0.519	40	XK10	0.291	61	TT12	-0.368
20	XK24	0.514	41	XK9	0.281	62	TT11	-0.406
21	XK12	0.498	42	NK5	0.267	63	TT1	-0.475

Khoảng cách tối thiểu trung bình giữa hai véc tơ dữ liệu trong tập dữ liệu đầu vào là 0.569, tức là $\rho_0^2 = 0.569$.

Tập dữ liệu của 63 biến trong Bảng 3.6 cũng được đặt tên là EXP và được sử dụng để thực nghiệm so sánh hiệu suất giảm chiều của phương pháp được đề xuất với các phương pháp giảm chiều PCA, SPCA, RSPCA, và ROBSPCA (xem Bảng 2.2).

3.3.4.3 Giai đoạn 3: Chiết xuất nhân tố bằng sử dụng phương pháp KTPCA LẶP

Với ngưỡng phần trăm giá trị riêng tích lũy là 75% và độ trễ tối ưu chung của các nhân tố trong mô hình ước lượng được xác định theo gợi ý trong [102] và là 6. Kết quả chiết xuất nhân tố bằng phương pháp KTPCA lặp được thể hiện trong Bảng 3.7, trong đó dòng thứ nhất là kết quả chiết xuất nhân tố sử dụng phương pháp PCA. Dòng này cho thấy số lượng nhân tố được chọn là 14. Bởi vì số lượng quan sát của tập dữ liệu đầu vào là 62, do đó không thể hồi quy biến *EX* trên 14 nhân tố đã chọn với độ trễ của chung đều là 6.

Bảng 3.7: Chiết xuất nhân tố bằng phương pháp KTPCA lặp

Hàm nhân $\boldsymbol{\kappa}$	*Các tham số*	*Số nhân tố*	*% Trị riêng tích lũy*	*RMSE*
(PCA)	$\kappa_0(,)$: d = 1, c = 0	14	76.72	Không tiếp tục
	$\kappa_1(,)$: d = 2, c = 0	5	76.02	0.0153
Đa thức	$\kappa_2(,)$: d = 3, c = 0	2	81.97	0.0270
	$\kappa_3(,)$; $\rho^2 = 0.569$	10	75.56	Không tiếp tục
Gauss	$\kappa_4(,)$: $\rho^2 = 0.833$	6	76.16	0.0104
	$\kappa_5(,)$: $\rho^2 = 0.500$	12	76.09	Không tiếp tục

Đối với các hàm nhân Gauss $\kappa_3(X_i, X_j)$ và $\kappa_5(X_i, X_j)$, ta cũng nhận được kết quả tương tự.

Bảng 3.7 cũng chỉ ra rằng hàm nhân $\kappa_4(X_i, X_j)$ là phù hợp nhất trong số các hàm nhân được thực nghiệm vì RMSE của mô hình dự báo biến *EX* trên các nhân tố được chọn bằng sử dụng phương pháp KTPCA với hàm nhân này là nhỏ nhất và bằng 0.0104 và tham số ρ^2 trong hàm nhân này không phải là khoảng cách trung bình tối thiểu của 2 véc tơ cột trong tập dữ liệu đầu vào.

Mô hình dự báo có điều kiện được sử dụng để dự báo kim ngạch xuất khẩu của Việt Nam theo tháng là mô hình ARDL nhân tố theo phương trình (1.33). Kết

thúc quá trình lặp, ta nhận được mô hình dự báo kim ngạch xuất khẩu có RMSE nhỏ nhất có dạng như sau:

$$EX = -0.111FC1^{***} + 0.023FC2^{**} - 0.029FC2(-4)^{***} - 0.017FC2(-5)^{**} + 0.030FC3(-1)^{**}$$
(0.015) (0.010) (0.008) (0.007) (0.013)

$$+ 0.045FC3(-2)^{***} - 0.034FC4^{***} + 0.020FC4(-3)^{**} - 0.044FC4(-6)^{***} - 0.030FC5(-3)^{***}$$
(0.013) (0.008) (0.009) (0.008) (0.007)

$$- 0.029FC5(-5)^{***} + 0.026FC6(-3)^{***} + 0.018FC6(-5)^{*}$$
(0.009) (0.010) (0.010) (3.13)

R^2*: 0.9068* D-W stat*: 2.3369* SMPL*: 56* sau khi điều chỉnh độ trễ các biến

3.3.4.4 Giai đoạn 4: Xây dựng mô hình dự báo phụ và thực hiện dự báo

Để chấp nhận mô hình dự báo biến *EX* theo mô hình (3.13), ta cần thực hiện dự báo kiểm định chấp nhận mô hình bằng cách sử dụng mô hình được xây dựng để dự báo kim ngạch xuất khẩu ở 3 tháng tiếp theo và so sánh kết quả dự báo với dữ liệu thống kê thực tế trong tập dữ liệu kiểm thử. Giai đoạn 4 của quy trình dự báo có điều kiện cho phép thực hiện nội dung như vậy. Theo đó trước hết cần xây dựng các mô hình phụ để dự báo các nhân tố ngoại sinh trong mô hình (3.13).

a. Dự báo các nhân tố trong mô hình dự báo được xây dựng

Mô hình dự báo phụ của các nhân tố trong mô hình (3.13) được xây dựng dựa vào mô hình AR(p) có xu thế theo phương trình (3.1). Bảng 3.8 dưới đây trình bày các kết quả dự báo của 06 nhân tố ở các tháng 4, 5 và 6 năm 2019.

Bảng 3.8: Kết quả dự báo 06 nhân tố

Tháng	*FC1F*	*FC2F*	*FC3F*	*FC4F*	*FC5F*	*FC6F*
04/2019	0.1315	-0.0287	0.1855	-0.2030	-0.0265	0.1604
05/2019	0.0124	-0.0039	-0.1255	0.0209	0.0106	-0.0695
06/2019	0.0543	0.0113	0.0113	0.0666	0.0016	-0.0419

b. Xây dựng mô hình cầu xuất khẩu và dự báo các biến ngoại sinh trong mô hình

Để so sánh, đánh giá độ chính xác dự báo kim ngạch xuất khẩu *EX* bằng sử dụng mô hình dự báo có điều kiện được đề xuất, luận án cũng thực hiện dự báo *EX*

bằng sử dụng mô hình dự báo được xây dựng dựa vào mô hình cầu xuất khẩu được giới thiệu ở mục 3.3.2.1.

Mô hình cầu xuất khẩu tổng quát có dạng như phương trình (3.9) [110]. Cụ thể, các biến giải thích được đưa vào mô hình gồm:

- Chỉ số ED_t được tính như sau: $ED_t = \sum n_k * lM_k$, trong đó n_k là tỷ trọng nhập khẩu của quốc gia thứ k về hàng hóa và dịch vụ của quốc gia được dự báo trong tổng kim ngạch xuất khẩu của quốc gia này, lM_k là tốc độ tăng nhập khẩu của nước thứ k.

- Do giá xuất khẩu tương đối là yếu tố quan trọng quyết định hoạt động xuất khẩu của nền kinh tế [109] nên giá xuất khẩu tương đối cần được thêm vào mô hình cầu xuất khẩu. Trong luận án này, $PEX/PWEX$ là giá xuất khẩu tương đối, trong đó *PEX* và *PWEX* lần lượt là chỉ số giá xuất khẩu của Việt Nam và thế giới. Véc tơ giá trong mô hình cầu xuất khẩu trong luận án này gồm giá thế giới về dầu thô (POIL) và gạo của Việt Nam (PRICE_VN).

Kiểm tra tính dừng của các biến ER, ED, *POIL*, *PRICE_VN*, $PEX/PWEX$ cho thấy rằng chúng đều là chuỗi thời gian dừng. Mô hình dự báo biến *EX* dựa vào mô hình cầu xuất khẩu (gọi tắt là mô hình cầu xuất khẩu) với độ trễ tối ưu chung là 6 [102] có dạng:

$$EX = -0.99\,EX(-1)^{***} - 0.77\,EX(-2)^{***} - 0.3\,EX(-3)^{***} - 0.20\,EX(-4)^{***} + 1.91\,ER(-1)^{**}$$
$$(0.124) \quad (0.168) \quad (1.180) \quad (1.099) \quad (0.816)$$
$$+2.36\,ER(-3)^{***} + 1.78\,ER(-4)^{**} + 1.56\,ER(-5)^{*} - 2.55\,ER(-6)^{***} - 0.17\,ED^{***} - 0.10\,ED(-1)^{*}$$
$$(0.848) \quad (0.798) \quad (0.927) \quad (0.915) \quad (0.047) \quad (0.055)$$
$$- 0.13\,ED(-2)^{***} + 0.07\,ED(-4)^{*} + 0.13\,ED(-5)^{***} + 0.58\,PEX/PWEX^{***}$$
$$(0.047) \quad (0.039) \quad (0.039) \quad (0.057)$$
$$- 0.31\,PEX/PWEX(-1)^{***} - 0.27\,PEX/PWEX(-3)^{***} + 0.14\,POIL^{***} + 0.16\,POIL(-1)^{***}$$
$$(0.093) \quad (0.075) \quad (0.078) \quad (0.041)$$
$$- 0.10\,POIL(-5)^{**} + 0.14\,PRICE\_VN(-3)^{*} + 0.34\,PRICE\_VN(-4)^{***} \quad (3.14)$$
$$(0.039) \quad (0.078) \quad (0.101)$$

R^2*: 0.9367* D-W stat*: 2.0113* SMPL*: 56* sau khi điều chỉnh độ trễ các biến

Các giá trị dự báo của các biến trong phương trình (3.14) ở 3 tháng 4, 5, và 6 năm 2019 được trình bày trong Bảng 3.9.

Bảng 3.9: Dự báo của các biến giải thích của mô hình cầu xuất khẩu

Tháng	ERF	EDF	POILF	PRICE_VNF	PEX/PWEX
04/2019	-0.0009	0.0011	-0.0067	-0.0169	0.9820
05/2019	-0.0002	-0.0018	0.0024	-0.00463	0.9461
06/2019	-0.0005	0.0009	0.0055	0.0045	0.9877

c. Thực hiện dự báo kiểm định và so sánh, đánh giá

Phân tích thống kê Jarque-Bera và Kurtosis của biến *EX*, 06 nhân tố trong mô hình dự báo có điều kiện và 05 biến ngoại sinh trong mô hình cầu xuất khẩu được thể hiện trong Bảng 3.10 ở dưới cho thấy các phân bố xác suất của *FC3, FC4, FC5, FC6* trong mô hình đề xuất và *POIL, PRICE_VN* trong mô hình cầu xuất khẩu có thể được xem là phân phối chuẩn, trong khi các nhân tố và biến còn lại thì không phải như vậy. Độ đo Skewness cũng cho thấy phân phối xác suất của các biến *EX* và *FC4, FC5, ER, POIL* và *PRICE_VN* lệch về phía trái trong khi của các nhân tố và các biến *FC1, FC2, FC3, FC6*, *ED* và *PEX/PWEX* lệch về phía phải. Mặt khác, tất cả nhân tố và các biến ngoại sinh tương ứng trong mô hình đề xuất và mô hình cầu xuất khẩu đều là chuỗi thời gian dừng. Điều này đảm bảo không có hiện tượng hồi quy giả mạo khi xây dựng các mô hình dự báo kim ngạch xuất khẩu (*EX)* theo hai cách tiếp cận khác nhau [51].

Bảng 3.10: Đặc trưng thống kê của các biến ngoại sinh

Mô hình đề xuất						
	FC1	*FC2*	*FC3*	*FC4*	*FC5*	*FC6*
Trung bình	0	0	0	0	0	0
Lớn nhất	0.90	1.17	0.50	0.49	0.55	0.49
Nhỏ nhất	-0.62	-0.99	-0.48	-0.64	-0.64	-0.52
Độ lệch chuẩn	0.22	0.35	0.2	0.22	0.26	0.19
Skewness	0.34	0.34	0.36	-0.39	-0.09	0.13
Kurtosis	8.33	7.04	3.31	2.9	2.64	2.93
Jarq-Bera	74.61	43.33	1.57	1.60	0.41	0.20
Xác suất	0.00	0.00	0.46	0.45	0.82	0.91
Số quan sát	62	62	62	62	62	62

Tính dừng	Có	Có	Có	Có	Có	Có
Mô hình cầu xuất khẩu						
	EX	*ER*	*ED*	*POIL*	*PRICE_VN*	*PEX/PWEX*
Trung bình	0	0	0	0	0	0.90
Lớn nhất	0.11	0.01	0.22	0.14	0.07	1.12
Nhỏ nhất	-0.11	-0.01	-0.24	-0.13	-0.06	0.83
Độ lệch chuẩn	0.04	0	0.06	0.05	0.02	0.05
Skewness	-0.14	-0.01	0.04	-0.11	-0.12	1.50
Kurtosis	4.27	6.02	12.92	3.46	3.66	6.58
Jarq-Bera	4.38	23.61	254.41	0.67	1.28	56.18
Xác suất	0.11	0	0	0.72	0.53	0
Số quan sát	62	62	62	62	62	62
Tính dừng	Có	Có	Có	Có	Có	Có

Ký hiệu *EXF* và *DEXF* lần lượt là các giá trị dự báo của biến *EX* bằng mô hình được đề xuất và mô hình cầu xuất khẩu. Giá trị của *EXF* và *DEXF* trong ba tháng 4, tháng 5 và tháng 6 năm 2019 được dự báo bằng sử dụng tương ứng các mô hình (3.12) và (3.13), ở đây giá trị của các biến ngoại sinh trong hai mô hình được trình bày trong Bảng 3.8 và Bảng 3.9.

Kết quả dự báo *EX* của tháng 4, 5 và 6 năm 2019 theo hai cách tiếp cận nêu trên với các giá trị thống kê thực tế được trình bày trong Bảng 3.11 dưới đây, trong đó *RMSE* được tính toán theo công thức (1.45).

Bảng 3.11 cho thấy rằng % sai số dự báo, *RMSE*, $RMSE_{OUT}$ của mô hình đề xuất luôn nhỏ hơn mô hình cầu xuất khẩu cho thấy độ chính xác dự báo trong mẫu cũng như ngoài mẫu của mô hình đề xuất là cao hơn so với mô hình cầu xuất khẩu. Mặt khác, xu hướng biến động của kim ngạch xuất khẩu (*EX*) thực tế và được dự báo theo mô hình đề xuất (*EXF*) là giống nhau, trong khi theo mô hình cầu xuất khẩu thì không như vậy. Với kết quả này ta có thể chấp nhận mô hình dự báo đề xuất và sử dụng nó để dự báo và hình thành các kịch bản dự báo về kim ngạch xuất khẩu của Việt Nam.

Bảng 3.11: So sánh kết quả dự báo kim ngạch xuất khẩu với thực tế

	Mô hình đề xuất			***Mô hình cầu xuất khẩu***	
Tháng	*EX*	*EXF*	% sai số dự báo	*DEXF*	% sai số dự báo
04/2019	20439.83	20051.57	1.90	19757.77	3.34
05/2019	21904.59	21603.89	1.37	21464.56	2.01
06/2019	21427.77	21203.48	1.05	22246.80	3.82
		% sai số dự báo TB = **1.44**		% sai số dự báo TB = **3.06**	
RMSE		0.0104		0.0261	
$RMSE_{OUT}$		0.0038		0.0296	

Tính trung bình % sai số dự báo kim ngạch xuất khẩu của 3 tháng 4, 5, và 6 năm 2019 bằng sử dụng mô hình dự báo có điều kiện được đề xuất và mô hình cầu xuất khẩu với cùng những điều kiện giả định (các yếu tố tác động đến xuất khẩu ở 3 tháng 4, 5, và 6 năm 2019 không có những biến động bất thường) thì độ chính xác dự báo của mô hình được xây dựng theo thuật toán có điều kiện có độ chính xác dự báo cao hơn độ chính xác dự báo của mô hình cầu xuất khẩu là 1.62 điểm %, cải thiệt độ chính xác dự báo lên đến 52.9%.

Ở đây xin nhấn mạnh thêm rằng mô hình phương trình (3.14) ở trên là một dạng riêng của mô hình ARDL nhân tố bằng cách loại bỏ các trễ của biến phụ thuộc *EX* . Mô hình này được sử dụng để đánh giá tác động của các biến giải thích đến biến phụ thuộc *EX*. Nếu bổ sung đầy đủ các trễ của biến phụ thuộc vào mô hình (3.14) để đạt độ dài trễ tối ưu thì độ chính xác dự báo của mô hình được xây dựng có thể còn cao hơn nữa.

3.3.4.5 Dự báo kim ngạch xuất khẩu và xây dựng các kịch bản dự báo

a. Dự báo ngoài mẫu kim ngạch xuất khẩu

Tương tự như dự báo ngoài mẫu kim ngạch xuất khẩu theo cách tiếp cận sử dụng mô hình dự báo không điều kiện, để dự báo có điều kiện biến này cũng cần thực hiện các nội dung sau:

- Cập nhật bổ sung các quan sát đến tháng 6/2019 cho 6 nhân tố được chiết xuất từ 63 biến giải thích bằng phương pháp KTPCA với hàm nhân Gauss có tham số $\rho^2 = e^{0.833}$.

- Ước lượng lại mô hình (3.14) với các nhân tố có số quan sát đến tháng 6/2019;

- Dự báo các nhân tố trong mô hình (3.14) cho 3 tháng tiếp theo.

- Sử dụng mô hình (3.14) vừa được cập nhật và kết quả dự báo của các nhân tố trong mô hình đó để dự báo kim ngạch xuất khẩu Việt Nam ở 3 tháng tiếp theo.

Như vậy để thực hiện các nội dung đó bằng mô hình dự báo có điều kiện ta chỉ cần thực hiện lại các Giai đoạn 3 và Giai đoạn 4 trong quy trình dự báo có điều kiện với hàm nhân Gauss có tham số $\rho^2 = e^{0.833}$. Vì thế luận án đã bỏ qua không trình bày lại những nội dung như vậy.

b. Xây dựng các kịch bản dự báo

Như đã biết dự báo bằng mô hình định lượng là thừa nhận rằng tương lai diễn ra gần giống như hiện tại và quá khứ. Nhưng thực tế cuộc sống không phải luôn như vậy nhất là trong bối cảnh toàn cầu hóa kinh tế như hiện nay. Có rất nhiều biến động khó lường tác động đến hoạt động xuất khẩu của Việt Nam. Để đối phó với thực tiễn ấy khi thực hiện dự báo có điều kiện, người ta thường thực hiện theo một trong 3 cách tiếp cận như sau:

- *Cách tiếp cận thứ nhấ*t: Nếu người dự báo cảm nhận rằng tương lai không có những sự kiện hoặc biến động đặc biệt có ảnh hưởng đến biến cần được dự báo thì kết quả dự báo được thực hiện bằng sử dụng thuật toán dự báo có điều kiện như được trình bày ở trên sẽ được sử dụng làm kết quả dự báo chính thức.

Khi người làm dự báo cho rằng tương lai có thể có những biến động bất thường làm cho biến cần được dự báo biến động không giống như quy luật của hiện tại và quá khứ thì người ta thường sử dụng một trong hai cách tiếp cận là: kết hợp dự báo bằng mô hình định lượng với phương pháp phán xử hoặc tiến hành dự báo theo các kịch bản.

- *Cách tiếp cận thứ hai: kết hợp dự báo bằng mô hình định lượng và phương pháp phán xử*: Cách tiếp cận này được đề xuất dựa vào thực tế là có nhiều yếu tố mới có tác động đến biến cần được dự báo, nhưng không thể đưa các yếu tố này vào mô hình định lượng vì không có hoặc thiếu dữ liệu về nó. Khi đó người làm dự báo sử

dụng tri thức miền ứng dụng để đánh giá các yếu tố đó tác động tích cực, tiêu cực thế nào đến biến cần được dự báo để điều chỉnh (lên hoặc xuống) kết quả dự báo. Việc điều chỉnh đó phụ thuộc chủ quan vào tri thức, kinh nghiệm của người làm dự báo. Trong lĩnh vực kinh tế - tài chính, cách tiếp cận này thực tế được ứng dụng nhiều nhất và phổ biến nhất trong các đơn vị chuyên về công tác phân tích dự báo kinh tế - tài chính. Có thể tham khảo chi tiết quy trình và nguyên tắc kết hợp dự báo bằng mô hình định lượng với phương pháp phán xử trong [78].

- *Cách tiếp cận dự báo theo kịch bản*: Khi việc đánh giá hiệu quả tác động của các yếu tố mới đến biến cần được dự báo gặp nhiều khó khăn, người ta thường đánh giá những tác động ấy thông qua việc đánh giá tác động của các yếu tố mới đến các biến ngoại sinh trong mô hình dự báo. Từ đó hình thành nhiều bộ giả định khác nhau về giá trị của các biến ngoại sinh trong tương lai. Mỗi bộ giả định của các biến ngoại sinh sẽ hình thành một kịch bản dự báo. Như vậy về nguyên tắc có rất nhiều kịch bản dự báo, nhưng trong thực tế ứng dụng người ta thường hình thành 3 kịch bản dự báo, kịch bản cơ sở, kịch bản tiêu cực và kịch bản tích cực.

Kịch bản cơ sở là kịch bản dự báo với giả định rằng trong giai đoạn dự báo, không có các yếu tố bất thường tác động đến biến cần được dự báo và như vậy quy luật biến động của các biến ngoại sinh trong mô hình dự báo là gần giống như quy luật của hiện tại và quá khứ. Khi đó các biến ngoại sinh được dự báo theo các mô hình dự báo phụ như được trình bày ở trên. Nói cách khác kịch bản cơ sở chính là kịch bản dự báo được thực hiện theo cách tiếp cận thứ nhất.

Kịch bản tiêu cực và tích cực tương ứng được hình thành trên cơ sở phân tích, đánh giá mức độ tác động tích cực và tiêu cực của các yếu tố bất thường có thể xảy ra trong tương lai đến tất cả các biến ngoại sinh trong mô hình.

Trong trường hợp dự báo kim ngạch xuất khẩu: Để hình thành các kịch bản dự báo thì các biến ngoại sinh trong mô hình dự báo có điều kiện kim ngạch xuất khẩu bây giờ là 63 biến trong Bảng 3.6 chứ không phải là 6 nhân tố trong Bảng 3.7.

Việc xác định các bộ giả định của 63 biến này phải được xuất phát từ phân tích bối cảnh của thời điểm thực hiện dự báo dựa vào lý thuyết kinh tế thương mại, phân tích xu thế phát triển địa - chính trị - kinh tế - tài chính ở các nước trong khu vực và

trên thế giới nhất là ở các nước đối tác thương mại chủ yếu của nền kinh tế Việt Nam, và nhất là phân tích các chính sách điều hành nền kinh tế có liên quan của chính phủ ở thời điểm hiện tại cũng như khả năng điều chỉnh các chính sách ấy nhằm đối phó với tình hình trong tương lai để từ đó: Xác định những biến nào trong 63 biến trên chịu hoặc không chịu sự tác động của những yếu tố mới và của những biến động bất thường. Những biến ngoại sinh không hoặc ít chịu sự tác động sẽ được dự báo bằng sử dụng mô hình dự báo phụ của riêng nó. Các bộ giả định sẽ được thực hiện cho các biến còn lại.

Trong mục này, luận án chỉ tập trung trình bày dự báo có điều kiện kim ngạch xuất khẩu theo cách tiếp cận thứ nhất, tức là với giả định rằng ở 3 tháng tiếp theo không có những yếu tố đặc biệt hoặc những biến động bất thường làm ảnh hưởng đến các hoạt động xuất khẩu. Phần này sẽ không trình bày sâu và cụ thể về việc thực hiện dự báo có điều kiện theo cách tiếp cận thứ hai và thứ ba nêu trên. Có thể tham khảo chi tiết nguyên lý kết hợp mô hình dự báo định lượng với phương pháp phán xử và phương pháp xây dựng các kịch bản dự báo tương ứng trong các nghiên cứu [78], [79], [80].

3.4 Kết luận Chương 3

Dựa vào quy trình mô hình hóa dự báo được trình bày trong Chương 1, Chương này đã đề xuất quy trình và thuật toán dự báo (không và có điều kiện) trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất ở Chương 2 . Độ phức tạp tính toán của thuật toán này cũng được ước lượng và nó là đa thức bậc 3.

Việc giảm chiều trong thuật toán được đề xuất sử dụng cả hai phương pháp lựa chọn thuộc tính và học thuộc tính. Phương pháp lựa chọn thuộc tính được xây dựng dựa vào quan hệ nhân quả Granger đối với thuật toán dự báo không điều kiện hoặc độ đo hệ số tương quan Pearson với thuật toán dự báo có điều kiện. Phương pháp học thuộc tính là KTPCA lặp .

Chương 3 cũng trình bày việc ứng dụng các thuật toán dự báo không và có điều kiện trên tập dữ liệu chuỗi thời gian lớn để dự báo kim ngạch xuất khẩu hàng tháng của Việt Nam. Độ chính xác dự báo (không và có điều kiện) kim ngạch xuất khẩu của Việt Nam là khá cao, trong đó mô hình dự báo kim ngạch xuất khẩu sử dụng

thuật toán dự báo không và có điều kiện làm tăng độ chính xác dự báo lần lượt là 63.6% và 52.9% so với các mô hình ARIMA và mô hình cầu xuất khẩu. Từ đó cho thấy có thể ứng dụng thuật toán dự báo sử dụng phương pháp giảm chiều được đề xuất để dự báo kim ngạch xuất khẩu cũng như dự báo các chỉ tiêu kinh tế - tài chính khác trên các tập dữ liệu chuỗi thời gian lớn.

Kết quả nghiên cứu liên quan đến Chương này được công bố trên Nghiên cứu [CT1], [CT2], [CT4], [CT5] phần danh mục Nghiên cứu của tác giả.

KẾT LUẬN

1. Kết quả nghiên cứu của luận án

Luận án tập trung nghiên cứu khắc phục nhược điểm của các phương pháp giảm chiều PCA và SPCA trên tập dữ liệu chuỗi thời gian lớn. Luận án có những đóng góp nghiên cứu chính như sau:

- Đề xuất phương pháp giảm chiều dựa vào kỹ thuật hàm nhân, gọi tắt là KTPCA. Nó là mở rộng tự nhiên của phương pháp PCA và khắc phục được hạn chế của phương pháp PCA trong việc giảm chiều các tập dữ liệu không xấp xỉ một siêu phẳng. Hiệu suất giảm chiều của phương pháp KTPCA dựa vào mô hình có RMSE tốt nhất (được gọi là KTPCA LẶP) là bằng hoặc cao hơn hiệu suất giảm chiều của các phương pháp giảm chiều PCA, SPCA, RSPCA, và ROBSPCA trên các tập dữ liệu có tần suất lấy mẫu giống nhau hoặc hỗn hợp. Luận án cũng cho thấy hiệu suất giảm chiều của phương pháp PCA và họ SPCA là cạnh tranh. Kết quả này là khác với niềm tin lâu nay rằng hiệu suất giảm chiều của phương pháp SPCA và các phiên bản phát triển của nó là bằng hoặc nổi trội hơn phương pháp PCA.

- Đề xuất quy trình và thuật toán dự báo không và có điều kiện trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều được đề xuất. Độ phức tạp tính toán của thuật toán này là đa thức bậc 3 của số quan sát và số biến của tập dữ liệu đầu vào. Kết quả so sánh Quy trình dự báo đó với cách tiếp cận dự báo 3 bước trong [17] (được xem là phương pháp dự báo nổi trội nhất hiện nay) cho thấy 2 bước đầu tiên ở Quy trình dự báo sử dụng phương pháp giảm chiều được đề xuất là nổi trội hơn tương ứng 2 bước đầu tiên trong cách tiếp cận dự báo 3 bước, bước thứ 3 còn lại hiện chưa được so sánh. Việc ứng dụng thuật toán dự báo sử dụng phương pháp giảm chiều được đề xuất để dự báo kim ngạch xuất khẩu hàng tháng của Việt Nam trên tập dữ liệu của 161 biến giải thích chuỗi thời gian cho thấy:

  - Phần trăm sai số dự báo của mô hình dự báo có điều kiện kim ngạch xuất khẩu theo thuật toán được đề xuất là thấp hơn phần trăm sai số dự báo của mô hình dự báo cầu xuất khẩu là 1.62 điểm phần trăm, cải thiện độ chính xác dự báo lên đến 52.9% so với mô hình dự báo cầu xuất khẩu.

- Phần trăm sai số dự báo của mô hình dự báo không điều kiện kim ngạch xuất khẩu theo thuật toán được đề xuất là thấp hơn phần trăm sai số dự báo của mô hình ARIMA (2,1,2) (mô hình dự báo kim ngạch xuất khẩu tốt nhất trong các mô hình đơn biến được xây dựng dựa vào mô hình ARIMA, AR(p), và Holt -Winter) là 2.38 điểm phần trăm, cải thiện độ chính xác dự báo lên đến 63.6% so với mô hình ARIMA(2,1,2).

Những kết quả trên cùng với độ phức tạp tính toán của các thuật toán dự báo là đa thức bậc 3 cho thấy triển vọng ứng dụng quy trình và thuật toán dự báo sử dụng phương pháp giảm chiều được đề xuất trong dự báo không chỉ kim ngạch xuất khẩu mà còn cho nhiều chỉ tiêu kinh tế - tài chính khác trên tập dữ liệu chuỗi thời gian lớn.

Các kết quả của luận án đã được công bố trên các tạp chí và hội nghị chuyên ngành trong nước, quốc tế có phản biện.

2. Hạn chế của luận án

Luận án có những hạn chế chính sau:

- Thứ nhất: Thuật toán dự báo không và có điều kiện sử dụng phương pháp giảm chiều được đề xuất và ứng dụng của nó mới chỉ được đề xuất đối với các tập dữ liệu có cùng tần suất lấy mẫu, chưa được đề xuất đối với các tập dữ liệu có tần suất lấy mẫu hỗn hợp.

- Thứ hai: Thuật toán dự báo dựa vào quy trình trên mới được tin học hóa một phần, chưa tin học hóa được toàn bộ làm hạn chế việc ứng dụng quy trình dự báo sử dụng phương pháp giảm chiều được đề xuất để dự báo các chỉ số kinh tế - tài chính trên các tập dữ liệu chuỗi thời gian lớn.

DANH MỤC CÁC NGHIÊN CỨU CỦA TÁC GIẢ

[CT1] Thanh, D. Van, Hai, N. M., & Hieu, D. D. Building unconditional forecast model of Stock Market Indexes using combined leading indicators and principal components: application to Vietnamese Stock Market. *Indian Journal of Science & Technology*, *11*(2), 2018. https://doi.org/10.17485/ijst/2018/v11i2/104908.

[CT2] Hai, N. M., Thanh, D. Van, & Dung, N. D. Building Export Forecast Model Using a Kernel-based Dimension Reduction Method. *Economic Computation and Economic Cybernetics Studies and Research*, *56*(1), pp.91–106, 2022. https://doi.org/10.24818/18423264/56.1.22.06.

[CT3] Thanh, D. Van, & Hai, N. M. The performance of a kernel-based variable dimension reduction method. *In Nature of Computation and Communication: 8th EAI International Conference, ICTCC 2022, Cham: Springer Nature Switzerland*, 2023. https://doi.org/10.1007/978-3-031-28790-9_4.

[CT4] Nguyễn Minh Hải, Đỗ Văn Thành và Nguyễn Đức Dũng. Xây Dựng Mô Hình Dự Báo Không Điều Kiện Sử Dụng Phương Pháp Giảm Chiều Biến Dựa Vào Thủ Thuật Kernel, *Proceedings of the 15th National Conference on Fundamental and Applied Information Technology*, pp. 211-218, 2022. https://doi.org/ *10.15625/vap.2022.0226*

[CT5] Thanh, D. Van, & Hai, N. M. Forecast of the VN30 Index by Day Using a Variable Dimension Reduction Method Based on Kernel Tricks. *In Nature of Computation and Communication: 7th EAI International Conference, ICTCC 2021, Virtual Event, October 28–29, 2021*, Proceedings 7, pp. 83-94. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-92942-8_8

[CT6] Đỗ Văn Thành và Nguyễn Minh Hải. Dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều dựa vào hàm kernel và ứng dụng. *Hội thảo quốc gia lần thứ 25: Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông*, pp. 48-54, 2022.

TÀI LIỆU THAM KHẢO

[1] C. Zhang, N. N. A. Sjarif, and R. Ibrahim, “Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020 - 2022,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 1, p. e1519, 2024.

[2] K. Sako, B. N. Mpinda, and P. C. Rodrigues, “Neural networks for financial time series forecasting,” *Entropy*, vol. 24, no. 5, p. 657, 2022.

[3] S. Zaheer *et al.*, “A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model,” *Mathematics*, vol. 11, no. 3, p. 590, 2023.

[4] D. Hopp, “Economic nowcasting with long short-term memory artificial neural networks (LSTM),” *Journal of Official Statistics*, vol. 38, no. 3, pp. 847–873, 2022.

[5] J. F. Torres, D. Hadjout, A. Sebaa, F. Mart\’\inez-Álvarez, and A. Troncoso, “Deep learning for time series forecasting: a survey,” *Big Data*, vol. 9, no. 1, pp. 3–21, 2021.

[6] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, R. P. Ramachandran, and G. Rasool, “Transformers in time-series analysis: A tutorial,” *Circuits, Systems, and Signal Processing*, vol. 42, no. 12, pp. 7433–7466, 2023.

[7] Q. Wen *et al.*, “Transformers in time series: A survey,” *arXiv Prepr. arXiv2202.07125*, 2022.

[8] A. Vaswani *et al.*, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] G. Kapetanios, F. Papailias, and others, “Big data & macroeconomic nowcasting: Methodological review,” *Economic Statistics Centre of Excellence (ESCoE) Discussion Papers ESCoE DP-2018-12*, 2018.

[10] A. Zeng, M. Chen, L. Zhang, and Q. Xu, “Are transformers effective for time series forecasting?,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11121–11128, 2023.

[11] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial time series forecasting with deep learning: A systematic literature review: 2005--2019,” *Applied Soft Computing*, vol. 90, p. 106181, 2020.

[12] H. H. Kim and N. R. Swanson, “Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods,” *International Journal of Forecasting*, vol. 34, no. 2, pp. 339–354, 2018.

[13] K. Chikamatsu, N. Hirakata, Y. Kido, and K. Otaka, “Mixed-frequency approaches to nowcasting GDP: An application to Japan,” *Japan and the World Economy*, vol. 57, p. 101056, Mar. 2021, doi: 10.1016/j.japwor.2021.101056.

[14] D. Bragoli, “Now-casting the Japanese economy,” *International Journal of Forecasting*, vol. 33, no. 2, pp. 390 – 402, 2017.

[15] S. Urasawa, “Real-time GDP forecasting for Japan: A dynamic factor model approach,” *Journal of the Japanese and International Economies*, vol. 34, pp. 116 –134, 2014.

[16] C. Jardet and B. Meunier, “Nowcasting world GDP growth with high-frequency data,” *Journal of Forecasting,* vol. 41, no. 6, pp. 1181 – 1200, 2022.

[17] M. D. Chinn, B. Meunier, and S. Stumpner, “Nowcasting world trade with machine learning: a three-step approach,” National Bureau of Economic Research Working Paper, 2023. **DOI** 10.3386/w31419

[18] J. Shlens, “A tutorial on principal component analysis,” *arXiv Prepr. arXiv1404.1100*, 2014.

[19] L. Van Der Maaten, E. Postma, and J. den Herik, “Dimensionality reduction: a comparative,” *Journal of Machine Learning Research*, vol. 10, no. 66 – 71, p. 13, 2009.

[20] X. Zhong and D. Enke, “Forecasting daily stock market return using dimensionality reduction,” *Expert Systems with Applications*, vol. 67, pp. 126 – 139, 2017.

[21] Y. Koren and L. Carmel, “Robust linear dimensionality reduction,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 4, pp. 459

– 470, 2004.

[22] *[Data]* M. H. Rafiei and H. Adeli, "A novel machine learning model for estimation of sale prices of real estate units," *Journal of Construction Engineering and Management*, vol. 142, no. 2, p. 4015066, 2016.

[23] *[Data]* E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," *Expert Systems with Applications*, vol. 129, pp. 273–285, 2019.

[24] *[Data]* S. De Vito, "Air Quality," UCI Machine Learning Repository, 2016

[25] *[Data]* L. Candanedo, "Appliances energy," UCI Machine Learning Repository, 2017.

[26] *[Data]* K. Hamidieh, "Superconductivity data," UCI Machine Learning Repository, 2018.

[27] D. Giannone, L. Reichlin, and D. H. Small, "Nowcasting GDP and inflation: the real-time informational content of macroeconomic data releases," ECB Working Paper, 2006.

[28] Y. Wang *et al.*, "Guidelines for nowcasting techniques," *World Meteorological Organization: Geneva, Switzerland,* 2017.

[29] B. Bok, D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti, "Macroeconomic nowcasting and forecasting with big data," *Available SSRN 3102227*, 2018.

[30] E. Baldacci *et al.*, "Big data and macroeconomic nowcasting: From data access to modelling," *Luxembourg: Eurostat*, 2016, doi: Doi: http://dx. doi. org/10.2785/360587.

[31] C. Foroni and M. G. Marcellino, "A survey of econometric methods for mixed-frequency data," *Available SSRN 2268912*, 2013, doi: https://doi.org/10.2139/ssrn.2268912.

[32] C. Foroni and M. Marcellino, "A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates," *International Journal of Forecasting*, vol. 30, no. 3, pp. 554–568, 2014.

[33] J. Castle, D. Hendry, O. Kitov, and others, “Forecasting and Nowcasting Macroeconomic Variables: A Methodological Overview,” *University of Oxford, Department of Economics Economics Series Working Papers*, no. 674, 2013.

[34] H. H. Kim and N. R. Swanson, “Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP,” *Journal of Forecasting*, vol. 37, no. 3, pp. 281–302, 2018, doi: 10.1002/for.2499.

[35] T. Shi and S. Horvath, “Unsupervised learning with random forest predictors,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

[36] L. Alzubaidi *et al.*, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[37] J. H. Stock and M. W. Watson, “Forecasting using principal components from a large number of predictors,” *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1167–1179, 2002.

[38] Đ. V. Thành, “Mô hình dự báo giá cổ phiếu trong ngữ cảnh dữ liệu số chiều cao,” in *Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ X, FAIR10, Đã Nẵng ngày 17-18/8/2017*, pp. 422–433, 2017, doi: DOI: 10.15625/vap.2017.00056.

[39] Z. Hajirahimi and M. Khashei, “Hybridization of hybrid structures for time series forecasting: A review,” *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1201–1261, 2023.

[40] F. Petropoulos *et al.*, “Forecasting: theory and practice,” *International Journal of Forecasting*, 2022.

[41] S. Velliangiri, S. Alagumuthukrishnan, and others, “A review of dimensionality reduction techniques for efficient computation,” *Procedia Computer Science*, vol. 165, pp. 104–111, 2019.

[42] S. F. Crone and N. Kourentzes, “Feature selection for time series prediction--A combined filter and wrapper approach for neural networks,” *Neurocomputing*, vol. 73, no. 10–12, pp. 1923–1936, 2010.

[43] Hà Văn Sáng, “Nghiên cứu cải tiến các kỹ thuật rút gọn đặc trưng cho phân lớp dữ liệu,” Luận án Tiến sĩ Công nghệ thông tin, Trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội, 2018.

[44] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “A review of unsupervised feature selection methods,” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, 2020, doi: 10.1007/s10462-019-09682-y.

[45] H. Xie, J. Li, and H. Xue, “A survey of dimensionality reduction techniques based on random projection,” *arXiv Prepr. arXiv1403.2877*, pp. 1–35, 2017, [Online]. Available: http://arxiv.org/abs/1706.04371

[46] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne),” *Computer Science Review*, vol. 40, p. 100378, 2021.

[47] B. Schölkopf, A. Smola, and K.-R. Müller, “Kernel principal component analysis,” in *International conference on artificial neural networks*, pp. 583–588, 1997.

[48] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation,*, vol. 10, no. 5, pp. 1299–1319, 1998.

[49] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[50] V. Vapnik and R. Izmailov, “Reinforced SVM method and memorization mechanisms,” *Pattern Recognition*, vol. 119, p. 108018, 2021.

[51] W. H. Greene, “Econometric Analysis, New York University, Seventh Edition,” *Stern School of Business, New York University*, 2012.

[52] J. Bai, E. Ghysels, and J. H. Wright, “State space models and MIDAS regressions,” *Econometric Reviews,* vol. 32, no. 7, pp. 779–813, 2013.

[53] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *Journal of Computational and Graphical Statistics,* vol. 15, no. 2, pp. 265–286, 2006.

[54] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of The American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[55] F. D’Amuri and J. Marcucci, “The predictive power of Google searches in forecasting US unemployment,” *International Journal of Forecasting*, vol. 33, no. 4, pp. 801–816, 2017.

[56] J. W. Galbraith and G. Tkacz, “Nowcasting with payments system data,” *International Journal of Forecasting*, vol. 34, no. 2, pp. 366–376, 2018.

[57] S. L. Heston and N. R. Sinha, “News vs. sentiment: Predicting stock returns from news stories,” *Financial Analysts Journal*, vol. 73, no. 3, pp. 67–83, 2017.

[58] J. A. Doornik and D. F. Hendry, “Statistical model selection with ‘Big Data,’” *Cogent Economics & Finance*, vol. 3, no. 1, p. 1045216, 2015.

[59] K. Chikamatsu, N. Hirakata, Y. Kido, K. Otaka, and others, *Nowcasting Japanese GDPs*. Bank of Japan Working Paper Series, 2018.

[60] R. E. Kalman, “A new approach to linear filtering and prediction problems,” 1960.

[61] G. Welch, G. Bishop, and others, “An introduction to the Kalman filter,” Chapel Hill, NC, USA, 1995.

[62] A. Panagiotelis, G. Athanasopoulos, R. J. Hyndman, B. Jiang, and F. Vahid, “Macroeconomic forecasting for Australia using a large number of predictors,” *International Journal of Forecasting*, vol. 35, no. 2, pp. 616–633, 2019.

[63] Y. Yu and R. J. Samworth, “Discussion of Large Covariance Estimation by Thresholding Principal Orthogonal Complements by Fan, Liao and Mincheva,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 4, pp. 603–680, 2013.

[64] A. Baffigi, R. Golinelli, and G. Parigi, “Bridge models to forecast the euro area GDP,” *International Journal of Forecasting*, vol. 20, no. 3, pp. 447–460, 2004.

[65] E. Ghysels, P. Santa-Clara, and R. Valkanov, “The MIDAS touch: Mixed data sampling regression models,” 2004.

[66] C. Schumacher, “A comparison of MIDAS and bridge equations,” *International Journal of Forecasting*, vol. 32, no. 2, pp. 257–270, 2016.

[67] E. Ghysels, V. Kvedaras, and V. Zemlys, “Mixed frequency data sampling regression models: the R package midasr,” *Journal of Statistical Software*, vol. 72, pp. 1–35, 2016.

[68] S. Ankargren and U. Lindholm, *Nowcasting Swedish GDP Growth*. National Institute of Economic Research, 2021.

[69] A. Timmermann, “Forecast combinations,” *Handbook of Economic Forecasting*, vol. 1, pp. 135–196, 2006.

[70] N. T. Hien, H. A. Tuan, D. T. Ha, L. M. Trang, T. K. Anh, and others, “Vietnamese Export Growth Prediction Applying MIDAS and MF-VAR on Mixed-Frequency Data,” in *International Conference on Nature of Computation and Communication*, pp. 1–19, 2021.

[71] Nguyễn Long Giang, “Nghiên cứu một số phương pháp khai phá dữ liệu theo tiếp cận lý thuyết tập thô,” Luận án Tiến sĩ, Viện Công nghệ thong tin, 2012.

[72] Nguyễn Văn Thiện, “Một số phương pháp lai ghép trong rút gọn thuộc tính theo tiếp cận tập thô mờ,” Luận án Tiến sĩ, Viện Công nghệ thông tin, 2018.

[73] Hồ Thị Phượng, “Phương pháp gia tăng rút gọn thuộc tính trong bảng quyết định thay đổi theo tiếp cận tập thô mờ,” Luận án Tiến sĩ ngành máy tính, Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2021.

[74] Nguyễn Anh Tuấn, “Rút gọn thuộc tính trong bảng quyết định không đầy đủ có dữ liệu thay đổi theo tiếp cận mô hình tập thô dung sai,” Luận án tiến sĩ khoa học máy tính, Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên, 2021.

[75] V. Bá Anh and C. Thu Thủy, “Xây dựng mô hình dự báo các chỉ tiêu kinh tế vĩ mô chủ yếu theo quý bằng kết hợp phương pháp lựa chọn thuộc tính và phương pháp chỉ số dẫn báo,” Hà Nội, 2018.

[76] G. Koop, *Analysis of economic data*. John Wiley & Sons, 2013.

[77] F. X. Diebold, “Elements of forecasting,” Thomson/South-Western, Mason, Ohio, 4, 2007.

[78] J. S. Armstrong and K. C. Green, “Forecasting methods and principles: Evidence-based checklists,” *Journal of Global Scholars of Marketing Science*, vol. 28, no. 2, pp. 103–159, 2018.

[79] T. Do Van, “Macro-econometric model for medium-term socio-economic development planning in Vietnam. Part 1: structure of the model,” *Экономика региона*, vol. 15, no. 1, 2019.

[80] T. Do Van, “Macro-econometric model for medium-term socio-economic development planning in vietnam. Part 2: application of the model.,” *Economy of Region/Ekonomika Regiona*, vol. 15, no. 3, 2019.

[81] G. Koop and R. Quinlivan, *Analysis of economic data*, vol. 2. Wiley Chichester, 2000.

[82] R. G. Brown, “Smoothing, forecasting and prediction of discrete time series,” 2004.

[83] W. Enders, “Applied econometric time series fourth edition,” *New York (US): University of Alabama*, 2015.

[84] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[85] A. De Mauro, M. Greco, and M. Grimaldi, “A formal definition of Big Data based on its essential features,” *Library Review*, 2016.

[86] L. Tang, J. Li, H. Du, L. Li, J. Wu, and S. Wang, “Big Data in Forecasting Research: A Literature Review,” *Big Data Research*, vol. 27, p. 100289, 2022, doi: https://doi.org/10.1016/j.bdr.2021.100289.

[87] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[88] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[89] Y. Zhang, S. Li, and Y. Teng, “Dynamic processes monitoring using recursive kernel principal component analysis,” *Chemical Engineering Science*, vol. 72, pp. 78–86, 2012.

[90] J. Shawe-Taylor, N. Cristianini, and others, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[91] N. B. Erichson, P. Zheng, K. Manohar, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin, “Sparse principal component analysis via variable projection,” *SIAM Journal on Applied Mathematics,* vol. 80, no. 2, pp. 977–1002, 2020.

[92] A. Garcia-González, A. Huerta, S. Zlotnik, and P. Diez, “A kernel Principal Component Analysis (kPCA) digest with a new backward mapping (pre-image reconstruction) strategy,” *arXiv Prepr. arXiv2001.01958*, 2020.

[93] B. Schölkopf and A. J. Smola, “A short introduction to learning with kernels,” in *Advanced lectures on machine learning*, Springer, pp. 41–64, 2003.

[94] S. Y. Kung, *Kernel methods and machine learning*. Cambridge University Press, 2014.

[95] V. Kvedaras, V. Zemlys, M. Imports, and M. numDeriv, “Package ‘midasr.’” 2021.

[96] T. Reinartz, “Stages of the discovery process,” in *Handbook of data mining and knowledge discovery*, Willi Klosgen and Jan M. Zytkow, Ed., Oxford, pp. 185–192, 2002.

[97] X. Ma and N. Zabaras, “Kernel principal component analysis for stochastic input model generation,” *Journal of Computational Physics*, vol. 230, no. 19, pp. 7311–7331, 2011.

[98] Z. Sun *et al.*, “A survey on dimension reduction algorithms in big data visualization,” in *Cloud Computing, Smart Grid and Innovative Frontiers in Telecommunications: 9th EAI International Conference, CloudComp 2019,*

and 4th EAI International Conference, SmartGIFT 2019, Beijing, China, December 4-5, 2019, and December 21-22, 2019 9, pp. 375–395, 2020.

[99] N. B. Erichson, P. Zheng, and S. Aravkin, "sparsepca: Sparse Principal Component Analysis (SPCA), R package version 0.1. 2." 2018.

[100] A. Karatzoglou, A. Smola, K. Hornik, and M. A. Karatzoglou, "Package 'kernlab,'" 2019.

[101] M. Kuhn, "The caret package," *R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret*, 2012.

[102] J. M. Wooldridge, *Introductory econometrics: A modern approach*. Nelson Education, 2016.

[103] J. De Winter, "Forecasting GDP growth in times of crisis: private sector forecasts versus statistical models," De Nederlandsche Bank Working Paper, 2011.

[104] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[105] M. Kuhn and others, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[106] V. K. P.M. Tan, M. Steibach, A.Karpactne, *Introduction to Data Mining*, 2nd editio. Addition Wesley, Pearson Publishing, 2018.

[107] R. Mehta and P. Mathur, *Short-term Forecasting of India's Export: Developing a Framework by Countries and Commodities*. Research and Information System for the Non-aligned and Other Developing Countries (RIS), 2003.

[108] M. Bussière, J. Fidrmuc, and B. Schnatz, "Trade integration of Central and Eastern European countries: Lessons from a gravity model," *ECB Working Paper No. 545*, 2005.

[109] E. Siggel, "International competitiveness and comparative advantage: a survey and a proposal for measurement," *Journal of Industry, Competition and Trade*, vol. 6, no. 2, pp. 137–159, 2006.

[110] G. Stoevsky, *Econometric Forecasting of Bulgaria's Export and Import Flows*. Bulgarian National Bank Discussion Papers DP/77/2009, 2009.

[111] R. Lehmann, "Survey-based indicators vs. hard data: What improves export forecasts in Europe?," Ifo Working Paper No. 196, 2015.

[112] V. Eskin and M. Gusev, "High-frequency Forecasting Model for the Russian Economy," *The Making of National Economic Forecasts,*, p. 93, 2009.

PHỤ LỤC

Bảng A1: Kết quả giảm chiều biến đổi với các tập dữ liệu

1. Tập EXP	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	GA_5	**GA_6**	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	5	2	10	12	**6**	14	10	10	10
Phần trăm trị riêng tích lũy (%)	76.2	81.97	75.56	76.16	76.09	76.72	75.6	75.6	75.6
Độ trễ tối đa của các biến	6	6	6	6	6	6	6	6	6
RMSE	0.0153	0.027	NA	NA	**0.0104**	NA	NA	NA	NA

2. *Tập VN30*	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	**GA_4**	GA_5	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	9	4	13	14	8	14	14	14	15
Phần trăm trị riêng tích lũy (%)	76.04	77.16	75.98	75.07	75.85	75.83	75.8	75.8	76.9
Độ trễ tối đa của các biến	5	5	5	5	5	5	5	5	5
RMSE	0.2863	0.5838	**0.1871**	0.1819	0.2128	0.1895	0.1968	0.1968	0.2054

3. *Tập CPI*	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	**GA_5**	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	1	1	1	6	1	4	4	4	4
Phần trăm trị riêng tích lũy (%)	97.27	99.82	76.76	75.36	82.90	78.72	77.80	77.80	76.70
Độ trễ tối đa của các biến	6	6	6	6	6	6	6	6	6
RMSE	1.7156	1.7134	3.1932	**0.4452**	3.0652	1.4836	1.0659	1.0673	1.0659

4. Tập VIP	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	**PL_2**	PL_3	GA_4	GA_5	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	4	2	2	7	1	4	4	4	4
Phần trăm trị riêng tích lũy (%)	77.51	75.76	76.38	75.72	76.92	76.19	75.20	75.20	76.00
Độ trễ tối đa của các biến	6	6	6	5	6	6	6	6	6
RMSE	**672.66**	1886.69	2634.23	999.90	3991.93	715.96	826.28	1373.57	2642.83

5. Tập Resid. Building	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	**GA_5**	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	1	1	1	2	1	1	1	1	1

Phần trăm trị riêng tích lũy (%)	98.76	98.40	91.01	79.55	91.37	99.28	99.00	99.00	98.30
Độ trễ tối đa của các biến	10	10	10	10	10	10	10	10	10
RMSE	1149.9	1148.6	1070.3	**919.9**	1069.7	1152.4	1152.5	1152.5	1151.2
6. Tập S&P500	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	$\boldsymbol{GA_5}$	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	1	1	1	2	1	1	1	1	1
Phần trăm trị riêng tích lũy (%)	99.99	99.99	98.08	76.87	98.08	99.96	99.80	99.80	99.80
Độ trễ tối đa của các biến	5	5	5	5	5	5	5	5	5
RMSE	161.442	161.442	149.98	**61.60**	172.907	161.415	161.441	161.441	161.441
7. Tập DJI	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	GA_5	GA_6	**PCA**	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	1	1	1	3	1	1	1	1	1
Phần trăm trị riêng tích lũy (%)	99.06	98.66	94.21	76.05	94.37	99.50	99.2	99.2	98.9
Độ trễ tối đa của các biến	5	5	5	5	5	5	5	5	5
RMSE	301.88	294.59	2319.55	695.42	2277.97	**91.82**	309.24	309.24	309.23
8. Tập NASDAQ	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	$\boldsymbol{PL_2}$	PL_3	GA_4	GA_5	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	1	1	1	3	1	1	1	1	1
Phần trăm trị riêng tích lũy (%)	99.12	99.12	94.62	76.98	94.81	99.67	99.4	99.4	99.1
Độ trễ tối đa của các biến	5	5	5	5	5	5	5	5	5
RMSE	**81.05**	193.25	887.21	183.77	817.75	365.97	85.47	85.47	85.46
9. Tập Air Quality	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	$\boldsymbol{GA_5}$	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	2	2	4	5	2	1	1	1	1
Phần trăm trị riêng tích lũy (%)	90.29	93.51	82.82	79.91	81.75	75.96	75.70	75.70	74.80
Độ trễ tối đa của các biến	12	12	12	12	12	12	12	12	12
RMSE	63.165	64.342	62.937	**50.297**	67.153	71.459	71.499	71.499	71.427
10. Tập Appliances Energy	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	$\boldsymbol{GA_5}$	GA_6	PCA	SPCA	RSPCA	ROBSPCA

Số lượng các nhân tố/PC	1	1	4	6	2	3	3	3	3
Phần trăm trị riêng tích lũy (%)	85.46	96.62	78.07	78.48	79.37	79.30	78.70	78.70	78.80
Độ trễ tối đa của các biến	6	6	6	6	6	6	6	6	6
RMSE	101.83	101.86	100.75	**98.81**	102.10	101.74	101.76	101.76	101.75
11.Tập SuperConductivity	*Phương pháp KTPCA lặp*					*Phương pháp họ SPCA*			
Phương pháp	PL_2	PL_3	GA_4	$\boldsymbol{GA_5}$	GA_6	PCA	SPCA	RSPCA	ROBSPCA
Số lượng các nhân tố/PC	2	2	1	2	1	2	2	2	2
Phần trăm trị riêng tích lũy (%)	90.32	90.32	88.84	76.34	89.28	92.30	91.90	91.90	90.50
Độ trễ tối đa của các biến	10	10	10	10	10	10	10	10	10
RMSE	27.164	27.161	26.316	**26.094**	26.403	27.314	27.332	27.332	27.319

Bảng A2: Giá trị của tham số ρ^2 trong các hàm nhân Gauss GA_5 và GA_6

Các tập dữ liệu	*DS1*	*DS2*	*DS3*	*DS4*	*DS5*	*DS6*	*DS7*	*DS8*	*DS9*	*S10*	*DS11*
$\boldsymbol{GA_5}$	$e^{-0.693}$	$e^{6.9}$	$e^{0.7}$	$e^{34.1}$	e^{22}	e^{8}	e^{8}	e^{8}	e^{18}	$e^{12.8}$	$e^{18.5}$
$\boldsymbol{GA_6}$	$e^{-0.183}$	$e^{7.5}$	e^{2}	$e^{35.3}$	$e^{27.2}$	$e^{15.7}$	$e^{15.5}$	$e^{13.5}$	e^{20}	$e^{24.5}$	$e^{22.8}$

Bảng B: Kết quả giảm chiều biến đối với tập dữ liệu lấy mẫu tần suất hỗn hợp

1. Các mô hình Nowcast được xây dựng dựa vào mô hình phương trình bắc cầu (BE)nhân tố						
Phương pháp giảm chiều	**PCA**	**SPCA**	**RSPCA**	**ROBSPCA**	**KTPCA**	**Hàm nhân k**(x, y) =
Tập dữ liệu RGDP						PL(1, 1, 0)
Số các nhân tố được chọn	2	2	2	2	2	2
% giá trị riêng tích lũy	77.50	77.00	77.10	77.00	77.50	77.50
Độ trễ đơn tối ưu của các nhân tố	3	3	3	3	3	
RMSE	**0.000493**	0.000788	0.000788	0.000788	**0.000493**	
Tập dữ liệu CPI						PL(1,1,0)
Số các nhân tố được chọn	4	4	4	4	4	4
% giá trị riêng tích lũy	78.72	77.80	77.80	76.70	78.72	78.72
Độ trễ đơn tối ưu của các nhân tố	2	2	2	2	2	
RMSE	**0.000183**	0.000485	0.00051	0.000485	**0.000183**	
Tập dữ liệu IIP						$GA_2(e^{8.7})$
Số các nhân tố được chọn	12	13	13	13	15	15
% giá trị riêng tích lũy	75.22	75.50	75.50	75.40	75.96	75.96
Độ trễ đơn tối ưu của các nhân tố	5	5	5	5	5	
RMSE	1.348981	1.203836	1.044371	1.545299	**0.569320**	
Tập dữ liệu Air Quality						$GA_2(e^{18.0})$

Số các nhân tố được chọn	1	1	1	1	5	5
% giá trị riêng tích lũy	75.96	75.70	75.70	74.80	79.91	79.91
Độ trễ đơn tối ưu của các nhân tố	15	15	15	15	15	
RMSE	0.615228	0.611051	0.610396	0.611060	**0.592861**	
Tập dữ liệu Appliances Energy						$GA_2(e^{12.8})$
Số các nhân tố được chọn	3	3	3	3	6	6
% giá trị riêng tích lũy	79.30	78.7	78.7	78.8	78.48	78.48
Độ trễ đơn tối ưu của các nhân tố	12	12	12	12	12	
RMSE	377.6252	377.2618	377.2618	377.0618	**360.1310**	
Tập dữ liệu Res.Building						$GA_2(e^{22.0})$
Số các nhân tố được chọn	1	1	1	1	2	2
% giá trị riêng tích lũy	99.28	99.00	99.00	98.30	79.55	79.55
Độ trễ đơn tối ưu của các nhân tố	3	3	3	3	3	
RMSE	565.5147	565.523	565.523	565.516	513.6189	
Tập dữ liệu S&P 500						PL(1, 1, 0)
Số các nhân tố được chọn	1	1	1	1	1	1
% giá trị riêng tích lũy	99.96	99.8	99.8	99.8	99.96	99.96
Độ trễ đơn tối ưu của các nhân tố	7	7	7	7	7	
RMSE	4.3074	4.3076	4.3076	4.3076	4.3074	
Tập dữ liệu DJI						PL(0.5, 2, 0.5)
Số các nhân tố được chọn	1	1	1	1	1	1
% giá trị riêng tích lũy	99.5	99.2	99.21	98.9	99.06	99.06
Độ trễ đơn tối ưu của các nhân tố	7	7	7	7	7	
RMSE	57.1033	56.4321	56.4321	56.4321	56.2975	
Tập dữ liệu NASDAQ						PL(0.5, 2, 0.5)
Số các nhân tố được chọn	1	1	1	1	1	1
% giá trị riêng tích lũy	99.67	99.4	99.4	99.1	99.12	99.12
Độ trễ đơn tối ưu của các nhân tố	6	6	6	6	6	
RMSE	18.5945	18.5941	18.5941	18.5489	18.3479	
Tập dữ liệu SuperConduct						PL(0.5, 2, 0.5)
Số các nhân tố được chọn	2	2	2	2	2	2
% giá trị riêng tích lũy	92.3	91.9	91.9	90.5	90.323	90.323
Độ trễ đơn tối ưu của các nhân tố	8	8	8	8	8	
RMSE	13.5381	13.5397	13.5425	13.5429	13.3662	

2. Dựa vào Mô hình MIDAS với hàm trọng số STEP bậc ba

Phương pháp giảm chiều	**PCA**	**SPCA**	**RSPCA**	**ROBSPCA**	**KTPCA**	**Hàm nhân k**(x, y) =
Tập dữ liệu RGDP						PL(1, 1, 0)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	9	9	9	9	9	2
Độ trễ riêng tối ưu của các nhân tố HF	9 8[10]	9 8	9 8	9 8	9 8	77.50
RMSE	**0.007440**	0.009727	0.009722	0.009727	**0.007440**	
Tập dữ liệu CPI						$GA_1(e^{1.461})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	6	6	6	6	10	1
Độ trễ riêng tối ưu của các nhân tố HF	5 4 5 4	5 5 4 4	5 5 4 4	5 5 4 4	9	76.76
RMSE	0.008236	0.004390	0.004387	0.004390	**0.003948**	
Tập dữ liệu IIP						$GA_2(e^{8.7})$

[10] : Hai con số này lần lượt là độ trễ tối ưu của hai nhân tố trong mô hình Nowcast.

Độ trễ đơn cố định cho các nhân tố HF	15	15	15	15	10	15
Độ trễ riêng tối ưu của các nhân tố HF						75.96
RMSE	26.522320	21.39361	28.868560	*28.133150*	**8.788050**	
Tập dữ liệu Air Quality						PL(1, 1, 0)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	7	7	7	7	7	1
Độ trễ riêng tối ưu của các nhân tố HF	5	5	5	5	5	75.96
RMSE	**0.630038**	0.63004	0.630040	***0.630038***	**0.630038**	
Tập dữ liệu Appliances Energy						PL(1, 1, 0)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	5	5	5	5	5	3
Độ trễ riêng tối ưu của các nhân tố HF	5 4 5	5 4 5	5 4 5	5 4 5	5 4 5	79.30
RMSE	**385.1972**	385.680	385.6800	385.3454	**385.1972**	
Tập dữ liệu Res.Building						PL(1, 1, 0)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	30	30	30	30	30	1
Độ trễ riêng tối ưu của các nhân tố HF	30	30	30	30	30	99.29
RMSE	430.8412	430.837	430.8373	430.8397	421.709	
Tập dữ liệu S&P 500						$GA_1(e^{15.426})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	98.07
RMSE	259.8844	259.808	257.6644	259.8065	**72.7871**	
Tập dữ liệu DJI						$GA_1(e^{15.171})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	94.21
RMSE	4101.593	4101.96	4101.958	4102.275	**1024.708**	
Tập dữ liệu Nasdaq						PL(0.5, 3, 0.5)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.12
RMSE	1419.767	1419.807	1419.807	1419.756	**687.2987**	
Tập dữ liệu SuperConductivity						PL(0.5, 2, 0.5)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	22	22	22	22	22	2
Độ trễ riêng tối ưu của các nhân tố HF	22	22	22	22	22	90.32
RMSE	14.3425	14.3462	14.3462	14.3431	13.9649	

3. Dựa vào mô hình MIDAS với hàm trọng số ALMON Đa thức bậc 2 (PAW2-MIDAS)

Phương pháp giảm chiều	**PCA**	**SPCA**	**RSPCA**	**ROBSPCA**	**KTPCA**	**Hàm nhân k(x, y) =**
Tập dữ liệu RGDP						PL(1, 1, 0)

Độ trễ đơn tối ưu cho tất cả các nhân tố HF	8	8	8	8	8	2
Độ trễ riêng tối ưu của các nhân tố HF	8 3	7 8	7 8	7 8	7 8	77.50
RMSE	**0.000026**	0.000208	0.00020	0.000208	**0.000026**	
Tập dữ liệu CPI						PL(1,1, 0)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	12	12	12	12	12	4
Độ trễ riêng tối ưu của các nhân tố HF	5 8 8 4	5 8 8 4	5 8 8 4	5 8 8 4	5 8 8 4	78.72
RMSE	**0.0014 7**	0.001833	0.00182	0.001833	**0.001473**	
Tập dữ liệu IIP						$GA_2(e^{8.7})$
Độ trễ cố định của các nhân tố HF	61	57	59	60	56	15
Độ trễ riêng tối ưu của các nhân tố HF						75.96
RMSE	1.1268 2	0.734195	0.75076	0.620822	**0.043310**	
Tập dữ liệu Air Quality						$GA_2(e^{18.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	5	5	5	5	5	5
Độ trễ riêng tối ưu của các nhân tố HF	2	2	2	2	5 2 2 2 2	79.91
RMSE	0.6297 8	0.629339	0.64020	0.629782	**0.617446**	
Tập dữ liệu Appliances Energy						$GA_2(e^{12.8})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	4	4	4	4	4	6
Độ trễ riêng tối ưu của các nhân tố HF	4 2 2	4 2 2	3 3 3	4 3 3	2 4 2 3 4 2	78.48
RMSE	384.40 1	384.4115	384.322	384.3270	**384.0171**	
Tập dữ liệu Res.Building						$GA_2(e^{22.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	30	30	30	30	30	2
Độ trễ riêng tối ưu của các nhân tố HF	30	30	30	30	30	79.55
RMSE	404.33 9	399.4798	399.480	399.48	**399.4498**	
Tập dữ liệu S&P 500						$GA_2(e^{8.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	2
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	76.87
RMSE	40.701 9	42.8444	42.8444	42.8444	**33.6159**	
Tập dữ liệu DJI						$GA_2(e^{8.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	3
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	76.05
RMSE	337.80 48	337.8025	337.802 5	337.8026	**311.3913**	

Tập dữ liệu Nasdaq						$GA_2(e^{8.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	3
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	76.98
RMSE	107.967	107.9666	107.9666	107.9666	**107.0302**	
Tập dữ liệu SuperConductivity						$GA_2(e^{18.5})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	22	22	22	22	22	2
Độ trễ riêng tối ưu của các nhân tố HF	22	22	22	22	22	76.34
RMSE	13.958	13.958	13.958	13.958	**13.9485**	

4. Dựa vào mô hình MIDAS với hàm trọng số Almon hàm mũ (EAW-MIDAS)

Phương pháp giảm chiều	**PCA**	**SPCA**	**RSPCA**	**ROBSPCA**	**KTPCA**	**Hàm nhân k(x, y) =**
Tập dữ liệu RGDP						$GA_3(e^{2.5})$
Độ trễ đơn cố định cho các nhân tố HF	7	7	7	7	7	1
Độ trễ riêng tối ưu của các nhân tố HF	2 7	2 7	2 7	2 7	7	90.47
RMSE	0.00523	0.005274	0.005277	0.005274	**0.004544**	
Tập dữ liệu CPI						$GA_1(e^{1.461})$
Độ trễ đơn cố định cho các nhân tố HF	5	5	5	5	5	1
Độ trễ riêng tối ưu của các nhân tố HF	3 2 2 3	2 3 2 3	2 3 2 3	2 3 2 3	2	76.75
RMSE	0.00691	0.005465	0.00742	0.005465	**0.00509**	
Tập dữ liệu IIP						$GA_3(e^{10.2})$
Độ trễ đơn cố định cho các nhân tố HF	4	4	4	4	6	4
Độ trễ riêng tối ưu của các nhân tố HF	2(12)	2(12)	2(12)	2(12)	5 5 4 4	75.23
RMSE	4.4983	4.7174	4.3561	4.3146	**4.1810**	
Tập dữ liệu Air Quality						$GA_3(e^{20.0})$
Độ trễ đơn cố định cho các nhân tố HF	8	8	8	8	8	2
Độ trễ riêng tối ưu của các nhân tố HF	8	8	8	8	7 2	81.75
RMSE	0.4762	0.4765	0.4765	0.4761	**0.4392**	
Tập dữ liệu Appliances Energy						$GA_1(e^{13.595})$
Độ trễ đơn cố định cho các nhân tố HF	5	5	5	5	5	4
Độ trễ riêng tối ưu của các nhân tố HF	3 4 5	5 4 5	5 4 5	3 4 5	2 3 5 4	78.07
RMSE	385.4549	385.4515	385.4515	385.4597	**385.0**	
Tập dữ liệu Res. Building						$GA_2(e^{22.0})$
Độ trễ đơn cố định cho các nhân tố HF	30	30	30	30	30	2
Độ trễ riêng tối ưu của các nhân tố HF	30	30	30	30	30	79.55

RMSE	504.9074	504.9076	504.9076	504.9069	**379.0157**	
Tập dữ liệu S&P 500						PL (1, 1, 0)
Độ trễ đơn cố định cho các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.96
RMSE	2.806	2.953	2.953	2.953	**2.806**	
Tập dữ liệu DJI						$GA_2(e^{8.0})$
Độ trễ đơn cố định cho các nhân tố HF	58	58	58	58	58	3
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	
RMSE	240	239.7	239.7	239.5	**118.9**	76.04
Tập dữ liệu Nasdaq						PL(0.5, 2, 0.5)
Độ trễ đơn cố định cho các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.12
RMSE	82.2279	82.1254	82.1254	82.0357	**36.3656**	
Tập dữ liệu SuperConduct						PL(0.5, 2, 0.5)
Độ trễ đơn cố định cho các nhân tố HF	22	22	22	22	22	2
Độ trễ riêng tối ưu của các nhân tố HF	22 20	22 20	22 20	22 20	22 20	90.32
RMSE	13.9322	13.931	13.931	13.9322	13.9302	

5. Dựa vào mô hình U-MIDAS

Phương pháp giảm chiều	**PCA**	**SPCA**	**RSPCA**	**ROBSPCA**	**KTPCA**	**Hàm nhân k**(x, y) =
Tập dữ liệu RGDP						$GA_2(e^{0.5})$
Độ trễ đơn cố định cho các nhân tố HF	7	7	7	7	7	6
Độ trễ riêng tối ưu của các nhân tố HF	7 3	7 3	7 3	7 3	1 1 0 0 0 0	76.14
RMSE	0.002040	0.000951	0.000919	0.000951	**0.000699**	
Tập dữ liệu CPI						PL(1, 1, 0)
Độ trễ đơn cố định cho các nhân tố HF	5	5	5	5	5	4
Độ trễ riêng tối ưu của các nhân tố HF	5 2 2 5	3 3 2 2	3 3 3 3	3 3 3 2	5 2 2 5	78.72
RMSE	0.000109	0.002515	0.002955	0.002512	**0.000109**	
Tập dữ liệu IIP						PL(1, 1, 0)
Độ trễ đơn cố định cho các nhân tố HF	6	6	6	6	6	12
Độ trễ riêng tối ưu của các nhân tố HF	6(7) 5(4)[11]	6(2) 5(11)	6(2) 5(11)	6(2) 5(11)	6(7) 5(4) 6	75.22
RMSE	**0.028318**	0.986031	0.310864	0.663163	**0.028318**	
Tập dữ liệu Air Quality						$GA_1(e^{19.977})$
Độ trễ đơn cố định cho các nhân tố HF	28	28	28	28	28	4

[11] : Ký hiệu này có nghĩa là có 12 nhân tố tần suất cao, trong đó 7 nhân tố đầu tiên có độ trễ tối ưu là 6, 4 nhân tố tiếp theo có độ trễ tối ưu là 5, nhân tố cuối cùng có độ trễ tối ưu là 6.

Độ trễ riêng tối ưu của các nhân tố HF	28	28	28	28	28 22 22 15	82.81
RMSE	0.405435	0.405821	0.405821	0.405494	**0.3330**	
Tập dữ liệu Appliances Energy						$GA_1(e^{13.595})$
Độ trễ đơn cố định cho các nhân tố HF	324	324	324	324	324	4
Độ trễ riêng tối ưu của các nhân tố HF	324 (3)	324 (3)	324 (3)	324 (3)	324(2) 315	78.07
RMSE	376.9851	377.4016	377.4016	376.8008	351,2	
Tập dữ liệu Res. Building						$GA_2(e^{22.0})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	30	30	30	30	30	2
Độ trễ riêng tối ưu của các nhân tố HF	30	30	30	30	30	79.55
RMSE	430.1182	430.1732	430.1732	430.1286	**389.1229**	
Tập dữ liệu S&P 500						PL(0.5, 2, 0.5)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.99
RMSE	0.00070	0.0000558	0.000056	0.0000587	**0.000054 6**	
Tập dữ liệu DJI						PL(0.5, 2, 0.5)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.06
RMSE	2.932	2.931	2.931	2.931	**2.930**	
Tập dữ liệu Nasdaq						PL(0.5, 2, 0.5)
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	58	58	58	58	58	1
Độ trễ riêng tối ưu của các nhân tố HF	58	58	58	58	58	99.12
RMSE	0.8993	0.8992	0.8992	0.8992	**0.8841**	
Tập dữ liệu SuperConduct						$GA_1(e^{22.353})$
Độ trễ đơn tối ưu cho tất cả các nhân tố HF	22	22	22	22	22	1
Độ trễ riêng tối ưu của các nhân tố HF	22 19	22 19	22 19	22 19	22 19	88.84
RMSE	14.0231	14.0219	14.0219	14.0231	**13.9115**	

Bảng C: Đánh giá việc lựa chọn biến phản ánh phía cung, phía cầu và sức mạnh của thị trường dựa vào các quy trình dự báo và mô hình cầu xuất khẩu

Phương pháp lựa chọn biến	**Mô hình EX sử dụng quy trình dự báo không điều kiện**	**Mô hình EX sử dụng quy trình dự báo có điều kiện**
Nhóm biến phản ánh phía cung (**70 biến**)	DN4, IIP, TK13, TK17, TK5, XK12, XK19, XK20, XK23, XK6, XK9 (**11 biến**)	IIP, IIP2, IIP4, IIP5, IIP6, IIP8, IIP9, INVES, XK1, XK10, XK11, XK12, XK14, XK15, XK16, XK17, XK18, XK19, XK2, XK20, XK21, XK22, XK23, XK24, XK3, XK4, XK5, XK6, XK7, XK8, XK9 (**31 biến**)
Nhóm biến phản ánh phía cầu (**78 biến**)	BL4, CPI3, CPI4, CPI5, CPI6, CPI8, IM_AUS, IM_IND, IM_INDO, IM_ITA, IM_MAL, IM_TAI, NK1, NK16, NK17, NK2, NK9, PEX/PWEX, PCOPP, POIL, PRUBB, TT17, TT3 (**23 biến**)	TBL, BL1, BL4, IM_AUS, IM_CHI, IM_GER, IM_MAL, IM_SIN, NK, NK10, NK11, NK12, NK13, NK14, NK15, NK18, NK2, NK3, NK4, NK5, NK7, NK9, TT1, TT11, TT12, TT13, TT14, TT19, TT2, TT5, TT6, TT8 (**32 biến**)
Nhóm biến phản ánh sức mạnh thị trường. **(14 biến)**	M2, M2DC, M2KT (**3 biến**)	
Nhóm biến trong mô hình cầu xuất khẩu	IM_AUS, IM_IND, IM_INDO, IM_ITA, IM_MAL, IM_TAI; POIL; PEX/PWEX.	IM_CHI, IM_GER, IM_MAL, IM_SIN