BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

ĐÀO XUÂN KỲ

ỨNG DỤNG MÔ HÌNH XÍCH MARKOV VÀ CHUỖI THỜI GIAN MỜ TRONG DỰ BÁO

LUẬN ÁN TIẾN SĨ TOÁN HỌC

Hà Nội, 2017

BỘ GIÁO DỤC VÀ ĐÀO TẠO **VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM**

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

ĐÀO XUÂN KỲ

ỨNG DỤNG MÔ HÌNH XÍCH MARKOV VÀ CHUỖI THỜI GIAN MỜ TRONG DỰ BÁO

LUẬN ÁN TIẾN SĨ TOÁN HỌC

Chuyên ngành: Cơ sở Toán học cho Tin học

Mã số: 62.46.01.10

Người hướng dẫn khoa học:

1. PGS.TS. Đoàn Văn Ban
2. TS. Nguyễn Văn Hùng

Hà Nội, 2017

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các kết quả được công bố với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố trong bất kỳ công trình nào khác.

Hà nội, ngày 01 tháng 12 năm 2017

NGHIÊN CỨU SINH

Đào Xuân Kỳ

LỜI CẢM ƠN

Luận án được thực hiện tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, dưới sự hướng dẫn của PGS.TS. Đoàn Văn Ban và TS. Nguyễn Văn Hùng. Tôi xin bày tỏ lòng biết ơn sâu sắc đến hai Thầy về định hướng khoa học, người đã động viên, trao đổi nhiều kiến thức và chỉ bảo tôi vượt qua những khó khăn để hoàn thành luận án này.

Tôi cũng xin gửi lời cảm ơn chân thành đến các nhà khoa học, tác giả của các công trình công bố đã được trích dẫn trong luận án, đây là những tư liệu quý, kiến thức liên quan quan trọng giúp Nghiên cứu sinh hoàn thành luận án; Xin cảm ơn đến các nhà khoa học đã phản biện các công trình nghiên cứu của Nghiên cứu sinh.

Tôi trân trọng cảm ơn Phòng Thống kê -tính toán và Ứng dụng, Viện Công nghệ Thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tạo điều kiện thuận lợi cho tôi trong suốt quá trình nghiên cứu thực hiện luận án.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc tới gia đình, bạn bè, những người đã luôn ủng hộ, giúp đỡ và hỗ trợ tôi về mọi mặt để tôi yên tâm học tập đạt kết quả tốt.

Hà nội, ngày 01 tháng 12 năm 2017

NGHIÊN CỨU SINH

Đào Xuân Kỳ

MỤC LỤC

Danh mục từ viết tắt

ACF	Autocorrelation Function
ANN	Artificial Neural Network
AIC	Akaike Information Criterion
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayessian Information Criterion
BPNN	Back Propagation Neural Network
BWP	Backward Probabilities
CMC	Comerical Higher Order Markov Chain
DJIA	Dow Jones Industrial Average Index
EM	Expectation-Maximization
FTS	Fuzzy Time Series
FWP	Forward Probabilities
GA	Genetic Algorithm
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GDP	Gross Domestic Product
GPS	Global Positioning System
HMM	Hidden Markov Model
HMMs	Hidden Markov Models
IMC	Improved Higher Order Markov Chain
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC	Markov Chain
MLE	Maximum Likelihood Estimation
PCA	Principle Component Analysis
RMSE	Root Mean Square Error
SSE	Shanghai Stock Exchange
STNN	Stochastic Time Neural Network

SVM	Support Vector Machine
TAIEX	Taiwan Exchange Index
VN-Index	Chỉ số chứng khoán Việt Nam

Các ký hiệu toán học

Ký hiệu, từ viết tắt	**Diễn giải**
Γ	Ma trận xác suất chuyển xích Markov
(C_t)	Xích Markov
γ_{ij}	Xác suất chuyển Markov
π	Vector phân phối dừng của xích Markov
$p_i(x)$	Phân phối trạng thái i trong HMM
λ	Tham số của phân phối Poisson
μ_i	Trung bình của các phân phối chuẩn
σ_i^2	Phương sai của các phân phối chuẩn
μ_A	Được gọi là hàm thuộc
U	Không gian nền
$Y_{(t)}$	Là chuỗi thời gian
"o"	Là toán tử thành phần Max–Min
(X_t)	Chuỗi dữ liệu quan sát
(x_t)	Chuỗi dữ liệu quan sát

Danh sách bảng

Danh sách hình vẽ

MỞ ĐẦU

1. Tính cấp thiết của luận án

Bài toán dự báo chuỗi thời gian với đối tượng dự báo là biến ngẫu nhiên X thay đổi theo thời gian nhằm đạt được độ chính xác dự báo cao luôn là thách thức đối với các nhà khoa học không chỉ trong nước mà còn đối với các nhà khoa học trên thế giới. Bởi lẽ, giá trị của biến ngẫu nhiên này tại thời điểm t sinh ra một cách ngẫu nhiên và việc tìm một phân phối xác suất phù hợp cho nó không phải lúc nào cũng dễ dàng. Muốn làm được điều này dữ liệu lịch sử cần được thu thập và phân tích, từ đó tìm ra phân phối ướm khít với nó. Tuy nhiên, một phân phối tìm được có thể phù hợp với dữ liệu ở một giai đoạn này, nhưng có thể sai lệch lớn so với giai đoạn khác. Do đó, việc sử dụng một phân phối ổn định cho đối tượng dự đoán là không phù hợp với bài toán dự báo chuỗi thời gian.

Chính vì lý do trên, để xây dựng mô hình dự báo chuỗi thời gian cần thiết phải có sự liên hệ, cập nhật dữ liệu tương lai với dữ liệu lịch sử, xây dựng mô hình phụ thuộc giữa giá trị dữ liệu có được tại thời điểm t với giá trị tại các thời điểm trước đó $t-1, t-2 \ldots$. Nếu xây dựng quan hệ

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

cho ta mô hình hồi quy tuyến tính ARIMA[11]. Trong đó α_i, θ_i là các hệ số hồi quy, ε_{t-i} là các biến ngẫu nhiên độc lập cùng phân phối chuẩn có kỳ vọng bằng 0. Mô hình này đã được áp dụng rộng rãi bởi cơ sở lý thuyết dễ hiểu và dễ thực hành, hơn nữa mô hình này đã được tích hợp vào hầu hết các phần mềm thống kê hiện nay như Eviews, SPSS, Matlab, R,…. Tuy nhiên, nhiều chuỗi thời gian thực tế cho thấy nó không biến đổi tuyến tính. Do đó mô hình tuyến tính như ARIMA không phù hợp. R. Parrelli đã chỉ ra trong [53], các chuỗi thời gian về độ dao động của chỉ số kinh tế hay tài chính thường có quan hệ phi tuyến, vậy dự báo chuỗi thời gian phi tuyến thì đối tượng phù hợp cho nó là dự báo độ dao động của sự biến đổi trong chuỗi thời gian làm sơ sở trong quản lý rủi ro. Mô hình phổ biến cho dự báo chuỗi

thời gian phi tuyến phải kể đến mô hình GARCH [49, 53]. Hạn chế của mô hình GARCH lại nằm ở việc phải giả sử dữ liệu dao động tuân theo một phân phối cố định (thường là phân phối chuẩn) trong khi dữ liệu thực tế cho thấy phân phối thống kê lại là phân phối nặng đuôi [66] (trong khi phân phối chuẩn có độ lệch cân đối). Với hi vọng xây dựng những mô hình dự báo có độ chính xác cao hơn, nhiều nhà nghiên cứu đã tiến hành áp dụng những kỹ thuật cũng như công nghệ mới nhất trong các lĩnh vực khác nhau (như mô hình mạng thần kinh nhân tạo (ANN) [41] hay véc tơ học máy hỗ trợ (SVM) [62] nhằm giải quyết bài toán và đạt được những kết quả nhất định.

Cho đến nay, mặc dù đã có nhiều mô hình mới được xây dựng theo hướng kết hợp các mô hình sẵn có nhằm cải thiện độ chính xác của dự báo nhưng mặc dù mô hình rất phức tạp trong khi độ chính xác dự báo cải thiện không đáng kể. Do đó một số hướng có thể thực hiện nhằm đơn giản hóa mô hình và đảm bảo hoặc tăng độ chính xác dự báo có thể được phát triển.

Một là: Xây dựng mô hình Markov ẩn (HMM) với những trạng thái ẩn là những phân phối xác suất nhất định (chẳng hạn phân phối chuẩn) để từ đó dự báo phân bố của giá trị tương lai. Chẳng hạn, chuỗi thời gian chỉ số chứng khoán thay đổi ngẫu nhiên ngày qua ngày với những trạng thái mà nhà đầu tư có thể hiểu là "tốt", "bình thường" và "xấu". Mỗi trạng thái này không thể định nghĩa bởi một hằng số vì có nhiều giá trị trong mỗi trạng thái. Do đó, coi mỗi trạng thái là một phân bố xác suất được đặc trưng bởi một bộ tham số là một suy diễn hợp lý.

Hai là: Kết hợp xích Markov và chuỗi thời gian mờ. Mỗi trạng thái "tốt", "xấu", "bình thường" như trên thay vì hiểu theo một phân bố xác suất (bởi thực tế có thể chưa chắc nó đã khớp với một phân bố xác suất) thì có thể hiểu theo nghĩa tập mờ, nghĩa là mỗi giá trị được coi là "tốt" hay "xấu" tùy thuộc vào quan điểm của mỗi cá nhân và có thể trong cái "tốt" có những giá trị "rất tốt" hay "rất rất tốt",v.v... Khi các trạng thái được định nghĩa theo cách mờ hóa ở những mức độ khác nhau, xích Markov có thể đóng vai trò tìm mối quan hệ giữa giá trị hiện tại và

giá trị tương lai (xích Markov bậc một) hoặc giữa giá trị lịch sử với giá trị tương lai (xích Markov bậc cao).

2. Mục tiêu của luận án: Trên cơ sở những hướng nghiên cứu có thể phát triển và mở rộng đã đề xuất trong mục tính cấp thiết, luận án đề xuất mô hình kết hợp (combining approach) mới trong dự báo nhằm đơn giản hóa mô hình đồng thời cải thiện độ chính xác trong dự báo.

Mục tiêu cụ thể: luận án tập trung vào *hai vấn đề*:

Thứ nhất, mô hình hóa chuỗi thời gian bởi những trạng thái mà trong đó mỗi trạng thái là một phân phối xác suất tất định (phân phối chuẩn đối với chuỗi thời gian có giá trị thực trong khoảng $(0;1)$ hoặc phân phối Poisson đối với chuỗi thời gian có giá trị là số tự nhiên). Việc lựa chọn phân phối xác suất này phụ thuộc vào đặc trưng của loại dữ liệu cũng như độ phức tạp của tính toán nhưng vẫn đáp ứng sai số dự báo. Dựa vào kết quả thực nghiệm để đánh giá sự phù hợp của mô hình.

Thứ hai, kết hợp xích Markov và chuỗi thời gian mờ thành mô hình mới nhằm cải thiện độ chính xác của dự báo. Hơn nữa, mở rộng mô hình với xích Markov bậc cao nhằm tương thích với những dữ liệu có tính chất thời vụ.

3. Đối tượng nghiên cứu của luận án: là **c**ác mô hình dự báo chuỗi thời gian trong tài chính cũng như những chỉ số kinh tế - xã hội.

4. Phạm vi nghiên cứu của luận án: mô hình Markov ẩn, mô hình kết hợp xích Markov và chuỗi thời gian mờ trong dự báo chuỗi thời gian. Luận án nghiên cứu làm tăng độ chính xác của mô hình dự báo mà không đề cập đến hiệu năng tính toán.

5. Phương pháp nghiên cứu

Từ các mô hình đã biết xây dựng mối quan hệ giữa chúng để chọn ra những mô hình tương hỗ lẫn nhau, khắc phục những nhược điểm của mỗi mô hình đã được chỉ ra để xây dựng mô hình kết hợp. Xây dựng thuật toán cho mô hình mới

dựa trên các mối quan hệ đã được thiết lập. Cài đặt chương trình thử nghiệm bằng ngôn ngữ lập trình R và chạy thử nghiệm trên các dữ liệu thực.

Lựa chọn dữ liệu huấn luyện và dữ liệu kiểm tra trùng khớp với các mô hình đã công bố trên thế giới. Chạy mô hình đề xuất trên cùng dữ liệu với các mô hình đã có để so sánh độ chính xác của dự báo. Khi so sánh với các mô hình dự báo chuỗi thời gian có kết quả tốt được công bố gần đây nhất.

6. Đóng góp của luận án các đóng góp của luận án tương ứng với hai mục tiêu nghiên cứu đã đề ra như sau:

Thứ nhất, mô hình hóa chuỗi thời gian bởi những trạng thái là những phân phối chuẩn. Liên kết các trạng thái hiện tại và tương lai bởi xích Markov. Cả hai công việc được thực hiện tự động dựa trên mô hình HMM.

Thứ hai, xây dựng thành công mô hình kết hợp xích Markov và chuỗi thời gian mờ trong dự báo chuỗi thời gian bao gồm cả phát triển mô hình cho xích Markov bậc cao.

Các công trình đã công bố liên quan đến luận án bao gồm: 01 bài báo công bố trên Tạp chí Tin học và Điều kiển học [A5]; 02 bài báo công bố trên tạp chí quốc tế (có chỉ số ESCI) [A3, A4]; 02 báo cáo công bố trong hội thảo quốc gia @ [A2, A1].

7. Bố cục của luận án gồm phần mở đầu và ba chương nội dung, phần kết luận và danh mục các tài liệu tham khảo.

Phần mở đầu trình bày tổng quan về các nội dung nghiên cứu của luận án bao gồm chỉ ra những hạn chế của các mô hình dự báo đã biết cũng như đề xuất mô hình mới, đồng thời giới thiệu những đóng góp đã đạt được của luận án. Các nội dung chính của luận án được trình bày trong 3 chương còn lại. Nội dung của mỗi chương có thể tóm tắt lại như sau:

Chương 1 trình bày những nghiên cứu liên quan đến luận án, phân tích những hạn chế của các mô hình hiện tại. Nghiên cứu tổng quan xích Markov và mô hình Marko ẩn cũng như chuỗi thời gian mờ. Các nghiên cứu tổng quan của chương này

tập trung đi vào khai thác cách mà xích Markov và mô hình HMM có thể ứng dụng trong dự báo chuỗi thời gian cũng như các ứng dụng tiềm năng khác. Để phục vụ nghiên cứu của luận án cho việc xây dựng mô hình mới, phương pháp ước lượng tham số của các mô hình được trình bày chỉ tiết. Chương này cũng chỉ ra kết quả của một số mô hình dự báo theo hướng kết hợp gần đây. Những kết quả mà luận án sẽ so sánh trên dữ liệu tương ứng.

Chương 2 trình bày lập luận dẫn đến đề xuất áp dụng mô hình HMM trong dự báo chuỗi thời gian. Cụ thể, mô hình hóa chuỗi thời gian thành những trạng thái trong đó: (1) mỗi trạng thái là một phân phối xác suất (việc lựa chọn phân phối xác suất này phụ thuộc vào đặc điểm của dữ liệu cần dự báo); (2) các trạng thái theo thời gian tuân theo một xích Markov rời rạc thuần nhất và chính quy. Sau đó, mô hình được thực nghiệm trên dữ liệu chỉ số VN-Index cũng như một số dữ liệu khác để đánh giá hiệu quả dự báo của mô hình. Cuối chương luận án phân tích những hạn chế và sự không phù hợp của mô hình dự báo với phân phối xác suất tất định làm động cơ cho mô hình kết hợp đề xuất ở Chương 3.

Chương 3 trình bày mô hình kết hợp xích Markov và chuỗi thời gian mờ trong dự báo chuỗi thời gian. Trong đó, mô hình chuỗi thời gian mờ làm mờ hóa tập nền của dữ liệu nhằm xác định các trạng thái của tập nền bởi những tập mờ theo thời gian. Giả sử rằng các trạng thái này tuân theo một xích Markov có phân phối dừng thì ma trận xác suất chuyển cho biết trạng thái dự báo tương lai. Tính ngược từ tập mờ trả về giá trị của chuỗi thời gian cần dự báo. Chương này cũng trình bày mô hình mở rộng cho xích Markov bậc cao với hai khái niệm xích Markov bậc cao cổ điển (CMC) và xích Markov bậc cao cải tiến (IMC). Mô hình sau đó thực nghiệm với các tập dữ liệu tương ứng chính xác với tập dữ liệu của các mô hình so sánh hiện có.

Cuối cùng, luận án tóm tắt lại những kết quả chính của nghiên cứu về ý nghĩa khoa học và thực tiễn. Đồng thời chỉ ra một số định hướng cho nghiên cứu tiếp theo trong tương lai.

Chương 1. BÀI TOÁN ĐỀ XUẤT VÀ KIẾN THỨC TỔNG QUAN

1.1. Mở đầu

Chương này luận án trình bày các kiến thức tổng quan phục vụ nghiên cứu của nghiên cứu sinh cũng như những kết quả trực tiếp được sử dụng cho nghiên cứu. Những tính chất của khái niệm mà không sử dụng cho nghiên cứu sẽ không được đề cập đến luận án này. Cụ thể, các nội dung tổng quan chính của chương như sau:

Thứ nhất, luận án trình bày các hướng nghiên cứu dự báo chuỗi thời gian gần đây nhất và phân tích những hạn chế của nó. Từ đó đưa ra đề xuất phát triển mô hình của nghiên cứu sinh.

Thứ hai, luận án trình bày các khái niệm về xích Markov, xích Markov thuần nhất và dừng cũng như phương pháp ước lượng ma trận xác suất chuyển.

Thứ ba, luận án trình bày mô hình Markov ẩn (HMM) và các vấn đề về ước lượng tham số cũng như dự báo.

Thứ tư, luận án tổng hợp các khái niệm về chuỗi thời gian mờ và một số vấn đề sử dụng chuỗi thời gian mờ trong dự báo.

Cuối cùng, luận án đưa ra một số kết quả của các nghiên cứu được công bố gần đây của các mô hình dự báo theo hướng kết hợp các mô hình dự báo sẵn có. Các kết quả này sẽ được nghiên cứu sinh so sánh với kết quả của nghiên cứu.

Toàn bộ luận án nghiên cứu về vấn đề dự báo chuỗi thời gian bằng các mô hình khác nhau hoặc các mô hình xây mới bằng phương pháp kết hợp mô hình. Do đó, khái niệm về chuỗi thời gian trước tiên có thể được phát biểu như sau:

Định nghĩa 1.1.1. *Chuỗi thời gian là một chuỗi có thứ tự của một biến ngẫu nhiên tại các thời điểm được chia thành những khoảng thời gian bằng nhau* $X_1, X_2, \cdots, X_t$.

Như vậy, chuỗi thời gian có thể được coi là một trường hợp đặc biệt của dãy biến ngẫu nhiên $X_1, X_2, \cdots, X_t$. Các $X_t, t = 1,...,T$ có thể là một biến ngẫu nhiên cũng có thể là các biến ngẫu nhiên khác nhau. Các giá trị quan sát được do biến ngẫu nhiên X_t sinh ra tại thời điểm t thường ký hiệu là x_t. Đôi khi để thuận lợi trong cách viết và biến đổi, nhiều sách vẫn giữ ký hiệu X_t mà vẫn hiểu là giá trị quan sát.

1.2. Các nghiên cứu liên quan và hướng phát triển của luận án

Như đã đề cập trong phần mở đầu, các phương pháp dự báo chuỗi thời gian truyền thống như ARIMA hay GARCH ít nhiều bộc lộ những hạn chế. Do đó, các hướng tiếp cận mới đã được phát triển mạnh mẽ. Một lựa chọn khác cho dự báo chuỗi thời gian được phát triển gần đây hơn là mô hình mạng thần kinh nhân tạo (ANN). Các mô hình ANN không dựa trên phân phối tất định cho dữ liệu mà nó hoạt động tương tự bộ não con người, cố gắng tìm ra quy luật và đường đi của dữ liệu huấn luyện, kiểm tra thực nghiệm và tổng quát hóa kết quả. Hơn nữa, bản chất của ANN là thực hiện thông qua các ràng buộc, vì vậy nó cần rất nhiều dữ liệu huấn luyện để dự báo chính xác và hiệu quả hơn. Với cách hoạt động của nó, các mô hình ANN thường sử dụng hiệu quả hơn cho mục đích phân lớp dữ liệu [41]. Gần đây hơn, lý thuyết mới về học máy thống kê đang được nhiều nhà khoa học chú ý là phương pháp vector học máy hỗ trợ (SVM) cho bài toán phân lớp và dự báo [62, 14, 56]. Phương pháp SVM cố gắng đi tìm quy tắc quyết định có tính khái quát cao thông qua một số các tập con của tập huấn luyện, được gọi là các vector hỗ trợ. Theo đó, một ánh xạ phi tuyến được thực hiện từ không gian đầu vào lên không gian có số chiều lớn hơn. Sau đó, một siêu phẳng tối ưu sẽ được dùng để phân lớp các vector hỗ trợ được thực hiện trước khi ánh xạ ngược trở lại không gian ban đầu. Để làm được điều này, phương pháp SVM dẫn đến giải bài toán hồi quy tuyến tính. Do đó, ban đầu phương pháp SVM được sử dụng trong các bài toán phân lớp. Về sau, SVM được áp dụng rộng rãi hơn trong nhiều lĩnh vực như xấp xỉ hàm, ước lượng hồi quy và dự báo [14, 56]. Tuy nhiên, hạn chế lớn nhất của SVM là khi tập huấn luyện lớn, nó đòi hỏi lượng tính toán khổng lồ cũng như độ phức tạp của bài toán hồi quy tuyến tính trong đó.

Để khắc phục các hạn chế và phát huy các điểm mạnh của các phương pháp đã có, mộ xu thế nghiên cứu đang trở nên thịnh hành gần đây là phương tiếp cận kết hợp (CA), nghĩa là kết hợp một số phương pháp không giống nhau để tăng độ chính xác của dự báo. Rất nhiều nghiên cứu đã được thực hiện và theo hướng này và rất nhiều các mô hình kết hợp mới đã được công bố [71, 2, 3]. Một số phương pháp trong đó sử dụng xích Markov (MC) cũng như mô hình Markov ẩn (HMM). Refiul Hassan [33] đã

phát triển một mô hình hợp nhất bằng cách kết hợp một HMM với logic mờ để tạo ra các dự báo trong một ngày-trước của giá cổ phiếu. Cụ thể như sau

- Dữ liệu đầu vào là vector $x_i =< x_{i,open}, x_{i,high}, x_{i,low}, x_{i,close} >$ tương ứng với các giá trị cổ phiếu mở cửa, cao nhất, thấp nhất và đóng của của ngày thứ i.
- Mô hình HMM với tham số λ được dùng để huấn luyện cho tập dữ liệu này và các giá trị $\log Pr(\vec{x_i} | \lambda)$ (gọi là log-likelihood) chia làm 7 khoảng bằng nhau gọi là các nhóm log-likelihood. Các nhóm này đóng vai trò là các tập mờ của dữ liệu. Hàm thành viên $M(x)$ cho mỗi phần tử trong các tập mờ này là phân phối chuẩn tự sinh ra trong mô hình HMM với phân phối chuẩn.
- Luật mờ được tính như sau: Nếu x_{open} có mức M_{open} với tham số p_1, x_{high} có mức M_{high} với tham số p_2, ... thì giá trị đóng cửa dự đoán $predict_{close} = p_1 \times x_{open} + p_2 \times x_{high} + p_3 \times x_{low} + p_4 \times x_{close}$ trong đó các tham số p_i được ước lượng bằng phương pháp bình phương tối thiểu từ tập huấn luyện tương ứng với các nhóm log-likelihood (trạng thái).

Tương tự mô hình này, mô hình Markov với trọng số (các tham số tuyến tính cho mỗi trạng thái) đã được Peng [52] áp dụng trong dự báo và phân tích tỷ lệ truyền nhiễm bệnh ở tỉnh Giang Tô, Trung Quốc. Yang [69] đã kết hợp mô hình HMM để phân cụm dữ liệu thời gian. Các mô hình kết hợp này đã mang lại những kết quả có ý nghĩa trong thực tiễn cũng nhưng tăng đáng kể độ chính xác trong dự báo so với các mô hình truyền thống. Tuy nhiên, xuất hiện những tồn tại trong và nghi vấn trong mô hình cần được giải quyết như:

1. Việc phân lớp dữ liệu sử dụng HMM cho log-likelihood có thực sự hiệu quả hơn so với việc thực hiện đơn giản hơn bằng cách chia trực tiếp chuỗi tăng trưởng thành các khoảng.

2. Mối quan hệ tuyến tính giữa giá đóng cửa hôm sau so với vector gồm giá mở cửa, cao nhất, thấp nhất, đóng cửa hôm trước có thực sự tồn tại hay chỉ đơn giản là

những biến ngẫu nhiên độc lập theo thời gian. Nếu chúng độc lập, chỉ cần chuỗi đóng cửa có thể dự báo được chính nó.

Luận án sẽ thực hiện áp dụng mô hình HMM với những phân phối cụ thể cho dữ liệu có giá trị là số tự nhiên (phân phối Poisson) và dữ liệu thực (phân phối chuẩn) cho dự báo chuỗi thời gian chỉ số chứng khoán trong Chương 2 để kiểm tra độ chính xác dự báo so với các mô hình cổ điển như ARIMA hay ANN.

Các dữ liệu chuỗi thời gian tài chính nói chung đều là các dữ liệu mờ. Nghĩa là ranh giới giữa các mức độ tăng trưởng không rõ ràng phụ thuộc vào cảm quan của người đánh giá. Do vậy, việc phân lớp dữ liệu để phân tích dự báo cần được mờ hóa. Để đối phó với những dữ liệu mờ, một hướng nghiên cứu mới trong dự báo chuỗi thời gian được mở ra gần đây là sử dụng mô hình chuỗi thời gian mờ (FTS). Kết quả đầu tiên cần được kể đến trong việc áp dụng lý thuyết này là Song and Chissom [60]. Những nghiên cứu tập trung theo hướng cải thiện các mô hình chuỗi thời gian mờ và tìm cách áp dụng vào bài toán dự báo. Jilani et al. and Nan et al.kết hợp mô hình Heuristic với chuỗi thời gian mờ để nâng cao độ chính xác của mô hình [46].Chen và Hwang mở rộng thêm các chuỗi thời gian mờ vào mô hình Binary [17] và sau đó Hwang and Yu phát triển thành mô hình N bậc để dự báo chỉ số chứng khoán [37].

Trong một bài báo gần đây [61], BaiQing Sun et al. đã mở rộng mô hình mờ cho chuỗi thời gian mờ đa biến để dự báo giá tương lai của thị trường chứng khoán. Mô hình chuỗi thời gian mờ của tác giả thực hiên trên 3 chuỗi gồm: chỉ số CSI300 (300 mã chứng khoán Trung Quốc); giá mua (spot price) và khối lượng giao dịch. Các chuỗi tăng trưởng tương ứng của 3 chuỗi này lần lượt được mờ hóa theo 6 tập $(A_1,...,A_6)$, 4 tập (B_1,B_2,B_3,B_4) và 3 tập (C_1,C_2,C_3). Mục tiêu của dự báo là các A_i. Luật mờ được phát hiện từ $A_{i_1},B_{i_2},C_{i_3}\rightarrow A_{j_1},A_{j_2},A_{j_k}$, trong đó $i_j=0$ có nghĩa là khuyết A_{i_j},B_{i_j} hoặc C_{i_j} tương ứng.

Giá trị dự báo của mô hình dựa vào các luật tính ngược từ các quan hệ mờ phát hiện ở trên. Cụ thể:

- Nếu không tồn tại quan hệ mờ của một nhóm nào đó, tức $A_{i_1},B_{i_2},C_{i_3}\rightarrow$, thì

giá trị dự báo

$$AF(t) = d_{i_1}$$

với d_{i_1} là giá trị chính giữa khoảng A_{i_1}.

- Nếu $A_{i_1}, B_{i_2}, C_{i_3} \rightarrow A_{j_1}$ thì giá trị dự báo

$$AF(t) = d_{j_1}.$$

- Nếu $A_{i_1}, B_{i_2}, C_{i_3} \rightarrow A_{j_1}, A_{j_2}, ..., A_{j_k}$ thì

$$AF(t) = \frac{\sum_{k}^{p} n_{j_k} d_{j_k}}{\sum_{k}^{p} n_{j_k}} \qquad (1.2.1)$$

với n_{j_k} là số lần xuất hiện quan hệ $A_{i_1}, B_{i_2}, C_{i_3} \rightarrow A_{j_k}$.

Như vậy, mô hình của Sun cần phải sử dụng đến những chuỗi phụ để dự báo chuỗi mục tiêu nhưng chưa chỉ ra được tương quan giữa các chuỗi theo thời gian. Thực tế, tổng giá trị dao dịch tăng nhưng chỉ số chứng khoán có khi tăng cũng có khi giảm. Vì vậy mối quan hệ mờ tìm được giữa chúng trong tập huấn luyện không hẳn sẽ phản ánh trong tương lai.

Hơn nữa, cách tính giá trị dự báo theo trung bình của tần số xuất hiện như trong (1.2.1) tương đương với kỳ vọng của một phân phối xác xuất. Điều này tương tự với cách dự báo trong một xích Markov nhưng thuật toán tìm kiếm và liệt kê phức tạp hơn. Do đó, mô hình có thể đơn giản hóa bằng cách kết hợp chuỗi thời gian mờ (nhằm phân nhóm dữ liệu) với một xích Markov (tương đương với tìm quan hệ mờ một cách tự động). Một khi mô hình thay thế được tính toán trên cũng dữ liệu, rõ ràng các tính toán sẽ đơn giản hơn trong khi có thể vẫn đảm bảo được độ chính xác dự báo. Mô hình như vậy luận án sẽ xây dựng trong Chương 3.

Một nghiên cứu khác của Qisen Cai et al.[13] đã kết hợp mô hình dự báo chuỗi thời gian mờ bậc cao với thuật toán tối ưu hóa đàn kiến và tự hồi quy để có được một kết quả tốt hơn. Cụ thể như sau

- Chuỗi tăng trưởng $\{(y_k)\}$ của dữ liệu được chia thành các tập mờ $A_i, i = 1,...,n$.
- Tìm các quan hệ mờ bậc cao cho chuỗi thời gian mờ $\{F(t)\}$ tương ứng dạng $F(t-k) \to F(t)$ nhằm dự báo các giá trị $\hat{y}_{t,k}$ tương ứng của chuỗi tăng trưởng.
- Giá trị dự báo cuối cùng được tính bởi

$$predicted_t = \phi_1 \hat{y}_{t,1} + \phi_2 \hat{y}_{t,2} + \cdots + \phi_k \hat{y}_{t,k}$$

trong đó các trọng số tuyến tính ϕ_i được ước lượng tối ưu bằng cách kết hợp mô hình tự hồi quy và thuật toán tối ưu hóa đàn kiến để tìm bộ tham số tốt nhất đối với dữ liệu huấn luyện.

Cũng như nghiên cứu của Sun, nghiên cứu của Cai cho thấy việc sử dụng quan hệ mờ bậc cao kết hợp với hồi quy tuyến tính tương ứng với một xích Markov bậc cao cải tiến mà thuật toán ước lượng tham số của nó tự động và đơn giản hơn nhiều. Chính vì vậy, mô hình dạng này có thể đề xuất thay thế bởi mô hình Markov bậc cao cải tiến mà luận án sẽ thực hiện và so sánh trong Chương 3.

Ở Việt Nam, mô hình chuỗi thời gian mờ gần đây cũng đã được áp dụng trong một số lĩnh vực cụ thể nhưng trong lĩnh vực dự báo chuỗi thời gian vẫn còn khá ít. Có thể kể đến nghiên cứu của Nguyễn Duy Hiếu và cộng sự [B2] trong phân tích ngữ nghĩa. Ngoài ra, các công trình của tác giả Nguyễn Công Điều [B3, B4] đã kết hợp mô hình chuỗi thời gian mờ với một số kỹ thuật điều chỉnh tham số trong thuật toán hay những đặc trưng riêng của dữ liệu để làm tăng độ chính xác của dự báo. Nghiên cứu của tác giả Nguyễn Cát Hồ [B1] đã ứng dụng đại số gia tử vào dự báo chuỗi thời gian mờ cho thấy độ chính xác dự báo cải thiện hơn một số mô hình hiện có.

Nghiên cứu của Nguyễn Công Điều chỉ dừng lại ở điều chỉnh thuật toán tối ưu hóa tham số từ dữ liệu huấn luyện nhằm tăng độ chính xác của mô hình chuỗi thời gian mờ cổ điển thực hiện trên chỉ 1 bộ dữ liệu. Do đó, tính ưu việt so với các mô hình khác trong dự báo chuỗi thời gian bất kỳ chưa được kiểm chứng. Đối với hương tiếp cận đại số gia tử (ĐSGT) vào dự báo chuỗi thời gian là một hướng đi không phổ biến bởi

ĐSGT phân tích cấu trúc ngữ nghĩa cho những biến ngôn ngữ. Trong nghiên cứu của các tác giả trong [B1] chỉ thực hiện mô hình trên 1 dữ liệu số lượng tiếp nhận sinh viên của trường đại học Mỹ, một dữ liệu mà có tính ổn định cao. Trong khi đó, độ chính xác của mô hình dự báo cho chuỗi thời gian bất kỳ, đặc biệt là chuỗi thời gian tài chính vẫn là một câu hỏi bởi các chuỗi thời gian này mang tính ngẫu nhiên cao hơn nhiều. Chính vì lẽ đó, luận án sẽ không đi theo hướng này để phát triển mô hình dự báo cho chuỗi thời gian nói chung.

Từ các phân tích trên, luận án sẽ chỉ ra ưu điểm và hạn chế của mô hình HMM trong dự báo chuỗi thời gian trong Chương 2 đồng thời tập trung xây dựng mô hình dự báo chuỗi thời gian dựa trên mô hình kết hợp xích Markov và chuỗi thời gian mờ nhằm đơn giản hóa những mô hình mang tính tương đương đã đề cập trước đó trong Chương 3. Các nghiên cứu được thực hiện trên nhiều tập dữ liệu tài chính khác nhau và so sánh với nhiều mô hình sẵn có.

Các mục tiếp theo, luận án trình bày các kiến thức tổng quan về xích Markov và chuỗi thời gian mờ gồm các phần kiến thức được sử dụng trong quá trình xây dựng các mô hình dự báo ở các chương tiếp theo.

1.3. Xích Markov

Trong lý thuyết xác suất và các lĩnh vực liên quan, quá trình Markov (đặt theo tên của nhà toán học người Nga Andrey Markov) là một quá trình ngẫu nhiên thỏa mãn một tính chất đặc biệt, gọi là tính chất Markov [29] (còn gọi là tính mất trí nhớ). Tính chất này giúp dự báo được tương lai chỉ dựa vào trạng thái hiện tại. Điều này cũng có nghĩa trạng thái tương lai và quá khứ là độc lập nhau. Tuy nhiên về sau, quá trình Markov được mở rộng thành Markov bậc cao [20], trong đó tương lai phụ thuộc vào hiện tại và một quãng thời gian nào đó trong quá khứ.

Xích Markov là quá trình Markov đặc biệt mà trong đó hoặc có trạng thái rời rạc hoặc thời gian rời rạc. Quá trình Markov được nhà toán học Markov bắt đầu nghiên cứu từ khoảng đầu thế kỷ 20 mặc dù có nhiều nghiên cứu hàng trăm năm trước đó về quá trình này nhưng dưới dạng các biến ngẫu nhiên phụ thuộc. Hai ví dụ quan trọng nhất của quá trình Markov là quá trình Wiener (hay chuyển động Brownian) và quá

trình Poisson [45]. Hai quá trình này được coi là quan trọng nhất và là trung tâm của lý thuyết quá trình ngẫu nhiên.

Xích Markov có rất nhiều ứng dụng với vai trò là các mô hình xác suất trong các quá trình thực tế [40, 31, 42]. Thuật toán được biết đến là PageRank được thực hiện khởi nguồn cho công cụ tìm kiếm của Google được dựa trên xích Markov [48].

Đối với các dữ liệu thống kê trong thực tế, các mô hình thường sử dụng các biến rời rạc thậm chí rời rạc hóa cho thực nghiệm. Đối với mỗi trạng thái kinh tế, nó xuất hiện một lần trong dữ liệu huấn luyện và không chuyển sang trang thái khác (trạng thái hấp thụ) không có nghĩa trong tương lai trạng thái đó mãi duy trì ở đó. Vì vậy, luận án chỉ nghiên cứu áp dụng mô hình đối với xích Markov cả thời gian rời rạc và trạng thái rời rạc, thuần nhất và chính quy.

1.3.1. Các định nghĩa

Ta xét một hệ thống kinh tế hoặc một hệ thống vật chất S với m trạng thái có thể, ký hiệu bởi tập I :

$$I = \{1, 2, ..., m\}.$$

hệ thống S tiến hóa ngẫu nhiên trong thời gian rời rạc ($t = 0,1,2,...,n,...$), và đặt C_n là biến ngẫu nhiên tương ứng với trạng thái của hệ thống S ở thời điểm $n(C_n \in I)$.

Định nghĩa 1.3.1. *Dãy biến ngẫu nhiên ($C_n, n \in \mathbb{N}$) là một xích Markov nếu và chỉ nếu với tất cả $c_0, c_1, ..., c_n \in I$:*

$$Pr(C_n = c_n \mid C_0 = c_0, C_1 = c_1, ..., C_{n-1} = c_{n-1}) = Pr(C_n = c_n \mid C_{n-1} = c_{n-1}) \quad (1.3.1)$$

(với điều kiện xác suất này có nghĩa)

Định nghĩa 1.3.2. *Một xích Markov được gọi là thuần nhất nếu chỉ nếu xác suất trong (1.3.1) không phụ thuộc vào n và không thuần nhất trong các trường hợp còn lại.*

Hiện tại, ta chỉ xét trường hợp thuần nhất mà với nó ta viết:

$$Pr(C_n = c_n \mid C_{n-1} = c_{n-1}) = \gamma_{ij},$$

và ta đưa ra ma trận $\mathbf{\Gamma}$ được định nghĩa:

$$\mathbf{\Gamma} = \left[\gamma_{ij}\right].$$

Các phần tử của ma trận $\mathbf{\Gamma}$ có các tính chất:

(i) $\gamma_{ij} \geq 0$, với mọi $i, j \in I$,

(ii) $\sum_{j \in I} \gamma_{ij} = 1$, với mọi $i \in I$.

Một ma trận $\mathbf{\Gamma}$ thỏa mãn 2 điều kiện này được gọi là một *ma trận Markov* hay *ma trận chuyển.*

Đối với mọi ma trận chuyển, ta có thể liên kết với một *đồ thị chuyển* với các đỉnh là các trạng thái. Tồn tại một *cung* giữa đỉnh i và j nếu và chỉ nếu $\gamma_{ij} > 0$.

Để định nghĩa đầy đủ sự tiến triển của một xích Markov, cần thiết phải cố định một *phân phối ban đầu* cho trạng thái C_0, chẳng hạn, một véc tơ:

$$\mathbf{p} = (p_1, p_2, ..., p_m),$$

sao cho:

$$p_i \geq 0, i \in I,$$

$$\sum_{i \in I} p_i = 1.$$

Với mọi i, p_i được hiểu là *xác suất đầu tiên* của sự bắt đầu có trạng thái i:

$$p_i = Pr(C_0 = i).$$

Vấn đề ở chương này ta chỉ dừng lại ở việc xem xét xích Markov thuần nhất mà được đặc trưng bởi cặp $(\mathbf{p}, \mathbf{\Gamma})$.

Nếu $C_n = i$ h.c.c (hầu chắc chắn), đó nghĩa là hệ thống bắt đầu với xác suất bằng 1 từ trạng thái i, thì véc tơ $\boldsymbol{p}$ sẽ là:

$$p_j = \delta_{ij}$$

Bây giờ ta giới thiệu *các xác suất chuyển của bậc* $\gamma_{ij}^{(n)}$, được định nghĩa:

$$\gamma_{ij}^{(n)} = Pr(C_{\nu+n} = j \mid C_\nu = i).$$

Từ tính chất Markov (1.3.1), dễ dàng có được cách tính toán điều này nhờ mối quan hệ với $C_{\nu+1}$, ta có:

$$\gamma_{ij}^{(2)} = \sum_k \gamma_{ik} \gamma_{kj}. \qquad (1.3.2)$$

Viết theo ký hiệu ma trận:

$$\mathbf{\Gamma}^{(2)} = \left[\gamma_{ij}^{(2)}\right],$$

Ta thấy quan hệ (1.3.2) tương đương với

$$\mathbf{\Gamma}^{(2)} = \mathbf{\Gamma}^2.$$

Sử dụng quy nạp, dễ dàng chứng minh được rằng, nếu viết

$$\mathbf{\Gamma}^{(n)} = \left[\gamma_{ij}^{(n)}\right],$$

ta đạt được với mọi $n \geq 1$:

$$\mathbf{\Gamma}^{(n)} = \mathbf{\Gamma}^n. \tag{1.3.3}$$

Chú ý rằng quan hệ (1.3.3) làm đơn giản hóa ma trận chuyển trong n bước thì bằng lũy thừa n lần ma trận $\mathbf{\Gamma}$.

Với các phân phối biên duyên có quan hệ với C_n, ta định nghĩa cho $i \in I$ và $n \geq 0$:

$$\gamma_i(n) = Pr(C_n = i).$$

Những xác suất này được tính toán theo:

$$\gamma_i(n) = \sum_j \gamma_j \gamma_{ji}^{(n)}, \ i \in I. \tag{1.3.4}$$

Nếu ta viết:

$$\gamma_{ji}^0 = \delta_{ji} \text{ hoặc } \mathbf{\Gamma}^{(0)} = \mathbf{I},$$

thì quan hệ (1.3.4) đúng với mọi $n \geq 0$. Nếu:

$$\boldsymbol{\gamma}(\mathrm{n}) = \left(\gamma_1(n), \gamma_2(n), ..., \gamma_m(n)\right),$$

thì quan hệ (1.3.4) có thể được tính qua ký hiệu ma trận: $\boldsymbol{\gamma}^{(n)} = \mathbf{p}\mathbf{\Gamma}^{\mathbf{n}}$.

Định nghĩa 1.3.3. *Một ma trận Markov* $\mathbf{\Gamma}$ *được gọi là chính quy nếu tồn tại một số nguyên dương* k *sao cho tất cả các phần tử của ma trận* $\mathbf{\Gamma}^{(k)}$ *là thực sự dương.*

Từ quan hệ (1.3.3), $\mathbf{\Gamma}$ là *chính quy* nếu và chỉ nếu tồn tại một số nguyên $k > 0$ sao cho tất cả các phần tử của ma trận lũy thừa bậc k của $\mathbf{\Gamma}$ là dương thực sự.

Ví dụ 1.3.1:

(i) Nếu:

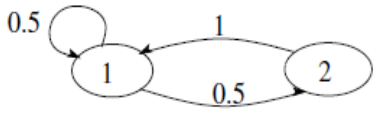
$$\mathbf{\Gamma} = \begin{bmatrix} 0,5 & 0,5 \\ 1 & 0 \end{bmatrix}$$

ta có

$$\mathbf{\Gamma}^2 = \begin{bmatrix} 0,75 & 0,25 \\ 0,5 & 0,5 \end{bmatrix}$$

vì vậy $\mathbf{\Gamma}$ là chính quy.

Sơ đồ chuyển liên kết với $\mathbf{\Gamma}$ được cho trong Hình (1.3.1).



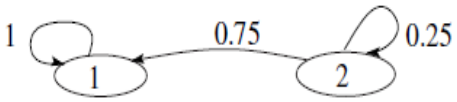
Hình 1.3.1. *Ví dụ ma trận Markov chính quy*

(ii) Nếu:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 0 \\ 0,75 & 0,25 \end{bmatrix}$$

$\mathbf{\Gamma}$ không chính quy bởi vì với mọi số nguyên k,

$$\gamma_{12}^{(k)} = 0.$$



Hình 1.3.2. *Ví dụ ma trận Markov không chính quy*

Đồ thị chuyển trong trường hợp này được mô tả trong Hình (1.3.2).

Cũng như vậy đối với ma trận:

$$\mathbf{\Gamma} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

(iii) Mọi ma trận $\mathbf{\Gamma}$ mà các phần tử của nó là thực sự dương thì chính quy.

Ví dụ:

$$\begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \begin{bmatrix} 0,7 & 0,2 & 0,1 \\ 0,6 & 0,2 & 0,2 \\ 0,4 & 0,1 & 0,5 \end{bmatrix}$$

1.3.2. Phân loại trạng thái xích Markov

Lấy $i \in I$ và đặt $d(i)$ là ước chung lớn nhất của tập các số nguyên n sao cho

$$\gamma_{ii}^{(n)} > 0.$$

Định nghĩa 1.3.4. *Nếu $d(i) > 1$, trạng thái i được gọi là tuần hoàn chu kỳ $d(i)$. Nếu $d(i) = 1$, thì trạng thái i không tuần hoàn.*

Dễ thấy, nếu $\gamma_{ii} > 0$ thì i là không tuần hoàn. Tuy nhiên, điều ngược lại chưa chắc đúng.

Chú ý 1.3.1. Nếu Γ là chính quy thì tất cả các trạng thái đều không tuần hoàn.

Định nghĩa 1.3.5. *Một xích Markov mà tất cả các trạng thái của nó không tuần hoàn được gọi là xích Markov không tuần hoàn.*

Từ đây, ta chỉ nghiên cứu loại xích Markov này.

Định nghĩa 1.3.6. *Một trạng thái i được gọi là vươn tới trạng thái j (viết là $i \triangleright j$) nếu tồn tại số nguyên dương n sao cho*

$$\gamma_{ij}^{n} > 0.$$

$i \not\to j$ nghĩa là i không vươn tới được j.

Định nghĩa 1.3.7. *Trạng thái i và j được gọi là liên thông nếu $i \triangleright j$ và $j \triangleright i$, hoặc nếu $i = j$. Ta viết $i \lhd\rhd j$.*

Định nghĩa 1.3.8. *Trạng thái i được gọi là cốt yếu nếu nó liên thông với mọi trạng thái mà nó vươn tới; trường hợp ngược lại gọi là không cốt yếu.*

Quan hệ $\lhd\rhd$ xác định một quan hệ tương đương trên không gian trạng thái I dẫn tới một sự chia lớp trên I. Lớp tương đương chứa i được ký hiệu bởi $Cl(i)$.

Định nghĩa 1.3.9. *Xích Markov được gọi là không khai triển được nếu chỉ tồn tại duy nhất một lớp tương đương trên nó.*

Dễ thấy, nếu Γ là chính quy, xích Markov vừa là không khai triển được, vừa không tuần hoàn. Xích Markov vừa không khai triển được (tức là chỉ có 1 lớp tương đương), vừa không tuần hoàn được gọi là xích Markov *ergodic*.

Dễ dàng chỉ ra rằng, nếu trạng thái i là cốt yếu (không cốt yếu) thì tất cả các phần tử của lớp $C(i)$ cũng cốt yếu (không cốt yếu) (xem Chung (1960)) [21].

Ta có thể gọi là lớp cốt yếu hoặc lớp không cốt yếu.

Định nghĩa 1.3.10. *Tập con E của không gian trạng thái I được gọi là đóng nếu:*

$$\sum_{j\in E}\gamma_{ij}=1, \text{với mọi } i\in E.$$

Có thể chỉ ra rằng mọi lớp cốt yếu là đóng nhỏ nhất. Xem Chung (1960) [21].

Định nghĩa 1.3.11. *Trạng thái $i\in I$ của xích Markov (C_t) được gọi là hồi quy nếu tồn trại trạng thái $j\in I$ và $n\in\mathbb{N}$ sao cho $\gamma_{ji}^{n}>0$. Ngược lại, i được gọi là trạng thái chuyển tiếp (dịch chuyển).*

Mệnh đề 1.3.1. (Định lý khai triển) [21]: *Không gian trạng thái I của mọi xích Markov đều có thể phân chia thành $r(r\geq 1)$ tập con $C_1, C_2, ..., C_r$, tạo thành một sự chia lớp, sao cho mỗi tập con C_i là một và chỉ một trong các loại:*

(i) một tập đóng cốt yếu hồi quy dương.

(ii) một tập không đóng, dịch chuyển không cốt yếu.

Chú ý 1.3.2.

(1) Nếu một lớp không cốt yếu giảm tới tập đơn $\{i\}$, thì có 2 khả năng:

a) Tồn tại một số nguyên dương N sao cho:

$$0<p_{ii}^{N}<1.$$

b) Số N trong a) không tồn tại. Trong trường hợp này, trạng thái i được gọi là *trạng thái không trở lại*.

(2) Nếu tập đơn $\{i\}$ lập thành một lớp cốt yếu, thì

$$p_{ii}=1$$

và trạng thái i được gọi là *trạng thái hấp dẫn*.

(3) Nếu $m=\infty$, có thể có 2 loại lớp khác nhau trong định đính phân ly:

a) đóng cốt yếu chuyển tiếp,

b) các lớp không đóng cốt yếu hồi quy.

Các tài liệu trên xích Markov đưa ra điều kiện cần và đủ cho sự hồi quy và sự chuyển tiếp [21].

Mệnh đề 1.3.2. [21]

(i) Trạng thái i là chuyển tiếp nếu và chỉ nếu

$$\sum_{n=1}^{\infty} \gamma_{ii}^{(n)} < \infty.$$

Trong trường hợp này, với mọi $k \in I$:

$$\sum_{n=1}^{\infty} \gamma_{ki}^{(n)} < \infty,$$

và đặc biệt:

$$\lim_{n\to\infty} \gamma_{ki}^{(n)} = 0, \forall k \in I.$$

(ii) Trạng thái i là hồi quy nếu và chỉ nếu

$$\sum_{n=1}^{\infty} \gamma_{ii}^{(n)} = \infty.$$

Trong trường hợp này:

$$k \lhd\rhd i \Rightarrow \sum_{n=1}^{\infty} \gamma_{ki}^{(n)} = \infty,$$

và

$$k \mathrel{C} i \Rightarrow \sum_{n=1}^{\infty} \gamma_{ki}^{(n)} = 0.$$

Các mô hình sử dụng xích Markov ở Chương 2 và Chương 3 được giả sử rằng các trạng thái là chuyển tiếp, có nghĩa là nó không dừng lại ở trạng thái nào nhằm đảm bảo với quy luật tiến triển của chuỗi thời gian trong thực tế. Trong thực tế, một trạng thái kinh tế bất kỳ không thể duy trì mãi mãi ở trạng thái đó. Trong trường hợp ước lượng ma trận xác suất chuyển từ tập huấn luyện có thể tồn tại một trạng thái không chuyển tiếp đến bất kỳ trạng thái khác (do tập huấn luyện là hữu hạn), ta cần hiệu chỉnh

xác xuất chuyển cho trạng thái đó bằng cách cố định cho nó một phân phối xác suất nhất định hoặc giảm số lượng tập huấn luyện đến khi nó không bị hấp thụ nữa.

1.3.3. Ước lượng ma trận Markov

Phần này luận án trình bày phương pháp ước lượng tham số của xích Markov đã được biết đến rộng rãi trong lĩnh vực thống kê. Trên cơ sở đó, phương pháp ước lượng sẽ được nhúng vào trong mô hình kết hợp mà luận án đề xuất.

Xét xích Markov (C_t), $t = 1, 2, ...$ và giả sử quan sát được n các trạng thái xảy ra $c_1, c_2, ..., c_n$. Ký hiệu $c^n \equiv c_1, c_2, ..., c_n$ sinh bởi cá biến ngẫu nhiên C^n thì hàm hợp lý của ma trận xác suất chuyển được cho bởi

$$Pr(C^n = c^n) = Pr(C_1 = c_1)\prod_{t=2}^{n} Pr\left(C_t = c_t \mid C^{t-1} = c^{t-1}\right)$$

$$= Pr(C_1 = c_1)\prod_{t=2}^{n} Pr\left(C_t = c_t \mid C_{t-1} = c_{t-1}\right)$$

$$= Pr(C_1 = c_1)\prod_{t=2}^{n} \gamma_{c_{t-1}c_t}$$

Định nghĩa số lần chuyển $n_{ij} \equiv$ số lần mà trạng thái i chuyển tiếp theo sau là trạng thái j trong dãy C^n, khi đó hàm hợp lý (likelihood) có dạng

$$L(p) = Pr(C_1 = c_1)\prod_{i=1}^{k}\prod_{j=1}^{k} \gamma_{ij}^{n_{ij}}$$

Ta cần tìm cực đại hàm hợp lý $L(p)$ với các ẩn là γ_{ij}. Để giải quyết bài toán này đơn giản, trước tiên ta lấy logarit của $L(p)$ để thành hàm tổng nhằm mục đích lấy đạo hàm dễ dàng.

$$\mathcal{L}(p) = \log L(p) = \log Pr(C_1 = c_1) + \sum_{i,j} n_{ij} \log \gamma_{ij}$$

Do ràng buộc

$$\sum_{j} \gamma_{ij} = 1,$$

nên với mỗi $i, \gamma_{i1} = 1 - \sum_{j=2}^{m} \gamma_{ij}$, lấy đạo hàm theo tham số

$$\frac{\partial \mathcal{L}}{\partial \gamma_{ij}} = \frac{n_{ij}}{\gamma_{ij}} - \frac{n_{i1}}{\gamma_{i1}}$$

Cho đạo hàm bằng 0 đạt được tại γ_{ij} ta có

$$\frac{n_{ij}}{\hat{\gamma}_{ij}} = \frac{n_{i1}}{\hat{\gamma}_{i1}}$$

vậy theo tính chất dãy tỉ số với mọi $j \neq 1$

$$\frac{n_{ij}}{n_{i1}} = \frac{\hat{\gamma}_{ij}}{\hat{\gamma}_{i1}} \tag{1.3.5}$$

Từ (1.3.5) nên

$$\hat{\gamma}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{m} n_{ij}} \tag{1.3.6}$$

1.3.4. Phân phối dừng của xích Markov

Xét một xích Markov không tuần hoàn, không phân tích được mà là hồi quy dương.

Giả sử giới hạn sau tồn tại:

$$\lim_{n \to \infty} \gamma_j(n) = \pi_j, \ j \in I \tag{1.3.7}$$

bắt đầu với $C_0 = i$.

Quan hệ

$$\gamma_j(n+1) = \sum_{k \in I} \gamma_k(n) \gamma_{kj} \tag{1.3.8}$$

trở thành:

$$\gamma_{ij}^{(n+1)} = \sum_{k \in I} \gamma_{ik}^{(n)} \gamma_{kj}, \tag{1.3.9}$$

vì

$$\gamma_j(n) = \gamma_{ij}^{(n)}.$$

Từ không gian trạng thái I là hữu hạn, từ (1.3.7) và (1.3.8) ta đạt được:

$$\pi_j = \sum_{k \in I} \pi_k \gamma_{kj}, \tag{1.3.10}$$

và từ (1.3.9)

$$\sum_{i \in I} \pi_i = 1. \tag{1.3.11}$$

Đẳng thức

$$\lim_{n \to \infty} \gamma_{ij}^{(n)} = \pi_j \tag{1.3.12}$$

được gọi là *đẳng thức ergodic*, do giá trị của giới hạn trong (1.3.12) độc lập với trạng thái ban đầu i.

Từ kết quả (1.3.12) và (1.3.4), ta thấy rằng với mọi phân phối ban đầu $\boldsymbol{\pi}$:

$$\lim_{n \to \infty} \pi_i(n) = \lim_{n \to \infty} \sum_j \pi_j \gamma_{ji}^{(n)},$$

$$= \sum_j \pi_j \pi_i,$$

vì vậy:

$$\lim_{n \to \infty} \gamma_i(n) = \pi_i.$$

Điều này chỉ ra rằng dáng điệu tiệm cận của xích Markov được cho bởi sự tồn tại (hoặc không tồn tại) của giới hạn của ma trận $\boldsymbol{\Gamma}^{\mathbf{n}}$.

Một kết quả chuẩn mực về dáng điệu tiệm cận của $\boldsymbol{\Gamma}^{\mathbf{n}}$ được đưa ra ở mệnh đề tiếp theo. Chứng minh của nó xem ở Chung (1960) [21], Parzen (1962) [50] hoặc Feller (1957) [28].

Trong [21] đã chỉ ra rằng, đối với một xích Markov hữu hạn trạng thái với ma trận xác chuyển chính quy luôn tồn tại duy nhất phân phối dừng duy nhất không phụ thuộc vào phân phối ban đầu. Đối với thực tiễn, nếu một quá trình kinh tế biến đổi quanh một số trạng thái theo một xích Markov chính quy, thì phân phối xác suất tại một thời điểm bất kỳ là ổn định. Điều này có ý nghĩa quan trọng trong dự báo cũng như

quản lý rủi ro trong tài chính cũng như trong bảo hiểm. Luận án cũng cho thấy điều này ở kết quả dự báo tiến tới phân phối ổn định trong Chương 2.

1.4. Mô hình Markov ẩn

Mô hình Markov ẩn (HMM) là một mô hình dùng để đặc tả một chuỗi thời gian trong đó giả sử các giá trị của chuỗi thời gian được sinh bởi m biến ngẫu nhiên khác nhau mà các biến ngẫu nhiên này phụ thuộc theo một xích Markov. Do đó, một mô hình HMM bao gồm hai thành phần cơ bản: chuỗi $X_t, t = 1,..., T$ gồm các quan sát nhìn thấy và $C_t = i, t = 1,.., T, i \in \{1, 2,..., m\}$ là các thành phần sinh ra từ các quan sát đó. Thực chất, mô hình HMM là một trường hợp đặc biệt của mô hình trộn phụ thuộc [24] và các C_t là các thành phần trộn.

1.4.1. Định nghĩa và ký hiệu

Ký hiệu $\mathbf{X}^{(t)}$ và $\mathbf{C}^{(t)}$ biểu diễn các dữ liệu lịch sử từ thời điểm 1 đến thời điểm t, ta có thể tóm tắt mô hình đơn giản nhất của HMM như sau:

$$Pr(C_t \mid \mathbf{C}^{(t-1)}) = Pr(C_t \mid C_{t-1}), t = 2, 3,..., T.$$

$$Pr(X_t \mid \mathbf{X}^{(t-1)}, C^{(t)}) = Pr(X_t \mid C_t), t \in \mathbb{N}$$

Như vậy, thành phần thứ nhất là quá trình tham số $\{C_t : t = 1, 2,...\}$ không quan sát được (ẩn) thỏa mãn tính chất Markov, thành phần thứ hai là quá trình trạng thái phụ thuộc (phân bố phụ thuộc vào mỗi trạng thái) $\{X_t, t = 1, 2,...\}$ sao cho, khi C_t xác định thì phân phối của X_t chỉ phụ thuộc vào trạng thái hiện tại C_t mà không phụ thuộc vào trạng thái hoặc quan sát trước đó. Nếu xích Markov có m trạng thái, ta nói X_t là mô hình HMM m trạng thái.

Bây giờ ta giới thiệu một số ký hiệu sử dụng trong nghiên cứu. Trong trường hợp quan sát rời rạc, ta định nghĩa

$$p_i(x) = Pr(X_t = x \mid C_t = i).$$

Đối với trường hợp liên tục, $p_i(x)$ là hàm mật độ xác suất của X_t nếu xích Markov nhận trạng thái i tại thời điểm t.

Ta ký hiệu ma trận xác suất chuyển của một xích Markov thuần nhất là $\mathbf{\Gamma}$ với các thành phần của nó là γ_{ij} được xác định bởi

$$\gamma_{ij} = Pr(C_t = j \mid C_{t-1} = i).$$

Từ bây giờ, m phân phối $p_i(x)$ được gọi là các phân phối trạng thái phụ thuộc của mô hình.

1.4.2. Likelihood và ước lượng cực đại likelihood

Đối với các quan sát rời rạc X_t, định nghĩa $u_i(t) = Pr(C_t = i)$ với $i = 1, 2, ..., T$, ta có

$$Pr(X_t = x) = \sum_{i=1}^{m} Pr(C_t = i) Pr(X_t = x \mid C_t = i)$$

$$= \sum_{i=1}^{m} u_i(t)\, p_i(x). \qquad (1.4.1)$$

Để thuận tiện trong tính toán, công thức (1.4.1) có thể được viết lại dưới dạng ma trận sau:

$$\Pr(X_t = x) = (u_1(t), ..., u_{(m)}(t)) = \begin{pmatrix} p_1(x) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_m(x) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$= \mathbf{u}(t)\mathbf{P}(x)\mathbf{1}'.$$

trong đó $\mathbf{P}(x)$ là ma trận đường chéo với phần tử thứ i trên đường chéo là $p_i(x)$. Mặt khác, theo tính chất của xích Markov thuần nhất, $\mathbf{u}(t) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}$ với $\mathbf{u}(1)$ là phân phối trạng thái ban đầu của xích Markov, thường được ký hiệu chung với phân phối dừng là $\boldsymbol{\delta}$. Và do vậy, ta có

$$Pr(X_t = x) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}\mathbf{P}(x)\mathbf{1}'. \qquad (1.4.2)$$

Bây giờ gọi L_T là hàm hợp lý (likelihood) của mô hình với T quan sát $x_1, x_2, ..., x_T$ thì $L_T = Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})$. Xuất phát từ công thức xác suất đồng thời

$$Pr(\mathbf{X}^{(T)}, \mathbf{C}^{(T)}) = Pr(C_1) \prod_{k=1}^{T} Pr(C_k \mid C_{k-1}) \prod_{k=1}^{T} Pr(X_k \mid C_k),$$

ta lấy tổng trên tất cả các trạng thái có thể có của C_k, sau đó sử dụng kỹ thuật như trong công thức (1.4.2), ta được

$$L_T = \delta \mathbf{P}(x_1)\mathbf{\Gamma P}(x_2)...\mathbf{\Gamma P}(x_T)\mathbf{1}'.$$

Nếu phân phối ban đầu $\boldsymbol{\delta}$ là phân phối dừng của xích Markov, thì

$$L_T = \delta \mathbf{\Gamma P}(x_1)\mathbf{\Gamma P}(x_2)...\mathbf{\Gamma P}(x_T)\mathbf{1}'.$$

Để có thể tính toán dễ dàng likelihood bằng thuật toán đồng thời giảm thiểu số phép toán mà máy tính cần thực hiện, ta định nghĩa vector $\boldsymbol{\alpha}_\mathbf{t}$ với $t = 1,..., T$ bởi

$$\alpha_t = \delta \mathbf{P}(x_1)\mathbf{\Gamma P}(x_2)...\mathbf{\Gamma P}(x_t) = \delta \mathbf{P}(x_1)\prod_{s=2}^{t}\mathbf{\Gamma P}(x_s), \tag{1.4.3}$$

thì lập tức ta có

$$L_T = \alpha_T \mathbf{1}', \quad \text{và} \quad \alpha_t = \alpha_{t-1}\mathbf{\Gamma P}(x_t), t \geq 2. \tag{1.4.4}$$

Từ đây, ta dễ dàng tính được L_T bằng thuật toán hồi quy. Để tìm bộ tham số thỏa mãn L_T lớn nhất, ta có thể thực hiện theo hai phương pháp:

Ước lượng trực tiếp cực trị hàm L_T (MLE): Trước tiên, từ phương trình (1.4.4) ta cần tính toán logarit của L_T một cách hiệu quả nhằm thuận lợi trong việc tìm cực đại dựa vào các xác suất lũy tiến $\boldsymbol{\alpha}_\mathbf{t}$. Với $t = 0,1,..., T,$ định nghĩa vector

$$\phi_t = \alpha_t / w_t$$

trong đó $w_t = \sum_i \alpha_t(i) = \alpha_t \mathbf{1}'$, và

$$\mathbf{B}_\mathbf{t} = \Gamma \mathbf{P}(x_t)$$

ta có

$$\begin{gathered} w_0 = \alpha_0 \mathbf{1}' = \delta \mathbf{1}' = \mathbf{1}; \\ \phi_0 = \delta; \\ w_t \phi_t = w_{t-1}\phi_{t-1}\mathbf{B}_t; \\ L_T = \delta_t \mathbf{1}' = w_T(\phi_\mathbf{T}\mathbf{1}') = w_T. \end{gathered} \tag{1.4.5}$$

Khi đó $L_T = w_T = \prod_{t=1}^{T}(w_t / w_{t-1})$. Từ (1.3.5) thấy rằng

$$w_t = w_{t-1}\left(\mathbf{B}_\mathbf{t}\mathbf{1}'\right),$$

dẫn đến

$$\log L_T = \sum_{t=1}^{T} \log\left(w_t / w_{t-1}\right) = \sum_{t=1}^{T} \log\left(\phi_{t-1} B_t 1'\right).$$

Tiếp theo, ta cần đổi biến số để loại bỏ các ràng buộc. Chẳng hạn, với tham số dương $\lambda_i > 0$ thì đổi biến thành $\eta_i = \log \lambda_i$. Sau khi ước lượng được $\hat{\eta}_i$ thì phép biến đổi ngược $\hat{\lambda}_i = \exp \hat{\eta}_i$ cho ta ước lượng của tham số ban đầu.

Việc tham số lại các tham số của ma trận chuyển $\mathbf{\Gamma}$ phức tạp hơn bởi $\mathbf{\Gamma}$ có m^2 tham số nhưng chỉ có $m(m-1)$ tham số tự do và tổng theo mỗi dòng thỏa mãn ràng buộc

$$\gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{im} = 1 \, (i = 1, ..., m).$$

Ví dụ, với $m = 3$ ta có thể tham số lại ma trận $\mathbf{\Gamma}$ như sau. Xét một ma trận có $m(m-1)$ tham số (ẩn) có dạng

$\mathbf{T} = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix}$, với các phần tử (là các tham số) $\tau_{ij} \in \mathbb{R}$.

Xét $g(x) = e^x$, ta định nghĩa

$$\rho_{ij} = \begin{cases} g(\tau_{ij}) \text{ voi } i \neq j \\ 1 \qquad \text{voi } i = j \end{cases}$$

sau đó đặt

$$\gamma_{ij} = \frac{\rho_{ij}}{\sum_{k=1}^{3} \rho_{ik}} \text{ (với } i, j = 1, 2, 3) \tag{1.4.6}$$

và $\Gamma = (\gamma_{ij})$.

Đến đây, ta tìm cực tiểu của hàm $-\log L_T$ với biến là các tham số tự do. Sau đó ta biến đổi ngược lại được tham số ban đầu.

Việc tìm cực tiểu của hàm với các biến tự do trong R dễ dàng thực hiện nhờ hàm *nlm*.

Tuy nhiên, phương pháp này đòi hỏi khối lượng tính toán lớn, nhất là khi phải thực hiện với nhiều các tham số ban đầu khác nhau để tránh trường hợp có nhiều cực trị.

Thuật toán EM: Thuật toán này còn được gọi là thuật toán Baum-Welch [7] áp dụng cho xích Markov thuần nhất (không nhất thiết là Markov dừng). Thuật toán sử dụng các xác suất lũy tiến (FWP) và xác suất lũy lùi (BWP) để tính L_T (tính từ 2 phía).

Ưu điểm lớn nhất của thuật toán này là tận dụng được các tính chất của FWP và BWP để tính toán các phân bố dự báo hay chỉ ra dãy trạng thái có khả năng cao nhất về sau.

Theo phương trình (1.4.3), các xác suất FWP đã được định nghĩa bởi

$$\alpha_t = \delta \mathbf{P}(x_1)\mathbf{\Gamma P}(x_2)...\mathbf{\Gamma P}(x_t) = \delta \mathbf{P}(x_1)\prod_{s=2}^{t}\mathbf{\Gamma P}(x_s), \tag{1.4.7}$$

Bây giờ, các vector BWP $\boldsymbol{\beta}_t$ được định nghĩa bởi

$$\boldsymbol{\beta}_{t'} = \Gamma \mathbf{P}(x_{t+1})\Gamma \mathbf{P}(x_{t+2})\cdots\Gamma \mathbf{P}(x_T)\mathbf{1}' = \left(\prod_{s=t+1}^{T}\Gamma \mathbf{P}(x_s)\right)\mathbf{1}'. \tag{1.4.8}$$

Luận án sẽ chỉ ra một số các tính chất của FWP và BWP mà sẽ được sử dụng trong dự báo chuỗi thời gian.

Mệnh đề 1.4.1. [73]: Với $t = 1,2,...,T$ và $j = 1,2,...,m$, thì

$$\alpha_t(j) = Pr\left(\mathbf{X}^{(t)} = \mathbf{x}^{(t),C_t=j}\right).$$

Mệnh đề 1.4.2. [73]: Với $t = 1,2,...,T-1$ và $j = 1,2,...,m$, thì

$$\beta_t(i) = Pr\left(X_{t+1} = x_{t+1}; X_{t+2} = x_{t+2},\cdots, X_T = x_T \mid C_t = i\right),$$

với điều kiện $Pr(C_t = i > 0)$. Hay viết ngắn gọn hơn

$$\beta_t(i) = Pr\left(\mathbf{X}_{t+1}^{T} = \mathbf{x}_{t+1}^{T} \mid C_t = i\right),$$

trong đó $\mathbf{X}_a^b = vector(X_a, X_{a+1},\cdots, X_b)$.

Mệnh đề 1.4.3. [73]: Với $t = 1,2,...,T-1$ và $j = 1,2,...,m$, thì

$$\alpha_t(i)\beta_t(i) = Pr\left(\mathbf{X}^T = \mathbf{x}^T, C_t = i\right),$$

dẫn tới

$$\alpha_t\beta_t^{'} = Pr\left(\mathbf{X}^T = \mathbf{x}^T\right) = L_T.$$

Bây giờ, luận án mô tả thuật toán EM trong mô hình HMM. Giả sử $c_1, c_2, ..., c_T$ là một xích Markov và các trạng thái c_i là của x_t tương ứng (lưu ý ở đây c_i, x_i là các

giá trị của các biến ngẫu nhiên C_i, X_i). Để thuận tiện trong tính toán, định nghĩa các biến ngẫu nhiên 0-1 như sau:

$$u_j(t) = 1 \text{ nếu } c_t = j, (t = 1, 2, ..., T)$$

và $v_{jk} = 1$ nếu $c_{t-1} = j$ và $c\_t = k\,(t = 2, 3, ..., T)$. Các biến ngẫu nhiên này thay thế cho các trường hợp xảy ra của c_t đối với các xác suất δ_{c_t}, p_{c_t} và $\gamma_{c_{t-1}c_t}$.

Với các ký hiệu này, log-likelihood đầy đủ (CLL) của mô hình HMM kể cả dữ liệu không đầy đủ được cho bởi

$$\log\left(Pr(x^{(T)}, c^{(T)})\right) = \log\left(\delta_{c_1}\prod_{t=2}^{T}\gamma_{c_{t-1},c_t}\prod_{t=1}^{T}p_{c_t}(x_t)\right)$$

$$= \log\delta_{c_1} + \sum_{t=2}^{T}\log\gamma_{c_{t-1},c_t} + \sum_{t=1}^{T}\log p_{c_t}(x_t).$$

Với định nghĩa của u_j và v_{jk} ta có

$$\log\left(Pr(x^{(T)}, c^{(T)})\right) = \sum_{j=1}^{m}u_j(1)\log\delta_j + \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\sum_{t=2}^{T}v_{jk}(t)\right)\log\gamma_{jk}$$

$$+\sum_{j=1}^{m}\sum_{t=1}^{T}u_j(t)\log p_j(x_t) \qquad (1.4.9)$$

$$= \text{thành phần 1 + thành phần 2 + thành phần 3}$$

Thuật toán EM cho mô hình HMM thứ tự như sau.

- **Bước E:** Thay thế tất cả các đại lượng v_{jk} và $u_j(t)$ bởi

$$\hat{u}_j(t) = Pr(C_t = j \mid x^{(T)}) = \alpha_t(j)\beta_t(j)/L_T;$$

và

$$\hat{v}_{jk}(t) = Pr(C_{t-1} = j, C_t = k \mid x^{(T)}) = \alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)/L_T.$$

trong đó α_t và β_t tương ứng là các to *FWP và BWP* như ở (1.4.7) và (1.4.8). Các thay thế này chính là các ước lượng thống kê cho v_{jk} và $u_j(t)$ với mẫu x^T.

- **Bước M:** Sau khi thay thế xong $v_{jk}(t)$ và $u_j(t)$ bởi $\hat{u}_j(t)$ và $\hat{v}_{jk}(t)$, tìm cực đại hàm CLL, phương trình (1.4.9), tương ứng với 3 bộ tham số:

Phân bố ban đầu δ, ma trận xác suất chuyển $\mathbf{\Gamma}$ và các tham số của phân bố xác suất trạng thái.

Đạo hàm hàm (CLL) và cho bằng 0 theo từng tham số tương tự trong (1.3.6), ta có ước lượng:

1. Với thành phần 1: Đặt

$$\delta_j = \hat{u}_j(1) / \sum_{j=1}^{m} \hat{u}_j(1) = \hat{u}_j(1). \tag{1.4.10}$$

2. Với thành phần 2: Đặt

$$\gamma_{jk} = f_{jk} / \sum_{k=1}^{m} f_{jk}, \tag{1.4.11}$$

trong đó $f_{jk} = \sum_{t=2}^{T} \hat{v}_{jk}(t).$

3. Với thành phần 3: Thành phần này có thể dễ xử lý hoặc khó tùy thuộc với phân phối được chọn. Đối với mô hình HMM có trạng thái là phân phối chuẩn với hàm mật độ có dạng $p_j(x) = (2\pi\sigma_j^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2 \right)$, và các cực đại của các tham số μ_j và σ_j^2 là

$$\hat{\mu}_j = \sum_{t=1}^{T} \hat{u}_j(t) x_t / \sum_{t=1}^{T} \hat{u}_j(t) \tag{1.4.12}$$

và

$$\hat{\sigma}_j^2 = \sum_{t=1}^{T} \hat{u}_j(t)(x_t - \hat{\mu}_j)^2 / \sum_{t=1}^{T} \hat{u}_j(t). \tag{1.4.13}$$

1.4.3. Phân phối dự báo

Đối với các quan sát có giá trị rời rạc, phân phối dự báo $Pr(X_{n+h} = x \mid X^{(n)} = x^{(n)})$ thực chất là một tỷ lệ của L_T dựa vào xác suất điều kiện:

$$Pr(X_{T+h} = x \mid X^{(T)} = x^{(T)}) = \frac{Pr(X^{(T)} = x^{(T)}, X_{T+h} = x)}{Pr(X^{(T)} = x^{(T)})}$$

$$= \frac{\delta \mathbf{P}(\mathbf{x}_1)\mathbf{B}_2\mathbf{B}_3 \ldots \mathbf{B}_T \mathbf{\Gamma}^{\mathbf{h}} \mathbf{P}(\mathbf{x}) \mathbf{1}'}{\delta \mathbf{P}(\mathbf{x}_1)\mathbf{B}_2\mathbf{B}_3 \ldots \mathbf{B}_T \mathbf{1}'}$$

$$= \frac{\alpha_T \mathbf{\Gamma^h P(x) 1'}}{\alpha_T \mathbf{1'}}.$$

Bằng cách viết $\phi_T = \alpha_T / \alpha_T 1'\$,$ ta có $Pr(X_{T+h} = x \mid X^{(T)} = x^{(T)}) = \phi_T \Gamma^h \mathbf{P(x)1}'.$

Phân phối dự báo từ đây có thể được viết như một phân phối xác suất trộn của các biến ngẫu nhiên phụ thuộc:

$$Pr(X_{T+h} = x \mid X^{(T)} = x^{(T)}) = \sum_{i=1}^{m} \xi_i(h)\, p_i(x).$$

trong đó trọng số $\xi_i(h)$ là thành phần thứ i của vector $\phi_T \Gamma^h$.

1.4.4. Thuật toán Viterbi

Mục tiêu của thuật toán Viterbi là đi tìm dãy trạng thái tốt nhất $i_1, i_2, ..., i_T$ tương ứng với dãy quan sát $x_1, x_2, ..., x_T$ mà làm cực đại hàm L_T.

Đặt

$$\xi_{1i} = Pr(C_1 = i, X_1 = x_1) = \delta_i p_i(x_1),$$

và với $t = 2,3,...,T$

$$\xi_{ti} = \max_{c_1, c_2, ..., c_{t-1}} Pr(C^{(t-1)} = c^{(t-1)}, C_t = i, X^{(T)} = x^{(T)}).$$

Khi đó có thể thấy xác suất ξ_{tj} thỏa mãn quá trình đệ quy sau đối với $t = 2,3,...,T$ và $i = 1,2,...,m$:

$$\xi_{tj} = \left(\max_i (\xi_{t-1,i} \gamma_{ij})\right) p_j(x_t).$$

Dãy trạng thái tốt nhất $i_1, i_2, ..., i_T$ do đó được xác định bằng hồi quy từ

$$i_T = \underset{i=1,..,m}{\operatorname{argmax}}\, \xi_{Ti}$$

và, với $t = T-1, T-2, ..., 1,$ thì từ

$$i_t = \underset{i=1,...,m}{\operatorname{argmax}} (\xi_{ti} \gamma_{i, i_{t+1}}).$$

1.4.5. Dự báo trạng thái

Đối với dự báo trạng thái, chỉ cần sử dụng công thức Bayes trong xác suất cổ điển.

Với $i = 1,2,..., m,$

$$Pr(C_{T+h} = i \mid X^{(T)} = x^{(T)}) = \boldsymbol{\alpha}_{\mathbf{T}}\boldsymbol{\Gamma}^{\mathbf{h}}(,\mathbf{i}) / L_T = \phi_T \Gamma^h(,i)$$

Lưu ý rằng, khi $h \to \infty$, $\phi_{\mathbf{n}}\boldsymbol{\Gamma}^{\mathbf{h}}$ tiến tới phân phối dừng của xích Markov.

1.5. Chuỗi thời gian mờ

1.5.1. Một số khái niệm

Giả sử U là không gian nền. không gian nền này xác định một tập hợp các đối tượng cần nghiên cứu. Nếu A là một tập con rõ của U thì ta có thể xác định chính xác một hàm đặc trưng:

$$\mu_A(X) = \begin{cases} 0 \text{ nếu } X \text{ nằm } trong\ A \\ 1 \text{ nếu } X \text{ năm } ngoài\ A \end{cases}$$

Nhưng với một tập mờ B trong không gian nền U thì phần tử x không xác định chính xác được. Khi đó ta có định nghĩa: $\mu_A : U \to [0,1]$, μ_A được gọi là hàm thuộc (Membership function). Còn với bất kỳ một phần tử u nào của A thì hàm $\mu_A(u)$ được gọi là độ thuộc của u vào tập mờ A.

Giả sử $Y(t)$ là chuỗi thời gian ($t = 0,1,2,...$), U là tập nền chứa các khoảng giá trị của chuỗi thời gian từ nhỏ nhất đến lớn nhất. Xác định hàm thuộc $\mu_A : U \to [0,1]$ của tập mờ A, còn tập A trên không gian nền U được viết như sau:

$$A = (\mu_A(u_1)/u_1, \mu_A(u_2)/u_2, ..., \mu_A(u_n)/u_n)),: u_i \in U; I = 1,2, ...,n \qquad (1.5.1)$$

$\mu_A(u_i)$ là độ thuộc của u_i vào tập A hay cách viết khác:

$$A=(A(u_1)/u_1, A(u_2)/u_2,...,A(u_n)/u_n)) \qquad (1.5.2)$$

Định nghĩa 1.5.1. [60]: *Giả sử U là không gian nền và $U = \{u_1, u_2, ..., u_n\}$. Tập mờ A trên không gian nền U được viết như sau:*

$$A=f_A(u_1)/u_1+f_A(u_2)/u_2+...+f_A(u_n)/u_n \qquad (1.5.3)$$

f_A là hàm thuộc của tập mờ A và $f_A : U \to [0;1]$, $f_A(u_i)$ là độ thuộc của u_i vào tập A.

Định nghĩa 1.5.2. [60]: *Cho $Y(t)(t = 0,1,2,...)$ là tập nền, là một tập con của R^1. Giả sử $f_i(t)(i = 0,1,2,...)$ được xác định trên $Y(t)$, và $F(t)$ chứa các tập $f_1(t), f_2(t),...,$ khi đó $F(t)$ được gọi là chuỗi thời gian mờ xác định trên tập $Y(t)$.*

Định nghĩa 1.5.3. [60]: *Giả sử rằng* $F(t)$ *chỉ được suy ra từ* $F(t-1)$, *kí hiệu là* $F(t-1) \to F(t)$, *mối quan hệ này có thể được diễn đạt như sau* $F(t) = F(t-1) o R(t,t-1)$, *trong đó* $F(t) = F(t-1) o R(t,t-1)$ *được gọi là mô hình bậc một của* $F(t), R(t,t-1)$ *là mối quan hệ mờ giữa* $F(t-1)$ *và* $F(t)$, *và "o" là toán tử thành phần Max–Min .*

Định nghĩa 1.5.4. [60]: *Cho* $R(t,t-1)$ *là mô hình bậc một của* $F(t)$. *Nếu mọi* $t, R(t,t-1) = R(t-1,t-2)$, *thì* $F(t)$ *được gọi là chuỗi thời gian mờ dừng. Trái lại* $F(t)$ *được gọi là chuỗi thời gian mờ không dừng.*

Quá trình dự báo chuỗi thời gian mờ cũng dựa trên các bước của phương pháp lập luận xấp xỉ mờ như sau:

1. Giải nghĩa các mệnh đề mờ điều kiện
2. Kết nhập các quan hệ mờ
3. Tính kết quả từ phép hợp thành
4. Khử mờ

1.5.2. Mô hình một số thuật toán dự báo trong chuỗi thời gian mờ

Mục này luận án trình bày 2 thuật toán nổi tiếng và được sử dụng nhiều nhất là của Song và Chissom (1993) [60] và của Huarng (2000) [38].

Mô hình thuật toán của Song và Chissom

Trong phần này, sử dụng khái niệm và phương pháp dự báo của chuỗi thời gian mờ được Song et. al. và Chissom đưa ra để xây dựng thuật toán dự báo cho chuỗi thời gian.

Giả sử U là không gian nền: $U = u_1, u_2, ..., u_n$. Tập A là mờ trên không gian nền U nếu A được xác định bởi hàm:

$$\mu_A : U \to [0.1].$$

Còn đối với bất kỳ một phần tử u nào của A thì hàm $\mu_A(u)$ được gọi là độ thuộc của u vào tập mờ A. Tập mờ A trên không gian nền U được viết như sau:

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + ... + \mu_A(u_n)/u_n \qquad (1.5.4)$$

Mô hình thuật toán gồm một số bước sau:

Bước 1: Xác định tập nền U trên đó các tập mờ được xác định

Bước 2: Chia các tập nền U thành một số các đoạn bằng nhau

Bước 3: Xác định các biến ngôn ngữ để diễn tả các tập mờ trên các khoảng đã chia của tập nền.

Bước 4: Mờ hoá các giá trị lịch sử của chuỗi thời gian.

Bước 5: Chọn tham số $w>1$ thích hợp và tính $R^w(t,t-1)$ và dự báo theo công thức sau:

$$F(t) = F(t-1) \text{oR}^{\text{w}} (\text{t, t-1}),$$

trong đó $F(t)$ là giá trị dự báo mờ tại thời điểm t còn $F(t-1)$ là giá trị dự báo mờ tại thời điểm $t-1$.

Mối quan hệ mờ được tính như sau:

$$R^w(t,t-1) = F^T(t-2) \times F(t-1) \cup F^T(t-2) \ldots \cup F^T(t-w) \times F(t-w+1),$$

với T là toán tử chuyển vị, dấu "$\times$" là toán tử tích Cartesian còn w được gọi là "mô hình cơ sở" mô tả số lượng thời gian trước thời điểm t.

Bước 6: Giải mờ giá trị dự báo mờ.

Mô hình Heuristic cho chuỗi thời gian mờ

Huarng đã sử dụng mô hình của Chen và đưa vào các thông tin có sẵn của chuỗi thời gian để cải tiến độ chính xác và giảm bớt các tính toán phức tạp của dự báo. Nhờ sử dụng những thông tin có trong chuỗi thời gian nên mô hình của Huarng được gọi là mô hình Heuristic.

Các bước thực hiện của mô hình Huarng cũng triển khai theo các bước trên. Điều khác biệt là sử dụng một hàm h để xác định mối quan hệ logic mờ. dưới đây là mô tả các bước thực hiện của mô hình Heuristic chuỗi thời gian mờ.

Bước 1: Xác định tập nền. Tập nền U được xác định như sau: lấy giá trị lớn nhất $f_{\max}$ và nhỏ nhất $f_{\min}$ của chuỗi thời gian $U = [f_{\max}, f_{\min}]$. Đôi khi có thể mở rộng khoảng này thêm một giá trị nào đó để dễ tính toán. Chia đoạn U thành m khoảng con bằng nhau $u_1, u_2, \ldots, u_n$.

Bước 2: Xác định tập mờ A_i và mờ hoá giá trị. Mỗi tập A_i gán cho một biến ngôn ngữ và xác định trên các đoạn đã xác định $u_1, u_2, ..., u_n$. Khi đó các tập mờ A có thể biểu diễn như sau:

$$A_i = \mu_A i(u_1)/u_1 + \mu_A i(u_2)/u_2 + ... + \mu_A i(u_n)/u_n$$

Bước 3: Thiết lập mối quan hệ mờ và nhóm các mối quan hệ mờ. Như định nghĩa ở trên, đối với chuỗi thời gian mờ ta có thể xác định được mối quan hệ mờ tại mỗi thời điểm t và qua đó ta xác định được nhóm các mối quan hệ mờ.

Bước 4: Sử dụng hàm h để thiết lập các nhóm mối quan hệ logic mờ Heuristic.

$$A_i \rightarrow h_j(x, A_p1, A_p2, ...,) = A_p1, A_p2, ..., A_pk$$

Bước 5: Dự báo. Từ các nhóm quan hệ logic mờ Heuristic. Các giá trị chủ yếu lấy từ điểm giữa hay trung bình các điểm giữa các khoảng cách trong nhóm quan hệ mờ heuristic.

1.6. Kết luận

Chương này luận án trình bày những kiến thức cơ sở được sử dụng cho các chương sau, bao gồm:

- Trình bày các hướng nghiên cứu dự báo chuỗi thời gian gần đây nhất và phân tích những hạn chế của nó. Từ đó đưa ra đề xuất phát triển mô hình của nghiên cứu sinh.
- Các khái niệm xích Markov, phân loại xích Markov và ước lượng tham số của xích Markov. Đặc biệt, xích Markov chính quy được sử dụng trong Chương 3.
- Mô hình Markov ẩn được trình bày chi tiết cùng các thuật toán ước lượng tham số. Đây là cơ sở cho mô hình Markov ẩn cho phân phối Poisson và phân phối chuẩn (Normal distribution) được thực hiện trong Chương 2.
- Xích Markov bậc cao và phương pháp ước lượng tham số được trình bày trong Mục 3.2. Lý thuyết về chuỗi thời gian mờ và một số thuật toán trong dự báo chuỗi thời gian sử dụng chuỗi thời gian mờ được trình bày

trong Mục 1.5. Đây là kiến thức cơ sở cho mô hinh kết hợp xích Markov và chuỗi thời gian mờ trong dự báo được luận án phát triển trong Chương 3.

Chương 2. MÔ HÌNH MARKOV ẨN TRONG DỰ BÁO CHUỖI THỜI GIAN

2.1. Mở đầu

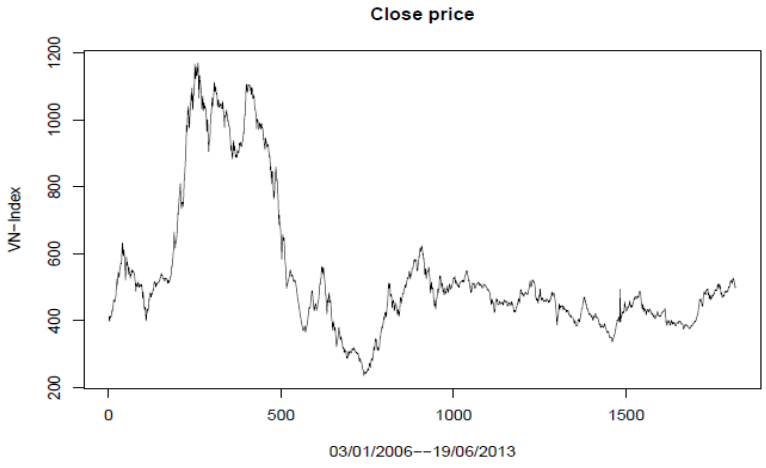
Mô hình Markov ẩn (HMMs) là một công cụ được sử dụng rộng rãi để phân tích và dự báo chuỗi thời gian. Các yếu tố toán học đằng sau mô hình HMM bắt đầu được phát triển bởi L. E. Baum và các công sự [5-9]. Mô hình HMMs đã được sử dụng thành công cho nhiều loại chuỗi thời gian bao gồm phân tích chuỗi DNA [18], nhận dạng giọng nói [67], phân tích ECG [22]. Trong lĩnh vực tài chính, Hassan and Nath, 2005 [36] đã sử dụng mô hình HMM để sinh ra dự báo từng ngày của giá cổ phiếu theo cách đặc biệt. Ta có thể đề cập đến nghiên cứu gần đây hơn của Rafiul Hassan [35] với sự kết hợp mô hình HMM và chuỗi thời gian mờ cho dự báo cổ phiếu. Ngoài các mô hình dự báo giá cổ phiếu, mô hình HMM còn được sử dụng trong các vấn đề khác của tài chính như mô hình lợi suất (returns) của cổ phiếu, mô hình sự biến động của tỉ lệ tăng trưởng của GDP thực tế, mối quan hệ giữa sản suất công nghiệp và thị trường chứng khoán, ... như được trình bày trong [10].

Ở Việt Nam, các nghiên cứu sâu về thị trường tài chính nói chung cũng như việc ứng dụng mô hình HMM trong dự báo nói riêng còn rất hạn chế. Các phân tích chủ yếu vào các mô hình hồi quy tuyến tính hay các mô hình dựa trên phân phối chuẩn như Neural-Fuzzy, ARMA, ... mà có thể kể đến như trong [23]. Điều hạn chế là, sự phụ thuộc tuyến tính hay tính chuẩn của phân phối đối với các số liệu trong tài chính đã được nhiều công trình nghiên cứu trên thế giới chỉ ra là bất hợp lý [66].

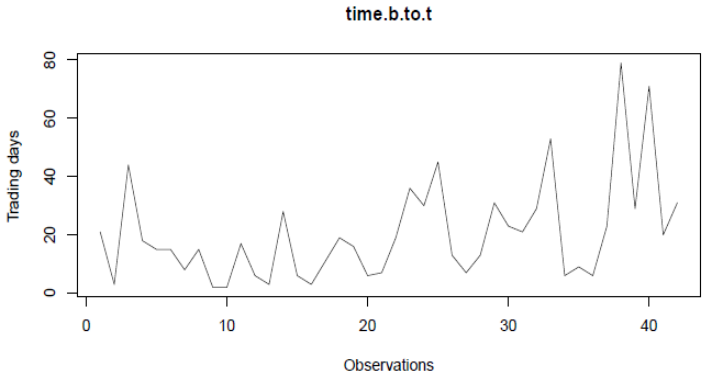
Để minh họa rằng, chuỗi thời gian chỉ số chứng khoán thường không thể ướm khít với một phân phối xác suất nhất định luận án tiến hành thử nghiệm trên dữ liệu

chỉ số VN-Index được thu thập từ 03/01/2006 đến 19/06/2013 và thống kê những thời điểm chính mà cổ phiếu lên đỉnh và xuống đáy. Sau đó, tập dữ liệu trực tiếp sử dụng trong nghiên cứu này là số phiên giao dịch mỗi lần cổ phiếu từ đáy lên đỉnh và được gói trong tập time.b.to.t (bottom to top). Đây là dữ liệu mà rất có ý nghĩa thực tế đối với những nhà đầu tư chứng khoán. Minh họa này sẽ chỉ ra rằng rất khó để ướm một dữ liệu kiểu như vậy với một phân phối xác suất ổn định. Cho dù ướm nó với phân phối trộn của nhiều phân phối, vấn đề tương quan giữa các thành phần trộn theo thời gian vẫn tồn tại. Do đó, phân bố xác suất của đối tượng dự đoán thay đổi theo thời gian. Từ đó đặt ra yêu cầu áp dụng mô hình dự báo chuỗi thời gian phù hợp mà mô tả được đặc điểm phụ thuộc này.

Cụ thể cho lập luận trên, hình 2.1.1 mô tả dao động của chỉ số VN-Index trong thời gian nói trên. Ta có thể thấy rất rõ những thời điểm mà cổ phiếu đạt đỉnh hay chạm đáy. Hình 2.1.2 biểu diễn dữ liệu của time.b.to.t.



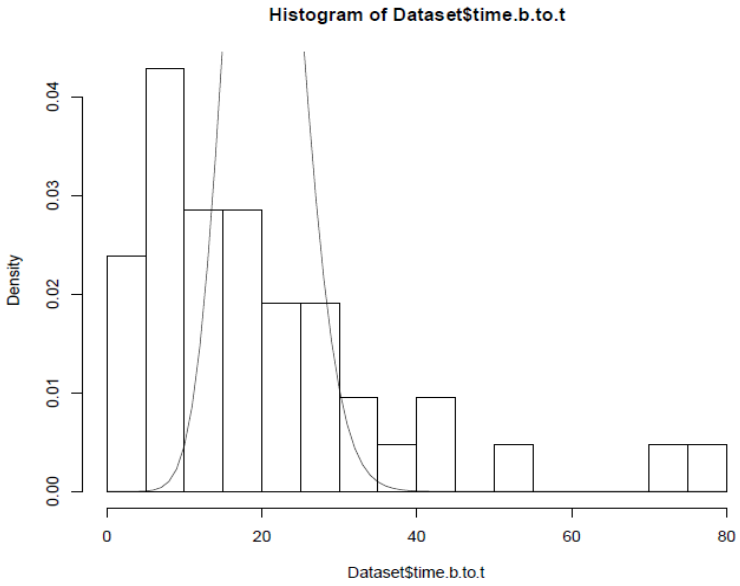
Hình 2.1.1. *Chỉ số đóng cửa của VN-Index từ 03/01/2006 đến 19/06/2013*



Hình 2.1.2. *Số phiên giao dịch mỗi lần chứng khoán từ đáy lên đỉnh*

Vì dữ liệu là các số tự nhiên nên trong xác suất, phân phối Poisson là lựa chọn phù hợp để kiểm tra xem dữ liệu có thể ướm khít với phân phối này không và ở bộ tham số nào. Hình 2.1.3 cho ta kết quả ướm phân phối Poisson vào phân phối thống kê của dữ liệu. Thực tế cho thấy, với phương mẫu $s^2 = 307,083$ lớn hơn nhiều so với trung bình mẫu $\bar{x} = 20,45238$ đủ thấy rằng mô hình một phân phối Poisson, phân phối mà phương sai và trung bình mẫu bằng nhau, là không phù hợp. Cũng từ hình 2.1.3 ta thấy rằng khó có thể tìm một phân phối xác suất cổ điển nào ướm khít phân phối thống kê của nó không chỉ phân phối Poisson. Do đó, ta nghĩ đến việc ướm dữ liệu bởi nhiều phân phối (nhiều trạng thái), nghĩa là các quan sát sinh ra có thể từ vài phân phối khác nhau, độc lập với nhau. Mô hình như vậy được gọi là mô hình trộn độc lập [43,73]. Mô hình trộn độc lập được giả thiết rằng, một quan sát X được sinh ra từ 1 trong m phân phối (trạng thái) độc lập với nhau $p_1(x), p_2(x),..., p_m(x)$ với các xác suất tương ứng là $\delta_1, \delta_2,..., \delta_m$. Dễ dàng chỉ ra rằng hàm mật độ hoặc xác suất của X được cho bởi

$$p(x) = \sum_{i=1}^{m} \delta_i p_i(x). \qquad (2.1.1)$$



Hình 2.1.3. *Phân phối mẫu (histogram) của time.b.to.t được ướm bởi phân phối Poisson*

Như vậy, đối với mô hình phân phối trộn Poisson, ta cần ước lượng $2m$ tham số gồm $\lambda_1, \lambda_2, ..., \lambda_m$ các trung bình của các phân phối thành phần và $\delta_1, \delta_2, ..., \delta_m$ các xác suất trộn với ràng buộc $\lambda_i > 0$ và $\sum_i \delta_i = 1$ với $i = 1, 2, ..., m$.

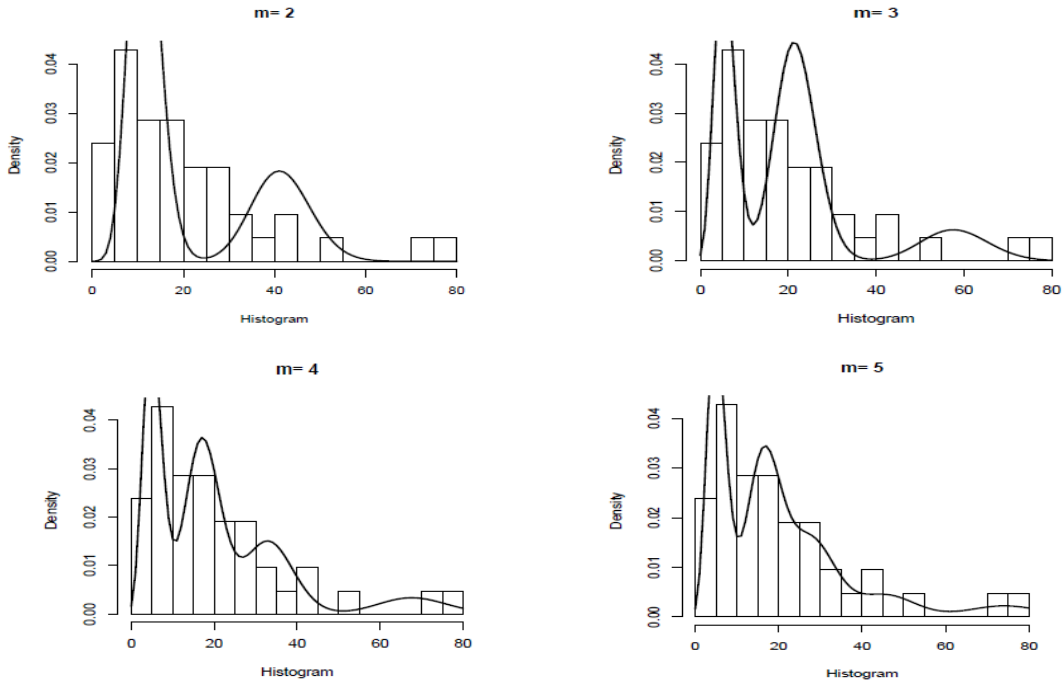
Mô hình trộn [43] ngày nay được sử dụng rất rộng rãi và hiệu quả đối với nhiều phân phối thống kê của các biến ngẫu nhiên trong thực tế. Trong [47], tác giả đã sử dụng mô hình trộn để nghiên cứu các trạng thái hoạt động của một số động vật hoang dã như nghỉ ngơi, săn mồi, di chuyển dựa trên sự di chuyển của các cá thể được gắn thiết bị GPS. Mô hình trộn cũng được sử dụng để phân loại chẳng hạn như phân loại bệnh thiếu máu [44], trong phân loại văn bản [32]. Trong kinh tế, mô hình trộn với phân phối chuẩn đã được sử dụng để mô hình sự biến động ngẫu nhiên trong hiệu ứng "nụ cười" (smile effects) [1]. Một nghiên cứu tương tự với mục đích mô hình sự biến động (volatility) của giá cổ phiếu từ đó cải tiến công thức Black-Scholes trong định giá quyền chọn được đề xuất bởi Damiano Brigo và Fabio Mercurio [12] bằng cách sử dụng phân phối log-normal.

Bằng phương pháp cực đại likelihood của dữ liệu time.b.to.t áp dụng cho mô hình với $m = 2,3,4,5$ trạng thái ứng với m phân phối Poisson độc lập, kết quả ước lượng tham số của mô hình trộn được cho ở Bảng 2.1.1

Bảng 2.1.1. *Ước lượng tham số của các mô hình trộn độc lập cho time.b.to.t*

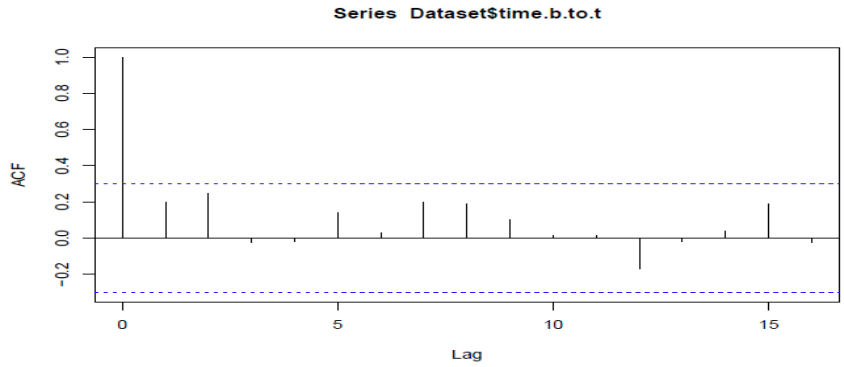
model	i	δ_i	λ_i	$-\log L$	Variance
$m = 2$	1	0,7048	11,621	217,9645	206,6592
	2	0,2951	41,538		
$m = 3$	1	0,3593	5,762	173,2254	269,9256
	2	0,5199	21,868		
	3	0,1207	58,077		
$m = 4$	1	0,333	5,371	165,5199	293,1273
	2	0,380	17,583		
	3	0,217	33,548		
	4	0,069	68,125		
$m = 5$	1	0,058	3,5	165,4839	293,3762
	2	0,279	5,839		
	3	0,378	17,664		
	4	0,216	33,579		
	5	0,069	68,140		

Rõ ràng ta thấy rằng, càng nhiều trạng thái thì $-\log L$ (log-likelihood) càng nhỏ, tức likelihood càng lớn. Điều này dẫn đến việc có thể nghĩ ràng càng nhiều trạng thái càng tốt. Tuy nhiên, việc đó sẽ dẫn đến việc gặp phải vấn đề về over-fitting, tức mất tính khái quát của mô hình. Do đó, việc chọn mô hình nào trong các mô hình ở trên sẽ được thực hiện qua các tiêu chuẩn chọn mô hình mà ta sẽ đề cập ở mô hình Markov ẩn HMM. Hình 2.1.4 minh họa histogram của dữ liệu time.b.to.t được ướm bởi phân phối trộn tương ứng với các mô hình trong Bảng 2.1.1. Ta có thể thấy mô hình với $m = 4$ hoặc $m = 5$ tốt hơn.



Hình 2.1.4. *Histogram được ướm với 4 mô hình trộn các phân phối Poisson độc lập với m=2,3,4,5*

Tuy nhiên, mô hình trộn độc lập thường được sử dụng cho các phân phối có tính ổn định, nó không thể hiện sự phụ thuộc theo thời gian giữa các quan sát. Hàm tự tương quan (ACF) của mẫu time.b.to.t được mô tả trong Hình 2.1.5 chỉ ra rằng các quan sát trong chuỗi có sự phụ thuộc.



Hình 2.1.5. *Hệ số tự tương quan của mẫu dữ liệu với 15 Lag*

Đến đây, ta cần đi tìm một phương pháp cho phép sự phụ thuộc trong chuỗi thời gian nhằm nới lỏng sự độc lập của các phân phối trong thành phần trộn. Nếu ta

giả sử các tham số của các trạng thái trong mô hình trộn tuân theo một xích Markov. Kết quả là ta có một mô hình được gọi là Hidden Markov (HMM). Vậy, mô hình HMM là mô hình tiềm năng cho việc phân tích chuỗi thời gian đảm bảo sự phụ thuộc theo thời gian của các trạng thái. Từ đó có thể cho những kết quả dự báo chính xác hơn các mô hình dự báo chuỗi thời gian cổ điển.

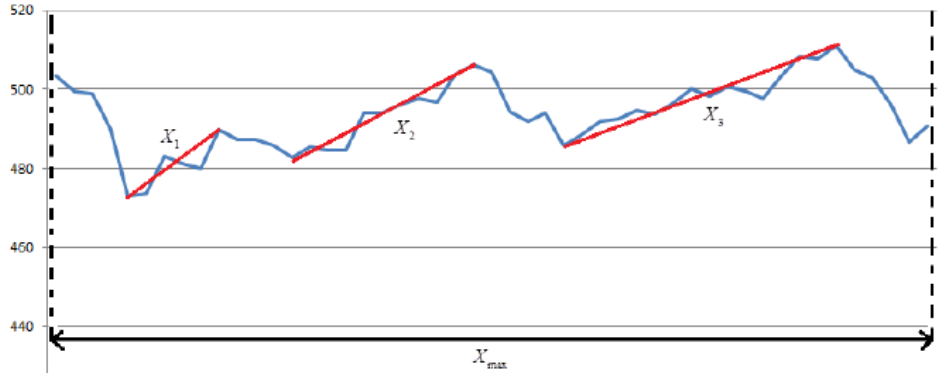
Mục tiếp theo luận án trình bày áp dụng mô hình HMM trong dự báo chuỗi thời gian bất kỳ. Kết quả thực nghiệm trên một vài dữ liệu tài chính cho thấy độ chính xác dự báo được cải thiện so với các mô hình cổ điển như ARIMA hay ANN.

2.2. Mô hình Markov ẩn trong dự báo chuỗi thời gian

Theo Chương 1, mô hình Markov ẩn (Hidden Markov Model) có thể coi là một mô hình đặc biệt của mô hình trộn phụ thuộc. Các thành phần trộn trong mô hình trộn bây giờ phụ thuộc theo một xích Markov. Do đó, một mô hình HMM bao gồm hai thành phần cơ bản: chuỗi $X_t, t = 1,...,T$ các quan sát và $C_t = i, t = 1,...,T, i \in \{1, 2,..., m\}$ thành phần trộn.

Bây giờ, để dễ minh họa cho mô hình HMM trong dự báo chuỗi thời gian, xét chuỗi thời gian time.b.to.t ở trên và ký hiệu là $X_t, t = 1,...,T$. Bài toán thực tế đối với nhà đầu tư là dự đoán giá trị của X_t trong tương lai để biết sau bao lâu chỉ số chứng khoán sẽ từ đáy lên đỉnh. Từ quan sát thực tế thấy rằng chỉ số chứng khoán khi đạt một đỉnh mới sẽ không thể ở giá trị đó (hoặc dao động nhẹ xung quanh giá trị đó) mãi mãi mà sẽ đi xuống sau một thời gian nào đó, tương tự đối với dao động từ đáy lên đỉnh. Vậy có thể quy định $X_{\max}$ là thời gian lâu nhất mà giá trị cổ phiếu từ đáy lên đỉnh. Khi đó, $0 < X_t \le X_{\max}$ (xem Hình 2.2.1). Nhà đầu tư muốn quy định các trạng thái xảy ra với X_t, chẳng hạn "chờ nhanh", "chờ khá nhanh", "chờ lâu", "chờ rất lâu" nhưng không biết phải định nghĩa như thế nào. Để giải quyết bài toán này, ta coi mỗi trạng thái trên là một phân phối Poisson với trung bình (cũng là phương sai) $\lambda_i, i = 1, 2, 3, 4$ và được "ẩn" trong chuỗi X_t. Nếu giả thiết thêm các trạng

thái này tuân theo một xích Markov, ta có mô hình Markov ẩn cho bài toán dự báo chuỗi thời gian.



Hình 2.2.1. *Định nghĩa chuỗi thời gian cần dự báo*

Để áp dụng mô hình HMM cho dự báo chuỗi thời gian, luận án minh họa cả hai phương pháp ước lượng tham số đã trình bày trong mục 1.4.2 của Chương 1.

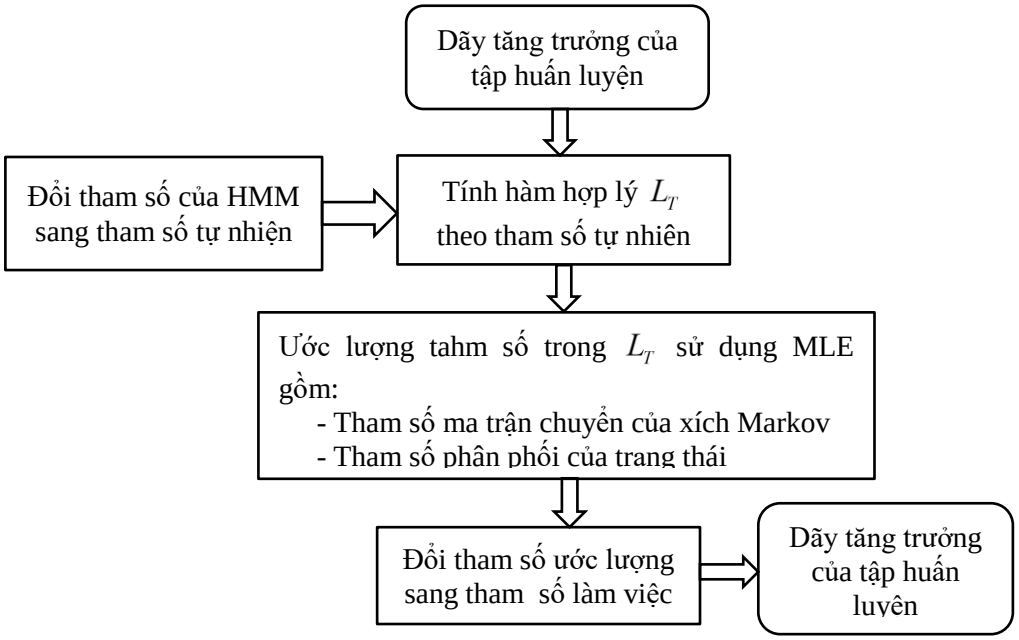
2.2.1. Mô hình HMM với phân phối Poisson

Đối với ước lượng MLE, luận án thực hiện cho mô hình HMM với trạng thái là các phân phối Poisson. Phân phối Poisson có tham số $\lambda > 0$ vừa là trung bình đồng thời là phương sai. Trong mô hình HMM gồm m trạng thái ứng với m phân phối Poisson, vậy có m tham số $\lambda_i, i = 1,...,m$ cần ước lượng. Cùng với m^2 tham số γ_{ij} trong ma trận xác suất chuyển $\mathbf{\Gamma}$ của xích Markov thì mô hình cần ước lượng $m^2 + m$ tham số.

Bản chất của phương pháp ước lượng cực đại hàm hợp lý là tìm cực đại của một hàm nhiều biến phi tuyến bằng phương pháp xấp xỉ trong giải tích. Do đó, các tham số cần phải loại bỏ các ràng buộc để trở thành các tham số tự do (gọi là pw). Đối với các phân phối Poisson, do $\lambda_i > 0$ nên tham số của mô hình (gọi là pn) được chuyển sang tham số tự nhiên bởi $\eta_i = \exp \lambda_i$.

Việc đổi tham số γ_{ij} sang tham số tự nhiên phức tạp hơn, đã được trình bày trong (1.4.6). Sau đó, hàm hợp lý L_T của mô hình được tính theo thuật toán trong (1.4.5).

Mô hình HMM ước lượng bởi MLE được mô tả trong Hình 2.2.2.



Hình 2.2.2. *Quá trình ước lượng tham số của mô hình HMM sử dụng MLE*

Cụ thể, luận án thực hiện phương pháp ước lượng MLE cho HMM với phân phối Poisson. Các bước tính làm hợp lý L_T và ước lượng tham số của L_T được tính bởi Thuật toán 2.1 và Thuật toán 2.2. Trong quá trình ước lượng tham số mô hình, thuật toán cho phép tính luôn tiêu chuẩn BIC và AIC cho lựa chọn mô hình về sau.Trong Thuật toán 2.1, các giá trị đầu vào *parvect*, *x, m* lần lượt là vector tham số, vector quan sát thống kê và số trạng thái của mô hình HMM, trong khi giá trị đầu ra *mllk* là − của logarit hàm hợp lý.

Trong thuật toán ước lượng cực đại hàm hợp lý, các tham số x, m vẫn được ký hiệu như trong thuật toán tính hợp hợp lý, còn *lambda0,gamma0* lần lượt là tham số của phân phối Poisson và ma trận xác suất chuyển ban đầu. Các tham số đầu ra có *lambda* và *gamma* là các tham số ước lượng tối ưu của phân phối Poisson và ma trận xác suất chuyển. Ngoài ra thuật toán tính thêm hai tiêu chuẩn BIC và AIC cho việc lựa chọn tối ưu số trạng thái m.

Thuận toán 2.1 Tính hàm hợp lý

Đầu vào: *parvect, x, m*

Đầu ra: *mllk*

1: Begin

2: **if** $m = 1$ **then** $mllk \leftarrow -\log \sum Poisson(x, \exp parvect)$ {Lấy giá trị của phân phối Poisson}

3: $n \leftarrow length(x)$ {Lấy độ dài dãy quan sát}

4: $pn \leftarrow HMM.pw2pn(m, parvect)$ {Đổi tham số tự do sang tham số mô hình}

// tính *mllk theo theo thuật toán mổ tả trong (1.4.5)*

5: $lscale \leftarrow 0$

6: $foo \leftarrow delta = 1$

7: *for i in 1:n do*

8: $foo \leftarrow foo * gamma * Poisson(x, exp parvect)$

9: $sumfoo \leftarrow sum(foo)$

10: $lscale \leftarrow lscale + log(sumfoo)$

11: $foo \leftarrow foo/sumfoo$

12: $mllk \leftarrow -lscale$

13: **return** *mllk*

14: End.

Thuận toán 2.2 Maximum hàm hợp lý

Đầu vào: *x*,*m*, *lambda*0,*gamma*0

Đầu ra: *m*, *lambda*0, *gamma*0, *BIC, AIC, mllk*

1: Begin

2: $parvect0 \leftarrow HMM.pn2pw(m, lambda0, gamma0)$ {Đổi tham số mô hình sang tham số tự do}

3: *mod* ←*nlm*(*HMM.mllk*, *parvect*0,*x* = *x*,*m* = *m*) {Ước lượng tham số làm cực đại hàm hợp lý}

4: *pn*← *HMM.pw2pn*(*m*,*mod*$*estimate*) {Đổi tham số tự do sang tham số mô hình pn}

5: *mllk* ←*mod*$*minimum* *{Lấy giá trị cực đại gán cho mllk}*

6: *np*←*length*(*parvect*0) {đếm số tham số mô hình}

7: *AIC* < −2 ∗ (*mllk*+*np*) {Tính tiêu chuẩn AIC}

8: *n* < −*sum*(!*is.na*(*x*)) {Tính số quan sát}

9: *BIC* < −2 ∗*mllk*+*np* ∗ *log*(*n*) {Tính tiêu chuẩn BIC}

10: return *m*, *lambda, gamma, AIC*, *BIC, mllk*

11: End.

Sau khi ước lượng được tham số của mô hình, các phân phối dự báo đều được tính toán theo hàm dự báo như đối với mô hình Normal-HMM được trình bày ở mục tiếp sau đây.

2.2.2. Mô hình HMM với phân phối chuẩn

Mục này xây dựng các hàm (trong ngôn ngữ lập trình R) để ước lượng tham số và dự báo cho mô hình HMM với phân phối chuẩn trong trường hợp sử dụng thuật toán EM. Thuật toán EM không cần phải đổi biến số thành biến tự do mà thực hiện tính toán trực tiếp thông qua FWP và BWP (xem 1.4.2).

Trong mô hình với phân phối chuẩn, các tham số của xích Markov vẫn là *gamma* nhưng tham số của phân phối trộn gồm trung bình *mu* và phương sai *sigma* trong khi *m* vẫn là số trạng thái của mô hình còn *delta* là phân phối dừng của xích Markov.

Hàm tính các FWP va BWP được thực hiện bởi hàm norm.HMM.lalphabeta (logarit của FWP va BWP) trong Thuận toán 2.3.

Thuận toán 2.3 Tính các xác suất lũy tiến và lùi của *LT*

Đầu vào: *x*,*m*,*mu*, *sigma*,*gamma*,*delta*

Đầu ra: *lalpha*, *lbeta*

1: Begin

2: **if** (*is*.*null*(*delta*)) **then** *delta*←*solve*(*t*(*diag*(*m*)−*gamma*+1), *rep*(1,*m*)) { Trong trường hợp không định trước được phân phối ban đầu của xích Markov}

3: Tính các xác suất FWP theo (1.4.7) cho lalpha

4: Tính các xác suất cho BWP theo (1.4.8) cho lbeta

5: return *lalpha*, *lbeta*

6: End.

Trong đó, *lalpha, lbeta* lần lượt là logarit của FWP và BWP.

Đến đây, theo thuật toán EM trong mục 1.4.2 của Chương 1 ta có thể thực hiện ngay ước lượng tham số bởi hàm norm.HMM.EM trong Thuật toán 2.4. Trong thuật toán này, các tham số *mu0, sigma0,gamma0,delta0* là các tham số khởi chạy (tham số ban đầu) của phân phối Normal, ma trận xác suất chuyển ban đầu và xác suất dừng ban đầu. Trong khi *maxiter, tol* lần lượt là số vòng lặp tối đa và độ chính xác ước lượng của thuật toán. Đầu ra là các tham số tối ưu của phân phối Normal và ma trận xác suất chuyển cũng như phân phối dừng. Đầu ra đồng thời tính các tiêu chuẩn BIC và AIC.

Thuận toán 2.4 Thuật toán EM cho Normal-HMM

Đầu vào: *x*,*m*,*mu()*, *sigma()*,*gamma()*,*delta()*,*maxiter*, *tol*

Đầu ra: *mu*, *sigma*, *gamma*, *delta*, *mllk*, *AIC, BIC*

1: **Begin**

2: *mu.next ←mu(); sigma ←sigma();delta ←delta()* *{Gán tham số cho giá trị ban đầu}*

3: **for** *iter* in 1 : *maxiter* **do**

4: *f b←norm.HMM.lalphabeta*(*x*,*m*,*mu*, *sigma*,*gamma*,*delta= delta*) {Tính FWP và BWP}

5: *llk ← gia 1trị hàm hợp lý*

6: **for** *j* in 1:m **do**

7: **for** *k* in 1:m **do**

8: Tính *gamma*[*j*,*k*] theo (1.4.11)

9: Tính *mu[j]* theo (1.4.12)

10: Tính *sigma* [*j*] theo (1.4.13)

11: Tính *delta* theo (1.4.10)

12: *crit ← sum*(*abs*(*mu[j] – mu()[j]*)) + *sum*(*abs*(*gamma[jk] – gamma()[jk]*)) + *sum*(*abs*(*delta[j] –delta()[j]*))+*sum(abs(sigma[j]−sigma()[j]))* {Tiêu chuẩn hội tụ}

13: **if** *crit* < *tol* **then**

14: *AIC*← -2 * (*llk*−*np*) {Tiêu chuẩn AIC}

15: *BIC*← -2 * *llk*+*np* * *log*(*n*) {Tiêu chuẩn BIC}

16: **return** (*mu*, *sigma*, *gamma*, *delta*, *mllk*, *AIC , BIC)*

17*:* **else** {Nếu chưa hội tụ}

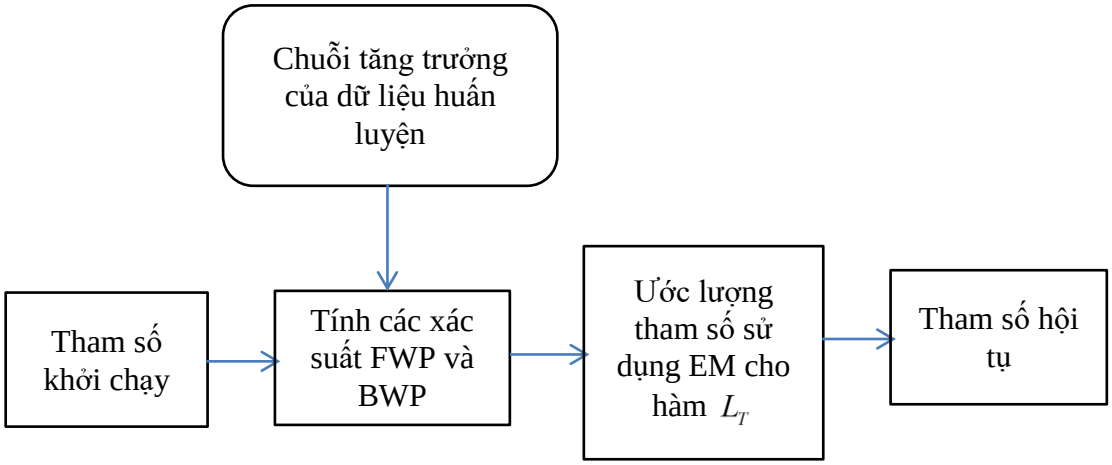
*mu*0←*mu*; *sigma*0←*sigma*; *gamma*0←*gamma*; *delta*0←*delta* *{Gán lại tham số ban đầu mới}*

18: Không hội tụ sau, "maxiter", vòng lặp

19: End.

Bộ tham số ước lượng được trong hàm này sẽ sử dụng để dùng cho các dự báo: phân phối giá trị dự báo và trạng thái dự báo theo công thức trong mục (1.4.3)

và (1.4.4) của Chương 1. Quá trình ước lượng tham số của mô hình HMM với phân phối chuẩn sử dụng thuật toán EM được minh họa trong Hình 2.2.3.



Hình 2.2.3. *Quá trình ước lượng tham số của mô hình HMM sử dụng EM*

2.3. Kết quả thực nghiệm cho HMM với phân phối Poisson

2.3.1. Ước lượng tham số

Trước tiên, luận án thử nghiệm mô hình trên tập dữ liệu time.b.to.t như đã mô tả ở trên, sau đó mô hình được áp dụng cho các tập dữ liệu khác nhằm so sánh với những mô hình sẵn có. Do dữ liệu của time.b.to.t là các số tự nhiên nên phân phối Poisson là lựa chọn thích hợp cho mô hình HMM. Luận án áp dụng mô hình HMM với phân phối Poisson đối với các quan sát từ tập dữ liệu time.b.to.t lần lượt với các trạng thái $m = 2,3,4,5$. Chẳng hạn, với $m = 4$ sử dụng hàm được minh họa dưới dạng

```
parameters <- pois.HMM.mle( b.to.t [1:43] , m=2,c (5 ,12 ,20 ,50) ,
        matrix ( rep (1/4 ,4^2) ,4 ,4))
```

trong đó m có thể được thay từ 2 đến $5, c(5,12,20,50)$ là tham số khởi chạy của $\lambda_i, i = 1,..., m$ và $matrix(rep(1/4, 4^2), 4, 4)$ là tham số khởi chạy của ma trận xác suất chuyển Γ. Lưu ý rằng các tham số ban đầu này có thể cho tùy ý trong miền xác định của nó do thuật toán EM hội tụ về cực đại mà không cần đổi tham số sang

dạng tự do. Vì vậy, ma trận $\mathbf{\Gamma}$ được chọn là ma trận phân phối đều cho tiện việc khai báo. Với $m=4$, ma trận $\mathbf{\Gamma}$ ban đầu cho bởi

$$\mathbf{\Gamma}=\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Đối với lựa chọn giá trị đầu cho tham số trạng thái $\lambda=(5,12,20,50)$ bằng cách lấy ngẫu nhiên trên khoảng $(\min_t x_t;\max_t x_t)$ với $\max$ và $\min$ lấy theo t trên tập huấn luyện. Mỗi giá trị tham số ban đầu khác nhau có thể dẫn tới giá trị cực đại khác nhau của hàm hợp lý do hàm hợp lý có thể có nhiều cực trị. Vì vậy, ta có thể thực hiện nhiều lần việc thay đổi giá trị này bất kỳ trong khoảng $(\min_t x_t;\max_t x_t)$ để kiểm tra cực đại toàn cục.

Trong luận án, chạy nhiều lần tham số ban đầu cho kết quả mà có $-L_T$ nhỏ nhất đối với mỗi m trạng thái cho trong Bảng 2.3.1 với $-l$ là $-\log L_T$.

Bảng 2.3.1. *Ước lượng tham số của mô hình Poisson-HMM cho time.b.to.t với các trạng thái m=2,3,4,5*

m	i	$\boldsymbol{\lambda}$	$\boldsymbol{\delta}$	$\mathbf{\Gamma}$	$-l$
2	1 2	11,46267 40,90969	0,6914086 0,3085914	$\begin{pmatrix} 0,8 & 0,2 \\ 0,51 & 0,49 \end{pmatrix}$	216,8401
3	1 2 3	5,78732 21,75877 57,17104	0,3587816 0,5121152 0,1291032	$\begin{pmatrix} 0,46 & 0,47 & 0,07 \\ 0,33 & 0,47 & 0,02 \\ 0,2 & 0,8 & 0 \end{pmatrix}$	171,1243

4	1 2 3 4	5,339722 16,943339 27,711948 58,394102	0,3189824 0,3159413 0,2301279 0,1349484	$\begin{pmatrix} 0,4 & 0,46 & 0,07 & 0,07 \\ 0,53 & 0,29 & 0,18 & 0 \\ 0 & 0 & 0,51 & 0,49 \\ 0,19 & 0,56 & 0,25 & 0 \end{pmatrix}$	159,898
5	1 2 3 4 5	5,226109 15,679316 25,435562 38,459987 67,708874	0,31513881 0,28158191 0,22224329 0,10376304 0,07727294	$\begin{pmatrix} 0,38 & 0,4 & 0,15 & 0,07 & 0 \\ 0,5 & 0,36 & 0 & 0,14 & 0 \\ 0,13 & 0 & 0,33 & 0,19 & 0,35 \\ 0 & 0,53 & 0,47 & 0 & 0 \\ 0,33 & 0 & 0,67 & 0 & 0 \end{pmatrix}$	154,6275

So sánh với các ước lượng của phân phối trộn độc lập, ta thấy các trạng thái là tương đối giống nhau. Tuy nhiên, mô hình HMM cho ta ma trận xác suất chuyển giữa các trạng thái $\mathbf{\Gamma}$, thể hiện sự phụ thuộc giữa các trạng thái đó. Điều hiển nhiên là, số trạng thái càng nhiều thì $-L_T$ càng nhỏ tức L_T càng lớn. Vậy có sở nào để chọn ra mô hình với số trạng thái là tốt nhất ? Trước tiên, ta tính trung bình và phương sai của mô hình để từ đó so sánh với trung bình và phương sai của mẫu dữ liệu. Kết quả được cho ở Bảng 2.3.2.

Bảng 2.3.2. *Trung bình và phương sai mô hình so với mẫu.*

m	Trung bình	Phương sai
1	20,45238	20,45238
2	20,45238	205,5624
3	20,45238	272,6776
4	20,45238	**303,7112**
5	20,45238	303,4568
Mẫu	20,45238	307,083

Kết quả cho thấy, mô hình Poisson-HMM với 4 trạng thái có phương sai gần với phương sai mẫu nhất. Tuy nhiên, điều đó không đủ bằng chứng để khẳng định mô hình 4 trạng thái là tốt nhất. Để có những phương pháp lựa chọn tốt hơn, ta cần có những tiêu chuẩn chọn mô hình theo nhiều cơ sở hơn.

2.3.2. Lựa chọn mô hình

Giả sử quan sát $x_1,...,x_T$ được sinh ra bởi mô hình "thật" f nào đó không biết và ta ướm mô hình bởi hai họ xấp xỉ khác nhau $\{g_1 \in G_1\}$ và $\{g_2 \in G_2\}$. Mục đích của chọn mô hình là xác định mô hình mà tốt nhất theo nghĩa nào đó.

Tồn tại ít nhất hai phương pháp để chọn mô hình. Phương pháp sử dụng nhiều nhất là chọn họ mô hình mà ước lượng gần nhất với mô hình thật. Với mục đích đó, ta định nghĩa sự sai lệch (đo sự không khít) giữa mô hình thật và mô hình được ướm, $\Delta(f,\hat{g}_1)$ và $\Delta(f,\hat{g}_2)$. Độ sai lệch này phụ thuộc vào mô hình thật f, mà ta không biết, cho nên không thể xác định độ lệch nào nhỏ hơn, nghĩa là mô hình nào sẽ được chọn. Thay vào đó, ta lựa chọn mô hình này dựa trên kỳ vọng của sự sai lệch, là $\hat{E}_f(\Delta(f,\hat{g}_1))$ và $\hat{E}_f(\Delta(f,\hat{g}_2))$, như là tiêu chuẩn để chọn mô hình. Bằng cách chọn độ lệch Kullback-Leibler, tiêu chuẩn chọn mô hình đơn giản hóa thành tiêu chuẩn thông tin Akaike (AIC):

$$\text{AIC} = -2\log L + 2p \qquad (2.3.1)$$

trong đó $\log L$ là log-likelihood của mô hình được ướm và p là số các tham số của mô hình. AIC có hai số hạng, số hạng thứ nhất là độ đo mức độ ướm vừa của mô hình, nó giảm theo số trạng thái m. Số hạng thứ hai là số hạng phạt, tăng lên theo m.

Ngoài ra, có thể dùng phương pháp Bayesian để lựa chọn mô hình mà được ước lượng giống thực tế nhất. Trong bước đầu tiên, trước khi xét dãy quan sát, ta xác định các phân phối tiên nghiệm (priors), nghĩa là các xác suất $P(f \in G_1)$ và $P(f \in G_2)$. Trong bước thứ hai, ta tính và so sánh các hậu nghiệm (posteriors), nghĩa là các xác suất mà f thuộc vào họ xấp xỉ, với điều kiện các quan sát, $P(f \in G_1 \mid x^{(T)})$ và $P(f \in G_2 \mid x^{(T)})$. Dưới điều kiện nhất định (xem Wasserman (2000)) [65], kết quả của phương pháp này trong thuật ngữ "tiêu chuẩn thông tin Bayesian (BIC)" chỉ khác AIC ở số hạng phạt:

$$\text{BIC} = -2\log L + p\log T, \qquad (2.3.2)$$

trong đó, L và p như trong AIC và T là số các quan sát. So với AIC, số hạng phạt trong BIC có trọng số lớn hơn khi $T > e^2$, điều này đúng trong hầu hết các ứng

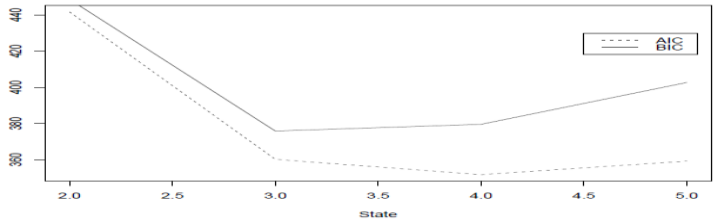
dụng. Do đó, BIC thường ưu tiên các mô hình có ít tham số hơn so với AIC.

Bây giờ, áp dụng hai tiêu chuẩn AIC và BIC đối với mô hình Poisson-HMM cho dữ liệu time.b.to.t, kết quả được liệt kê trong Bảng 2.3.3.

Bảng 2.3.3. *Tiêu chuẩn AIC và BIC*

m	2	3	4	5
AIC	441,6803	360,2486	**351,7961**	359,2551
BIC	448,6309	375,8876	**379,5988**	402,6968

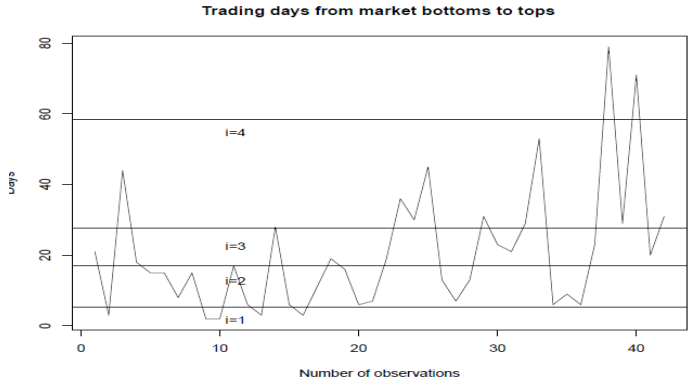
Để trực quan, Bảng 2.3.3 được minh họa bởi Hình 2.3.1.



Hình 2.3.1. *Minh họa AIC và BIC*

Như vậy, cả hai tiêu chuẩn AIC và BIC đều chọn mô hình 4 trạng thái là mô hình tốt nhất cho dữ liệu.

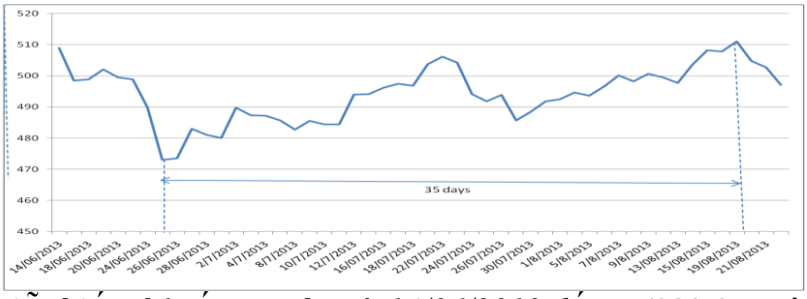
Xét sự tiến triển theo thời gian của chuỗi dữ liệu, mô hình 4 trạng thái được mô tả trong Hình 2.3.2, trong đó 4 đường kẻ ngang tương ứng với trung bình của 4 trạng thái i = 1,2,3,4 tính từ dưới lên.



Hình 2.3.2. *Mô hình Poisson-HMM với 4 trạng thái*

2.3.3. Phân phối dự báo

Như đã đề cập ở trên, dữ liệu huấn luyện đối với mô hình HMM được lấy từ 03/01/2006 đến 19/06/2013. Ta sẽ lấy dữ liệu tiếp theo từ 14/06/2013 đến 22/08/2013 để so sánh với kết quả dự báo của mô hình. Hình 2.3.3 mô tả diễn biến của chỉ số đóng của VN-Index trong khoảng thời gian này. Ta thấy rằng, số phiên dao dịch để chỉ số VN-Index từ đáy (26/06/2013) lên đỉnh (19/08/2013) là 35 ngày. Như vậy, giá trị này ứng với trạng thái 3 của mô hình (phân phối Poisson với trung bình 27,711948). Ta sẽ chờ xem kết quả dự báo của mô hình ra sao.



Hình 2.3.3. *Diễn biến chỉ số Vn-Index từ 14/06/2013 đến 22/08/2013 và thời gian chờ từ đáy lên đỉnh*

Bây giờ, ta cần tìm công thức xác định phân phối dự báo $Pr(X_{T+h} = x \mid \mathbf{X^{(T)}} = \mathbf{x^{(T)}})$. Với các ký hiệu dạng ma trận như đã trình bày ở các mục trước, phân phối này không khó để có thể tính được.

$$P\left(X_{T+h} = x|\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}\right) = \frac{\mathrm{P}\left(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}, X_{T+h} = x\right)}{P\left(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}\right)}$$

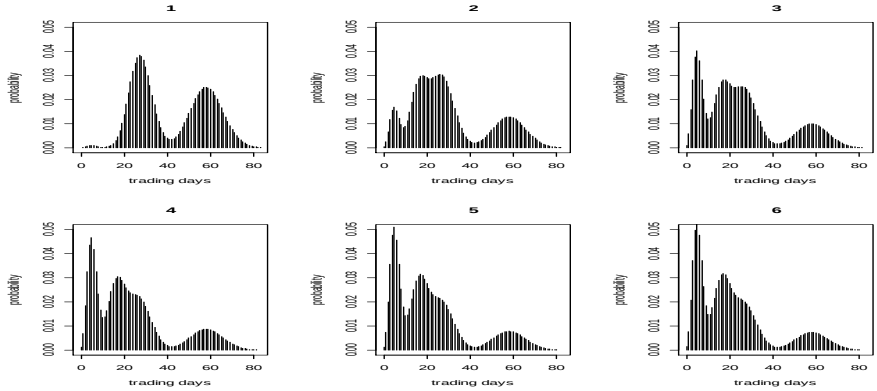
$$= \frac{\boldsymbol{\delta P}(x_1)\boldsymbol{\Gamma P}(x_2)\boldsymbol{\Gamma P}(x_3)\ldots\boldsymbol{\Gamma P}(x_T)\boldsymbol{\Gamma}^h\boldsymbol{P}(x)1'}{\boldsymbol{\delta P}(x_1)\boldsymbol{\Gamma P}(x_2)\boldsymbol{\Gamma P}(x_3)\ldots\boldsymbol{\Gamma P}(x_T)1'}$$

$$= \frac{\alpha_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x)1'}{\alpha_T 1'} \tag{2.3.3}$$

Viết $\phi_T = \boldsymbol{\alpha}_T / \boldsymbol{\alpha}_T 1'$, ta có

$$P\left(X_{T+h} = x|\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}\right) = \phi_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x)1'. \tag{2.3.4}$$

Hình 2.3.4 minh họa phân phối dự báo của số ngày giao dịch (khoảng thời gian chờ) mà cổ phiếu từ đáy lên đỉnh của 6 lần tiếp theo kể từ lần quan sát cuối cùng.



Hình 2.3.4. *Phân phối dự báo time.b.to.t cho 6 lần cổ phiếu từ đáy lên đỉnh tiếp theo*

Các phân phối này được tóm tắt trong Bảng 2.3.4

Bảng 2.3.4. *Thông tin phân phối dự báo và khoảng dự báo.*

h	1	2	3	4	5	6
Mode dự báo	27	26	5	5	5	5
Trung bình dự báo	42,30338	30,16801	25,53973	23,68432	22,48149	21,91300
Khoảng ước lượng với xác suất trên 90%						
Khoản dự báo	[20,70]	[8, 70]	[5 ,70]	[5 ,70]	[5 ,70]	[5 ,70]
Xác suất	0,9371394	0,9116366	0,9342868	0,9279009	0,9237957	0,9215904
Thực tế	35	-	-	-	-	-

Rõ ràng giá trị 35 ngày rơi vào khoảng tin cậy trên 90% dự báo của mô hình với 1 bước tiếp theo ($h=1$). Tuy nhiên, từ đồ thị phân phối của các dự báo ta thấy rằng các trạng thái tương đối phân hóa, nghĩa là phân phối có nhiều mode. Do đó, các khoảng ước lượng trở nên rộng hơn, đồng thời giá trị trung bình hay mode không phải đại lượng hợp lý sử dụng cho dự báo. Vì vậy, hợp lý hơn cả là dự báo trạng thái thay vì dự báo giá trị.

2.3.4. Trạng thái dự báo

Ở phần trước ta đã tìm ra phân phối điều kiện của trạng thái C_t cho trước

quan sát $X^{(T)}$. Làm như vậy ta chỉ xét trạng thái hiện tại và các trạng thái quá khứ. Tuy nhiên, cũng có thể tính được phân phối điều kiện cho trạng thái tương lai C_{T+h}, việc này gọi là dự báo trạng thái.

Không mấy khó khăn, ta có thể tính được

$$Pr(C_{T+h} = i \mid \mathbf{X^{(T)}} = \mathbf{x^{(T)}}) = \frac{\boldsymbol{\alpha}_\mathbf{T}\boldsymbol{\Gamma}^\mathbf{h}(,\mathbf{i})}{L_T} = \boldsymbol{\phi}_\mathbf{T}\boldsymbol{\Gamma}^\mathbf{h}(,\mathbf{i}) \tag{2.3.5}$$

với $\boldsymbol{\phi}_\mathbf{t} = \boldsymbol{\alpha}_\mathbf{T} / \boldsymbol{\alpha}_\mathbf{T}\mathbf{1}'$.

Ta tiến hành dự báo trạng thái của mô hình Poisson-HMM 4 trạng thái của dữ liệu time.b.to.t với 6 lần tiếp theo, kết quả được chỉ ra ở Bảng 2.3.5.

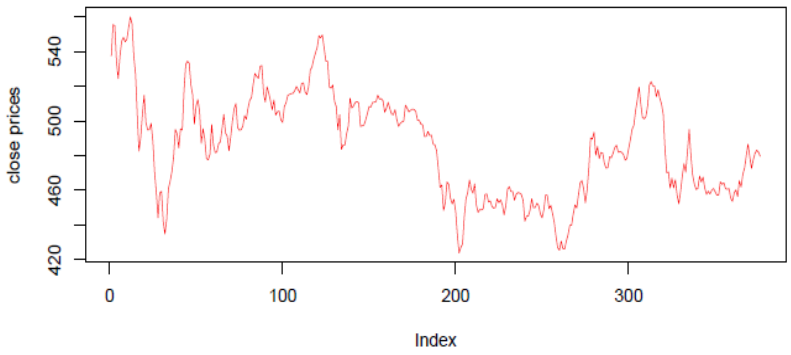
Bảng 2.3.5. *Dự báo trạng thái 6 lần tiếp theo cho time.b.to.t.*

h	1	2	3	4	5	6
1	0,006577011	0,09686901	0,2316797	0,2688642	0,2934243	0,3060393
2	0,003744827	0,27624774	0,2658957	0,2931431	0,3048425	0,3098824
3	**0,506712945**	0,37858412	0,3104563	0,2698832	0,2508581	0,2407846
4	0,482965217	0,24829913	0,1919683	0,1681095	0,1508750	0,1432937

Từ kết quả dự báo ta thấy, ở lần quan sát tiếp theo thứ nhất kể từ quan sát cuối cùng, mô hình dự báo rơi vào trạng thái 3 hoặc 4 với xác suất 0.507 và 0.483 tương ứng trong khi xác suất rơi vào trạng thái 1 và 2 xấp xỉ bằng 0. So với kết quả thực tế (trạng thái 3), kết quả dự báo này là chấp nhận được. Và dễ thấy, phân phối xác suất dự báo trạng thái hội tụ về phân phối dừng $\boldsymbol{\delta}$.

2.4. Kết quả thực nghiệm mô hình HMM với phân phối chuẩn

Mục này, luận án trình bày các kết quả đạt được khi áp dụng mô hình HMM với phân phối chuẩn để ướm vào dữ liệu VN-Index với 376 giá trị đóng cửa từ 11/4/2009 đến 13/5/2011. Các giá trị này được minh họa trên Hình 2.4.1.



Hình 2.4.1. *Hình ảnh của VN-Index với 376 giá đóng cửa từ 11/4/2009 đến 13/5/2011*

2.4.1. Ước lượng tham số

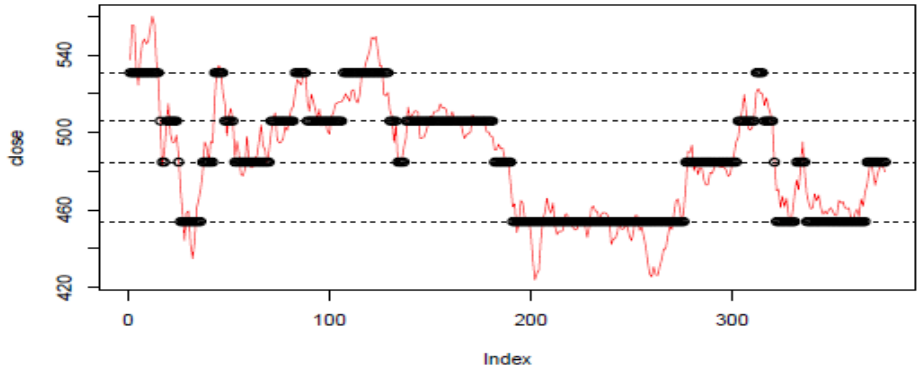
Với phân phối ban đầu bất kỳ (thực tế luận án cho khởi chạy với nhiều phân phối ban đầu khác nhau và chọn kết quả cho $-L_T$ nhỏ nhất ở mỗi m trạng thái), ước lượng bằng EM cho cùng một kết quả tối ưu:

$$\Gamma = \begin{pmatrix} 0,9717 & 0,0283 & 0,0000 & 0,0000 \\ 0,0927 & 0,8106 & 0,0804 & 0,0163 \\ 0,0000 & 0,0748 & 0,8624 & 0,0628 \\ 0,0000 & 0,0000 & 0,0818 & 0,9182 \end{pmatrix}$$

$$\mu = (453,9839; 484,6801; 505,9007; 530,8300)$$

$$\sigma = (10,6857; 7,1523; 6,4218; 13,0746)$$

Hình 2.4.2 mô tả giá trị của VNIndex cùng với dãy trạng thái tốt nhất tính theo thuật toán Viterbi. Các đường nét đứt biểu diễn 4 trạng thái trong khi các chấm đen đậm thể hiện trạng thái tốt nhất cho giá trị tại mỗi thời điểm.



Hình 2.4.2. *Dữ liệu VN-Index: dãy trạng thái tốt nhất*

2.4.2. Lựa chọn mô hình

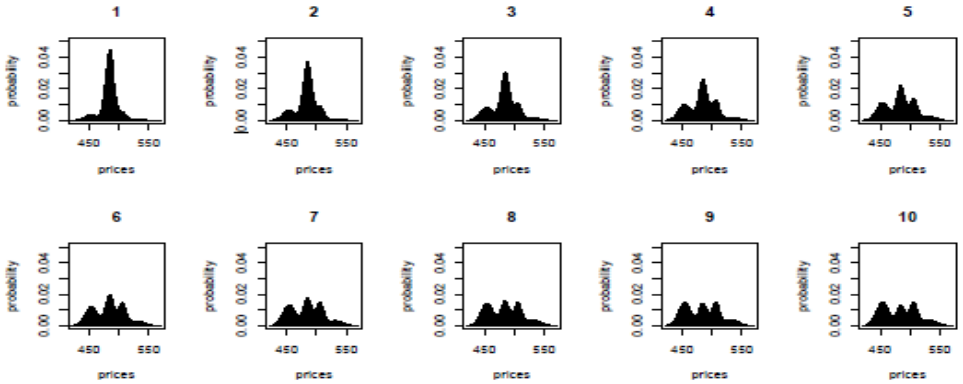
Theo lý thuyết chọn mô hình HMM trên tiêu chuẩn BIC và AIC cho chuỗi chỉ số VN-index, AIC và BIC đều chọn 4 trạng thái. Các giá trị của tiêu chuẩn cho trong Bàng 2.4.1.

Bảng 2.4.1. *Dữ liệu VN-Index: chọn số trạng thái*

Model	-log*L*	AIC	BIC
2-state HM	1.597,832	3.205,664	3.225,312
3-state HM	1.510,989	3.043,978	3.087,204
4-state HM	1.439,179	**2.916,358**	**2.991,02**
5-state HM	không hội tụ		

2.4.3. Phân phối dự báo

Như trình bày trong mục 1.4.3 trong Chương 1, Hình 2.4.3 biểu diễn 10 phân phối dự báo cho giá trị của VNIndex. Ta thấy phân phối dự báo tiến tới phân phối dừng khá nhanh.



Hình 2.4.3. *Dữ liệu VN-Index data: phân phối dự báo của 10 ngày tiếp theo.*

Như vậy, mô hình HMM với phân phối nhất định phù hợp với dự báo trong một số trường hợp, nhất là đối với dữ liệu mà nó thực sự khít với phân phối lựa chọn trong mô hình. Tuy nhiên, chuỗi thời gian sinh ra bởi một biến ngẫu nhiên có ướm khít với phân phối chuẩn (hoặc trộn các phân phối chuẩn) hay phân phối nào khác được chọn hay không là câu hỏi sẽ quyết định đến sự phù hợp cũng như độ chính xác của dự báo. Mục tiếp theo sẽ trình bày một số cơ sở cho thấy sự cần thiết phải xây dựng mô hình phi phân phối trong dự báo.

2.4.4. Trạng thái dự báo

Bảng 2.4.2 dự báo khả năng (xác suất) cao nhất đối với mỗi trạng thái cho 30 ngày tiếp theo kể từ ngày cuối cùng là 13/05/2011.

Bảng 2.4.2. *Dự báo khả năng (xác suất) cao nhất đối với mỗi trạng thái cho 30 ngày tiếp theo kể từ ngày cuối cùng là 13/05/2011*

Days	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
State=[1,]	0,0975	0,1695	0,2261	0,2709	0,3065	0,3350
[2,]	0,8062	0,6622	0,5517	0,4665	0,4005	0,3492
[3,]	0,0799	0,1351	0,1724	0,1971	0,2128	0,2223
[4,]	0,0162	0,0330	0,0496	0,0653	0,0800	0,0933

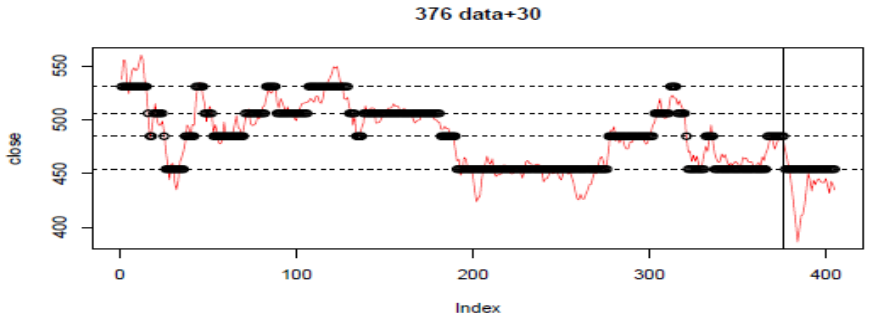
	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	0,3579	0,3764	0,3915	0,4039	0,4141	0,4225	0,4296
[2,]	0,3092	0,2778	0,2530	0,2334	0,2177	0,2052	0,1951
[3,]	0,2274	0,2296	0,2298	0,2288	0,2270	0,2248	0,2224

[4,]	0,1053	0,1160	0,1255	0,1338	0,1410	0,1473	0,1527
	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
[1,]	0,4355	0,4405	0,4448	0,4484	0,4515	0,4542	0,4565
[2,]	0,1870	0,1803	0,1749	0,1705	0,1669	0,1639	0,1614
[3,]	0,2200	0,2176	0,2154	0,2133	0,2113	0,2096	0,2080
[4,]	0,1573	0,1613	0,1647	0,1676	0,1701	0,1722	0,1739
	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]
[1,]	0,4586	0,4604	0,4619	0,4633	0,4646	0,4657	0,4667
[2,]	0,1593	0,1576	0,1561	0,1549	0,1539	0,1530	0,1523
[3,]	0,2066	0,2053	0,2041	0,2031	0,2022	0,2014	0,2007
[4,]	0,1754	0,1766	0,1776	0,1784	0,1791	0,1797	0,1801
		[,28]		[,29]		[,30]	
[1,]		0,4676		0,4684		0,4692	
[2,]		0,1517		0,1512		0,1507	
[3,]		0,2000		0,1995		0,1990	
[4,]		0,1805		0,1807		0,1809	

Ta thấy khả năng cao nhất trong 7 ngày đầu rơi vào trạng thái 2 và các ngày sau rơi vào trạng thái 1. Do đó, mô hình không hiệu quả trong dài hạn nhưng tốt cho ngắn hạn. Tuy nhiên, ta có thể dự báo bằng cách cập nhật liên tục dữ liệu một cách tự động.

Bây giờ luận án cập nhật tiếp dữ liệu từ 14/5/2011 đến 23/6/2011 với 30 giá đóng của của cổ phiếu nhằm so sánh giá trị dự báo với giá trị thực của dữ liệu.

Hình 2.4.4 cho thấy rằng giá trị của 30 ngày này hầu hết ở trạng thái 1. Điều này chứng tỏ dự báo là đúng đắn.



Hình 2.4.4. *Dữ liệu VNIndex: So sánh trạng thái dự báo với trạng thái thực tế.*

2.5. Một số kết quả so sánh

Mục này luận án trình bày kết quả dự báo của mô hình HMM với một số mô hình đã có [33] trên một số dữ liệu là các chuỗi chỉ số chứng khoán. Do đặc điểm giá trị của chuỗi thời gian tăng trưởng nhận các giá trị thực nên mô hình HMM với phân phối chuẩn được lựa chọn. Mô hình luận án đề xuất và mô hình so sánh được thực hiện trên cùng một tập huấn luyện và trên cùng một tập kiểm tra nhằm đảm bảo chính xác của phép so sánh. Độ đo độ chính xác được sử dụng là trung bình phần trăm sai số (MAPE) được tính bởi:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - p_i}{a_i}\right| * 100\%$$

Trong đó n số các giá trị cần test, a_i và p_i tương ứng là giá trị thực tế và giá trị dự báo của ngày thứ i của tập kiểm tra. Như vậy, MAPE càng bé nghĩa là độ chính xác càng cao.

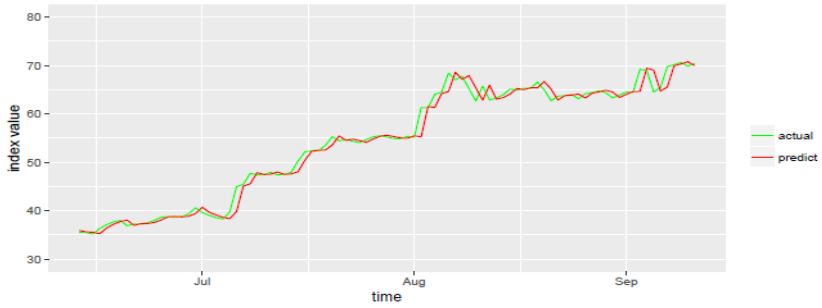
Đối với dữ liệu chỉ số cổ phiếu Apple Computer inc. từ ngày 10/01/2003 đến 21/01/2005 gồm 492 chỉ số, luận án thử nghiệm mô hình HMM với phân phối chuẩn trên tập huấn luyện gồm 400 quan sát, còn lại là tập kiểm tra các dự báo. Kết quả các sai số dự báo MAPE khi chạy mô hình với 4 trạng thái 16 lần với các tham số ban đầu ngẫu nhiên cho kết quả bởi Bảng 2.5.1.

Bảng 2.5.1. *MAPE nhiều lần chạy HMM cho dữ liệu Apple*

1,812	1,778	1,790	1,784	1,815	1,777	1,812	1,794
1,779	1,788	1,802	1,816	1,778	1,800	1,790	1,789

Trung bình: **1,795**.

Độ chính xác trung bình **1,795** và giá trị dự báo trung bình minh họa bởi Hình 2.5.1.



Hình 2.5.1. *Dự báo HMM cho giá cổ phiếu apple:actual-giá thật; predict-giá dự báo*

Tương tự độ với các dữ liệu cổ phiếu Ryanair Airlines từ 06/01/2003 đến 17/01/2005; IBM Corporation. từ 10/01/2003 đến 21/01/2005 và Dell Inc. từ 10/01/2003 đến 21/01/2005. Kết quả so sánh độ đo độ chính xác MAPE với 400 quan sát huấn luyện được chỉ ra trong Bảng 2.5.2.

Bảng 2.5.2. *So sánh độ chính xác của mô hình HMM với một số mô hình khác*

Dữ liệu	Mô hình ARIMA	Mô hình ANN	Mô hình HMM
Apple	1,801	1,801	**1,795**
Ryanair	1,504	1,504	**1,306**
IBM	0,660	0,660	**0,660**
Dell	0,972	0,972	**0,863**

Từ kết quả trong Bảng 2.5.2 ta thấy mô hình HMM với phân phối chuẩn cho độ chính xác dự báo cao hơn so với mô hình cổ điểm là ARIMA và mô hình ANN.

2.6. Hạn chế của mô hình dự báo với phân phối tất định

Mục này luận án sử dụng một số thống kê mô tả để chỉ ra rằng phân phối chuẩn (thường dùng trong các mô hình dự báo) không phản ánh đúng thực tế của chuỗi thời gian chẳng hạn như chuỗi tài chính.

Ta sẽ tập trung vào hai vấn đề chính.

- Ta thấy rằng logarit lợi suất thực tế không thể hiện theo một phân phối chuẩn.
- Các dao động hay các tham số dùng để ước lượng sự không chắc chắn của nó (hay nói chung là môi trường) thay đổi ngẫu nhiên trên toàn bộ

thời gian và được cộng gộp.

2.6.1. Phân phối chuẩn

Định nghĩa 2.6.1. [66] *Phân phối chuẩn,* $Normal(\mu,\sigma^2)$ *là phân phối có trung bình* $\mu\in\mathbb{R}$ *và phương sai* $\sigma^2>0$. *Hàm đặc trưng của nó được cho bởi*

$$\phi_{Normal}(u;\mu,\sigma^2)=e^{iu\mu}e^{-\frac{1}{2}\sigma^2u^2}$$

và hàm mật độ là

$$f_{Normal}(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Tính chất [66]

Phân phối chuẩn $Normal(\mu,\sigma^2)$ đối xứng quanh trung bình của nó và có độ nhọn bằng 3:

	$Normal(\mu,\sigma^2)$
Trung bình	μ
Phương sai	σ^2
Độ lệch đối xứng	0
Độ nhọn	3

2.6.2. Các tham số tương ứng từ dữ liệu thực

Không đối xứng và độ nhọn vượt chuẩn

Với một biến ngẫu nhiên X, ta ký hiệu $\mu_X=\mu=E[X]$ là trung bình của nó và $\text{var}[X]=E[(X-\mu_X)^2]\geq 0$ là phương sai. Căn bậc hai của phương sai $\sqrt{\text{var}[X]}$ được gọi là độ lệch chuẩn (SD). Chú ý rằng độ lệch chuẩn của một biến ngẫu nhiên có phân phối chuẩn $N(\mu,\sigma^2)$ bằng $\sigma>0$.

Bảng 2.6.1. *Trung bình, độ lệch chuẩn, độ lệch đối xứng, độ nhọn của một số chỉ số có VN-index*

Chỉ số	Trung bình	SD	Độ lệch đối xứng	Độ nhọn
VN-index (2009-2010)	-0,000786	0,012378	-0,133837	4,658174

S&P 500 (1997-1999)	0,0009	0,0119	-0,4409	6,94
Nasdaq-Composite	0,0015	0,0154	-0,5439	5,78
DAX	0,0012	0,0157	-0,4314	4,65
SMI	0,0009	0,0141	-0,3584	5,35
CAC-40	0,0013	0,0143	-0,2116	4,64

Trong bảng 2.6.1 ta tóm tắt theo thống kê trung bình, độ lệch chuẩn cho một tập hợp chỉ số bao gồm chỉ số VN-index bộ dữ liệu 2009-2010. Ta thấy một cách khái quát về sự không đối xứng và độ nhọn cao của phân phối thống kê.

Độ lệch đối xứng

Độ lệch đối xứng đo mức độ đối xứng của một phân phối. Được tính bởi:

$$\frac{E[(X-\mu_X)^3]}{(\text{var}[X])^{3/2}}.$$

Với một phân phối đối xứng (như phân phối chuẩn $N(\mu,\sigma^2)$), độ lệch đối xứng bằng 0.

Nếu ta nhìn vào loga lợi suất hàng ngày của các chỉ số trên, ta thấy một số độ lệch đối xứng âm đáng kể. Trong bảng 2.6.1, ta thấy độ lệch đối xứng theo thống kê của loga lợi suất hàng ngày của chỉ số VN-index là âm, trong khi phân phối chuẩn có độ lệch đối xứng bằng không.

Độ nhọn vượt mức

Tiếp theo, ta cũng chỉ ra rằng các biến động lớn trong giá tài sản xảy ra thường xuyên hơn trong một mô hình với số gia là phân phối chuẩn. Mức độ thường xuyên của sự biến động này thường được xác định khi độ nhọn vượt mức cũng như phần đuôi của phân phối bành trướng bất thường.

Cách để đo độ nhọn vượt mức này là sử dụng công thức $\frac{E[(X-\mu_X)^4]}{\text{var}[X]^2}$.

Với phân phối chuẩn, độ nhọn (mesokurtic) bằng 3. Nếu phân phối có đỉnh phẳng hơn (platykurtic), thì độ nhọn nhỏ hơn 3. Nếu phân phối có chỏm cao (leptokurtic), thì độ nhọn lớn hơn 3.

Trong bảng 2.6.1, ta dễ thấy rằng dữ liệu của ta luôn luôn có độ nhọn lớn hơn 3,

cho thấy rằng phần đuôi của phân phối chuẩn đi về 0 nhanh hơn nhiều các dữ liệu thống kê và điều đó có nghĩa phân phối thống kê có chóp cao hơn nhiều phân phối chuẩn. Thực tế những phân phối có chỏm cao hơn phân phối chuẩn (leptokurtic) đã được chú ý bởi Fama (1965) [26].

Sự ước lượng mật độ

Cuối cùng, ta sẽ nhìn vào bức tranh về mật độ theo thống kê và so sánh với mật độ phân phối chuẩn.

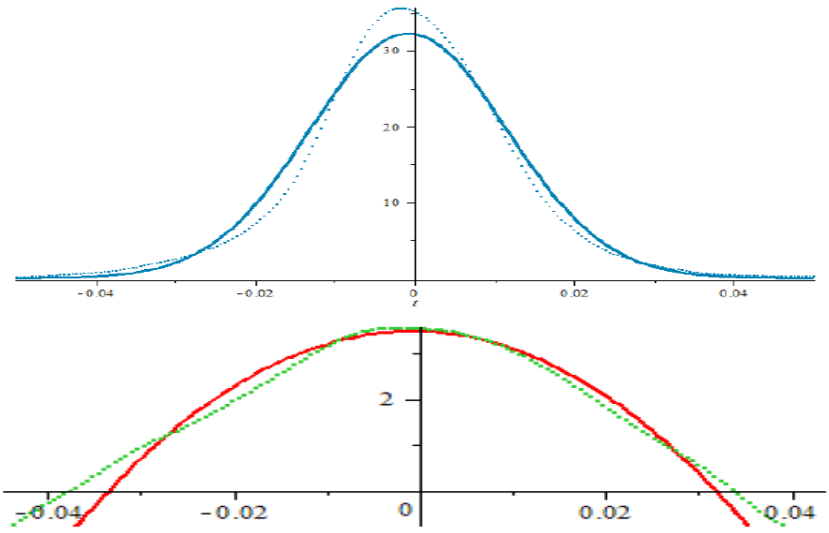
Hạt nhân ước lượng mật độ

Để ước lượng mật độ theo lối thực nghiệm, ta sử dụng hạt nhân ước lượng mật độ. Mục đích của sự ước lượng mật độ là xấp xỉ hàm mật độ xác suất $f(x)$ của biến ngẫu nhiên X. Giả sử rằng ta có n quan sát độc lập $x_1, x_2, \cdots, x_n$ từ biến ngẫu nhiên X. *Hạt nhân nước lượng mật độ* $f_h(x)$ cho sự ước lượng của hàm mật độ $f(x)$ tại điểm x được định nghĩa là

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

trong đó $K(x)$ gọi là hàm hạt nhân và h là băng thông. Trong lậun án này ta sử dụng hạt nhân Gauss: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Tuy nhiên, còn có các hàm hạt nhân tiêu biểu khác là hạt nhân đồng dạng, hạt nhân tam giác, toàn phương và hạt nhân cosin. Trong công thức ở trên ta cũng phải chọn băng thông h. Với hạt nhân Gauss, ta sử dụng giá trị "quy tắc ngón tay cái" của Silverman $h = 1.06\sigma n^{-1/5}$ (xem Silverman 1986)[59].

Hạt nhân ước lượng mật độ Gauss dựa trên loga lợi suất hàng ngày của VN-Index suốt thời gian từ 2009 đến 2010 được chỉ ra trong hình 2.6.1. Ta thấy đó là một phân phối có chóp nhọn. Điều này cho ta thấy, hầu hết các thời điểm, giá cổ phiếu không thay đổi nhiều; có một khối lượng đáng kể trong tổng số mật độ chuyển động quanh 0. Cùng được vẽ trong hình 2.6.1 là mật độ chuẩn với trung bình $\mu = -0.000786321$ và $\sigma = 0.01237793$, tương ứng với trung bình và độ lệch chuẩn của loga lợi suất hàng ngày.



Hình 2.6.1. *(a) Hạt nhân ước lượng mật độ Gauss và phân phối chuẩn và (b) loga các mật độ của loga lợi suất hàng ngày của VN-Index*

Phân phối nửa nặng đuôi (Semi-heavy tailed distribution)

Đồ thị hàm mật độ tập trung vào trung tâm; tuy nhiên, dáng vẻ của phần đuôi cũng rất quan trọng. Do đó, ta nhìn vào hình 2.6.1 loga mật độ, nghĩa là $\log f_h(x)$ và tương ứng là loga của mật độ chuẩn. Loga mật độ của phân phối chuẩn có sự phân tán bậc hai, trong khi loga mật độ thống kê dường như có nhiều sự phân tán tuyến tính hơn. Đặc trưng này điển hình cho dữ liệu tài chính và thường được đề cập đến như là sự "nửa nặng đuôi (semi-heavy tail)".

2.7. Kết luận

Chương này đã luận án đã trình bày giải pháp ứng dụng mô hình HMM trong dự báo chuỗi thời gian với phân phối tất định.

Thứ nhất, luận án chỉ ra rằng một phân phối xác suất đơn lẻ khó có thể ướm khít phân phối thống kê của một chuỗi thời gian sinh bởi biến ngẫu nhiên trong tài chính. Điều này mở ra yêu cầu cần được ướm bởi trộn của nhiều phân phối. Thực nghiệm cũng chỉ ra các phân phối này cần phải phụ thuộc bởi có sự tương quan giữa dữ liệu lịch sử và hiện tại như được chỉ ra trong Hình 2.1.5. Kết quả là mô hình HMM là một lựa chọn tự nhiên.

Thứ hai, luận án đã mô hình hóa một bài toán dự báo thực tế vào mô hình

HMM với các lựa chọn phân phối trộn khác nhau (Poisson và phân phối chuẩn). Kết quả cho thấy, mô hình HMM phù hợp cho dự báo chuỗi thời gian so với thực tế trong ngắn hạn. Hơn nữa, thực hiện trên các tập khác nhau và so sánh với các mô hình dự báo cổ điển như ARIMA và ANN cho thấy mô hình HMM có độ chính xác cao hơn.

Thứ ba, luận án chỉ ra bằng thống kê mô tả rằng phân phối chuẩn (dùng trong nhiều mô hình dự báo) không phù hợp với giả thiết về phân phối của biến ngẫu nhiên sinh ra chuỗi thời gian. Nhận định này là cơ sở để luận án phát triển mô hình dự báo theo hướng kết hợp các mô hình hiện có cùng với những mô hình dự báo phi phân phối.

Kết quả nghiên cứu này đã được nghiên cứu sinh công bố trong bài báo [A5].

Chương 3. MỞ RỘNG MÔ HÌNH XÍCH MARKOV BẬC CAO VÀ CHUỖI THỜI GIAN MỜ TRONG DỰ BÁO

3.1. Mở đầu

Như đã đề cập trong Chương 1 cũng như những hạn chế chỉ ra của mô hình HMM trong chương 2, chương này luận án trình bày mô hình kết hợp chuỗi thời gian mờ và xích Markov bậc cao nhằm cải thiện độ chính xác của dự báo cho các tập dữ liệu khác nhau. Một xích Markov bậc m (hay xích Markov với trí nhớ m) là một quá trình trong đó trạng thái tương lai phụ thuộc vào m trạng thái liên tiếp trong quá khứ, nghĩa là $Pr(C_{t+1} \mid C_t, C_{t-1}, ..., C_1) = Pr(C_{t+1} \mid C_t, C_{t-1}, ..., C_{t-m+1}),$ với $\{C_t\}$ là dãy trạng thái. Trong thị trường chứng khoán, sự thay đổi của chỉ số chứng khoán phụ thuộc vào rất nhiều nhân tố không rõ ràng trong hệ thống kinh tế. Những nhân tố này lặp lại ngày qua ngày theo một cách ngẫu nhiên nào đó. Vì vậy, những thay đổi trong quá khứ có thể xuất hiện lại trong tương lai khi các điều kiện tác động được lặp lại. Do đó, sử dụng mô hình Markov bậc cao có thể là một hướng tiếp cận hợp lý để cấu trúc mô hình dự báo chuỗi thời gian (đặc biệt là chuỗi thời gian chứng khoán) cho mục đích tái hiện lại viễn cảnh đã từng xuất hiện khá xa trong quá khứ và có thể xuất hiện lại ở tương lai. Mặt khác, chỉ số tăng trưởng của dãy chỉ số chứng khoán thay đổi liên tục ở những mức độ khác nhau. Điều này cho thấy chuỗi thời gian mờ có thể được sử dụng để mờ hóa chuỗi tăng trưởng thành những nhãn ngữ nghĩa như "tăng mạnh", "ổn định" hay "giảm mạnh" hay thậm chí còn nhiều hơn thế. Những trạng thái này theo một cách tự nhiên trở thành những trạng thái của một xích Markov bậc cao nếu ta giả thiết chúng tuân theo một xích Markov. Dựa vào tính dự báo của xích Markov có thể dễ dàng dự báo được trạng thái tương lai của chuỗi tăng trưởng, từ đó ước lượng được giá trị dự báo.

Chính vì vậy, luận án đề xuất mô hình kết hợp xích Markov bậc cao với chuỗi thời gian mờ được trình bày trong chương này với những nội dung chính:

Trước hết, mô hình chuỗi thời gian mờ được sử dụng để phân vùng dữ liệu

lịch sử thành các trạng thái (mục 1.5 chương 1). Sau đó sử dụng mô hình Markov dự báo trạng thái tương lai. Dựa vào lý thuyết mờ ta có được kết quả dự báo.

Tiếp theo luận án mở rộng mô hình kết hợp được đề xuất với xích Markov bậc cao cho cả xích Markov bậc cao cổ điển và xích Markov bậc cao cải tiến (mục 3.2).

Kết quả thực nghiệm cho thấy độ chính xác của kết quả dự báo được cải thiện đáng kể so với các mô hình trước đó như: ARIMA, mô hình ANN, mô hình (HMM), mô hình kết hợp HMM-Fuzzy. Hơn nữa, mô hình Markov bậc cao có độ chính xác cao hơn bậc nhất đối với dữ liệu có tính mùa vụ.

3.2. Xích Markov bậc cao

Giả sử rằng mỗi điểm dữ liệu C_t trong một dãy dữ liệu được phân loại lấy giá trị trong tập $I = 1, 2, ..., m$ và m là hữu hạn, nghĩa là dãy có m loại hoặc trạng thái. Một xích Markov bậc k là một chuỗi biến ngẫu nhiên mà

$$Pr(C_n = c_n \mid C_{n-1} = c_{n-1}, ..., C_1 = c_1) = Pr(C_n = c_n \mid C_{n-1} = c_{n-1}, ..., C_{n-k} = c_{n-k})$$

Ước tính một mô hình chuỗi Markov bậc k có $(m-1)m^k$ tham số mô hình. Vấn đề lớn trong việc sử dụng mô hình này là số lượng các tham số (các xác suất chuyển) tăng theo cấp số nhân theo bậc của mô hình. Vì số lượng các tham số quá lớn dẫn đến việc ít sử dụng trực tiếp chuỗi Markov bậc cao vào các bài toán thực tế. Trong [55], Raftery đã đề xuất một mô hình chuỗi Markov bậc cao (CMC). Mô hình này có thể được viết như sau:

$$P(C_n = c_n \mid C_{n-1} = c_{n-1}, ..., C_{n-k} = c_{n-k}) = \sum_{i=1}^{k} \lambda_i q_{c_n c_i} \quad (3.2.1)$$

Trong đó

$$\sum_{i=1}^{k} \lambda_i = 1$$

và $Q = [q_{ij}]$ là ma trận chuyển với tổng cột bằng 1, như vậy:

$$0 \le \sum_{i=1}^{k} \lambda_i q_{c_n c_i} \le 1, c_n, c_i \in I \tag{3.2.2}$$

Điều kiện (3.2.2) để đảm bảo rằng vế bên phải trong (3.2.1) là một phân phối xác suất. Tổng số lượng các tham số độc lập trong mô hình này là $(k+m^2)$. Raftery đã chứng minh được rằng (3.2.1) tương đương với mô hình chuẩn $AR_{(n)}$. Hơn nữa, các tham số $c_{j_n c_i}$ và λ_i có thể ước lượng bằng cách cực đại hàm log-likelihood của phương trình (3.2.1) với ràng buộc (3.2.2). Tuy nhiên, phương pháp này lại gặp phải vấn đề giải phương trình phi tuyến. Các phương pháp số đề xuất không đảm bảo hội tụ và cũng không phải là cực đại toàn cục ngay cả khi nó hội tụ.

3.2.1. Mô hình Markov bậc cao mới (IMC)

Trong tiểu mục này, luận án trình bày việc mở rộng mô hình Raftery [55] thành một mô hình chuỗi Markov bậc cao tổng quát hơn bằng cách cho phép Q để thay đổi theo độ trễ khác nhau. Ở đây chúng ta giả định rằng trọng số λ_i không âm thỏa mãn:

$$\sum_{i=0}^{k} \lambda_i = 1 \tag{3.2.3}$$

Ta có (3.2.1) có thể được viết lại như sau:

$$C_{n+k+1} = \sum_{i=1}^{k} \lambda_i Q C_{n+k+1-i} \tag{3.2.4}$$

Trong đó $C_{n+k+1-i}$ là phân phối xác suất của các trạng thái tại thời điểm $(n+k+1-i)$. Sử dụng (3.2.3) và Q là một ma trận xác suất chuyển, chúng ta có mỗi phần tử C_{n+k+1} nằm giữa 0 và 1, và tổng tất cả phần tử bằng 1. Trong mô hình Raftery, không giả sử λ không âm nên các điều kiện (3.2.2) được bổ sung vào để đảm bảo rằng C_{n+k+1} là phân phối xác suất của các trạng thái.

Mô hình trong (3.2.4) có thể được khái quát như sau:

$$C_{n+k+1} = \sum_{i=1}^{k} \lambda_i Q_i C_{n+k+1-i} \tag{3.2.5}$$

Tổng số lượng tham số độc lập trong mô hình mới là $(k+km^2)$.

Chúng ta lưu ý rằng nếu $Q_1 = Q_2 = ... = Q_k$ thì (3.2.5) trở thành mô hình của Raftery trong (3.2.4). Trong mô hình chúng ta giả sử rằng C_{n+k+1} phụ thuộc vào $C_{n+i}(i = 1,2,3,...,k)$ thông qua ma trận Q_i và trọng số λ_i. Chúng ta có mối quan hệ ma trận chuyển i bước Q_i của quá trình và chúng ta sử dụng quan hệ này để ước lượng Q_i. Ở đây chúng ta giả sử rằng mỗi Q_i là một ma trận ngẫu nhiên không âm với tổng cột bằng 1. Trước khi trình bày phương pháp ước lượng các tham số cho mô hình chúng ta cùng thảo luận về một số tính chất của mô hình đề xuất.

Mệnh đề 3.2.1. [72] *Nếu* Q_k *là tối giản và* $\lambda_k > 0$ *sao cho* $0 \le \lambda_i \le 1$ *và* $\sum_{i=1}^{k} \lambda_i = 1$ *thì mô hình trong (3.2.5) có một phân phối ổn định* $\bar{C}$ *khi* n *tiến đến* ∞ *không phụ thuộc vào các Vector trạng thái ban đầu* $C_0, C_1, ..., C_{k-1}$. *Phân phối ổn định* $\bar{C}$ *cũng là nghiệm duy nhất của hệ phương trình tuyến tính sau đây:*

$$(I - \sum_{i=1}^{n} \lambda_i Q_i)\bar{X} = 0 \text{ và } 1^T \bar{C} = 1 \qquad (3.2.6)$$

Trong đó I là ma trận mật độ dạng $m \times m$ (m là số trạng thái có thể được cho bởi mỗi điểm dữ liệu) và 1 là một Vector $m \times 1$ toàn số 1.

3.2.2. Ước lượng tham số

Trong mục này, tác giả trình bày các phương pháp hiệu quả để ước lượng các tham số Q_i và λ_i với $i = 1, 2, ..., k$ Để ước lượng Q_i, chúng ta có thể coi Q_i như là một ma trận chuyển i bước của dãy dữ liệu phân loại C_n. Cho dãy dữ liệu phân loại C_n, ta có thể đếm tần số chuyển $f_{jl}^{(i)}$ trong dãy từ trạng thái l đến trạng thái j sau i bước. Hơn nữa, chúng ta có thể xây dựng ma trận chuyển i bước cho dãy C_n như sau:

$$F^{(i)} = \begin{bmatrix} f_{11}^{(i)} & \cdots & \cdots & f_{m1}^{(i)} \\ f_{12}^{(i)} & \cdots & \cdots & f_{m2}^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1m}^{(i)} & \cdots & \cdots & f_{mm}^{(i)} \end{bmatrix}$$

Từ $F^{(i)}$, chúng ta nhận được các ước tính cho $Q_i = [q_{lj}^{(i)}]$ như sau:

$$\hat{Q}_i = \begin{pmatrix} \hat{q}_{11}^{(i)} & \cdots & \cdots & \hat{q}_{m1}^{(i)} \\ \hat{q}_{12}^{(i)} & \cdots & \cdots & \hat{q}_{m2}^{(i)} \\ \vdots & \cdots & \cdots & \vdots \\ \hat{q}_{1m}^{(i)} & \cdots & \cdots & \hat{q}_{mm}^{(i)} \end{pmatrix}$$

Ở đó

$$\hat{q}_{lj}^{(i)} = \begin{cases} \dfrac{f_{lj}^{(i)}}{\sum_{l=1}^{m} f_{lj}^{(i)}} \text{ neu } \sum_{l=1}^{m} f_{lj}^{(i)} \neq 0 \\ 0 \text{ truong hop khác} \end{cases}$$

Chúng ta lưu ý rằng các tính toán phức tạp của việc xây dựng $F^{(i)}$ là của phép tính $O(L^2)$, trong đó L là chiều dài của dãy dữ liệu. Vì thế tổng số tính toán phức tạp của việc xây dựng $F^{(i)k}_{i=1}$ là của phép tính $O(kL^2)$. Ở đây k là số độ trễ.

Bây giờ ta trình bày rõ các bước ước lượng các tham số λ_i như sau [19] mà luận án sẽ dùng để nhúng vào mô hình kết hợp đề xuất.

Giả sử $C_n \to \bar{C}$ khi n tiến đến vô cùng, khi đó $\bar{C}$ có thể được ước lượng từ dãy C_n bằng cách tính tỷ lệ sự xuất hiện của mỗi trạng thái trong dãy và chúng ta đặt bằng $\hat{C}$.

Từ (3.2.6) ta hy vọng rằng:

$$\sum_{i=1}^{k} \lambda_i Q_i \hat{C} \approx \hat{C}$$

Điều này cho chúng ta một cách ước lượng các tham số $\lambda = (\lambda_1, ..., \lambda_k)$ như sau. Chúng ta xét bài toán cực tiểu sau đây:

$$\min_{\lambda} \| \sum_{i=1}^{k} \lambda_i Q_i \hat{C} - \hat{C} \|$$

với điều kiện

$$\sum_{i=1}^{k} \lambda_i = 1, \text{và } \lambda_i \geq 0, \forall \text{i}$$

Ở đây $|.||$ là chuẩn Vector. Trường hợp đặc biệt, nếu chọn $\|.\|_\infty$, chúng ta có bài toán cực tiểu sau:

$$\min_\lambda \max_l |[\sum_{i=1}^{k} \lambda_i Q_i \hat{C} - \hat{C}]_l|$$

với điều kiện

$$\sum_{i=1}^{k} \lambda_i = 1, \text{và } \lambda_i \geq 0, \forall \text{i}$$

Ở đây $[.]_l$ xác định phần tử thứ l của Vector. Vấn đề khó khăn ở đây là việc tối ưu hóa để đảm bảo sự tồn tại của phân phối ổn định C. Tiếp theo, chúng ta xem bài toán cực tiểu ở trên được xây dựng như một bài toán tuyến tính:

$$\min_\lambda \omega$$

với điều kiện

$$\begin{pmatrix} \omega \\ \omega \\ \vdots \\ \omega \end{pmatrix} \geq \hat{C} - [\hat{Q}_1\hat{C} \,|\, \hat{Q}_2\hat{C} \,|\, ... \,|\, \hat{Q}_n\hat{C}] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}$$

$$\begin{pmatrix} \omega \\ \omega \\ \vdots \\ \omega \end{pmatrix} \geq -\hat{C} + [\hat{Q}_1\hat{C} \,|\, \hat{Q}_2\hat{C} \,|\, ... \,|\, \hat{Q}_n\hat{C}] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}$$

$$\omega \geq 0, \sum_{i=1}^{k} \lambda_i = 1, \text{và } \lambda_i \geq 0, \forall \text{i}$$

Chúng ta có thể giải bài toán tuyến tính ở trên và có được tham số λ_i. Thay vì giải một bài toán min-max, chúng ta cũng có thể chọn $\|.\|_1$ và xây dựng bài toán cực tiểu sau đây:

$$\min_{\lambda} \sum_{l=1}^{m} |[\sum_{i=1}^{k} \lambda_i \hat{Q}_i \hat{C} - \hat{C}]_l|$$

với điều kiện

$$\sum_{i=1}^{k} \lambda_i = 1, \text{và } \lambda_i \geq 0, \forall \text{i}$$

Bài toán tuyến tính tương ứng được đưa ra như sau:

$$\min_{\lambda} \sum_{l=1}^{m} \omega_l$$

với điều kiện

$$\begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{pmatrix} \geq \hat{C} - [\hat{Q}_1 \hat{X} \,|\, \hat{Q}_2 \hat{C} \,|\, ... \,|\, \hat{Q}_n \hat{C}] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix}$$

$$\begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{pmatrix} \geq -\hat{C} + [\hat{Q}_1 \hat{C} \,|\, \hat{Q}_2 \hat{C} \,|\, ... \,|\, \hat{Q}_n \hat{C}] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix}$$

$$\omega_i \geq 0, \forall i, \sum_{i=1}^{k} \lambda_i = 1, \text{và } \lambda_i \geq 0, \forall \text{i}$$

Trong việc xây dựng các bài toán tuyến tính ở trên, số lượng các biến là bằng nhau đều bằng k và số lượng điều kiện bằng $(2m+1)$. Sự phức tạp của việc giải các bài toán tuyến tính là việc tính toán $O(k^3 L)$, ở đây n là số biến và L là số bit nhị phân cần thiết để lưu trữ tất cả các dữ liệu (các điều kiện và hàm mục tiêu) [27].

Ví dụ 3.2.1. Xét một dãy C_n có 3 trạng thái $(m=3)$ cho bởi

1, 1, 2, 2, 1, 3, 2, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 1, 2 (3.2.7)

Ước lượng các tham số trong mô hình Markov bậc hai *

Dãy X_n có thể được viết dưới dạng vector

$$X_1 = (1,0,0)^T, \text{X}_2 = (1,0,0)^T, \text{X}_3 = (0,1,0)^T, ..., \text{X}_{20} = (0,1,0)^T$$

Chúng ta xét $k=2$ khi đó từ (3.2.7), chúng ta có các ma trận chuyển tần suất:

$$F^1 = \begin{pmatrix} 1 & 3 & 3 \\ 6 & 1 & 1 \\ 1 & 3 & 0 \end{pmatrix} \qquad F^2 = \begin{pmatrix} 1 & 4 & 1 \\ 3 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix} \tag{3.2.8}$$

Do đó từ (3.2.8), chúng ta có ma trận xác suất chuyển i bước $(i = 1,2)$ như sau:

$$\hat{Q}_1 = \begin{pmatrix} \frac{1}{8} & \frac{3}{7} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{7} & \frac{1}{4} \\ \frac{1}{8} & \frac{3}{7} & 0 \end{pmatrix} \qquad \hat{Q}_2 = \begin{pmatrix} \frac{1}{7} & \frac{4}{7} & \frac{1}{4} \\ \frac{3}{7} & \frac{2}{7} & \frac{3}{4} \\ \frac{3}{7} & \frac{1}{7} & 0 \end{pmatrix}$$

Hơn nữa chúng ta có:

$$\hat{Q}_1\hat{C} = (\frac{13}{35}, \frac{57}{140}, \frac{31}{140})^T$$

và

$$\hat{Q}_2\hat{C} = (\frac{47}{140}, \frac{61}{140}, \frac{8}{35})^T$$

Để ước lượng λ_i chúng ta có thể xét bài toán tối ưu:

$$\min_{\lambda_1,\lambda_2} \omega$$

với điều kiện

$$\begin{cases} \omega \geq \frac{2}{5} - \frac{13}{35}\lambda_1 - \frac{47}{140}\lambda_2 \\ \omega \geq -\frac{2}{5} + \frac{13}{35}\lambda_1 + \frac{47}{140}\lambda_2 \\ \omega \geq \frac{2}{5} - \frac{57}{140}\lambda_1 - \frac{61}{140}\lambda_2 \\ \omega \geq -\frac{2}{5} + \frac{57}{140}\lambda_1 + \frac{61}{140}\lambda_2 \\ \omega \geq \frac{1}{5} - \frac{31}{140}\lambda_1 - \frac{8}{35}\lambda_2 \\ \omega \geq -\frac{1}{5} + \frac{31}{140}\lambda_1 + \frac{8}{35}\lambda_2 \\ \omega \geq 0, \lambda_1 + \lambda_2 = 1, \lambda_1, \lambda_2 \geq 0. \end{cases}$$

Nghiệm tối ưu

$$(\lambda_1^*, \lambda_2^*, \omega^*) = (1, 0, 0.0286)$$

và chúng ta có mô hình

$$C_{n+1} = \hat{Q}_1 C_n$$

Lưu ý rằng nếu chúng ta không cho λ_1 và λ_2 không âm, các nghiệm tối ưu trở thành:

$$(\lambda_1^{**}, \lambda_2^{**}, \omega^{**}) = (1.80, -0.80, 0.0157)$$

Mô hình tương ứng

$$C_{n+1} = 1.80\hat{Q}_1 C_n - 0.80\hat{Q}_2 C_{n-1} \tag{3.2.9}$$

Mặc dù ω^{**} nhỏ hơn ω^*, nhưng mô hình (3.2.9) là không phù hợp.

Dễ dàng để kiểm tra:

$$1.80\hat{Q}_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - 0.80\hat{Q}_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.2321 \\ 1.1214 \\ 0.1107 \end{pmatrix}$$

Do đó λ_1^{**}và λ_2^{**} không là tham số hợp lệ.

Chúng ta lưu ý rằng nếu chúng ta xét bài toán cực tiểu:

$$\min_{\lambda_1, \lambda_2} (\omega_1 + \omega_2 + \omega_3)$$

với điều kiện

$$\begin{cases} \omega_1 \geq \frac{2}{5} - \frac{13}{35}\lambda_1 - \frac{47}{140}\lambda_2 \\ \omega_1 \geq -\frac{2}{5} + \frac{13}{35}\lambda_1 + \frac{47}{140}\lambda_2 \\ \omega_2 \geq \frac{2}{5} - \frac{57}{140}\lambda_1 - \frac{61}{140}\lambda_2 \\ \omega_2 \geq -\frac{2}{5} + \frac{57}{140}\lambda_1 + \frac{61}{140}\lambda_2 \\ \omega_3 \geq \frac{1}{5} - \frac{31}{140}\lambda_1 - \frac{8}{35}\lambda_2 \\ \omega_3 \geq -\frac{1}{5} + \frac{31}{140}\lambda_1 + \frac{8}{35}\lambda_2 \\ \omega_1, \omega_2, \omega_3 \geq 0, \lambda_1 + \lambda_2 = 1, \lambda_1, \lambda_2 \geq 0. \end{cases}$$

Nghiệm tối ưu giống như nghiệm tối ưu của bài toán min-max xây dựng trước và kết quả là:

$$(\lambda_1^*, \lambda_2^*, \omega_1^*, \omega_2^*, \omega_3^*) = (1, 0, 0.0286, 0.0071, 0.0214)$$

3.3. Lựa chọn chuỗi thời gian mờ trong mô hình kết hợp

Xét chuỗi thời gian có các quan sát $x_1, x_2, \cdots, x_T, \cdots$ với chuỗi tăng trưởng $y_1, y_2, \cdots, y_T, \cdots$ (được định nghĩa ngay ở ngay mục dưới đây). Ta muốn phân loại mức độ tăng trưởng thành những trạng thái khác nhau như "chậm", "bình thường", "nhanh" hay thậm chí nhiều mức độ hơn nữa. Tuy nhiên, mỗi y_t tại thời điểm t sẽ không rõ ràng thuộc mức độ nào cho dù ta định nghĩa rõ các mức độ. Nghĩa là, y_t có thể vừa thuộc mức độ này vừa thuộc mức độ khác với độ rõ ràng (membership) khác nhau. Chính vì vậy, lý thuyết chuỗi thời gian mờ ở mục 1.5 chương 1 có thể thực hiện điều này nhằm phân lớp tập nền của y_t (định nghĩa ở mục sau) thành các trạng thái mà các y_t là thành viên. Giả sử rằng các trạng thái này tuân theo một xích Markov chính quy (mục 1.3.3) thì mô hình Markov cho ta kết quả dự báo trạng thái tương lai. Từ trạng thái tương lai, giá trị dự báo của x_t được tính ngược từ định nghĩa chuỗi thời giam mờ trước đó.

3.3.1. Định nghĩa và phân vùng tập nền

Xét tập huấn luyện của $\{y_t\}_{t=1}^N$, ta có thể định nghĩa tập nền cho không gian tăng trưởng bởi

$$U = \left(\min_{t\in\{1,\ldots,N\}} y_t - \delta; \max_{t\in\{1,\ldots,N\}} y_t + \delta\right)$$

với $\delta > 0$ là một số dương được lựa chọn sao cho mức tăng trưởng trong tương lai không vượt quá được $\max_{t\in\{1,\ldots,N\}} y_t + \delta$. Tùy từng dữ liệu có thể chọn δ khác nhau. Tuy nhiên, chọn $\delta = 1$ thõa mãn cho mọi dãy tăng trưởng chứng khoán.

Để mờ hóa tập U thành các nhãn tăng trưởng như "tăng nhanh", "tăng chậm", "tăng đều", hoặc thậm chí k mức độ, tập nền U được chia thành k khoảng

(đơn giản nhất là chia thành các khoảng bằng nhau liên tiếp) $u_1, u_2, ..., u_k$. Ví dụ, nếu phân vùng của chỉ số VN-Index (chỉ số chứng khoán Việt Nam) là:

$$U = [-0.0449, -0.0150] \cup [-0.0150, 0.0149] \cup [0.0149, 0.0448]$$

thì các kết quả VN-Index được mã hóa như trong Bảng 3.3.1

Bảng 3.3.1. *Mờ hóa chuỗi tăng trưởng*

Ngày	x_i	chỉ số	tăng trưởng (y_i)	mã hóa
04/11/2009	537,5	-0,015997	NA	NA
05/11/2009	555,5	-0,031866	0,0334883	3
06/11/2009	554,9	-0,026580	-0,0010801	2
09/11/2009	534,1	0,054237	-0,0374842	1
10/11/2009	524,4	0,020036	-0,0181613	1
11/11/2009	537,6	0,002917	0,0251716	3
...	...	...	...	...

3.3.2. Quy luật mờ của chuỗi thời gian

Bây giờ ta xác định các tập mờ A_i, mỗi tập A_i gán cho một nhãn tăng trưởng và xác định trên các đoạn đã xác định $u_1, u_2, ..., u_k$. Khi đó các tập mờ A_i có thể biểu diễn như sau:

$$A_i = \mu_{Ai}(u_1)/u_1 + \mu_{Ai}(u_2)/u_2 + ... + \mu_{Ai}(u_k)/u_k$$

trong đó μ_{Ai} là hàm thành viên của mỗi $u_j, j = 1, ..., k$ trong $A_i, i = 1, ..., k$.

Mỗi giá trị mờ của chuỗi thời gian y_t được tính rõ lại dựa vào quy luật mờ hóa μ_{Ai}. Chẳng hạn như cách mờ hóa sau:

$$A_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + ... + 0/u_k$$

$$A_2 = 0.5/u_1 + 1/u_2 + 0.5/u_3 + ... + 0/u_k$$

...

$$A_k = 0/u_1 + 0/u_2 + 0/u_3 + ... + 1/u_k.$$

Khi đó với $y_t \in A_2$ là một giá trị chưa rõ, thì giá trị rõ được tính ngược theo quy luật mờ này bởi:

$$y_t = \frac{1}{2}\{0.5m_1 + m_2 + 0.5m_3\},$$

trong đó m_1, m_2, m_3 lần lượt là trung điểm của đoạn u_1, u_2, u_3.

Đối với các quy luật mờ hóa khác nhau thì quy tắc tính ngược cũng khác nhau.

3.4. Mô hình kết hợp xích Markov và chuỗi thời gian mờ

3.4.1. Mô hình kết hợp với xích Markov bậc nhất

Trong phần này, mô tả chi tiết việc kết hợp mô hình Markov- chuỗi thời gian mờ. Việc kết hợp này được minh họa trong Hình 3.4.1. Chi tiết của từng bước được thể hiện như sau:

Hình 3.4.1. *Cấu trúc của mô hình Markov- chuỗi thời gian mờ*

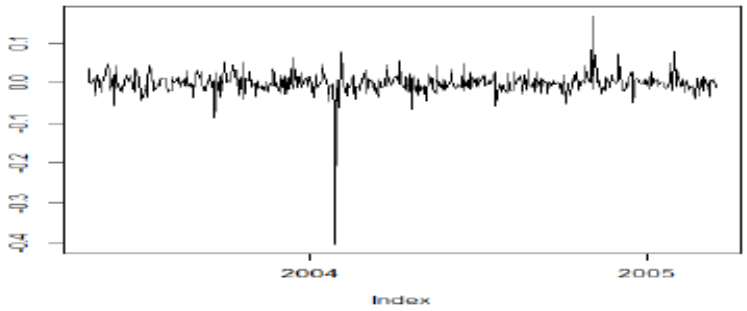
Bước 1:

Cho dữ liệu quan sát của một chuỗi thời gian $\{x_1, x_2, ..., x_t\}$ chuỗi tăng trưởng của dữ liệu huấn luyện được tính như sau:

$$y_t = \frac{x_{t+1} - x_t}{x_t},$$

Ta có $$x_{t+1} = (1 + y_t).x_t$$

Một số dữ liệu có sự thay đổi lớn (giá trị ngoại lai) không có vai trò quan trọng trong dự báo (Hình 3.4.2). Nó không đại diện cho dữ liệu, nhưng nó lại là nguyên nhân gây nên sự thiếu chính xác của mô hình. Do đó, việc làm trước tiên là chúng ta phải loại bỏ giá trị ngoại lai đi.



Hình 3.4.2. *Chuỗi tăng trưởng của Ryanair Airlines data*

Cho D_{max} và D_{min} là giá trị lớn nhất và giá trị nhỏ nhất chuỗi tăng trưởng sau khi bỏ đi giá trị ngoại lai , khi đó tập nền $U = [D_{min} - \delta, D_{max} + \delta]$ ở đó $\delta > 0$ có thể được thiết lập như một ngưỡng cho sự gia tăng của những thay đổi .

Bước 2:

Phân vùng tập nền theo cách đơn giản nhất là chia khoảng $[D_{max}, D_{min}]$ thành $k-2$ khoảng bằng nhau. Khi đó tập nền $U = u_1 \cup u_2 \cup ... \cup u_k$ trong đó $u_1 = [D_{min} - \delta, D_{min}]$ và $u_k = [D_{max}, D_{max} + \delta]$. Ví dụ minh họa như trong 3.3.1

Bước 3:

Như đã trình bày trong mục 3.3.2 ở trên, các tập mờ $A_1, A_2, A_3, ..., A_k$ của chuỗi thời gian được định nghĩa một cách đơn giản như sau:

$$A_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + ... + 0/u_k$$

$$A_2 = 0.5/u_1 + 1/u_2 + 0.5/u_3 + ... + 0/u_k$$

...

$$A_k = 0/u_1 + 0/u_2 + 0/u_3 + ... + 1/u_k$$

Sau đó mỗi A_i được mã hóa bởi i với $i \in \{1, 2, ..., k\}$. Vì vậy, một dữ liệu của chuỗi thời gian thuộc về u_i, nó được mã hóa bởi i ($i \in \{1, 2, ..., k\}$). Chúng ta có được một chuỗi thời gian mã hóa $\{c_t\}_{t=1}^{T}, c_t \in \{1, 2, ..., k\}$.

Bước 4:

Bước này giải thích làm thế nào các chuỗi Markov được áp dụng trong các

chuỗi thời gian mã hóa. Theo phần 3.2, chúng ta giả sử rằng chuỗi thời gian mã hóa $\{c_t\}$ là một chuỗi Markov như trong Định nghĩa 1.3.1. Ước lượng tham số của xích Markov như Mục 1.3.3, ta dễ dàng ước lượng được ma trận xác suất chuyển $\mathbf{\Gamma} = [\gamma_{ij}], i, j \in \{1, 2, ..., k\}$, trong đó:

$$\gamma_{ij} = Pr(c_{t+1} = j \mid c_t = i)$$

Trường hợp nếu tồn tại trạng thái $c_t = i$ là trạng thái hấp thụ (xem 1.3.1), để đảm bảo tính chính quy của $\mathbf{\Gamma}$ quy ước $Pr(c_{t+1} = j \mid c_t = i) = \frac{1}{k}$ với mọi $j = 1, 2, ..., k.$ Nghĩa là, xác suất chuyển từ i sang trạng thái bất kỳ là như nhau.

Bước 5:

Chúng ta dự báo một bước về phía trước cho chuỗi thời gian mã hóa và từ đó xác định giá trị dự báo. Cho c_t, cột $\mathbf{\Gamma}[, c_t]$ là phân phối xác suất của $c_{t+1} = j, j = 1, 2, ..., k$. Gọi

$$M = (\frac{2}{3}(m_1 + 0.5m_2), \frac{1}{2}(0.5m_1 + m_2 + 0,5m_3), \cdots, \frac{2}{3}(m_{k-1} + 0.5m_k))$$

trong đó m_i là giá trị trung bình của khoảng u_i khi đó kết quả dự báo ở thời điểm $t+1$ được tính như sau:

$$\hat{y}_{t+1} = \mathbf{\Gamma}[, c_t] * M = \sum_{j=1}^{k} a_{jc_t} m_j$$

Ở bước này, vectơ M có thể được chọn khác nhau tùy theo phương án mờ hóa ở Bước 2.

Cuối cùng, giá trị x dự báo được tính như sau:

$$\hat{x}_{t+1} = (\hat{y}_t + 1) * x_t$$

3.4.2. Mở rộng với xích Markov bậc cao

Mô hình kết hợp xích Markov bậc cao với chuỗi thời gian mờ chỉ khác mô hình xích Markov bậc một ở **Bước 4** và **Bước 5**.

Bước 1:

Cho dữ liệu quan sát của một chuỗi thời gian $\{x_1, x_2, ..., x_t\}$ chuỗi tăng trưởng của dữ liệu huấn luyện được tính như sau:

$$y_t = \frac{x_{t+1} - x_t}{x_t},$$

Ta có

$$x_{t+1} = (1 + y_t).x_t$$

Một số dữ liệu có sự thay đổi lớn (giá trị ngoại lai) không có vai trò quan trọng trong dự báo (Hình 3.4.2). Nó không đại diện cho dữ liệu, nhưng nó lại là nguyên nhân gây nên sự thiếu chính xác của mô hình. Do đó, việc làm trước tiên là chúng ta phải loại bỏ giá trị ngoại lai đi.

Cho D_{max} và D_{min} là giá trị lớn nhất và giá trị nhỏ nhất chuỗi tăng trưởng sau khi bỏ đi giá trị ngoại lai , khi đó tập nền $U = [D_{min} - \delta, D_{max} + \delta]$ ở đó $\delta > 0$ có thể được thiết lập như một ngưỡng cho sự gia tăng của những thay đổi .

Bước 2:

Phân vùng tập nền theo cách đơn giản nhất là chia khoảng $[D_{max}, D_{min}]$ thành $k-2$ khoảng bằng nhau. Khi đó tập nền $U = u_1 \cup u_2 \cup ... \cup u_k$ trong đó $u_1 = [D_{min} - \delta, D_{min}]$ và $u_k = [D_{max}, D_{max} + \delta]$. Ví dụ minh họa như trong 3.3.1

Bước 3:

Như đã trình bày trong mục 3.3.2 ở trên, các tập mờ $A_1, A_2, A_3, ..., A_k$ của chuỗi thời gian được định nghĩa một cách đơn giản như sau:

$$A_1 = 1/u_1 + 0.5/u_2 + 0/u_3 + ... + 0/u_k$$

$$A_2 = 0.5/u_1 + 1/u_2 + 0.5/u_3 + ... + 0/u_k$$

...

$$A_k = 0/u_1 + 0/u_2 + 0/u_3 + ... + 1/u_k$$

Sau đó mỗi A_i được mã hóa bởi i với $i \in \{1,2,...,k\}$. Vì vậy, một dữ liệu của chuỗi thời gian thuộc về u_i, nó được mã hóa bởi i ($i \in \{1,2,...,k\}$). Chúng ta có được một chuỗi thời gian mã hóa $\{c_t\}_{t=1}^{T}, c_t \in \{1,2,...,k\}$.

Bước 4:

Đối với mô hình Markov bậc cao cổ điển kết hợp với chuỗi thời gian mờ (gọi là CMC-Fuz), bằng cách cực đại hoá tương tự trong mô hình Markov bậc nhất, ta dễ dàng ước lượng ma trận xác suất chuyển $l+1$ chiều $\mathbf{\Gamma} = [\gamma_{i_{l+1} i_l ... i_1}], i_j \in \{1,2,...,k\}$. Theo nghĩa của xích Markov bậc cao, $\gamma_{i_{l+1} i_l ... i_1}$ là xác suất quan sát được c_{t+1} với điều kiện đã biết $c_t,...,c_{t-l+1}$:

$$\gamma_{i_{l+1} i_l ... i_1} = Pr(c_{t+1} = i_{l+1} \mid c_t = i_l,...,c_{t-l+1} = i_1)$$

Đối với mô hình Markov bậc cao mới kết hợp (gọi là IMC-Fuz), ma trận chuyển $m \times m\, \Gamma = \sum_{i=1}^{l} \lambda_i Q_i$ như trong (3.2.4).

Bước 5:

Tiếp theo ta tạo ra dự báo một bước cho chuỗi thời gian mã hoá dựa vào ma trận xác suất chuyển và tính ngược lại giá trị dự báo của chuỗi thời gian gốc.

Đối với mô hình CMC-Fuz, cho trước $c_t,...,c_{t-l+1}$, cột $\Gamma[,c_t,...,c_{t-l+1}]$ là phân bố xác suất của $c_{t+1} = j$ trên khắp k giá trị mã hoá $j = 1,2,...,k$. Giá trị tăng trưởng dự báo tại thời điểm $t+1$ khi đó được tính bởi:

$$\hat{y}_{t+1} = \Gamma[,c_t,...,c_{t-l+1}] * M = \sum_{j=1}^{k} \gamma_{j c_t ... c_{t-l+1}} m_j$$

Đối vơi IMC-Fuz, Giá trị tăng trưởng dự báo tại thời điểm $t+1$ được tính bởi:

$$\hat{y}_{t+1} = \sum_{i=1}^{l} \lambda_i Q_i[,c_{t-i+1}]$$

Cuối cùng, giá trị x_{t+1} dự báo được tính bởi:

$$\hat{x}_{t+1} = (\hat{y}_t + 1) * x_t$$

Mã giả của mô hình được minh hoạ bởi thuật toán Thuật toán 3.1. Trong thuật toán này, tham số đầu vào của mô hình bao gồm dữ liệu quan sát *Data*, phân phối dừng ban đầu *delta = 1*, còn lại các tham số *nTrain, nOrder, nStates* lần lượt là số lượng quan sát trong tập đào tạo, số bậc của xích Markov trong mô hình và số trạng thái được chia ra tương ứng với số tập mờ. Tham số đầu ra bao gồm các giá trị dự báo *predict*, các tiêu chuẩn đánh giá độ chính xác phổ biến bao gồm *RMSE,MAPE,MAE*. Trong đó, *nTrain* là số quan sát trong tập huấn luyện; *nOrder* là bậc của xích Markov bậc cao và *nStates* là số trạng thái (các A_k) của mô hình.

Như vậy, mô hình CMC-Fuz và IMC-Fuz với bậc $nOrder = 1$ trùng với mô hình kết hợp bậc 1 như trong mục 3.4.1. Do đó, các kết quả thực nghiệm cho mô hình xích Markov bậc nhất thực hiện đồng thời trong mô hình xích Markov bậc cao.

Thuật toán 3.1 Thuật toán Markov - Fuzzy kết hợp

Đầu vào: $Data, \delta = 1, nTrain, nOrder, nStates$

Đầu ra: *predict*, RMSE, MAPE, MAE

1: $y_t \leftarrow \frac{Data_{t+1} - Data_t}{Data_t}, t = 2, ..., nTrain$

2: *Train* <- Bỏ phần tử ngoại lai của y_t

3: Chia khoản $[\min(Train) - \delta; \max(Train) + \delta]$ thành *nStates* khoảng bằng nhau *Ak*

4: **if** x_t in $A_k c$ **then** $encoded_t \leftarrow k$

6: **if** *Model* = CMC-Fuz **then** Ước lượng ma trận chuyển của mô hình CMC_Fuz.

7: **for** i in 1:nOrder **do** Ước lượng ma trận Q_i

8: **if** *Model* = IMC-Fuz **then**

9: $C \leftarrow counts(encoded) / sum(counts)$

10: Giải bài toán tối ưu min-max $\min_{\lambda} \max_{k} \left\| \left[\sum_{i=1}^{n} \lambda_i Q_i C - C \right] \right\| for \lambda_i$

12: $IMC.Fuz1.Mat \leftarrow \sum_{i=1}^{nOrder} \lambda_i Q_i$ Ước lượng ma trận chuyển của IMC-Fuz dựa trên phân phối dừng

13: **for** $close_t$ in $testset$ **do**

14: **if** $close_t$ in A_k **then** $encoded_{t-1} \leftarrow k$ { mã hoá quan sát mới, $t > nTrain$}

15: $M \leftarrow vector(2/3(mid(A_1) + 0.5mid(A_2)), 1/2(0.5mid(A_1) + mid(A_2)$

$+0,5mid(A_3)), ..., 2/3(0.5mid(A_{k-1}) + mid(A_k)))$ {tính ngược quy luật mờ}

16:

$predict_t \leftarrow (transition.Mats[, encoded_{t-1}, encoded_{t-2}, ..., encoded_{t-nOrder+1}]\%*\%M + 1) * Data_t$

17: *errors* (RMSE, MAPE, MAE) $\leftarrow f(predict_t - actual_t)$ {tính toán độ đo độ chính xác}.

18: **return** *predict*,*RMSE*,*MAPE*,*MAE*

3.4.3. Kết quả thực nghiệm

Lựa chọn dữ liệu

Nhằm so sánh kết quả với [33, 34, 25, 51, 64, 58], ta sử dụng dữ liệu tương tự lấy trong [68, 54, 4, 63]. Hơn nữa, nhiều dữ liệu khác nhau cũng được sử dụng để kiểm tra độc chính xác của mô hình. Chi tiết cho trong bảng 3.4.1

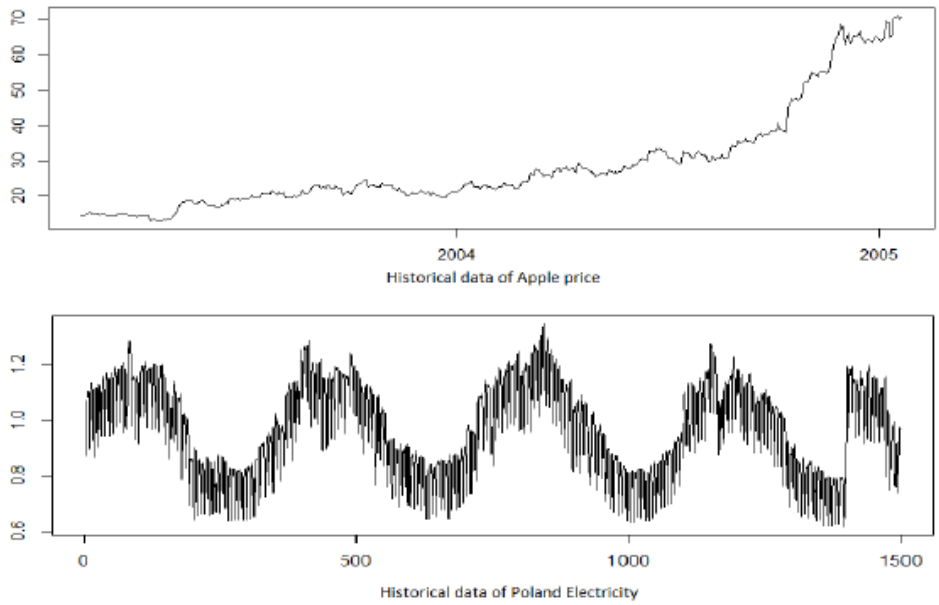
Bảng 3.4.1. *Các tập dữ liệu so sánh*

Tên dữ liệu	từ ngay	đến ngay	tần suất
Apple Computer Inc.	10/01/2003	21/01/2005	Daily
IBM Corporation	10/01/2003	21/01/2005	Daily
Dell Inc.	10/01/2003	21/01/2005	Daily
Ryanair Airlines	06/01/2003	17/01/2005	Daily
TAIEX (Taiwan exchange index)	01/01/2001	31/12/2009	Daily

SSE(Shanghai Stock Exchange)	21/06/2006	31/12/2012	Daily
DJIA(Dow Jones Industrial Average Index)	04/08/2006	31/08/2012	Daily
S&P500	04/08/2006	31/08/2012	Daily
Unemployment rate	01/01/1948	01/12/2013	Monthly
Australian electricity	01/01/1956	01/08/1995	Monthly
Poland Electricity Load From	1990's	1500 values	Daily

Nghiên cứu này không cố định tập huấng luyện và tập test và do đó cho phép độc giả thay đổi phù hợp khi áp dụng vào dữ liệu cụ thể. Trong nhiều trường hợp, kết quả thực nghiệm cho thấy rằng dữ liệu huấn luyện vào khoảng 75% đến 85% cho kết quả dự báo tốt nhất.

Hình 3.4.3 minh hoạ dữ liệu lịch sử của chỉ số cổ phiếu Apple và lượng tiêu thụ điện của Ba Lan. Từ hình ảnh cho thấy rõ ràng dữ liệu sử dụng điện có tính mùa vụ, tức lặp lại theo chu kỳ ở mức độ nào đó. Do vậy, về trực quan mô hình Markov bậc cao có thể cho kết quả tốt hơn bậc 1 thông thường.



Hình 3.4.3. *Chuỗi giá cổ phiếu lịch sử của Apple và chỉ số thiêu thụ điện của Ba Lan*

Kết quả so sánh với các mô hình khác

Độ đo tính chính xác của mô hình trong nghiên cứu này là trung bình phần

trăm sai số (MAPE), căn bậc hai trung bình bình phương sai số (RMSE) và trung bình sai số (MAE). Công thức được cho bởi 3.4.1.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - p_i}{a_i}\right| * 100\%;$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{n}};$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i|. \qquad (3.4.1)$$

trong đó n số các giá trị cần test, a_i và p_i tương ứng là giá trị thực tế và giá trị dự báo của ngày thứ i của tập kiểm tra.

Mô hình đầu tiên được so sánh là mô hình được đề cập trong [33]. Tập huấn luyện và tập test của các dữ liệu Apple inc., Dell comp., IBM cor., Ryanair Airlines được sử dụng hoàn toàn tương tự (nTrain = 400). British Airlines và Delta Airlines không được so sánh do cơ sở dữ liệu trên http://finance.yahoo.com//. không đầy đủ tương ứng với [33].

Bảng 3.4.2. *So sánh MAPEs cho các mô hình khác nhau.*

Stock	HMM-based forecasting model	Fusion HMM-ANN-GA with weighted average (MAPE)	Combination of HMM-fuzzy model(MAPE)	CMC-Fuz model nStates =6 nOrder =1	IMC-Fuz model nStates =6 nOrder =2
Ryanair Air.	1,928	1,377	1,356	**1,275**	**1,271**
Apple	2,837	1,925	1,796	**1,783**	**1,783**
IBM	1,219	0,849	0,779	**0,660**	**0,656**
Dell Inc.	1,012	0,699	0,405	**0,837**	**0,823**

Từ Bảng 3.4.2, cùng với nStates =6, ta có thể thấy mô hình IMC-Fuz với nOrder = 1 tốt hơn mô hình CMC-Fuz với nOrder = 1. Cả hai mô hình tốt hơn các mô hình được so sánh với 4 dữ liệu như trong [33].

Một mô hình HMM khác thực hiện dự báo chỉ số đóng cửa của chỉ số chứng khoán được thực hiện bởi Gupta [30] cho thấy độ chính xác cao hơn của Hassan [33]. Tuy nhiên, mô hình của Gupta sử dụng chỉ số cổ phiếu trong ngày gồm giá mở cửa, giá cao nhất, giá thấp nhất để dự báo giá đóng cửa trong khi Hassan cũng như trong luận án này chỉ sử dụng giá đóng cửa của những ngày trước đó dự báo cho ngày tiếp theo. Do đó, việc so sánh trong mô hình của Gupta không trên cùng một dữ liệu mặc dù cùng cơ sở dữ liệu. Hơn nữa, việc sử dụng các giá trị trong ngày để dự báo chính giá trị trong ngày đó bao giờ cũng có độ chính xác cao hơn do độ dao động thấp hơn. Tuy nhiên, điều này không phù hợp với thực tế trong giao dịch mua bán cổ phiếu.

Mô hình thứ hai được so sánh là các mô hình trong [64], trong đó mạng nơ-ron thời gian ngẫu nhiên (STNN) được kết hợp với thành phần phân tích chính (PCA) nhằm so sánh với mạng nơ-ron cổ điển (BPNN), PCA-BPNN, STNN và vector học máy (SVM). Các mô hình này thực hiện đánh giá dự báo cho các chỉ số chứng khoán SSE, S&P500 và DJIA trong Bảng 3.4.1. Tất cả các mô hình sử dụng 1300 dữ liệu huấn luyện và phần còn lại sử dụng cho kiểm chứng. Mô hình chứng tôi xây dựng sử dụng 6 trạng thái và bậc 2 cho xích Markov. Kết quả so sánh của mô hình IMC-Fuz và CMC-Fuz chỉ ra trong Bảng 3.4.3 có tốt hơn với các mô hình khác cho dữ liệu SSE và tốt hơn rất nhiều cho dữ liệu DJIA và S&P500.

Bảng 3.4.3. *So sánh các mô hình khác nhau cho dữ liệu SSE, DJIA và S\&P500*

Dữ liệu	Độ đo	IMC-Fuz	CMC-Fuz	BPNN	STNN	SVM	PCA-BPNN	PCA-STNN
SSE								
	MAE	20,5491	20,4779	24,4385	22,8295	27,8603	22,4485	22,0844
	RMSE	27,4959	27,4319	30,8244	29,0678	34,5075	28,6826	28,2975
	MAPE	0,8750	0,8717	1,0579	0,9865	1,2190	0,9691	0,9540
DJIA								
	MAE	90,1385	90,4159	258,4801	230,7871	278,2667	220,9163	192,1769
	RMSE	123,2051	123,2051	286,6511	258,3063	302,793	250,4738	220,4365

S&P500	MAPE	0,7304	0,7304	2,0348	1,8193	2,2677	1,7404	1,5183
	MAE	10,4387	10,4387	24,7591	22,1833	22,9334	16,8138	15,5181
	RMSE	14,2092	14,2092	28,1231	25,5039	25,9961	20,5378	19,2467
	MAPE	0,8074	0,8074	1,8607	1,6725	1,7722	1,282	1,1872

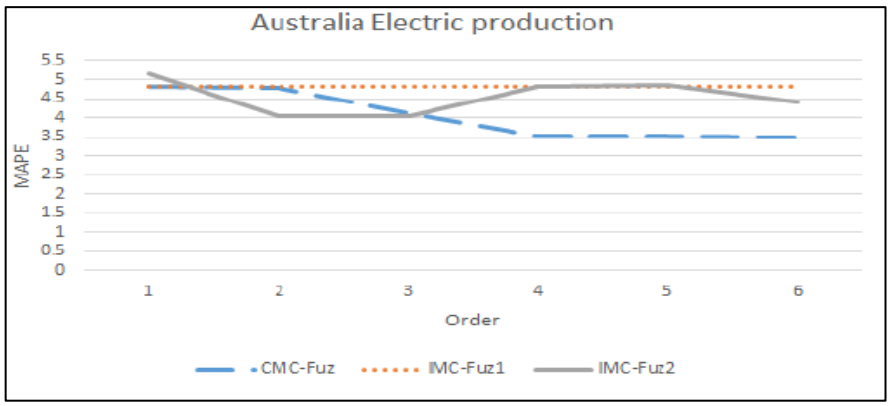
Trong công trình mới đây [58], các tác giả đã đề xuất một mô hình dự báo thời gian mờ mới và so sánh với các phương pháp khác nhau trong dự báo chỉ số TAIEX từ 2001 đến 2009. Dữ liệu từ tháng Một đến tháng Mười của mỗi năm sử dụng làm dữ liệu huấn luyện và phần còn lại từ tháng 11 đến tháng 12 để dự báo và tính độ chính xác. Bảng 3.4.4 chỉ ra rằng mô hình của chúng tôi với nStates = 6 và nOrder =1,2 tốt hơn tất cả các mô hình được đề cập.

Bảng 3.4.4. *So sánh RMSEs của TAIEX cho các năm từ 2001 đến 2009 nStates = 6*

Method	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**Average**
Chen 1996[15]	104,25	119,33	68,06	73,64	60,71	64,32	171,62	310,52	92,75	118,36
ARIMA	97,43	121,23	71,23	70,23	58,32	64,43	169,33	306,11	94,39	116,97
Yu 2005[70]	100,54	119,33	65,35	71,50	57,00	63,18	168,76	310,09	91,32	116,34
ETS	96,80	119,43	68,01	72,33	54,70	63,72	165,04	303,39	95,60	115,45
Yu 2005 [70]	98,69	119,18	63,66	70,88	54,69	60,87	167,69	308,40	89,78	114,87
Huarng 2006[39]	97,86	116,85	61,32	70,22	52,36	58,37	167,69	306,07	87,45	113,13
Chen 2011[16]	96,39	114,08	61,38	66,75	52,18	55,83	165,48	304,35	85,06	111,28
ARFIMA	95,18	115,13	59,43	58,47	50,78	51,23	163,77	315,17	89,23	110,93
Javedani 2014 [57]	94,80	111,70	59,00	64,10	49,80	55,30	163,10	301,70	84,80	109,37
Sadaei2016 [58]	89,47	104,37	49,67	59,43	37,80	47,30	154,43	294,37	78,80	101,74
Sadaei2016 [58]	86,67	101,62	45,04	55,80	34,91	45,14	152,88	293,96	74,98	99,00
IMC-Fuz										
Order=1	**117,73**	**68,44**	**55,96**	**56,58**	**55,97**	**51,87**	**159,36**	**106,9**	**71,51**	**82,7**
Order=2	**115,75**	**67,5**	**53,75**	**56,58**	**55,97**	**51,73**	**159,36**	**105,12**	**71,51**	**81,92**
CMC-Fuz										
Order 1	**116,52**	**68,45**	**55,97**	**56,58**	**55,97**	**51,87**	**159,37**	**106,9**	**71,51**	**82,57**
Order 2	**119,42**	**71,51**	**54,81**	**56,93**	**60,12**	**53,57**	**164,32**	**106,97**	**82,03**	**85,52**

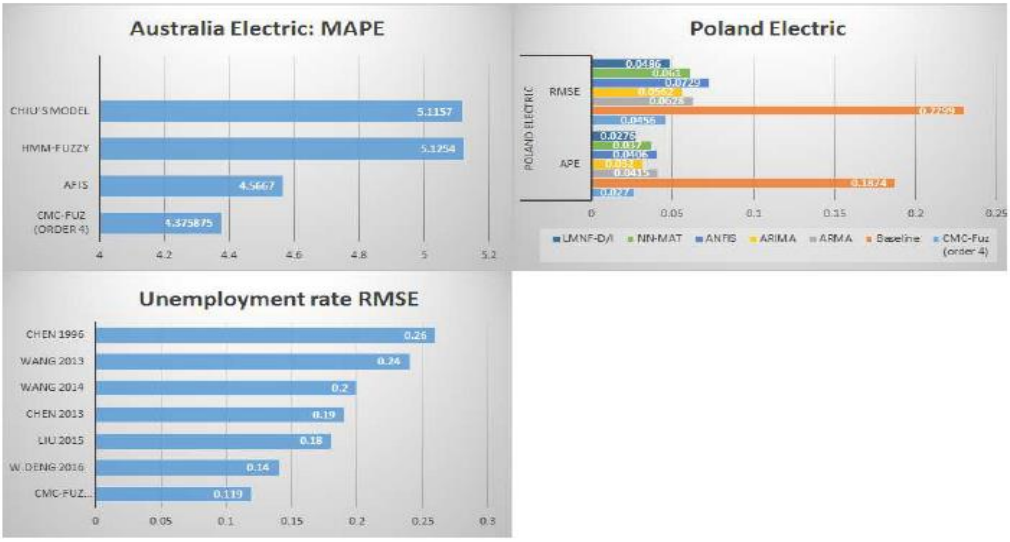
Cuối cùng, mô hình chúng tôi đề xuất được so sánh với các mô hình khác đối với các dữ liệu có tính mùa vụ như lượng tiêu thụ điện hay tỉ lệ thất nghiệp. Mô

hình CMC-Fuz cho kết quả tốt nhất đối với dữ liệu lại này. Hình 3.4.4 chỉ ra MAPE của dữ liệu tiêu thụ điện của Australia với 1000 dữ liệu huấn luyện và 500 dữ liệu còn lại cho kiểm tra, nStates = 4 đối với tất cả các bậc. Kết quả cho thấy rằng mô hình CMC-Fuz dự báo chính xác hơn tất cả các mô hình.



Hình 3.4.4. *MAPEs của dữ liệu tiêu thụ điện của Australia với các bậc khác nhau của mô hình đề xuất*

Hình 3.4.5 minh hoạ sự so sánh giữa mô hình CMC-Fuz các mô hình khác mới đây cho dự báo sản lượng tiêu thụ điện và tỉ lệ thất nghiệp. Tập huấn luyện và tập test là hoàn toàn giống nhau đối với tất cả các mô hình. Mô hình đề xuất sử dụng 7 trạng thái với 4 bậc cho xích Markov. Ta có thể thấy mô hình CMC-Fuz tốt hơn tất cả các mô hinh đề cập đến trong [34](nTrain = 200), và trong [51](nTrain=1000), thậm chí cả trong [25](nTrain=780).



Hình 3.4.5. *So sánh mô hình CMC-Fuz (7states, 4 bậc) và một số mô hình gần đây*

Từ các kết quả so sánh trên thấy rằng mô hình mà nghiên cứu sinh đề xuất không chỉ tốt hơn tất cả các mô hình đề cập đến mà còn mở ra một hướng mới trong việc phát triển các công cụ dự báo hiệu quả hơn.

3.5. Kết luận

Chương này luận án trình bày mô hình kết hợp xích Markov (cả bậc 1 và bậc cao) và chuỗi thời gian mờ trong dự báo chuỗi thời gian.

Thứ nhất, đề xuất được phương pháp mờ hóa chuỗi thời gian mà các tập mờ trở thành những trạng thái của một xích Markov. Sau khi xích Markov dự báo trạng thái, quy tắc tính ngược từ các tập mờ cho kết quả dự báo của chuỗi thời gian.

Thứ hai, mở rộng mô hình cho xích Markov bậc cao cổ điển và xích Markov bậc cao cải tiến tương ứng với các thuật toán ước lượng tham số của xích Markov bậc cao.

Thứ ba, thực hiện thực nghiệm trên cùng một tập đào tào và tập kiểm tra đối với các mô hình dự báo gần đây cho thấy mô hình đề xuất có độ chính xác cao hơn đáng kể mặc dù thuật toán đơn giản hơn. Hơn nữa, mô hình xích Markov bậc cao cho thấy hiệu quả hơn hẳn đối với dữ liệu có tính chất mùa vụ.

Kết quả nghiên cứu của chương này đã được công bố trong bài báo [A3] và [A4].

KẾT LUẬN

Kết quả

Với mục tiêu phát triển mô hình dự báo theo hướng kết hợp các mô hình sẵn có thành mô hình mới nhằm cải thiện độ chính xác dự báo, luận án đã thực hiện được các nội dung nghiên cứu:

Nghiên cứu tổng quan về xích Markov, xích Markov bậc cao và các phương pháp ước lượng tham số của xích Markov. Phân tích các ứng dụng tiềm tàng của xích Markov trong bài toán dự báo chuỗi thời gian. Luận án nhận thấy mô hình chuỗi thời gian mờ trong dự báo chuỗi thời gian khắc phục hạn chế về mặt dữ liệu không rõ ràng của chuỗi thời gian, do đó một số lý thuyết về chuỗi thời gian mờ cũng như một vài thuật toán dự báo sử dụng chuỗi thời gian mờ được khái quát lại. Từ cơ sở những ưu điểm và hạn chế của các mô hình dự báo hiện có, luận án đề xuất mô hình dự báo kết hợp mới cải thiện độ chính xác dự báo.

Nội dung nghiên cứu chuyên sâu của luận án tập trung vào hai nội dung chính:

Thứ nhất, áp dụng mô hình Markov ẩn (HMM) đối với phân phối Poisson và phân phối chuẩn (Normal) cho mô hình dự báo đối với chuỗi thời gian cụ thể dựa trên phân tích về sự tương thích của dữ liệu với mô hình (Mục 2.1). Một loạt các thuật toán được thực hiện và chạy trên dữ liệu thực cho thấy sự hợp lý của dự báo đối với thời gian ngắn hạn.

Thứ hai, để khắc phục nhược điểm của mô hình HMM (dựa vào phân phối xác suất tất định mà phân phối thực nghiệm không tuân theo) và khắc phục tính mờ (không rõ ràng) của dữ liệu chuỗi thời gian, luận án đề xuất mô hình kết hợp xích Markov và chuỗi thời gian mờ trong dự báo chuỗi thời gian. Các thuật toán kết hợp giữa hai mô hình đã được thiết lập và thực nghiệm trên một loạt các dữ liệu so với những mô hình dự báo gần đây cho thấy kết quả dự báo có độ chính xác cải thiện đáng kể. Đặc biệt, mô hình Markov bậc cao kết hợp chuỗi thời gian mờ có tiềm năng lớn áp dụng cho dự báo chuỗi thời gian có tính thời vụ.

Các đóng góp của luận án đều đã được cài đặt và chạy thử nghiệm trên ngôn ngữ lập trình R.

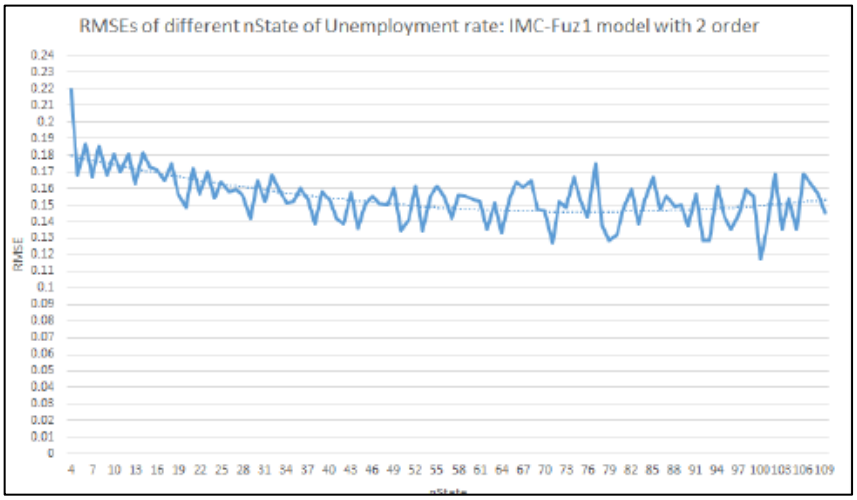
Hướng phát triển của đề tài luận án

Các nội dung nghiên cứu của luận án vẫn có thể tiếp tục được phát triển và hoàn thiện hơn. Cụ thể một số hướng phát triển như sau:

Kết hợp xích Markov với các luật mờ phức tạp hơn nhằm xác định chính xác hơn vai trò của mỗi giá trị trong chuỗi thời gian đối với một tập mờ. Từ đó có thể cải thiện thêm độ chính xác của dự báo.

Mở rộng mô hình cho chuỗi thời gian đa biến, trong đó các chuỗi thời gian thành phần phụ thuộc nhau. Chuỗi thời gian mục tiêu (đối tượng dự báo) liên quan đến các chuỗi khác (chuỗi tác động) theo các trạng thái Markov được xác định trên các chuỗi tác động này. Từ nhiều chuỗi tác động, có thể kết hợp với mô hình ANN để xây dựng được mô hình dự báo có tính đến các yếu tố phụ thộc bên ngoài. Điều này phù hợp với thực tế.

Vấn đề tối ưu hóa các tham số vẫn là một hướng mở. Cụ thể, mô hình luận án đề xuất thực hiện với $nOrder = 2$ và $nStates = 7$ đủ để so sánh với các mô hình khác. Tuy nhiên, chúng chưa phải là tham số tốt nhất (như Hình 3.5.1). Do đó, việc xây dựng một cơ sở suy luận và thuật toán xác định tham số tốt nhất cho mô hình cũng là vấn đề có thể mở rộng nghiên cứu.



Hình 3.5.1. *RMSEs dự báo tỷ lệ thất nghiệp với các nStates khác nhau, nOrder = 2*

Các công trình khoa học của nghiên cứu sinh

[A1] **Đào Xuân Kỳ**, Lục Trí Tuyen, va Phạm Quốc Vương. A combination of higher order markov model and fuzzy time series for stock market forecasting". In *Hội thảo lần thứ 19: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, Hà Nội*, pages 1–6, 2016.

[A2] **Đào Xuân Kỳ,** Lục Trí Tuyen, Phạm Quốc Vương, va Thạch Thị Ninh. Mô hinh markov-chuỗi thời gian mờ trong dự báo chứng khoán. In *Hội thảo lần thứ 18: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, TP HCM*, pages 119–124, 2015.

[A3] **Dao Xuan Ky** and Luc Tri Tuyen. A markov-fuzzy combination model for stock market forecasting. *International Journal of Applied athematics and StatisticsTM*, 55(3):109–121, 2016.

[A4] **Dao Xuan Ky** and Luc Tri Tuyen. A Higher order Markov model for time series forecasting. International Journal of Applied athematics and StatisticsTM, vol 57(3), 2018.

[A5] Lục Trí Tuyen, Nguyễn Văn Hung, Thạch Thị Ninh, Phạm Quốc Vương, Nguyễn Minh Đức, va **Đào Xuân Kỳ.** A normal-hidden markov model model in forecasting stock index. *Journal of Computer Science and Cybernetics*, 28(3):206–216, 2012.

Tài liệu tiếng việt

[B1] Nguyễn Cát Hồ, Điều Nguyễn Công, và Lân Vũ Như. Ứng dụng của đại số gia tử trong dự báo chuỗi thời gian mờ. *Journal of Science and Technology*, 54(2):161, 2016.

[B2] Nguyễn Duy Hiếu, Lân Vũ Như, và Hồ Nguyễn Cát. Dự báo chuỗi thời gian mờ dựa trên ngữ nghĩa. *PROCEEDING of Publishing House for Science and Technology*, 2016.

[B3] Nguyễn Công Điều. Một thuật toán mới cho mô hình chuỗi thời gian mờ heurictic trong dự báo chứng khoán. *Journal of Science and Technology*, 49(4), 2012.

[B4] Nguyễn Công Điều và Tính Nghiêm Văn. Dự báo chuỗi thời gian mờ dựa trên nhóm quan hệ mờ phụ thuộc thời gian và tối ưu bầy đàn. *PROCEEDING of Publishing House for Science and Technology*, 2017.

Tài liệu tiếng anh

[1] Carol Alexander. Normal mixture diffusion with uncertain volatility: Modelling short- and long-term smile effects. *Journal of Banking & Finance*, 28(12):2957– 2980, 2004.

[2] J Scott Armstrong. Combining forecasts. In *Principles of forecasting*, pages 417–439. Springer, 2001.

[3] J Scott Armstrong. Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, 22(3):583–598, 2006.

[4] Monthly australia electricity data. http://datamarket.com/data/set/22l0/monthly-electricity-production-in-australia-millionkilowatt-hours-jan-1956-aug-1995#!display=line&ds=22l0. Accessed: 2016-05-07.

[5] L. E Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3(1):1–8, 1972.

[6] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

[7] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 7(6):1554–1563, 1966.

[8] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[9] L. E. Baum and R. G. Sell. Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, 72(2):211–227, 1968.

[10] Ramaprasad Bhar and Shigeyuki Hamori. *Hidden Markov Models: Applications to Financial Economics*. Advanced Studies in Theoretical and

Applied Econometrics, Volume 40, Springer, 2004.

[11] George EP Box and Gwilym M Jenkins. Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 17(2):91–109, 1968.

[12] Damiano Brigo and Fabio Mercurio. Lognormal-mixture dynamics and calibration to market volatility smiles. *International Journal of Theoretical and Applied Finance*, 5(4):427–451, 2002.

[13] Qisen Cai, Defu Zhang, Wei Zheng, and Stephen CH Leung. A new fuzzy time series forecasting model combined with ant colony optimization and autoregression. *Knowledge-Based Systems*, 74:61–68, 2015.

[14] Li-Juan Cao and Francis Eng Hock Tay. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on neural networks*, 14(6):1506–1518, 2003.

[15] Shyi-Ming Chen. Forecasting enrollments based on fuzzy time series. *Fuzzy sets and systems*, 81(3):311–319, 1996.

[16] Shyi-Ming Chen and Chao-Dian Chen. Handling forecasting problems based on high-order fuzzy logical relationships. *Expert Systems with Applications*, 38(4):3857–3864, 2011.

[17] Shyi-Ming Chen and Jeng-Ren Hwang. Temperature prediction using fuzzy time series. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 30(2):263–275, 2000.

[18] L. W. K. Cheung. Use of runs statistics for pattern recognition in genomic dna sequences. *Journal of Computational Biology*, 11(1):107–124, 2004.

[19] Wai-Ki Ching, Ximin Huang, Michael K Ng, and Tak-Kuen Siu. Higher-order markov chains. In *Markov Chains*, pages 141–176. Springer, 2013.

[20] Wai-Ki Ching, Ximin Huang, Michael K Ng, and Tak-Kuen Siu. Higher-order markov chains. In *Markov Chains*, pages 141–176. Springer, 2013.

[21] Kai Lai Chung. *Markov Chains with Stationary Transition Probabilities: 2d Ed*. Springer, 1967.

[22] D. A. Coast, R.M. Stern, G.G. Cano, and S.A. Briller. An approach to cardiac arhythmia analysis using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 37(9):826–836, 1990.

[23] B.C. Cuong and P.V. Chien. An experiment result based on adaptive neuro-fuzzy inference system for stock price. *Journal of Computer science and cybernetics*, 27(1):51–60, 2011.

[24] E. Demidenko. *Mixed Models: Theory and Applications with R*. Wiley Series in Probability and Statistics. Wiley, 2013.

[25] Weihui Deng, Guoyin Wang, Xuerui Zhang, Ji Xu, and Guangdi Li. A multigranularity combined prediction model based on fuzzy trend forecasting and particle swarm techniques. *Neurocomputing*, 173:1671–1682, 2016.

[26] Eugene F Fama. The behavior of stock-market prices. *The journal of Business*, 38(1):34–105, 1965.

[27] Shu-Cherng Fang and Sarat Puthenpura. *Linear optimization and extensions: theory and algorithms*. Prentice-Hall, Inc., 1993.

[28] William Feller. *An introduction to probability theory and its applications, Vol. 1*. John Wiley, 1957.

[29] P.A. Gagniuc. *Markov Chains: From Theory to Implementation and Experimentation*. Wiley, 2017.

[30] Aditya Gupta and Bhuwan Dhingra. Stock market prediction using hidden markov models. In Engineering and Systems (SCES), 2012 Students Conference on, pages 1–4. IEEE, 2012.

[31] B. Hajek. *Random Processes for Engineers*. Cambridge University Press, 2015.

[32] Li Hang and Kenji Yamanishi. Document classification using a finite mixture model. In *EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Stroudsburg 1997, PA, USA*, pages 39–47, 1997.

[33] Md Rafiul Hassan. A combination of hidden markov model and fuzzy model

for stock market forecasting. *Neurocomputing*, 72(16):3439–3446, 2009.

[34] Md Rafiul Hassan, Kotagiri Ramamohanarao, Joarder Kamruzzaman, Mustafizur Rahman, and M Maruf Hossain. A hmm-based adaptive fuzzy inference system for stock market forecasting. *Neurocomputing*, 104:10–25, 2013.

[35] M.D.R Hassan. A combination of hidden markov model and fuzzy model for stock market forecasting. *Neurocomputing*, 72:3439–3446, 2009.

[36] M.D.R. Hassan and B. Nath. Stock market forecasting using hidden markov model: a new approach. In *Proceedings of 5th international conference on intelligent system design and application, ISDA 2005, Wroclaw, Poland*, pages 192–196, 2005.

[37] Kunhuang Huamg and Hui-Kuang Yu. N-th order heuristic fuzzy time series model for taiex forecasting. *International Journal of Fuzzy Systems*, 5(4):247– 253, 2003.

[38] Kunhuang Huarng. Heuristic models of fuzzy time series for forecasting. *Fuzzy sets and systems*, 123(3):369–386, 2001.

[39] Kunhuang Huarng and Tiffany Hui-Kuang Yu. Ratio-based lengths of intervals to improve fuzzy time series forecasting. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):328–340, 2006.

[40] S. Karlin and H.E. Taylor. *A First Course in Stochastic Processes*. Elsevier Science, 2012.

[41] J Kihoro, R Otieno, and C Wafula. Seasonal time series forecasting: A comparative study of arima and ann models. *AJST*, 5(2), 2004.

[42] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, 1999.

[43] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley, New York, 2000.

[44] C. EMcLaren, I. V. Cadez, P. Smyth, and G. J.McLachlan. Multivariate mixture models for classification of anemias. In *2000 Proceedings of the Biometrics Section of the American Statistical Association, Virginia 2000, USA*, pages 112–117, 2000.

[45] E.W. Montroll, M.F. Shlesinger, and G.H. Weiss. *The Wonderful world of stochastics: a tribute to Elliott W. Montroll*. Studies in statistical mechanics. North-Holland, 1985.

[46] Guofang Nan, Shuaiyin Zhou, Jisong Kou, and Minqiang Li. Heuristic bivariate forecasting model of multi-attribute fuzzy time series based on fuzzy clustering. *International Journal of Information Technology & Decision Making*, 11(01):167–195, 2012.

[47] Norman Owen-Smith, Victoria Goodall, and Paul Fatti. Applying mixture models to derive activity states of large herbivores from movement rates obtained using gps telemetry. *Wildlife Research*, 39(5):452–462, 2012.

[48] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[49] H Park. Forecasting three-month treasury bills using arima and garch models, 1999.

[50] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[51] Hung-Wen Peng, Shen-Fu Wu, Chia-Ching Wei, and Shie-Jue Lee. Time series forecasting with a neuro-fuzzy modeling scheme. *Applied Soft Computing*, 32:481–493, 2015.

[52] Zhihang Peng, Changjun Bao, Yang Zhao, Honggang Yi, Letian Xia, Hao Yu, Hongbing Shen, and Feng Chen. Weighted markov chains for forecasting and analysis in incidence of infectious diseases in jiangsu province, china. *Journal of biomedical research*, 24(3):207–214, 2010.

[53] Roberto Perrelli. Introduction to arch & garch models. *University of Illinois*

Optional TA Handout, pages 1–7, 2001.

[54] Poland electricity load from 1990's. http://research.ics.aalto.fi/eiml/datasets.shtml. Accessed: 2016-05-07.

[55] Adrian E Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 528–539, 1985.

[56] Thanapant Raicharoen, Chidchanok Lursinsap, and Paron Sanguanbhokai. Application of critical support vector machine to time series prediction. In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*, volume 5, pages V–V. IEEE, 2003.

[57] Hossein Sadaei and Muhammad Hisyam Lee. Multilayer stock forecasting model using fuzzy time series. *The Scientific World Journal*, 2014, 2014.

[58] Hossein Javedani Sadaei, Rasul Enayatifar, Frederico Gadelha Guimaraes,Maqsood Mahmud, and Zakarya A Alzamil. Combining arfima models and fuzzy time series for the forecast of long memory time series. *Neurocomputing*, 175:782–796, 2016.

[59] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[60] Qiang Song and Brad S Chissom. Fuzzy time series and its models. *Fuzzy sets and systems*, 54(3):269–277, 1993.

[61] BaiQing Sun, Haifeng Guo, Hamid Reza Karimi, Yuanjing Ge, and Shan Xiong. Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series. *Neurocomputing*, 151:1528–1536, 2015.

[62] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[63] Civilian unemployment rate. http://www.forecasts.org/data/employment-data.htm. Accessed: 2016-05-07.

[64] Jie Wang and Jun Wang. Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks. *Neurocomputing*, 156:68–78, 2015.

[65] L. Wasserman. Bayesian model selection and model averaging. . *J. Math. Psychology*, 44:92–107, 2000.

[66] Schoutens Wim. *Levy Processes in Finance: Pricing Financial Derivatives*. John Wiley & Sons, Ltd., West Sussex PO19 8SQ, England, 2003.

[67] H. Xie, P. Andreae, M. Zhang, and P. Warren. Learning models for english speech recognition. In *Proceedings of the 27th conference on Australasian computer science, ACSC 2004, Darlinghurst, Australia*, pages 323–329, 2004.

[68] getting stock index from yahoo. http://finance.yahoo.com/. Accessed: 2016-05-07.

[69] Yun Yang and Jianmin Jiang. Hmm-based hybrid meta-clustering ensemble for temporal data. *Knowledge-Based Systems*, 56:299–310, 2014.

[70] Hui-Kuang Yu. Weighted fuzzy time series models for taiex forecasting. *Physica A: Statistical Mechanics and its Applications*, 349(3):609–624, 2005.

[71] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[72] Weigang Zhao, Jianzhou Wang, and Haiyan Lu. Combining forecasts of electricity consumption in china with time-varying weights updated by a high-order markov chain model. *Omega*, 45:80–91, 2014.

[73] W. Zucchini and I. L. Macdonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall, New York, 2009.