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

TÓM TẮT LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

Mã số: 9 48 01 04

Hà Nội - 2024

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ,
Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:
Người hướng dẫn 1: PGS.TS. Đỗ Văn Thành, Khoa CNTT, Đại Học Duy Tân
Người hướng dẫn 2: PGS.TS. Nguyễn Đức Dũng, Viện Công nghệ thông tin

Phản biện 1: PGS.TS.
Phản biện 2: PGS.TS.
Phản biện 3: PGS.TS.

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi … giờ …, ngày … tháng … năm 2024.

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN
LIÊN QUAN ĐẾN LUẬN ÁN

1. 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.
2. 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.
3. 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.
4. 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*
5. 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
6. 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.

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) (C. Zhang et al., 2024), (Sako et al., 2022), (Zaheer et al., 2023), (Hopp, 2022), mô hình mạng các đơn vị định kỳ kiểm soát (GRU) (Torres et al., 2021), và mô hình transformer chuỗi thời gian (Ahmed et al., 2023), (Wen et al., 2022). Các mô hình học sâu LSTM và GRU 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 (gradient) (Vaswani et al., 2017). Theo nghiên cứu (Kapetanios et al., 2018), các mô hình LSTM và GRU phù hợp với những bài toán dự báo trên tập dữ liệu ở đó số lượng quan sát lớn nhưng số lượng các biến *không quá lớn.* Mô hình học sâu Transformers có ưu điểm nắm bắt được sự phụ thuộc và tương tác ở phạm vi dài giữa các biến nên đang thu hút nghiên cứu sử dụng mô hình này trong dự báo chuỗi thời gian. Các kết quả đạt được của mô hình transformer chuỗi thời gian mới ở mức ban đầu (Wen et al., 2022). Thông qua nghiên cứu thực nghiệm, nghiên cứu (Zeng et al., 2023) cho thấy mô hình dựa trên mạng nơtron đa lớp đơn giản vẫn có thể đạt được kết quả dự báo tốt hơn so với mô hình Transformer chuỗi thời gian. 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 còn hạn chế (Hopp, 2022), (Sezer et al., 2020; Torres et al., 2021). Theo (Hopp, 2022), việc ứng dụng các phương pháp học sâu trong việc dự báo kinh tế-xã hội vẫn còn sơ khai một phần do còn có những hạn chế khi thực hiện chúng.

Nghiên cứu (Kim & Swanson, 2018b) 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 (Chikamatsu et al., 2021), (Bragoli, 2017), (Urasawa, 2014), (Jardet & Meunier, 2022), (Chinn et al., 2023) 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 (Chinn et al., 2023) 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 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 (Shlens, 2014) 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 (Van Der Maaten et

al., 2009), (Zhong & Enke, 2017) 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 (Koren & Carmel, 2004) 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 (Van Der Maaten et al., 2009), (Zhong & Enke, 2017) 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 (Chikamatsu et al., 2021), (Bragoli, 2017), (Urasawa, 2014), (Jardet & Meunier, 2022) và nhất là (Van Der Maaten et al., 2009), (Zhong & Enke, 2017), và (Chinn et al., 2023) 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 (Koren & Carmel, 2004) 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) 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; 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 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. Phương pháp giảm chiều được đề xuất không chỉ khắc phục được nhược điểm mà còn có hiệu suất giảm chiều không 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.

3. Bố cục của luận án

Cấu trúc 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 của các tập dữ liệu chuỗi thời gian lớn dựa vào thủ thuật hàm nhân, gọi là KTPCA, và so sánh hiệu suất giảm chiều biến của phương pháp KTPCA dựa vào mô hình 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.

- **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 LỚN CHUỖI THỜI GIAN

1.1. Tổng quan các nghiên cứu trong và ngoài nước

Nội dung tổng quan các nghiên cứu trong và ngoài nước được trình bày trong 17 trang, tham khảo chi tiết từ trang 9 – 24 trong Luận án.

1.2 Các vấn đề còn tồn tại

Từ những phân tích, đánh giá các công trình liên quan trong và ngoài nước ở trên, Luận án tập trung nghiên cứu giải pháp để khắc phục tồn tại trên. 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

Nội dung phần này trình bày các kiến thức cơ sở phục vụ cho luận án bao gồm 20 trang, tham khảo từ trang 28 – 48 trong luận án.

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. PHƯƠNG PHÁP GIẢM CHIỀU BIẾN DỰA VÀO THỦ THUẬT HÀM NHÂN

Chương này sẽ đề xuất phương pháp giảm chiều mới dựa vào thủ 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 RMSE tốt nhất (gọi tắt là KTPCA#) 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 thủ thuật hàm nhân

Giả sử $\mathbf{X} = [X_1, X_2, \dots, X_m]_{N\times m}$ 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, $\mathbf{X}$ 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 thủ 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 vào hàm nhân và được gọi là KTPCA để phân biệt nó với phương pháp KPCA. Phương pháp này khác với phương pháp KPCA, xem trang 49 – 50 Luận án.\

- Ma trận hàm nhân xác định bởi $\mathbf{K}=[\boldsymbol{\kappa}(\mathrm{X_i},\mathrm{X_j})] \equiv [\boldsymbol{\Phi}(\mathrm{X}_i).\boldsymbol{\Phi}(\mathrm{X}_j)]$, ở đây X_i là *véc tơ dữ liệu đầu vào*. Như vậy ma trận hàm nhân trong phương pháp này khác với ma trận hàm nhân trong phương pháp KPCA như được xác định bởi công thức (1.29).

- Thay vì *chiếu tập dữ liệu* $\boldsymbol{\Phi(X)}$ *được cân chỉnh trung bình* lên các véc tơ riêng của ma trận hàm nhân trong không gian đặc trưng $\boldsymbol{\mathcal{H}}$, 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 **K**.

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 **K.** Nói cách khác thuật toán giảm chiều bằng sử dụng phương pháp KTPCA có thể được viết dưới dạng giả code như sau:

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 $\boldsymbol{\kappa}(X_i,X_j) = \langle X_i,X_j\rangle$ thì ma trận hàm nhân **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.

Thuật toán giảm chiều bằng sử dụng phương pháp KTPCA có thể được viết dưới dạng giả code như sau:

Thuật toán KTPCA

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

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

1. Xây dựng ma trận hàm nhân $K=[\kappa(X_i,X_j)] \equiv [\Phi(X_i).\boldsymbol{\Phi}(X_j)]$
2. Tìm giá trị riêng và véc tơ riêng của ma trận hàm nhân
3. Sắp xếp các véc tơ riêng theo các giá trị riêng theo thứ tự giảm dần
4. Xây dựng ma trận $\tilde{\mathbf{E}}_{m\times p}$ với p vectơ riêng đầu tiên
5. Biến đổi X sử dụng $\tilde{\mathbf{E}}_{m\times p}$ để thu được không gian con mới Y = X. $\tilde{\mathbf{E}}_{m\times p}$

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 RMSE tốt nhất. Phương pháp KTPCA dựa vào mô hình RMSE tốt nhất được gọi là KTPCA#.

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, xem trang 49 – trang 53 trong Luận án.

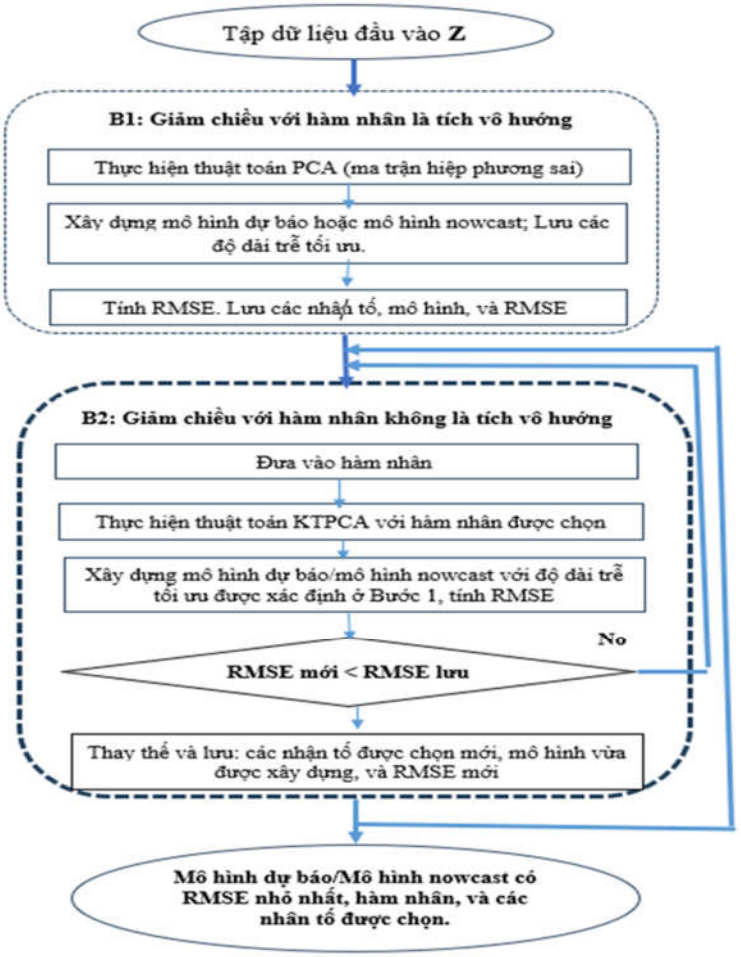
Bảng 2.1: Sự khác nhau của các phương pháp PCA, KPCA, và KTPCA

PCA (Shlens, 2014)	**KPCA** (Schölkopf et. al. 1998)	**KTPCA**
- Tập dữ liệu $\mathbf{X} \in \mathbb{R}^{N\times m}$ được cân chỉnh trung bình - Tìm trị riêng và véc tơ riêng của ma trận hiệp	- Tập dữ liệu $\mathbf{X} \in \mathbb{R}^{N\times m}$ - Xác định ma trận hàm nhân $\mathbf{K} = [\boldsymbol{\kappa}(\chi_i,\chi_j)]$, χ_i là véc tơ điểm dữ liệu của **X** và ma trận Gramm cấp $N \times N$:	- Tập dữ liệu $\mathbf{X} \in \mathbb{R}^{N\times m}$ được cân chỉnh trung bình - Xác định ma trận hàm nhân $\boldsymbol{K}_{\boldsymbol{m}\times\boldsymbol{m}} = [\boldsymbol{\kappa}(X_i,X_j)]$, X_i là véc

phương sai của **X** - Sắp véc tơ riêng theo giá trị riêng - p nhân tố đầu tiên được xác định bởi: $\boldsymbol{PC}_{N\times \mathrm{p}} = \boldsymbol{X}_{\mathrm{N}\times \mathrm{m}}.\boldsymbol{E}_{\mathrm{m}\times \mathrm{p}}$	- $\boldsymbol{K_c} = \boldsymbol{K} - \boldsymbol{K}.\boldsymbol{1_N} - \boldsymbol{1_N}.\boldsymbol{K} + \boldsymbol{1_N}.\boldsymbol{K}.\boldsymbol{1_N}$ - Tìm trị riêng, véctơ riêng của $\mathbf{K_c}$ - 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**.	tơ dữ liệu của **X.** - Tìm trị riêng và véc tơ của ma trận **K** ứng với hàm nhân k; - p nhân tố được xác định bởi: $\boldsymbol{PC}_{N\times p} = \boldsymbol{X}_{\mathrm{N}\times \mathrm{m}}.\widetilde{\boldsymbol{E}}_{\mathrm{m}\times \mathrm{p}}$

2.1.2. Giảm chiều biến sử dụng phương pháp KTPCA#

Việc giảm chiều biến bằng sử dụng phương pháp KTPCA# được trình bày trong Hình 2.1 bên dưới.



Hình 2.1: Lưu đồ của phương pháp KTPCA dựa trên mô hình tốt nhất RMSE

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# 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.

2.3. Hiệu suất giảm chiều biến của phương pháp KTPCA#

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#. Và RMSE càng nhỏ, hiệu suất của phương pháp giảm chiều càng cao, xem chi tiết ở trang 55 – 56 trong Luận án.

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

2.2.1.1 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 UCI-Machine Learning Repository được trình bầy trong Bảng 2.2 ở dưới, xem trang 56 – 57 trong Luận án.

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

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	Time series	Real	60	63	No	Kim ngạch xuất khẩu	Tháng
VN30	Time series	Real	366	34	No	Chỉ số VN30	Ngày
CPI	Time series	Real	72	102	No	Chỉ số CPI	Tháng
VIP	Time Series	Real	60	265	No	Giá trị sản xuất các ngành	Tháng
Residential Building	Multivariate	Real	371	27	No	Giá bán	
S&P500	Time series	Real	1760	52	Yes	Chỉ số S&P500	Ngày
DJI	Time series	Real	1760	81	Yes	Chỉ số Dow Jones	Ngày
NASDAQ	Time series	Real	1760	81	Yes	Chỉ số Nasdaq	Ngày
Air Quality	Time series	Real	9348	12	Yes	Khí CO	Giờ
Appliances Energy	Time series	Real	19704	23	No	Sử dụng năng lượng của thiết bị (wh)	Mỗi 10 phút
SuperConduct.	Multivariate	Real	21263	81	No	Nhiệt độ tới hạn	

2.2.1.2. Phương pháp thực nghiệm

Để so sánh hiệu suất giảm chiều biến của phương pháp KTPCA# 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. 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.

2.2.1.3 Kết quả

a. Hiệu suất của KTPCA# so với các phương pháp 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#, 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, 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.

Bảng 2.4: Hiệu suất giảm chiều dữ liệu của phương pháp KTPCA#

Datasets	Phương pháp	KTPCA#	PCA	SPCA	RSPCA	ROBSPCA
	Số lượng nhân tố	$GA_6, 6$	14	10	10	10
EXP	RMSE	**0.0104**	NA	NA	NA	NA
	Số lượng nhân tố	$GA_4, 14$	14	14	14	15
VN30	RMSE	**0.1819**	0.1895	0.1968	0.1968	0.2054
	Số lượng nhân tố	$GA_5, 6$	4	4	4	4
CPI	RMSE	**0.4452**	1.4836	1.0659	1.0673	1.0659
	Số lượng nhân tố	$PL_2, 4$	4	4	4	4
VIP	RMSE	**672.66**	715.96	826.28	1373.57	2642.83
	Số lượng nhân tố	$GA_5, 2$	1	1	1	1
Res. Building	RMSE	**919.9**	1152.4	1152.5	1152.5	1151.2
	Số lượng nhân tố	$GA_5, 2$	1	1	1	1
S&P500	RMSE	**61.60**	161.415	161.441	161.441	161.441
	Số lượng nhân tố	$PL_1, 1$	1	1	1	1
DJI	RMSE	**91.82**	**91.82**	309.24	309.24	309.23
	Số lượng nhân tố	$PL_2, 1$	1	1	1	1
NASDAQ	RMSE	**81.05**	365.97	85.47	85.47	85.46
	Số lượng nhân tố	$GA_5, 5$	1	1	1	1
Air Quality	RMSE	**50.297**	71.459	71.499	71.499	71.427
	Số lượng nhân tố	$GA_5, 6$	3	3	3	3
App. Energy	RMSE	**98.81**	101.74	101.76	101.76	101.75
	Số lượng nhân tố	$GA_5, 2$	2	2	2	2
SuperCon.	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.

Từ phân tích trên Bảng 2.4, có thể kết luận rằng hiệu suất giảm chiều biến của phương pháp KTPCA# là bằng hoặc cao hơn so với các phương pháp PCA và họ SPCA.

b. Hiệu suất của phương pháp PCA so với phương pháp SPCA

Bảng 2.5 (ngoại trừ dữ liệu liên quan đến phương pháp KTPCA#) bên dưới và Hình 2.2 cũng cho thấy 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, xem trang 62-63 trong Luận án.

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

Các phương pháp	*DS2*	*DS3*	*DS4*	*DS5*	*DS6*
KTPCA#	**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#	**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.

2.2.2 Đối với 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 Các 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 kho UCI - Machine Learning 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, xem trang 64 – 65 trong luận án.

Bảng 2.6: Các đặc điểm 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	Time-series	Time-series	Time-series	Time-series	Time-series
Thuộc tính biến	Real	Real	Real	Real	Real
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 số thấp [1]	3	3	20	12	6
Dữ liệu khuyết	No	No	Yes	Yes	No
Biến phụ thuộc	Tốc độ tăng trưởng GDP	Lạm phát giá tiêu dùng	Chỉ số sản xuất công nghiệp	Khi CO	Sử dụng năng lượng của thiết bị
Các đặc điểm thống kê	Res. Build.	S&P 500	DJI	NASDAQ	SuperCond.
Dặc điểm của tập dữ liệu	cross data	Time-series	Time-series	Time-series	cross data
Thuộc tính biến	Real	Real	Real	Real	Real
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 số thấp	3	20	20	20	20
Dữ liệu khuyết	No	Yes	Yes	Yes	No
Biến phụ thuộc	Giá bán	Chỉ số S&P500	Chỉ số DJI	Chỉ số NASDAQ	Nhiệt độ tới hạn

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 (Zhang et al., 2012). 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 có thể xem trang 66-67 trong Luận án.

Việc so sánh hiệu suất giảm chiều biến của phương pháp KTPCA# 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

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 mới 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 bầy trong Bảng B (phần Phụ lục).

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

a. *Hiệu suất của KTPCA# so với các phương pháp 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#.

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# 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.

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

3a. BE	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#	**3b.STEP**	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#
SET1	**0.000493**	0.000788	0.00079	0.000788	**0.000493**	SET1	**0.00744**	0.009727	0.009722	0.009727	**0.00744**
SET2	**0.000183**	0.000485	0.00051	0.000485	**0.000183**	SET2	0.008236	0.00439	0.004387	0.00439	**0.003948**
SET3	1.348981	1.203836	1.04437	1.545299	**0.56932**	SET3	26.52232	21.39361	28.86856	28.13315	**8.78805**
SET4	0.615228	0.611051	0.6104	0.61106	**0.592861**	SET4	**0.630038**	0.63004	0.63004	**0.630038**	**0.630038**
SET5	377.6252	377.2618	377.262	377.0618	**360.131**	SET5	**385.1972**	385.68	385.68	385.3454	**385.1972**
SET6	565.5147	565.523	565.523	565.516	**513.6189**	SET6	430.8412	430.8373	430.8373	430.8397	**421.709**
SET7	**4.3074**	4.3076	4.3076	4.3076	**4.3074**	SET7	259.8844	259.8083	257.6644	259.8065	**72.7871**
SET8	57.1033	56.4321	56.4321	56.4321	**56.2975**	SET8	4101.593	4101.958	4101.958	4102.275	**1024.708**
SET9	18.5945	18.5941	18.5941	18.5489	**18.3479**	SET9	1419.767	1419.807	1419.807	1419.756	**687.2987**
SET10	13.5381	13.5397	13.5425	13.5429	**13.3662**	SET10	14.3425	14.3462	14.3462	14.3431	**13.9649**
3c.PAW	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA#**	**3d.EAW**	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA#**
SET1	**0.000026**	0.000208	0.000197	0.000208	**0.000026**	SET1	0.005232	0.005274	0.005277	0.005274	**0.004544**
SET2	**0.001473**	0.001833	0.001819	0.001833	**0.001473**	SET2	0.006911	0.005465	0.007418	0.005465	**0.00509**
SET3	1.1268	0.7342	0.7508	0.6208	**0.0433**	SET3	4.4983	4.7174	4.3561	4.3146	**4.1810**
SET4	0.6298	0.6293	0.6402	0.6298	**0.6174**	SET4	0.4762	0.4765	0.4765	0.4761	**0.4392**
SET5	384.4007	384.4115	384.3218	384.3270	**384.0171**	SET5	385.4549	385.4515	385.4515	385.4597	**385.000**
SET6	404.3389	399.4798	399.4798	399.4800	**399.3498**	SET6	504.9074	504.9076	504.9076	504.9069	**379.0157**
SET7	40.7019	42.8444	42.8444	42.8444	**33.6159**	SET7	**2.806**	2.953	2.953	2.953	**2.8060**
SET8	337.8048	337.8025	337.8025	337.8026	**311.3913**	SET8	240.0	239.7	239.7	239.5	**118.900**
SET9	107.9667	107.9666	107.9666	107.9666	**107.0302**	SET9	82.2279	82.1254	82.1254	82.0357	**36.3656**
SET10	13.9580	13.9580	13.9580	13.9580	**13.9485**	SET10	13.9322	13.931	13.931	13.9322	**13.9302**
3e.U	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#	**3e.U**	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#
SET1	0.00204	0.000951	0.000919	0.000951	**0.000699**	SET6	430.1182	430.1732	430.1732	430.1286	**389.1229**
SET2	**0.000109**	0.002515	0.002955	0.002512	**0.000109**	SET7	0.000701	5.58E-05	0.0000558	0.0000587	**0.0000546**
SET3	**0.0283**	0.9860	0.3109	0.6632	**0.0283**	SET8	2.932	2.931	2.931	2.931	**2.9300**
SET4	0.4054	0.4058	0.4058	0.4055	**0.3330**	SET9	0.8993	0.8992	0.8992	0.8992	**0.8841**
SET5	376.9851	377.4016	377.4016	376.8008	**351.2000**	SET10	14.0231	14.0219	14.0219	14.0231	**13.9115**

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

b. Hiệu suất của phương pháp PCA so với các phương pháp SPCA, RSPCA và ROBSPCA

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 và Bảng 2.8 ở dưới cho thấy 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 nhau, xem trang 70 – 72 trong luận án.

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

2.4 Kết luận Chương 2

Chương này đề xuất phương pháp giảm chiều dựa vào thủ thuật hàm nhân (gọi tắt KTPCA). Sự khác biệt của phương pháp này so với các phương pháp KPCA và PCA cũng được làm rõ. Phương pháp KTPCA sẽ 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 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.

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.

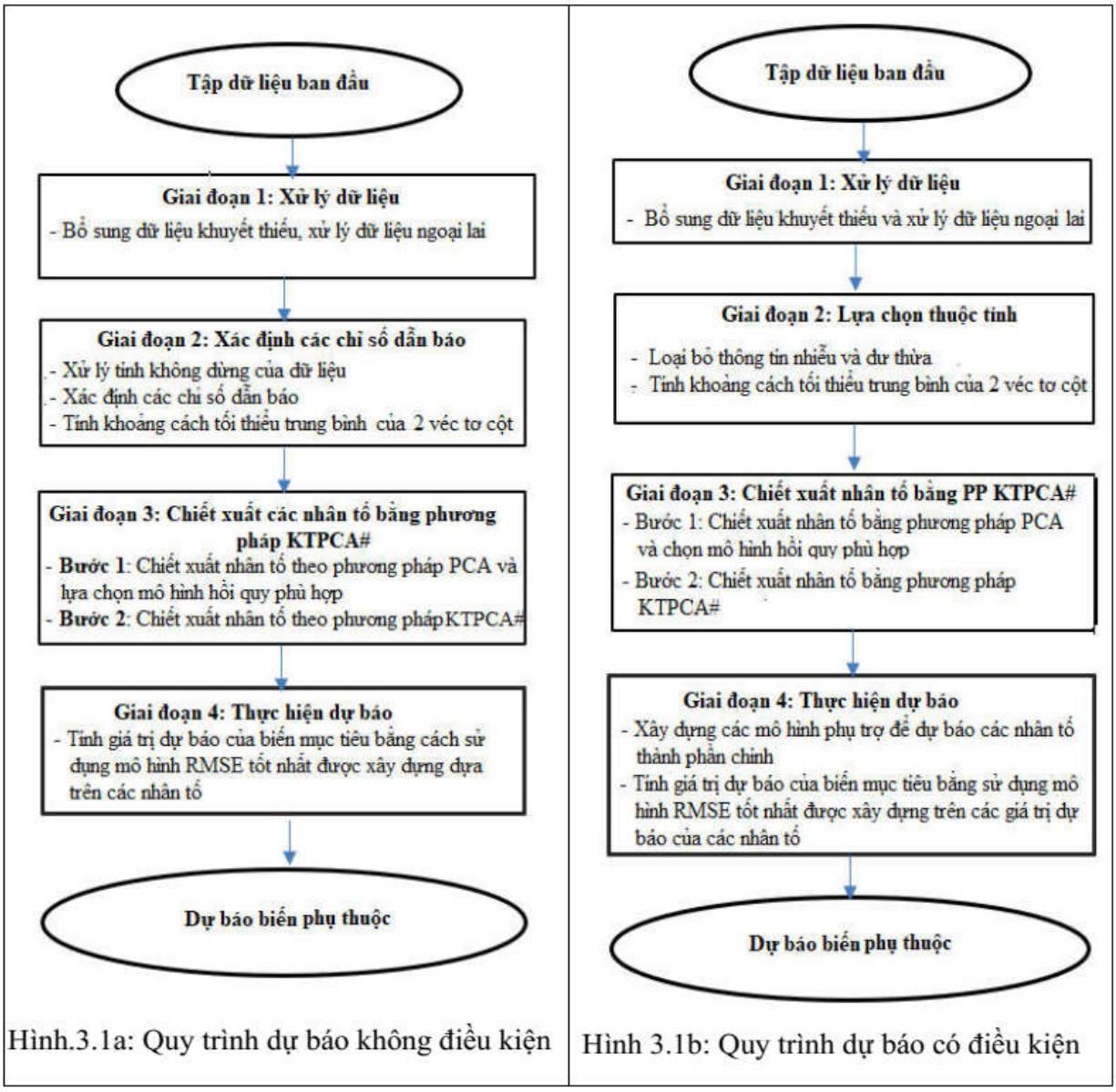
CHƯƠNG 3. DỰ BÁO TRÊN TẬP DỮ LIỆU LỚN CHUỖI THỜI GIAN SỬ DỤNG PHƯƠNG PHÁP GIẢM CHIỀU DỰA VÀO THỦ THUẬT KERNEL

Chương 3 đề xuất thuật toán dự báo không và có điều kiện trên tập dữ liệu lớn sử dụng phương pháp giảm chiều KTPCA# 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#. 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#

Quy trình dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều KTPCA# 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 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#. 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 này được trình bày chi tiết từ trang 73 – trang 79 trong Luận án



Hình 3.1: Lưu đồ quy trình dự báo không điều kiện và dự báo có điều kiện

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 (Chinn et al., 2023), xem trang 78-79 trong Luận án.

Bảng 3.4: So sánh hai cách tiếp cận xây dựng mô hình dự báo có điều kiện

Luận án so với nghiên cứu (Chinn et al., 2023)	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 trễ phân bố tự hồi quy ARDL trên các nhân tố được chiết xuất từ tập dữ liệu của tất các các biến đầu vào. Đánh giá kém hơn.
Nghiên cứu (Chinn et al., 2023)	Sử dụng phương pháp hồi quy góc nhỏ, nhưng không xử lý dữ liệu dư thừa. Đánh giá: kém hơn	Sử dụng phương pháp giảm chiều PCA (là trường hợp riêng của phương pháp giảm chiều trong luận án) cho cả các tập dữ liệu không xấp xỉ siêu phẳng. Đánh giá: kém hơn	Hồi quy rừng ngẫu nhiên kinh tế. Bản chất của nó là phân các biến giải thích thành các nhóm con, xây dựng mô hình dự báo biến phụ thuộc trên các nhóm con bằng sử dụng mô hình trễ phân bố tự hồi quy ARDL, sau đó kết hợp các kết quả dự báo biến phụ thuộc của các mô hình thành phần. Đánh giá tốt hơn

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

Các thuật toán này được xây dựng theo quy trình được đề xuất trong Hình 3.1. Giả sử $\mathbf{X_t} = [X_{1,t}, X_{2,t}, \dots, X_{m,t}] \in \mathbb{R}^{\mathbf{t\times m}}$ 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 và không 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 và không 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**	THUẬT TOÁN 1b: **UNCONF 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 và dư thừa, 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. Xác định h - thời điểm xa nhất của dự báo; 2. *Repetition* ← "Yes"; 3. *FeatureSelection* ($\mathbf{X_t}$, Y_t); 4. Center $\mathbf{X_t}$;	**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; …… . 3. *LeadingIndicatorSelection* ($\mathbf{X_t}$, Y_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***)*; 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 $\boldsymbol{K}$; 12. *FeatureLearning* (***K***); 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(*$\widehat{\mathrm{Y}}_{\mathrm{t+h}}$, Mô hình dự báo biến $\mathrm{Y_t}$*);* **End.**	…… …… …… 16. *Calculate(*$\widehat{\mathrm{Y}}_{\mathrm{t+h}}$, Mô hình dự báo biến Y_t*);* **End.**

Các hàm *FeatureSelection*, *LeadingIndicatorSelection,* thủ tục *FeatureLearning* và thủ tục *Forecast, Calculate* được giới thiệu chi tiết hơn bên dưới.

THUẬT TOÁN 2a: *FeatureSelection* Algorithm	THUẬT TOÁN 2b: *LeadingIndicatorSelection* Algorithm
Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$. **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**;	**Input**: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, p là độ trễ chung. **Output**: Tập các chỉ số dẫn báo của Y_t với trễ p trong $\mathbf{X_t}$; α (%) – mức ý nghĩa thống kê; **begin** 1. Chuyển biến Y_t và các biến trong $\mathbf{X_t}$ thành các chuỗi thời gian dừng; 2. **for** mỗi biến trong $\mathbf{X_t}$ thực hiện 3. Xây dựng mô hình dự báo biến Y_t theo biến này dựa vào mô hình (2.2) Tính xác suất của thống kê F trong mô hình dự báo; 4. **if** xác suất đó $<$ α **then** biến giải thích đó là chỉ số dẫn báo; 5. **end for**

	end;
THUẬT TOÁN 3a: *FeatureLearning* Procedure **Input**: Ma trận $\mathbf{K}_{m\times m}$. **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 3b: *FeatureLearning* Procedure **Input**: Ma trận $\mathbf{K}_{N\times g}$ là ma trận hàm nhân của tập gồm *g* chỉ số dẫn báo; **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** …… …… 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 trễ phân bố tự hồi quy ARDL ở đó độ trễ của biến phụ thuộc và biến giải thích đã được xác định trước. …… . **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
end;
	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.
	end;

Cụ thể ý nghĩa các dòng lệnh của các thuật toán, hàm và thủ tục được trình bày ở các trang 80 - trang 86 trong Luận án.

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 xây dựng mô hình dự báo trên các nhân tố được chiết suất 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, xem chi tiết ở trang 86 – 88 trong Luận án:

- Độ phức tạp tính toán của thuật toán *FeatureSelection* là: $O(m^2)$ (3.2)
- Độ phức tạp tính toán của các dòng lệnh 7 và 8 là: $O(N.m^2 + N^3)$ (3.3)
- Độ phức tạp tính toán của dòng lệnh 12 và 13 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 17 của thuật toán CONF (P.M. Tan, M. Steibach, A.Karpactne, 2018) thì 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 (P.M. Tan, M. Steibach, A.Karpactne, 2018). Do vậy độ phức tạp tính toán của thuật toán *LeadingIndicatorSelection* là:

$$O\big(m.O(N)\big) = O(m.N) \quad (3.7)$$

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:

$$q.O(N.m^2 + N^3 + m^3). \quad (3.8)$$

3.3 Dự báo kim ngạch xuất khẩu bằng phương pháp giảm chiều biến KTPCA#

3.3.1 Xác định vấn đề dự báo

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 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 thủ 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 giả 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

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 sản xuất. Trong các 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 được trình bầy chi tiết theo khung lý thuyết trong nghiên cứu (Siggel, 2006). Cụ thể, nghiên cứu (Siggel, 2006) đề xuất mô hình dự báo tổng kim ngạch xuất khẩu có dạng tổng quát như sau:

$$X_t = \boldsymbol{f}\left(X_{t-i,} ED_{t-i+1,} ER_{t-i+1}, \boldsymbol{P_{t-i+1}}\right), i \geq 1 \qquad (3.2)$$

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.

3.3.2.2 Tập dữ liệu phục vụ dự báo

Các nghiên cứu (Siggel, 2006), (Stoevsky, 2009), (Lehmann, 2015) 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 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, được gọi là **EXP** bao gồm các yếu tố tác động đến kim ngạch xuất khẩu trong mô hình cầu xuất khẩu (Siggel, 2006), (Stoevsky, 2009) được trình bày trong Bảng 3.1, trang 92 – 93.

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 tiền dữ liệu

Khắc phục giá trị khuyết thiếu, chuyển đổi số liệu - xử lý tính không dừng theo công thức (3.10) (Eskin & Gusev, 2009). 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 thực hiện dự báo kiểm định chấp nhận mô hình được xây dựng, luận án chia tập dữ liệu đầu vào gồm 65 quan sát thành 02 tập bao gồm tập huấn luyện có 62 quan sát bắt đầu từ tháng 2/2014 đến tháng 3/2019 và tập kiểm thử có 03 quan sát từ tháng 4 /2019 đến tháng 6/2019. Trước hết mô hình dự báo không điều kiện được xây dựng trên tập dữ liệu huấn luyện.

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.

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 (Wooldridge, 2016). Với ngưỡng $\alpha < 0.1$, nghĩa là xác suất bác bỏ là < 0.1, luận án chọn được 37 biến là các chỉ số dẫn báo của biến phụ thuộc. Bảng 3.2 ở dưới là danh sách các chỉ số dẫn báo có ý nghĩa thống kê đối với biến kim ngạch xuất khẩu *EX*, xem trang 98 trong Luận án.

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

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# theo 06 hàm 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 (Wooldridge, 2016) 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 $\boldsymbol{\kappa_1}(x_i, x_j) = \langle x_i, x_j \rangle$ cho thấy 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.

Bảng 3.4: Kết quả chiết xuất nhân tố bằng phương pháp KTPCA#

Hàm kernel	*Dạng hàm kernel*	*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
	$\boldsymbol{\kappa_2}(x_i, x_j) = \langle x_i, x_j \rangle^{\boldsymbol{2}}$	2	83,34	0.0228
	$\boldsymbol{\kappa_3}(x_i, x_j) = \langle x_i, x_j \rangle^{\boldsymbol{3}}$	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

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. 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 phương trình (3.11) trang 100.

3.3.3.4 Giai đoạn 4: Thực hiện dự 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.

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 100*(giá trị thống kê thực tế - giá trị dự báo)/giá trị thống kê thực tế. 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, xem trang 101 – 102 trong Luận án.

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 đoạn 3 và Giai đoạn 4 với hàm hàm nhân Gauss có tham số $\rho^2 = e^{-0.4}$ theo cách tương tự như đã được trình bầy ở trên.

3.3.4 Dự báo kim ngạch xuất khẩu sử dụng quy trình dự báo có điều kiện

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 3 tháng.

3.3.4.1 Giai đoạn 1: Xử lý tiền dữ liệu

Tương tự như trường hợp dự báo không điều kiện.

3.3.4.2 Giai đoạn 2: Loại bỏ biến nhiễu và dư thừa

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*, chỉ còn 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, trang 104.

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

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 (Wooldridge, 2016) và là 6. Kết quả chiết xuất nhân tố bằng phương pháp KTPCA# được trình bầy trong Bảng 3.7.

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

Kernel $\boldsymbol{\kappa}$	*Các tham số*	*Các nhân tố*	*% Trị riêng tích lũy*	*RMSE*
(PCA)	$\kappa_0(,)$: c = 0, d = 1	14	76.72	Không tiếp tục
	$\kappa_1(,)$: c = 0, d = 2	5	76.02	0.0153
Đa thức	$\kappa_2(,)$: c = 0, d = 3	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

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. Kết thúc quy trình lặp, ta thu được mô hình dự báo kim ngạch xuất khẩu tối ưu nhất có dạng theo phương trình (3.13) trang 106.

3.3.4.4 Giai đoạn 4: Xây dựng mô hình dự báo các biến ngoại sinh và thực hiện dự báo

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, xem trang 107 trong Luận án.

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.

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 kim ngạch xuất khẩu dựa vào mô hình cầu xuất khẩu theo phương trình (3.9) với độ trễ tối ưu chung là 6 (Wooldridge, 2016). Kết quả dự báo kim ngạch xuất khẩu theo mô hình cầu xuất khẩu ở 3 tháng 4, 5, và 6 năm 2019 được trình bầy trong Bảng 3.11.

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

Ký hiệu *EXF* và *DEXF* lần lượt là các giá trị dự báo của *EX* theo mô hình nhân tố và mô hình cầu xuất khẩu. 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 bầy trong Bảng 3.10 dưới đây.

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
	Abs(% sai số dự báo) TB = **1.44**			Abs(% sai số dự báo) TB = **3.06**	
RMSE	0.0104			0.0261	
$RMSE_{OUT}$	0.0038			0.0296	

Tính trung bình giá trị tuyệt đối của phần trăm 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%, xem trang 106 – 110.

.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 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 hoá 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, xem trang 111 – 113 trong Luận án.

3.4 Kết luận chương 3

Dựa vào quy trình mô hình hoá dự báo chuỗi thời gian được trình bầy trong Chương 1, Chương này đã đề xuất quy trình và thuật toán dự báo (không và có điều kiện) trên tập dữ liệu lớn chuỗi thời gian 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.

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# .

Chương 3 cũng trình bày việc ứng dụng các thuật toán dự báo không và có điều kiện trên tập dữ liệu lớn chuỗi thời gian 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 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 - xã hội khác trên các tập dữ liệu lớn chuỗi thời gian.

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 PCA và SPCA trên tập dữ liệu chuỗi thời gian lớn. Luận án có những đóng góp nghiên cứu chính như sau:

1. *Về mặt lý thuyết*

- Đề xuất phương pháp giảm chiều dựa vào thủ 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 RMSE tốt nhất (được gọi là KTPCA#) 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 (Chinn et al., 2023) (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.

2. *Về mặt ứng dụng thực tiễn*

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. Trong lĩnh vực kinh tế - tài chính, mô hình cầu xuất khẩu đang được sử dụng phổ biến và được đánh giá là nổi trội nhất để dự báo kim ngạch xuất khẩu ở các quốc gia cũng như ở Việt Nam.
- 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 hoá 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.