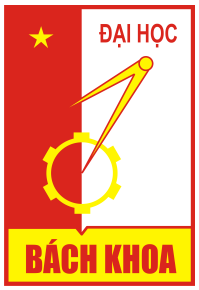
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

*



Xử lý chuỗi thời gian có yếu tố mùa

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Chuyên ngành: TOÁN TIN

Giảng viên:	**PGS. TS. Tống Đình Quỳ**
Sinh viên:	**Nguyễn Nhật Anh**
MSSV:	**20130154**
Lớp:	**KSTN - Toán Tin - K58**

HÀ NỘI - 2018

Mục lục

Mở đầu

Trong thực tế, có rất nhiều bộ dữ liệu được quan sát, đo đạc trong một khoảng thời gian liên tục gần nhau: chẳng hạn khi phân tích kĩ thuật, dự đoán giá cổ phiếu trong chứng khoán, dữ liệu sẽ là giá đóng cửa từng ngày của cổ phiếu đó và được theo dõi trong một khoảng thời gian nhất định (có thể là một vài tháng/quý/năm); nhiệt độ, độ ẩm, số giờ nắng, v.v. tại một trạm quan trắc hay đặc điểm thủy văn (lượng mưa, dòng chảy, mực nước trung bình, v.v.) của một hồ nước, con sông theo tháng/năm được đo đạc trong nhiều năm liên tiếp. Với những bộ dữ liệu đo đạc theo thời gian như vậy, những giá trị quan sát tại những thời điểm liên tiếp gần nhau sẽ có mối tương quan nào đó (chẳng hạn, giá cổ phiếu tại một ngày thường sẽ liên quan tới giá cổ phiếu của những ngày ngay trước đó; hay các đặc điểm khí hậu, khí tượng thủy văn thường biến đổi tuần hoàn theo chu kì năm, v.v.). Trong những tình huống này, phân tích chuỗi thời gian là một công cụ hợp lý và hữu hiệu để xử lý, phân tích và dự báo các bộ dữ liệu như trên.

Một trong những mô hình chuỗi thời gian phổ biến và hiệu quả là mô hình ARIMA (Autoregressive Integrated Moving Average) do George E.P.Box và Gwilym M.Jenkins đề xuất [4]. Đây là mô hình tuyến tính và được ứng dụng rộng rãi trong nhiều lĩnh vực bởi có thể áp dụng được với cả những chuỗi thời gian không dừng có yếu tố xu thế (trend). Ý tưởng chính khi xây dựng mô hình ARIMA là đầu tiên, ta thực hiện lấy sai phân chuỗi thời gian ban đầu cho tới khi thu được một chuỗi thời gian dừng, sau đó mô hình hóa chuỗi thời gian đó bởi một quá trình tự hồi quy trung bình trượt ARMA (Autoregressive Moving Average).

Đối với những chuỗi thời gian có cả yếu tố mùa (seasonality), chẳng hạn như nhiệt độ, lượng mưa, độ ẩm, v.v., George E.P.Box và Gwilym M.Jenkins đã đưa ra mô hình SARIMA (Seasonal ARIMA). Mô hình này là mở rộng của mô hình ARIMA, áp dụng được với những chuỗi thời gian không dừng có cả yếu tố xu thế lẫn yếu tố mùa, với chu kì s đã biết trước (ví dụ chuỗi thời gian đo theo tháng có chu kì bằng 1 năm thì $s = 12$, nếu đo theo từng quý thì $s = 4$). Do vậy, mô hình SARIMA rất thích hợp ứng dụng để dự báo chuỗi thời gian trong nhiều lĩnh vực khác nhau, đặc biệt là các lĩnh

vực như thủy văn, khí tượng, nông nghiệp, v.v. nhờ đặc trưng yếu tố mùa khá rõ ràng.

Trong đồ án tốt nghiệp, em xin trình bày ứng dụng của mô hình SARIMA phân tích, xử lý và dự báo lượng mưa hàng tháng tại trạm quan trắc Hà Nội. Đây là chuỗi thời gian có yếu tố mùa khá rõ ràng nên mô hình SARIMA rất thích hợp để mô hình hóa chuỗi thời gian này. Nội dung chính của đồ án được chia làm hai chương:

- Chương 1: Trình bày những khái niệm cơ bản của chuỗi thời gian: định nghĩa chuỗi thời gian, chuỗi ồn trắng, quá trình dừng, những đặc trưng thống kê cơ bản của chuỗi thời gian dừng như hàm tự hiệp phương sai, tự tương quan, tự tương quan riêng. Tiếp theo, em trình bày những khái niệm quá trình tự hồi quy (Autoregressive - AR), quá trình trung bình trượt (Moving Average - MA), quá trình tự hồi quy trung bình trượt (Autoregressive Moving Average - ARMA), quá trình ARIMA và cuối cùng là quá trình SARIMA.

- Chương 2: Tiến hành phân tích, xử lý dữ liệu lượng mưa hàng tháng tại trạm quan trắc Hà Nội sử dụng mô hình SARIMA. Quy trình mô hình hóa chuỗi thời gian được tiến hành theo các bước do George E.P.Box và Gwilym M.Jenkins đề xuất:

  – Model identification: Chọn mô hình phù hợp cho chuỗi thời gian.

  – Model estimation: Ước lượng các tham số của mô hình đã chọn.

  – Model diagnostic checking: Kiểm tra tính phù hợp của mô hình.

  Để phân tích, xử lý dữ liệu trong phần này, em sử dụng công cụ phần mềm R. R là một ngôn ngữ lập trình, môi trường phần mềm nhằm mục đích phục vụ tính toán thống kê và đồ họa thống kê, được phát triển, hỗ trợ bởi R Foundation [12]. Việc xây dựng mô hình SARIMA được hỗ trợ bởi gói thư viện forecast của tác giả Rob Hyndman và cộng sự [9].

Cuối cùng, em xin chân thành cảm ơn PGS. TS. Tống Đình Quỳ đã tận tình hướng dẫn em hoàn thành đồ án tốt nghiệp. Vì điều kiện thời gian cũng như hạn chế trong kinh nghiệm, năng lực nên đồ án này khó tránh khỏi thiếu sót. Rất mong thầy cô và các bạn thông cảm và đóng góp ý kiến để đề tài được hoàn thiện hơn.

Chương 1

Cơ sở toán học

1.1 Chuỗi thời gian

1.1.1 Khái niệm chuỗi thời gian

Chuỗi thời gian là một dãy các giá trị quan sát $\{x_t\}$ của một dãy biến ngẫu nhiên tương ứng (gọi là một *quá trình ngẫu nhiên*) $\{X_t\}$, $t \in T$ với T là một tập chỉ số nào đó. Chỉ số t thể hiện thời điểm mà giá trị quan sát đó được ghi lại. Nếu tập chỉ số thời gian T là rời rạc, chẳng hạn $T = \mathbb{Z}$ thì ta gọi chuỗi thời gian đó là rời rạc. Ngược lại, nếu tập T liên tục, chẳng hạn $T = [0, 1]$ thì chuỗi thời gian đó được gọi là liên tục. Ở đây, ta chỉ xét các chuỗi thời gian rời rạc với $T = \mathbb{Z}$.

Đôi khi, người ta cũng dùng thuật ngữ *chuỗi thời gian* để chỉ cả $\{x_t\}$ và $\{X_t\}$.

1.1.2 Toán tử lùi và toán tử sai phân

Xét $\{X_t\}$ là một chuỗi thời gian. *Toán tử lùi* B được định nghĩa là

$$BX_t = X_{t-1}.$$

Toán tử sai phân ∇ được định nghĩa là

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}.$$

Luỹ thừa của các toán tử này được định nghĩa một cách tự nhiên

$$B^j X_t = B(B^{j-1}X_t) = X_{t-j},$$
$$\nabla^j X_t = \nabla(\nabla^{j-1}X_t).$$

Đa thức của B hoặc ∇ được tính toán giống như đa thức bình thường. Chẳng hạn,

$$\begin{aligned}\nabla^2 X_t &= \nabla(\nabla X_t) \\ &= (1-B)(1-B)X_t \\ &= (1-2B+B^2)X_t \\ &= X_t - 2X_{t-1} + X_{t-2}.\end{aligned}$$

Toán tử sai phân trễ-d ∇_d

$$\nabla_d X_t = (1-B^d)X_t = X_t - X_{t-d}.$$

1.1.3 Ồn trắng

Gọi $\{X_t\}$ là một dãy các biến ngẫu nhiên cùng có kì vọng bằng 0, phương sai hữu hạn σ^2 và không tương quan (tức là $\mathrm{Cov}(X_t, X_s) = \mathrm{E}[X_t X_s] = 0, \forall s, t$). Khi đó chuỗi thời gian này được gọi là một *ồn trắng* (hoặc *nhiễu trắng*) với phương sai σ^2, kí hiệu $\{X_t\} \sim WN(0, \sigma^2)$.

Ví dụ: Chuỗi $\{X_t\}$ có các X_t độc lập với nhau và có cùng phân phối (i.i.d) chuẩn $\mathcal{N}(0, \sigma^2)$ là một ồn trắng. Chuỗi này thường được gọi là *ồn trắng Gauss*.

1.2 Quá trình dừng

Trong thực tế, hầu hết các chuỗi thời gian chưa qua xử lý ban đầu đều mang những dấu hiệu xu thế hoặc mùa tương đối rõ ràng (chẳng hạn giá cổ phiếu của một công ty đang làm ăn có lời có xu thế tăng, hay nhiệt độ, lượng mưa trung bình đo được tại một địa điểm có tính mùa với chu kì 1 năm, số vết đen quan sát được trên bề mặt Mặt Trời có chu kì xấp xỉ 11 năm, v.v.). Do vậy, những tính chất thống kê (kì vọng, phương sai, v.v.) của chuỗi biến động không ngừng phụ thuộc thời gian. Đối với mô hình ARMA(p, q), những thành phần xu thế hay mùa sẽ được khử trước khi tiến hành áp dụng mô hình, bởi ta chỉ quan tâm tới những chuỗi thời gian có tính dừng. Tính dừng của chuỗi thời gian là một tính chất quan trọng, nó đảm bảo những tính chất thống kê như kì vọng, phương sai, v.v. không phụ thuộc vào thời gian.

1.2.1 Dừng ngặt

Cho $\{X_t\}$ là một chuỗi thời gian, kí hiệu $F(x_{t_1}, \ldots, x_{t_k})$ là hàm phân phối tích lũy đồng thời của $X_{t_1}, \ldots, X_{t_k}$. Khi đó chuỗi $\{X_t\}$ được gọi là *dừng ngặt* nếu với $k, t_1, \ldots, t_k$ bất

kì,

$$F(x_{t_1}, \ldots, x_{t_k}) = F(x_{t_1+h}, \ldots, x_{t_k+h}) \quad \forall h,$$

tức là, phân phối đồng thời của $X_{t_1+h}, \ldots, X_{t_k+h}$ không phụ thuộc vào h.

1.2.2 Dừng yếu

Điều kiện *dừng ngặt* là điều kiện khá mạnh. Nó yêu cầu các biến X_t phải có cùng phân phối xác suất, và do đó moment mọi cấp của chuỗi thời gian là không đổi. Trên thực tế, ta thường sử dụng một điều kiện ít ngặt hơn, được gọi là *dừng yếu*. Một chuỗi thời gian $\{X_t\}$ với $\mathrm{E}[X_t^2] < \infty$ là *dừng yếu* nếu nó thỏa mãn các điều kiện sau:

(i) Hàm kì vọng của chuỗi thời gian $\mu(t) = \mathrm{E}[X_t]$ không phụ thuộc vào thời gian t.

(ii) Hàm hiệp phương sai của chuỗi thời gian

$$\begin{aligned} \gamma(t, t+h) &= \mathrm{Cov}(X_t, X_{t+h}) \\ &= \mathrm{E}[(X_t - \mu(t))(X_{t+h} - \mu(t+h))] \end{aligned} \tag{1.1}$$

không phụ thuộc vào t với mỗi h. Tức là hiệp phương sai của chuỗi thời gian tại hai thời điểm chỉ phụ thuộc vào khoảng thời gian chênh lệch h (hay còn gọi là *trễ* h) mà không quan tâm tới thời điểm cụ thể.

Lưu ý. Khi có chuỗi thời gian dừng mà không nói gì thêm dừng yếu hay dừng ngặt thì ta mặc định coi đó là dừng yếu.

Nếu chuỗi thời gian dừng $\{X_t\}$ có kì vọng μ khác 0 thì ta đặt $X'_t = X_t - \mu$ và xét chuỗi thời gian dừng $\{X'_t\}$ thay cho $\{X_t\}$. Như vậy khi làm việc với chuỗi thời gian dừng ta chỉ cần xét những chuỗi thời gian có kì vọng bằng 0.

1.3 Hàm tự hiệp phương sai và hàm tự tương quan

1.3.1 Hàm tự hiệp phương sai

Với điều kiện dừng của chuỗi $\{X_t\}$, hàm hiệp phương sai ban đầu $\gamma(\cdot, \cdot)$ vốn phụ thuộc vào hai biến thì giờ ta có thể coi nó chỉ phụ thuộc vào một biến

$$\gamma(h) := \gamma(0, h) = \gamma(t, t+h) \quad \forall t,$$

và hàm $\gamma(\cdot)$ được gọi là *hàm tự hiệp phương sai* của chuỗi thời gian dừng $\{X_t\}$.

Hàm tự hiệp phương sai có một số tính chất cơ bản sau:

(i) $\gamma(0) = \mathrm{Var}(X_t) \geq 0$,

(ii) $|\gamma(h)| \leq \gamma(0) \quad \forall h$,

(iii) γ là hàm chẵn

$$\gamma(h) = \gamma(-h) \quad \forall h,$$

(iv) γ là hàm xác định không âm: Với mỗi số nguyên dương k, kí hiệu $\mathbf{\Gamma}_k = [\gamma(i-j)]_{i,j=1}^k$ thì ta luôn có $\mathbf{\Gamma}_k$ là ma trận xác định không âm, tức là với mỗi vector $\boldsymbol{a} = (a_1, \ldots, a_k)^T$ có các thành phần a_i là các số thực bất kì, ta luôn có

$$\boldsymbol{a}^T \mathbf{\Gamma}_k \boldsymbol{a} \geq 0.$$

Cho một dãy các quan sát $\{x_t\}_{t=1}^n$ của chuỗi thời gian $\{X_t\}$. Ta định nghĩa *hàm tự hiệp phương sai mẫu* $\hat{\gamma}$ như sau

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}) \text{ với } -n < h < n, \tag{1.2}$$

với $\overline{x} = \dfrac{1}{n} \sum_{t=1}^{n} x_t$.

Hàm tự hiệp phương sai mẫu cũng có các tính chất tương tự như hàm tự hiệp phương sai:

(i) $\hat{\gamma}(0) \geq 0$,

(ii) $|\hat{\gamma}(h)| \leq \hat{\gamma}(0) \quad -n < h < n$,

(iii) $\hat{\gamma}$ là hàm chẵn

$$\hat{\gamma}(h) = \hat{\gamma}(-h) \quad -n < h < n,$$

(iv) Với mỗi số nguyên dương $k < n$, kí hiệu $\hat{\mathbf{\Gamma}}_k = [\hat{\gamma}(i-j)]_{i,j=1}^k$ thì ta luôn có $\hat{\mathbf{\Gamma}}_k$ là ma trận xác định không âm, tức là với mỗi vector $\boldsymbol{a} = (a_1, \ldots, a_k)^T$ với các thành phần a_i là các số thực bất kì ta luôn có

$$\boldsymbol{a}^T \hat{\mathbf{\Gamma}}_k \boldsymbol{a} \geq 0.$$

1.3.2 Hàm tự tương quan

Cho chuỗi thời gian dừng $\{X_t\}$, ta định nghĩa *hàm tự tương quan* ρ

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{1.3}$$

Hàm tự tương quan ρ cũng có các tính chất tương tự như hàm tự hiệp phương sai γ, ngoài ra ta còn có $\rho(0) = 1$ và do đó các tự tương quan đều có giá trị tuyệt đối nhỏ hơn hoặc bằng 1.

Cho một dãy các quan sát $\{x_t\}_{t=1}^n$ của chuỗi thời gian $\{X_t\}$. Ta định nghĩa *hàm tự tương quan mẫu* $r(h)$ như sau

$$r(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2} \text{ với } -n < h < n. \tag{1.4}$$

1.3.3 Hàm tự tương quan riêng

Xấp xỉ tuyến tính tốt nhất

Gọi $\{X_t\}$ là một chuỗi thời gian dừng, Y là một biến ngẫu nhiên có phương sai hữu hạn. Xét bài toán tìm tổ hợp tuyến tính của $1, X_n, \ldots, X_1$ để dự đoán giá trị của Y sao cho trung bình của bình phương sai số là nhỏ nhất, tức là

$$S(a_0, \ldots, a_n) = \mathrm{E}[(Y - a_0 - a_1X_n - \ldots - a_nX_1)^2]$$

đạt giá trị nhỏ nhất.

Ta có thể xem bài toán trên dưới góc độ tìm hình chiếu của một điểm lên một không gian con như sau: Gọi $L^2(\Omega, \mathcal{F}, P)$ là không gian tất cả các biến ngẫu nhiên X xác định trên không gian xác suất $(\Omega, \mathcal{F}, P)$ thỏa mãn $\mathrm{E}[X^2] < \infty$ (nếu X có kì vọng bằng 0 thì $\mathrm{E}[X^2]$ chính là phương sai của X). Đặt $\langle X, Y\rangle = \mathrm{E}[XY]$ và coi hai biến ngẫu nhiên là bằng nhau nếu chúng bằng nhau hầu khắp nơi (tức $P(X = Y) = 1$). Khi đó, $L^2(\Omega, \mathcal{F}, P)$ cùng với tích vô hướng $\langle\cdot,\cdot\rangle$ tạo thành một không gian Hilbert [6, tr. 46-47]. Gọi $\mathcal{M} = \overline{sp}\{1, X_n, X_{n-1}, \ldots, X_1\} = \{a_0 + a_1X_n + a_2X_{n-1} + \ldots + a_nX_1 | a_0, \ldots, a_n \in \mathbb{R}\}$ là một không gian con đóng của không gian Hilbert $L^2(\Omega, \mathcal{F}, P)$. Xét hình chiếu $\mathbb{P}_{\mathcal{M}}Y$ của Y lên $\mathcal{M} = \overline{sp}\{1, X_n, X_{n-1}, \ldots, X_1\}$. Theo định lý về phép chiếu Hilbert, tồn tại duy nhất phần tử $\mathbb{P}_{\mathcal{M}}Y \in \mathcal{M}$ thỏa mãn

$$\|Y - \mathbb{P}_{\mathcal{M}}Y\| \leq \|Y - X\| \quad \forall X \in \mathcal{M}$$

được gọi là hình chiếu của Y lên $\mathcal{M}$ [6, tr. 51-54]. Như vậy, $\mathbb{P}_{\mathcal{M}}Y$ chính là xấp xỉ tuyến tính tốt nhất của Y theo $1, X_n, X_{n-1}, \ldots, X_1$. Hình chiếu $\mathbb{P}_{\mathcal{M}}Y$ có tính chất là $Y - \mathbb{P}_{\mathcal{M}}Y$ vuông góc với không gian con $\mathcal{M}$, hay

$$\begin{aligned} \langle Y - \mathbb{P}_{\mathcal{M}}Y, 1\rangle &= 0 \\ \langle Y - \mathbb{P}_{\mathcal{M}}Y, X_n\rangle &= 0 \\ &\cdots \\ \langle Y - \mathbb{P}_{\mathcal{M}}Y, X_1\rangle &= 0. \end{aligned} \tag{1.5}$$

Như vậy, ta có thể tìm được các hệ số $a_0, a_1, \ldots, a_n$ của $\mathbb{P}_{\mathcal{M}}Y$ bằng cách giải hệ phương trình tuyến tính (1.5).

Tự tương quan riêng

Tự tương quan riêng $\alpha(h)$ được định nghĩa là hệ số tương quan giữa X_t và X_{t-h} với điều kiện không xét tới ảnh hưởng của các giá trị trung gian $X_{t-1}, \ldots, X_{t-h+1}$. Cụ thể, với chuỗi thời gian dừng $\{X_t\}$, kí hiệu $\mathbb{P}X_t$ và $\mathbb{P}X_{t-h}$ lần lượt là xấp xỉ tuyến tính tốt nhất của X_t và X_{t-h} theo $X_{t-1}, \ldots, X_{t-h+1}$. Khi đó *hàm tự tương quan riêng* của $\{X_t\}$, kí hiệu $\alpha(h)$, được xác định theo công thức sau

$$\alpha(h) = \rho(X_t - \mathbb{P}X_t, X_{t-h} - \mathbb{P}X_{t-h}), h > 1. \tag{1.6}$$

Khi $h = 1$ thì ta định nghĩa $\alpha(1) = \rho(X_t, X_{t-1}) = \rho(1)$.

Hàm tự tương quan riêng $\alpha(h)$ của một chuỗi thời gian dừng $\{X_t\}$ có một số đặc trưng cơ bản như sau:

(i) $|\alpha(h)| \leq 1 \quad \forall h$,

(ii) $\alpha(h)$ chính là hệ số thứ h trong xấp xỉ tuyến tính tốt nhất của X_t theo $X_{t-1}, \ldots, X_{t-h}$. Tức là, nếu xấp xỉ tuyến tính tốt nhất của X_t theo $X_{t-1}, \ldots, X_{t-h}$ có dạng

$$\phi_{h1}X_{t-1} + \phi_{h2}X_{t-2} + \ldots + \phi_{hh}X_{t-h}$$

thì $\alpha(h) = \phi_{hh}$ ([6, tr. 102]).

(iii) Đối với quá trình tự hồi quy bậc p AR(p), $\alpha(h) = 0$ với $h > p$. Do đó hàm tự tương quan riêng đóng vai trò quan trọng trong việc xác định bậc của mô hình AR(p).

1.4 Các mô hình ARIMA

1.4.1 Quá trình AR(p)

Xét chuỗi thời gian dừng $\{X_t\}_{t\in\mathbb{Z}}$ có kì vọng bằng 0. $\{X_t\}$ được gọi là một *quá trình tự hồi quy bậc p*, kí hiệu AR(p) nếu nó có dạng

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t, \tag{1.7}$$

trong đó $\{\epsilon_t\}$ là một ồn trắng với phương sai σ^2, $\{\epsilon_t\} \sim WN(0, \sigma^2)$.

Quá trình AR(p) viết ngắn gọn dưới dạng toán tử lùi

$$\phi(B)X_t = \epsilon_t, \tag{1.8}$$

với $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$.

1.4.2 Quá trình MA(q)

Xét chuỗi thời gian dừng $\{X_t\}_{t\in\mathbb{Z}}$ có kì vọng bằng 0. $\{X_t\}$ được gọi là một *quá trình trung bình trượt bậc q*, kí hiệu MA(q) nếu nó có dạng

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}, \tag{1.9}$$

trong đó $\{\epsilon_t\} \sim WN(0, \sigma^2)$.

Quá trình MA(q) viết ngắn gọn dưới dạng toán tử lùi

$$X_t = \theta(B)\epsilon_t, \tag{1.10}$$

với $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$.

1.4.3 Quá trình ARMA(p, q)

Xét chuỗi thời gian dừng $\{X_t\}_{t\in\mathbb{Z}}$ có kì vọng bằng 0. $\{X_t\}$ được gọi là một quá trình *tự hồi quy trung bình trượt bậc p, q*, kí hiệu ARMA(p, q) nếu nó thỏa mãn

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \tag{1.11}$$

trong đó $\{\epsilon_t\} \sim WN(0, \sigma^2)$. Viết ngắn gọn dưới dạng toán tử lùi

$$\phi(B)X_t = \theta(B)\epsilon_t, \tag{1.12}$$

với

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q.$$

Điều kiện dừng

Xét phương trình (1.11). Khi đó tồn tại quá trình dừng $\{X_t\}$ thỏa mãn phương trình (1.11) nếu và chỉ nếu $\phi(z) = 0$ không có nghiệm nào nằm trên đường tròn đơn vị ([7, tr. 85]), hay nói cách khác,

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p \neq 0 \quad \forall |z| = 1. \tag{1.13}$$

Ví dụ: Xét quá trình AR(1)

$$X_t = \phi_1 X_{t-1} + \epsilon_t.$$

- Trường hợp $|\phi_1| < 1$ (nghiệm của phương trình $\phi(z) = 1 - \phi_1 z = 0$ có trị tuyệt đối lớn hơn 1):

$$\begin{aligned} X_t &= \epsilon_t + \phi_1 X_{t-1} \\ &= \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 X_{t-2} \\ &= \ldots \\ &= \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}. \end{aligned}$$

Do $|\phi_1| < 1$ nên chuỗi $\sum_{j=0}^{\infty} |\phi_1^j|$ hội tụ, dẫn tới $\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ hội tụ theo bình phương trung bình (mean-square convergent) (tức là hội tụ theo chuẩn $\|\cdot\|$ với $\|X\|^2 = \langle X, X \rangle = \mathrm{E}[X^2]$).

Từ đó ta tính được $\mathrm{E}[X_t] = \mathrm{E}[\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}] = 0$ và

$$\begin{aligned}
\mathrm{Cov}(X_t, X_{t+h}) &= \mathrm{E}[X_t X_{t+h}] \\
&= \mathrm{E}\Big[\Big(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}\Big)\Big(\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t+h-j}\Big)\Big] \\
&= \sigma^2 \phi_1^h (1 + \phi_1^2 + \phi_1^4 + \ldots) \\
&= \sigma^2 \frac{\phi_1^h}{1 - \phi_1^2}
\end{aligned}$$

không phụ thuộc vào t. Vậy với $|\phi_1| < 1$ thì tồn tại quá trình dừng $\{X_t\}$ được xác định theo công thức $X_t = \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ thỏa mãn phương trình AR(1) $X_t = \phi_1 X_{t-1} + \epsilon_t$.

- Trường hợp $|\phi_1| = 1$ (nghiệm của phương trình $\phi(z) = 1 - \phi_1 z = 0$ có trị tuyệt đối bằng 1):

  Giả sử tồn tại quá trình dừng $\{X_t\}$ thỏa mãn $X_t = \phi_1 X_{t-1} + \epsilon_t$. Khi đó ta có

$$\begin{aligned}
X_t &= \phi_1 X_{t-1} + \epsilon_t \\
&= \phi_1^2 X_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \ldots \\
&= \phi_1^{h+1} X_{t-(h+1)} + \phi_1^h \epsilon_{t-h} + \ldots + \phi_1 \epsilon_{t-1} + \epsilon_t \\
\Rightarrow \mathrm{Var}(X_t - \phi_1^{h+1} X_{t-(h+1)}) &= \mathrm{Var}(\phi_1^h \epsilon_{t-h} + \ldots + \phi_1 \epsilon_{t-1} + \epsilon_t) \\
&= \sigma^2(\phi_1^{2h} + \phi_1^{2(h-1)} + \ldots + \phi_1^2 + 1) \\
&= h\sigma^2.
\end{aligned}$$

  Đồng thời do tính dừng nên ta có

$$\begin{aligned}
\mathrm{Var}(X_t - \phi_1^{h+1} X_{t-(h+1)}) &= \mathrm{Var}(X_t) + \phi_1^{2(h+1)} \mathrm{Var}(X_{t-(h+1)}) - 2\phi_1^{h+1} \mathrm{Cov}(X_t, X_{t-(h+1)}) \\
&= 2\gamma(0) - 2\phi_1^{h+1} \gamma(h+1) \\
&\leq 4\gamma(0) \quad (\text{Do } |\phi_1| = 1 \text{ và } |\gamma(h+1)| \leq \gamma(0)).
\end{aligned}$$

  Vậy ta luôn có $4\gamma(0) \geq h\sigma^2 \quad \forall h$, vô lý. Vậy không tồn tại quá trình dừng nào thỏa mãn $X_t = \phi_1 X_{t-1} + \epsilon_t$ với $|\phi_1| = 1$.

- Trường hợp $|\phi_1| > 1$ (nghiệm của phương trình $\phi(z) = 1 - \phi_1 z = 0$ có trị tuyệt đối nhỏ hơn 1):

Trong trường hợp này, $\sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ không hội tụ nên không thể biểu diễn $X_t = \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ như ở trường hợp đầu được. Tuy nhiên, ta vẫn có thể áp dụng ý tưởng đó một cách tương tự bằng cách biểu diễn

$$\begin{aligned} X_t &= -\phi_1^{-1}\epsilon_{t+1} + \phi_1^{-1} X_{t+1} \\ &= -\phi_1^{-1}\epsilon_{t+1} - \phi_1^{-2}\epsilon_{t+2} + \phi_1^{-2} X_{t+2} \\ &= \dots \\ &= -\sum_{j=1}^{\infty} \phi_1^{-j} \epsilon_{t+j}. \end{aligned}$$

Do $|\phi_1^{-1}| < 1$ nên chuỗi $-\sum_{j=1}^{\infty} \phi_1^{-j}\epsilon_{t+j}$ hội tụ theo bình phương trung bình. Trung bình $\mathrm{E}[X_t] = 0$ và hiệp phương sai

$$\begin{aligned} \mathrm{Cov}(X_t, X_{t+h}) &= \mathrm{E}[X_t X_{t+h}] \\ &= \mathrm{E}\Big[\Big(-\sum_{j=1}^{\infty} \phi_1^{-j}\epsilon_{t+j}\Big)\Big(-\sum_{j=1}^{\infty} \phi_1^{-j}\epsilon_{t+h+j}\Big)\Big] \\ &= \sigma^2 \phi_1^{-h}(\phi_1^{-2} + \phi_1^{-4} + \dots) \\ &= \sigma^2 \frac{1}{(\phi_1^2 - 1)\phi_1^h} \end{aligned}$$

không phụ thuộc vào t. Vậy $X_t = -\sum_{j=1}^{\infty} \phi_1^{-j}\epsilon_{t+j}$ là một quá trình dừng thỏa mãn $X_t = \phi_1 X_{t-1} + \epsilon_t$ với $|\phi_1| > 1$.

Tuy nhiên trong biểu diễn $X_t = -\sum_{j=1}^{\infty} \phi_1^{-j}\epsilon_{t+j}$, ta có thể nhận thấy X_t phụ thuộc vào các giá trị tương lai của ồn trắng $\epsilon_{t+1}, \epsilon_{t+2}, \dots$ Những quá trình như vậy không thỏa mãn tính nhân quả và không có nhiều giá trị, ý nghĩa trong thực tế. Do vậy, khi ước lượng tìm tham số cho mô hình ARMA ta chỉ xét những mô hình thỏa mãn tính nhân quả. Khái niệm tính nhân quả sẽ được trình bày cụ thể trong phần tiếp theo.

Tính nhân quả

Một quá trình ARMA(p, q) $\{X_t\}$ được gọi là có *tính nhân quả* nếu tồn tại dãy hằng số $\{\psi_j\}_{j=0}^{\infty}$ sao cho chuỗi tương ứng hội tụ tuyệt đối ($\sum_{j=0}^{\infty} |\psi_j| < \infty$) và

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \quad \forall t. \tag{1.14}$$

Điều kiện nhân quả được thỏa mãn khi và chỉ khi mọi nghiệm của $\phi(z) = 0$ đều nằm ngoài hình tròn đơn vị ([7, tr. 85]), hay nói cách khác,

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p \neq 0 \quad \forall |z| \leq 1. \tag{1.15}$$

Lưu ý.

- Phương trình $X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ có thể được viết ngắn gọn dưới dạng

  $$X_t = \psi(B)\epsilon_t$$

  với $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$. Mà đồng thời từ phương trình ARMA(p, q)

  $$\begin{aligned} \phi(B)X_t &= \theta(B)\epsilon_t \\ \Rightarrow X_t &= \phi(B)^{-1}\theta(B)\epsilon_t. \end{aligned}$$

  Như vậy ta có thể tìm được các ψ_j bằng cách đồng nhất hệ số $\psi(z) = \phi(z)^{-1}\theta(z)$.

- Như đã phân tích ở phần trên, đối với những chuỗi thời gian không thỏa mãn tính nhân quả, X_t được biểu diễn qua những giá trị tương lai $\epsilon_{t+1}, \epsilon_{t+2}, \ldots$ Những quá trình như vậy không phù hợp ứng dụng trong thực tế. Vì vậy, từ nay ta chỉ xét những quá trình thỏa mãn điều kiện nhân quả.

Tính khả nghịch

Một quá trình ARMA(p, q) $\{X_t\}$ được gọi là có *tính khả nghịch* nếu tồn tại dãy hằng số $\{\pi_j\}_{j=0}^{\infty}$ sao cho chuỗi tương ứng hội tụ tuyệt đối ($\sum_{j=0}^{\infty} |\pi_j| < \infty$) và

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \quad \forall t. \tag{1.16}$$

Điều kiện khả nghịch được thỏa mãn khi và chỉ khi mọi nghiệm của $\theta(z) = 0$ đều nằm ngoài hình tròn đơn vị ([7, tr. 86]), hay nói cách khác,

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q \neq 0 \quad \forall |z| \leq 1. \tag{1.17}$$

Lưu ý.

- Phương trình $\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ có thể được viết ngắn gọn dưới dạng

$$\epsilon_t = \pi(B) X_t$$

với $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$. Các π_j có thể tìm được bằng cách đồng nhất hệ số $\pi(z) = \theta(z)^{-1}\phi(z)$.

- Tương tự như tính nhân quả, đối với những chuỗi thời gian không thỏa mãn tính khả nghịch, ϵ_t được biểu diễn qua những giá trị tương lai $X_{t+1}, X_{t+2}, \ldots$ Do đó ta chỉ xét những quá trình ARMA thỏa mãn điều kiện nhân quả và điều kiện khả nghịch.

Tự tương quan, tự tương quan riêng

- Trường hợp $q = 0$, tức quá trình AR(p) (1.7)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t.$$

Nhân hai vế của phương trình (1.7) với $X_{t-h}, h = 1, 2, \ldots$ rồi lấy kì vọng và chia cho $\gamma(0)$, ta thu được phương trình sai phân tuyến tính thuần nhất bậc p với hệ số hằng theo ρ

$$\rho(h) = \phi_1 \rho(h-1) + \ldots + \phi_p \rho(h-p), h = 1, 2, \ldots \tag{1.18}$$

Theo lý thuyết phương trình sai phân tuyến tính, giả sử phương trình đặc trưng

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

có các nghiệm phân biệt $\xi_1, \xi_2, \ldots, \xi_r$ với bội $m_1, m_2, \ldots, m_r$ tương ứng ($m_1 + m_2 + \ldots + m_r = p$) thì nghiệm tổng quát của phương trình sai phân (1.18) có dạng

$$\rho(h) = \sum_{i=1}^{r} \sum_{n=0}^{m_i - 1} c_{in} t^n \xi_i^{-h}$$

với các c_{in} là các hằng số [6, tr. 104-108]. Như vậy, hàm tự tương quan $\rho(h)$ là tổng của các hàm mũ, sin, cos với biên độ tắt dần.

Việc xác định chính xác bậc p của mô hình AR(p) dựa vào dáng điệu của hàm tự tương quan $\rho(h)$ là rất khó. Trên thực tế, để xác định bậc của AR(p) ta thường sử dụng hàm tự tương quan riêng $\alpha(h)$. Cụ thể, xét hệ số tương quan $\alpha(h)$ với $h > p$ thì mối tương quan giữa X_t và X_{t-h} thông qua các giá trị trung gian $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$ đã bị khử và X_t và X_{t-h} không tương quan trực tiếp với nhau trong mô hình AR(p) nên $\alpha(h) = 0$.

Ví dụ:

– Quá trình AR(1):

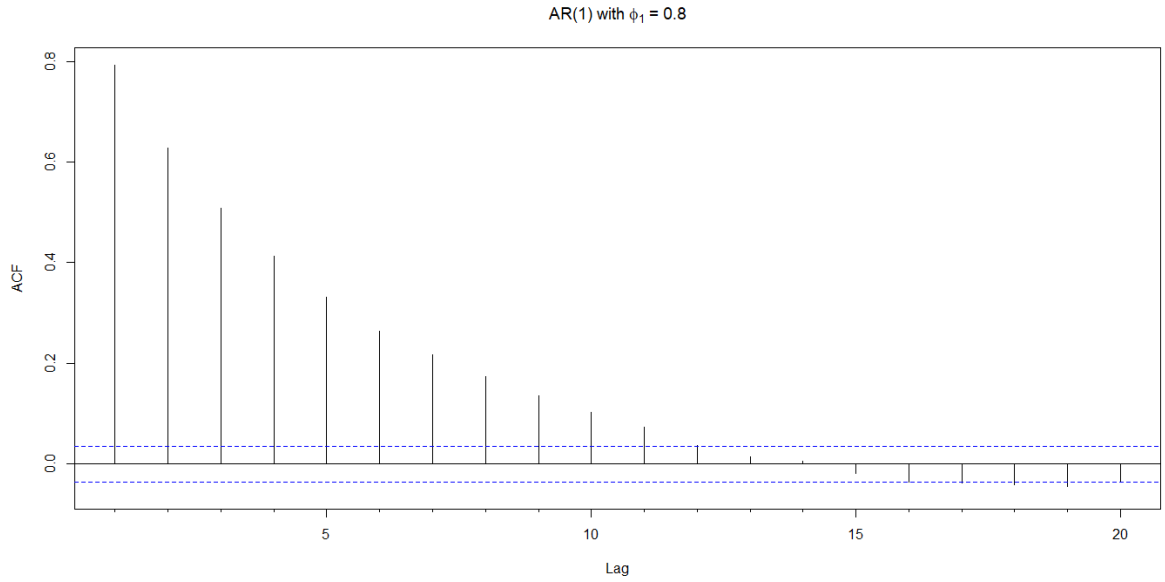
$$X_t = \phi_1 X_{t-1} + \epsilon_t,$$

với $|\phi_1| < 1$. Phương trình sai phân theo ρ tương ứng với quá trình AR(1) là

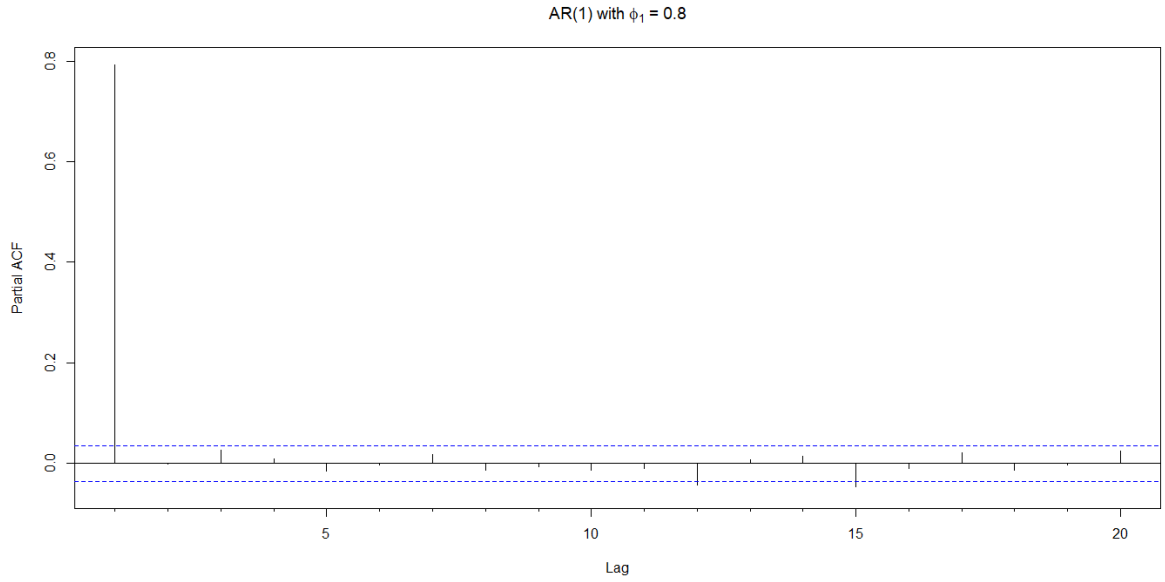
$$\rho(h) = \phi_1 \rho(h-1), h = 1, 2, \ldots$$

Từ đó ta tính được

$$\begin{aligned} \rho(0) &= 1 \\ \rho(h) &= \phi_1 \rho(h-1) \\ &= \phi_1^2 \rho(h-2) \\ &= \ldots \\ &= \phi_1^h \rho(0) = \phi_1^h. \end{aligned}$$



Hình 1.1: Tự tương quan mẫu của một chuỗi thời gian AR(1) với $\phi_1 = 0.8$.



Hình 1.2: Tự tương quan riêng mẫu của một chuỗi thời gian AR(1) với $\phi_1 = 0.8$.

- Trường hợp $p = 0$, tức quá trình MA(q) (1.9)

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \ldots + \theta_q\epsilon_{t-q}.$$

Ngược lại với quá trình AR(p), đối với quá trình MA(q) thì hàm tự tương quan $\rho(h)$ lại cho thấy ngay dấu hiệu nhận biết bậc q của mô hình bởi ta có $\rho(h) = 0$

với $h > q$. Thật vậy, nhân hai vế của phương trình (1.9) với $X_{t-h}, h = 0, 1, 2, \ldots$ rồi lấy kì vọng ta có

$$\gamma(h) = \mathrm{E}[X_{t-h}\epsilon_t] + \theta_1 \mathrm{E}[X_{t-h}\epsilon_{t-1}] + \ldots + \theta_q \mathrm{E}[X_{t-h}\epsilon_{t-q}].$$

$\theta_i \mathrm{E}[X_{t-h}\epsilon_{t-i}] = \theta_i \mathrm{E}[(\epsilon_{t-h} + \theta_1\epsilon_{t-h-1} + \ldots + \theta_q\epsilon_{t-h-q})\epsilon_{t-i}] = \sigma^2\theta_i\theta_{i-h}$ nếu $h \leq i \leq h+q$ và bằng 0 trong các trường hợp còn lại (ta coi $\theta_i = 0$ nếu $i = 0$ hoặc $i > q$). Vậy

$$\gamma(h) = \begin{cases} \sigma^2(\sum\limits_{i=h}^{q} \theta_i\theta_{i-h}) = \sigma^2(\sum\limits_{j=0}^{q-h} \theta_j\theta_{j+h}) \text{ nếu } 0 \leq h \leq q \\ 0 \text{ nếu } h > q \end{cases}$$

Từ đó

$$\rho(h) = \begin{cases} \dfrac{\theta_0\theta_h + \theta_1\theta_{h+1} + \ldots + \theta_{q-h}\theta_q}{\theta_0^2 + \theta_1^2 + \ldots + \theta_q^2} \text{ nếu } 1 \leq h \leq q \\ 0 \text{ nếu } h > q \end{cases} \tag{1.19}$$

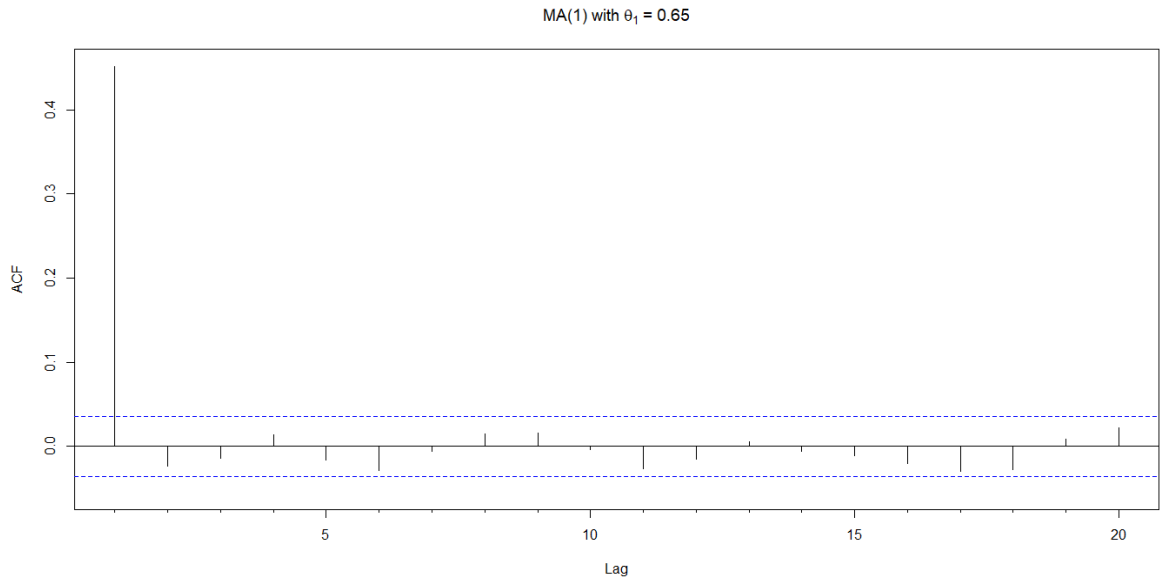
Ví dụ:

– Quá trình MA(1):

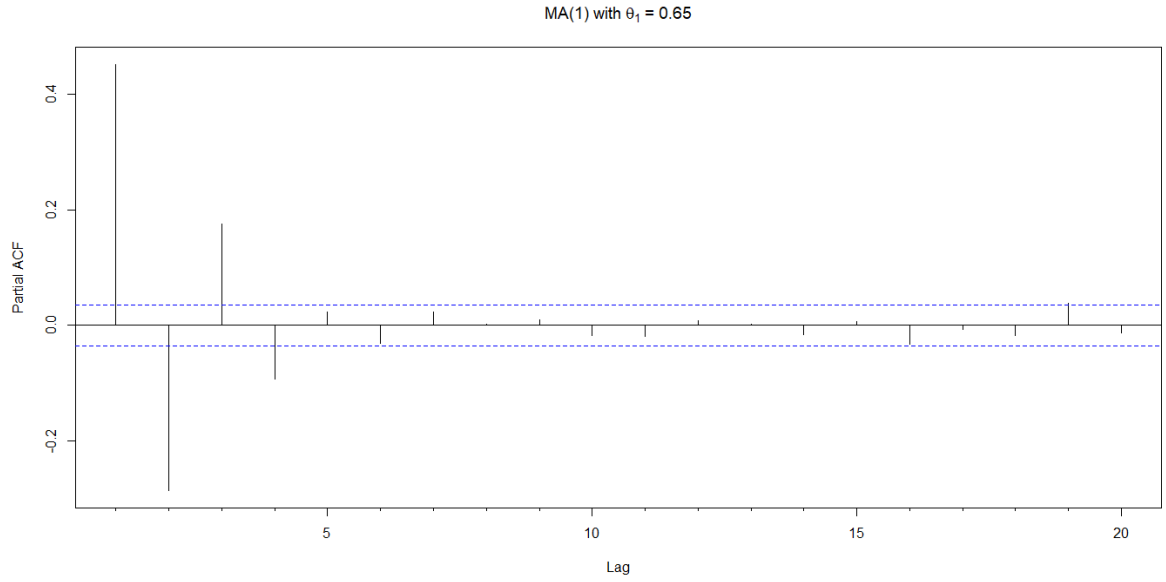
$$X_t = \epsilon_t + \theta_1\epsilon_{t-1},$$

với $|\theta_1| < 1$. Hệ số tự tương quan được tính theo công thức (1.19)

$$\begin{aligned} \rho(0) &= 1 \\ \rho(1) &= \frac{\theta_0\theta_1}{\theta_0^2 + \theta_1^2} = \frac{\theta_1}{1 + \theta_1^2} \\ \rho(h) &= 0, h > 1. \end{aligned}$$



Hình 1.3: Tự tương quan mẫu của một chuỗi thời gian MA(1) với $\theta_1 = 0.65$.



Hình 1.4: Tự tương quan riêng mẫu của một chuỗi thời gian MA(1) với $\theta_1 = 0.65$.

- Trường hợp $p, q \neq 0$:

  Trong trường hợp mô hình ARMA(p, q) tổng quát, cả hàm tự tương quan và tự tương quan riêng đều có dáng điệu tắt dần phức tạp chứ không tắt đột ngột như tự tương quan của mô hình MA(q) hay tự tương quan riêng của mô hình AR(p).

Do đó việc xác định bậc của mô hình ARMA(p, q) cần dựa thêm vào những kĩ thuật, tiêu chuẩn khác.

Ví dụ: Xét quá trình ARMA$(1, 1)$ nhân quả và khả nghịch

$$X_t = \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t. \tag{1.20}$$

Nhân hai vế phương trình (1.20) với $X_{t-h}, h = 0, 1, 2, \ldots$ và lấy kì vọng ta thu được

– Với $h = 0$:

$$\begin{aligned}\gamma(0) &= \phi_1\gamma(1) + \theta_1 \mathrm{E}[X_t\epsilon_{t-1}] + \mathrm{E}[X_t\epsilon_t] \\ &= \phi_1\gamma(1) + \theta_1 \mathrm{E}\left[\left(\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j}\right)\epsilon_{t-1}\right] + \mathrm{E}\left[\left(\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j}\right)\epsilon_t\right] \quad \text{(Tính nhân quả)} \\ &= \phi_1\gamma(1) + \sigma^2(\theta_1\psi_1 + \psi_0).\end{aligned}$$

ψ_0, ψ_1 có thể tính được nhờ đồng nhất hệ số

$$\begin{aligned}\psi(z)\phi(z) &= \theta(z) \\ (\psi_0 + \psi_1 z + \ldots)(1 - \phi_1 z) &= 1 + \theta_1 z.\end{aligned}$$

Từ đó ta có

$$\begin{aligned}\psi_0 &= 1 \\ \psi_1 &= \theta_1 + \phi_1\psi_0 = \theta_1 + \phi_1.\end{aligned}$$

Vậy

$$\gamma(0) = \phi_1\gamma(1) + \sigma^2(\theta_1^2 + \theta_1\phi_1 + 1). \tag{1.21}$$

– Với $h = 1$:

$$\begin{aligned}\gamma(1) &= \phi_1\gamma(0) + \theta_1 \mathrm{E}[X_{t-1}\epsilon_{t-1}] + \mathrm{E}[X_{t-1}\epsilon_t] \\ &= \phi_1\gamma(0) + \theta_1 \mathrm{E}\left[\left(\sum_{j=0}^{\infty}\psi_j\epsilon_{t-1-j}\right)\epsilon_{t-1}\right] \\ &= \phi_1\gamma(0) + \sigma^2\theta_1\psi_0 = \phi_1\gamma(0) + \sigma^2\theta_1.\end{aligned} \tag{1.22}$$

Từ (1.21) và (1.22) ta tính được

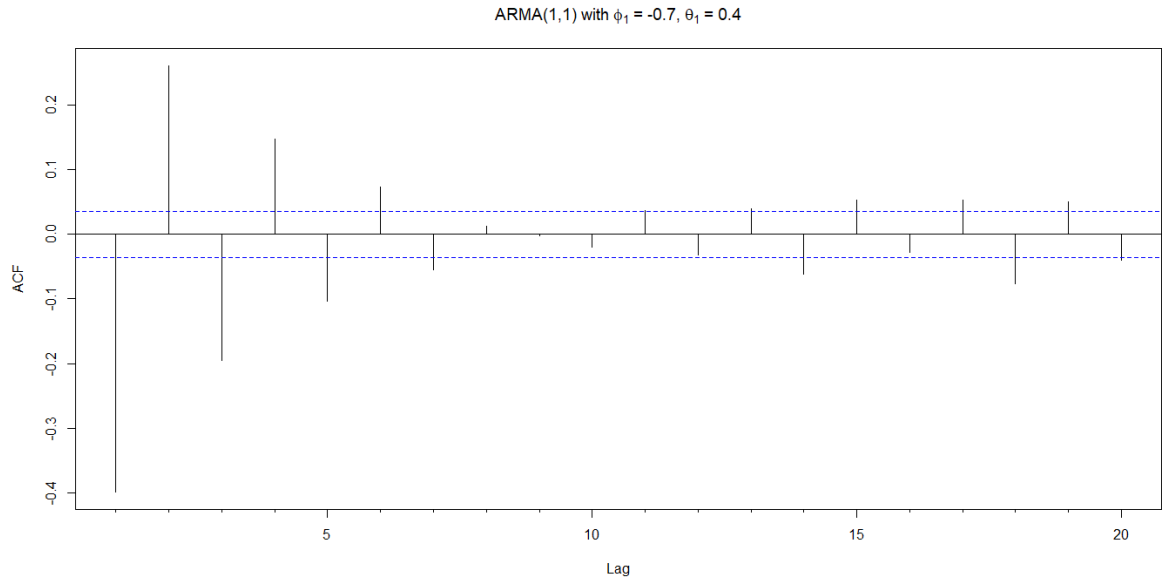
$$\gamma(0) = \sigma^2 \frac{1 + 2\phi_1\theta_1 + \theta_1^2}{1 - \phi_1^2} \text{ và } \gamma(1) = \sigma^2 \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 - \phi_1^2}.$$

– Với $h > 1$:

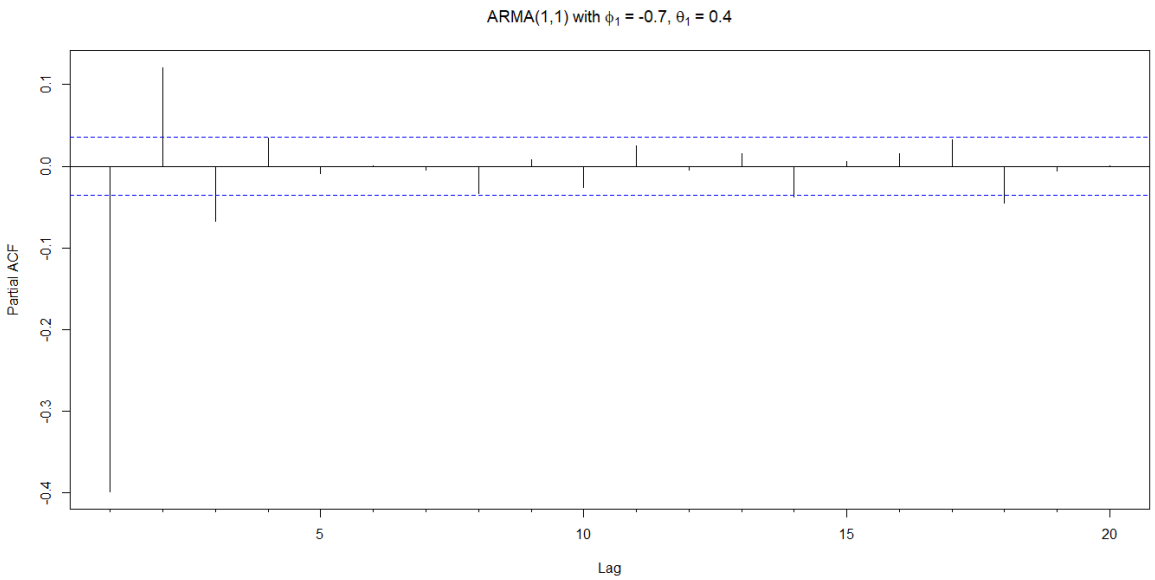
$$\begin{aligned}
\gamma(h) &= \phi_1\gamma(h-1) + \theta_1 \mathrm{E}[X_{t-h}\epsilon_{t-1}] + \mathrm{E}[X_{t-h}\epsilon_t] \\
&= \phi_1\gamma(h-1) \\
&= \dots \\
&= \phi_1^{h-1}\gamma(1) \\
&= \phi_1^{h-1}\sigma^2 \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 - \phi_1^2}.
\end{aligned}$$

Tóm lại ta có

$$\begin{aligned}
\rho(0) &= 1 \\
\rho(1) &= \frac{\gamma(1)}{\gamma(0)} = \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 + 2\phi_1\theta_1 + \theta_1^2} \\
\rho(h) &= \frac{\gamma(h)}{\gamma(0)} = \phi_1^{h-1} \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 + 2\phi_1\theta_1 + \theta_1^2}, h > 1.
\end{aligned}$$



Hình 1.5: Tự tương quan mẫu của một chuỗi thời gian ARMA$(1, 1)$ với $\phi_1 = -0.7, \theta_1 = 0.4$.



Hình 1.6: Tự tương quan riêng mẫu của một chuỗi thời gian ARMA$(1,1)$ với $\phi_1 = -0.7, \theta_1 = 0.4$.

Ước lượng tham số cho mô hình ARMA(p,q)

- Hệ phương trình Yule-Walker (phương pháp ước lượng moment)

  Xét quá trình ARMA(p,q) (1.11)

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

  với chuỗi $\{X_t\}$ có kì vọng 0 và $\{\epsilon_t\} \sim WN(0, \sigma^2)$.

  Nhân hai vế của phương trình (1.11) với với $X_{t-k} (k = 0, 1, 2, \ldots)$ và lấy kì vọng

$$\begin{aligned} &\mathrm{E}[X_{t-k}X_t] + \phi_1 \mathrm{E}[X_{t-k}X_{t-1}] + \ldots + \phi_p \mathrm{E}[X_{t-k}X_{t-p}] \\ =&\mathrm{E}[X_{t-k}\epsilon_t] + \theta_1 \mathrm{E}[X_{t-k}\epsilon_{t-1}] + \ldots + \theta_q \mathrm{E}[X_{t-k}\epsilon_{t-q}]. \end{aligned}$$

  Do chuỗi có kì vọng 0 nên $\mathrm{E}[X_{t-k}X_{t-i}]$ chính là tự hiệp phương sai $\gamma(k-i)$. Để tính $\mathrm{E}[X_{t-k}\epsilon_{t-i}]$, ta sử dụng tính nhân quả của chuỗi để phân tích $X_{t-k} =$

$\sum_{j=0}^{\infty} \psi_j \epsilon_{t-k-j}$. Từ đó,

$$\begin{aligned}
\mathrm{E}[X_{t-k}\epsilon_{t-i}] &= \mathrm{E}\left[\left(\sum_{j=0}^{\infty} \psi_j \epsilon_{t-k-j}\right)\epsilon_{t-i}\right] \\
&= \psi_{i-k}\mathrm{E}[\epsilon_{t-i}^2] \\
&= \psi_{i-k}\sigma^2 \quad (\psi_{i-k} := 0 \text{ nếu } i-k<0).
\end{aligned}$$

Như vậy ta có

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = \sigma^2 \sum_{j=0}^{q} \theta_j \psi_{j-k} \quad (\theta_0 := 1).$$

Xét $k = 0, 1, \ldots, p+q$, thay các tự hiệp phương sai γ bởi tự hiệp phương sai mẫu $\hat{\gamma}$, biểu diễn các hệ số ψ_j theo $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ nhờ đồng nhất hệ số của đẳng thức $\psi(z) = \phi(z)^{-1}\theta(z)$ và giải hệ $p+q+1$ phương trình

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \ldots - \phi_p\hat{\gamma}(k-p) = \sigma^2 \sum_{j=0}^{q} \theta_j \psi_{j-k}, \quad k = 0, 1, \ldots, p+q, \tag{1.23}$$

ta tìm được các tham số $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ và σ^2.

Nhìn chung, hệ phương trình (1.23) là phi tuyến. Do đó, trên thực tế việc giải hệ phương trình để tìm ra các hệ số của mô hình là rất khó khăn và ta thường phải dùng các cách tiếp cận khác để ước lượng các hệ số trong mô hình (chẳng hạn phương pháp ước lượng hợp lý nhất). Tuy nhiên, trong trường hợp đặc biệt $q = 0$, tức mô hình tự hồi quy AR(p) thì hệ phương trình (1.23) là tuyến tính. Hệ p phương trình tuyến tính với $k = 1, 2, \ldots, p$ được viết ngắn gọn dưới dạng ma trận

$$\hat{\boldsymbol{\Gamma}}_p \boldsymbol{\phi}_p = \hat{\boldsymbol{\gamma}}_p \tag{1.24}$$

với $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)^T, \hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))^T$ và

$$\begin{aligned}
\hat{\boldsymbol{\Gamma}}_p &= [\hat{\gamma}(i-j)]_{i,j=1}^{p} \\
&= \begin{bmatrix}
\hat{\gamma}(0) & \hat{\gamma}(-1) & \ldots & \hat{\gamma}(-(p-1)) \\
\hat{\gamma}(1) & \hat{\gamma}(0) & \ldots & \hat{\gamma}(-(p-2)) \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\gamma}(p-1) & \hat{\gamma}(p-1) & \ldots & \hat{\gamma}(0)
\end{bmatrix}.
\end{aligned}$$

Nhờ cấu trúc Toeplitz đặc biệt của ma trận $\hat{\mathbf{\Gamma}}_p$, có thể giải hệ phương trình (1.24) khá nhanh chóng bằng phương pháp truy hồi (thuật toán Levinson - Durbin [7, tr. 69-71]). Sau khi đã ước lượng được các tham số tự hồi quy $\phi_1, \ldots, \phi_p$, có thể tính được phương sai của ồn trắng σ^2 nhờ phương trình đầu tiên ($k = 0$)

$$\begin{aligned}\hat{\gamma}(0) - \phi_1\hat{\gamma}(1) - \ldots - \phi_p\hat{\gamma}(p) &= \mathrm{E}[X_t\epsilon_t] \\ &= \mathrm{E}[(\epsilon_t + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p})\epsilon_t] \\ &= \mathrm{E}[\epsilon_t^2] \quad \text{(Do tính nhân quả)} \\ &= \sigma^2.\end{aligned}$$

- Ước lượng hợp lý nhất (Maximum likelihood)

  Giả sử $\{X_t\}$ là một chuỗi thời gian Gauss (Gaussian time series, là chuỗi thời gian thỏa mãn với mọi $i_1, \ldots, i_n$, vector $\boldsymbol{X} = (X_{i_1}, \ldots, X_{i_n})^T$ có phân phối chuẩn n chiều) với kì vọng 0 và hàm tự hiệp phương sai $\gamma(i, j) = \mathrm{E}[X_i X_j]$. Gọi $\mathbf{\Gamma}_n = \mathrm{E}[\boldsymbol{X}_n \boldsymbol{X}_n^T]$ là ma trận hiệp phương sai của $\boldsymbol{X}_n$.

  Khi đó hàm hợp lý của $\boldsymbol{X}_n$ là

$$L(\mathbf{\Gamma}_n) = (2\pi)^{-n/2}|\mathbf{\Gamma}_n|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{X}_n^T\mathbf{\Gamma}_n^{-1}\boldsymbol{X}_n\right). \tag{1.25}$$

  Để tính hàm hợp lý, ta sẽ biểu diễn $|\mathbf{\Gamma}_n|$ và $\mathbf{\Gamma}_n^{-1}$ thông qua các phần dư $U_j = X_j - \hat{X}_j$ và phương sai $v_{j-1} = \mathrm{Var}(U_j)$ của chúng, trong đó $\hat{X}_j$ là xấp xỉ tuyến tính tốt nhất của X_j theo $X_{j-1}, \ldots, X_1$. Viết $\hat{X}_j$ dưới dạng

$$\hat{X}_j = -a_{j-1,1}X_{j-1} - a_{j-1,2}X_{j-2} - \ldots - a_{j-1,j-1}X_1, \quad j \geq 2.$$

  Đặt $\boldsymbol{U}_n = (U_1, \ldots, U_n)^T$ thì ta có

$$\boldsymbol{U}_n = \boldsymbol{A}_n\boldsymbol{X}_n,$$

  trong đó

$$\boldsymbol{A}_n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ a_{11} & 1 & 0 & \ldots & 0 \\ a_{22} & a_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1,n-1} & a_{n-1,n-2} & a_{n-1,n-3} & \ldots & 1 \end{pmatrix}.$$

  A_n có dạng ma trận tam giác dưới nên ma trận nghịch đảo $\boldsymbol{C}_n = \boldsymbol{A}_n^{-1}$ cũng là

ma trận tam giác dưới:

$$\boldsymbol{C}_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta_{11} & 1 & 0 & \dots & 0 \\ \theta_{22} & \theta_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \dots & 1 \end{pmatrix}$$

và $\boldsymbol{X}_n = \boldsymbol{C}_n \boldsymbol{U}_n = \boldsymbol{C}_n(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)$.

Người ta chứng minh được các phần dư U_j là không tương quan với nhau nên ma trận hiệp phương sai D_n của vector $\boldsymbol{U}_n$ có dạng đường chéo

$$\boldsymbol{D}_n = \begin{pmatrix} v_0 & & 0 \\ & \ddots & \\ 0 & & v_{n-1} \end{pmatrix}.$$

Mà $\boldsymbol{X}_n = \boldsymbol{C}_n \boldsymbol{U}_n$ nên

$$\boldsymbol{\Gamma}_n = \boldsymbol{C}_n \boldsymbol{D}_n \boldsymbol{C}_n^T.$$

Từ đó ta tính được

$$\begin{aligned} |\boldsymbol{\Gamma}_n| &= |\boldsymbol{C}_n||\boldsymbol{D}_n||\boldsymbol{C}_n^T| \\ &= v_0 v_1 \dots v_{n-1} \end{aligned}$$

và

$$\begin{aligned} \boldsymbol{X}_n^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n &= \boldsymbol{U}_n^T \boldsymbol{D}_n^{-1} \boldsymbol{U}_n \\ &= \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{v_{j-1}}. \end{aligned}$$

Vậy

$$L(\boldsymbol{\Gamma}_n) = (2\pi)^{-n/2}(v_0 \dots v_{n-1})^{-1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{v_{j-1}} \right).$$

Đặt $r_j = \dfrac{v_j}{\sigma^2}$ thì

$$L(\boldsymbol{\Gamma}_n) = (2\pi\sigma^2)^{-n/2}(r_0 \dots r_{n-1})^{-1/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \right).$$

Các r_j và $\hat{X}_j$ có thể biểu diễn qua các hệ số $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T, \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ mà không phụ thuộc σ^2 nhờ thuật toán Innovations [7, tr. 71-73, 100-101]

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-n/2}(r_0(\boldsymbol{\phi}, \boldsymbol{\theta}) \ldots r_{n-1}(\boldsymbol{\phi}, \boldsymbol{\theta}))^{-1/2} \exp\left(- \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2} \right)$$

với $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{n} \dfrac{(X_j - \hat{X}_j)^2}{r_{j-1}}$. Để tìm cực đại, ta lấy ln của hàm hợp lý

$$\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = -\frac{n}{2} \ln (2\pi\sigma^2) - \frac{1}{2} \sum_{j=1}^{n} \ln (r_{j-1}) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2}.$$

Xét đạo hàm riêng theo σ^2

$$\begin{aligned} 0 = \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2(\sigma^2)^2} \\ \Rightarrow \sigma^2 &= \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n}. \end{aligned} \tag{1.26}$$

Để tìm các hệ số $\boldsymbol{\phi}, \boldsymbol{\theta}$ ta thay $\sigma^2 = \dfrac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n}$ vào hàm log-likelihood

$$\begin{aligned} \ln L(\boldsymbol{\phi}, \boldsymbol{\theta}) &= -\frac{n}{2} \ln \left(2\pi \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right) - \frac{1}{2} \sum_{j=1}^{n} \ln r_{j-1} - \frac{n}{2} \\ &= -\frac{n}{2} \underbrace{\left(\ln \left(\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right) + \frac{\sum_{j=1}^{n} \ln r_{j-1}}{n} \right)}_{l(\boldsymbol{\phi}, \boldsymbol{\theta})} \underbrace{- \frac{n}{2} \ln (2\pi) - \frac{n}{2}}_{\text{const}}. \end{aligned}$$

Như vậy để tìm cực đại hàm $\ln L(\boldsymbol{\phi}, \boldsymbol{\theta})$ ta tìm cực tiểu của hàm $l(\boldsymbol{\phi}, \boldsymbol{\theta})$

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}) = \ln \left(\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right) + \frac{\sum_{j=1}^{n} \ln r_{j-1}}{n}. \tag{1.27}$$

Trong thực tế, việc tìm nghiệm cực tiểu $l(\boldsymbol{\phi}, \boldsymbol{\theta})$ dựa vào các phương pháp số.

1.4.4 Quá trình ARIMA(p, d, q)

Xét chuỗi thời gian $\{X_t\}_{t \in \mathbb{Z}}$ có kì vọng bằng 0. $\{X_t\}$ được gọi là một quá trình ARIMA(p, d, q) nếu nó thỏa mãn

$$\phi(B) \nabla^d X_t = \theta(B) \epsilon_t, \tag{1.28}$$

trong đó $Y_t = \nabla^d X_t$ là một chuỗi thời gian dừng, $\{\epsilon_t\} \sim WN(0, \sigma^2)$,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q.$$

Mô hình ARIMA(p, d, q) là mở rộng của mô hình ARMA(p, q) áp dụng trong trường hợp chuỗi thời gian không là chuỗi dừng. Mục đích của việc lấy sai phân cấp d ∇^d là để khử tính xu thế, làm cho chuỗi thời gian thỏa mãn tính chất dừng.

Một chuỗi thời gian có thành phần xu thế là một đa thức bậc d theo thời gian t thì sau khi lấy sai phân ∇^d, đa thức bậc d đó sẽ được khử về hằng số. Thật vậy, trước tiên dễ nhận thấy rằng $\nabla^k t^d = 0$ nếu $k > d$. Tiếp theo ta sẽ chỉ ra $\nabla^d t^d = d!$ bằng quy nạp. Với $d = 1$, $\nabla t = t - (t-1) = 1 = 1!$. Giả sử $\nabla^d t^d = d!$,

$$\begin{aligned}
\nabla^{d+1} t^{d+1} &= \nabla^d (t^{d+1} - (t-1)^{d+1}) \\
&= \nabla^d \left[t^{d+1} - \sum_{k=0}^{d+1} \binom{d+1}{k} (-1)^{d+1-k} t^k \right] \\
&= \nabla^d \left[\sum_{k=0}^{d} \binom{d+1}{k} (-1)^{d-k} t^k \right] \\
&= \nabla^d \left[\binom{d+1}{d} t^d \right] \\
&= (d+1) d! = (d+1)!,
\end{aligned}$$

như vậy, $\nabla^d t^d = d!$ với mọi số nguyên dương d. Xét chuỗi thời gian $X_t = m_t + Y_t$, với thành phần xu thế m_t là một đa thức bậc d theo thời gian t, Y_t là một chuỗi thời gian dừng có kì vọng 0. m_t có dạng

$$m_t = c_0 + c_1 t + c_2 t^2 + \ldots + c_d t^d.$$

Lấy sai phân cấp d ta thu được

$$\nabla^d X_t = d! c_d + \nabla^d Y_t,$$

là một quá trình dừng với trung bình $d! c_d$.

Lưu ý. Từ quá trình sau khi đã lấy sai phân $Y_t = \nabla X_t$ ta có thể khôi phục lại chuỗi

ban đầu X_t

$$\begin{aligned}
&\nabla X_t = Y_t \\
\Rightarrow &X_t = X_{t-1} + Y_t \\
&= X_{t-2} + Y_{t-1} + Y_t \\
&= X_{t-3} + Y_{t-2} + Y_{t-1} + Y_t \\
&= \dots
\end{aligned}$$

Trên thực tế, ta chỉ có thể quan sát một chuỗi thời gian bắt đầu từ thời điểm t_0 nào đó. Khi đó,

$$X_t = X_{t_0} + Y_{t_0+1} + Y_{t_0+2} + \ldots + Y_t. \tag{1.29}$$

1.5 Quá trình SARIMA$(p, d, q) \times (P, D, Q)_s$

Một chuỗi thời gian dừng $\{X_t\}_{t\in\mathbb{Z}}$ được gọi là một quá trình SARIMA$(p, d, q) \times (P, D, Q)_s$ nếu nó thỏa mãn

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D X_t = \theta(B)\Theta(B^s)\epsilon_t, \tag{1.30}$$

trong đó $Y_t = \nabla^d\nabla_s^D X_t$ là chuỗi dừng và $\{\epsilon_t\} \sim WN(0, \sigma^2)$,

$$\begin{aligned}
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q, \\
\Phi(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}, \\
\Theta(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs}.
\end{aligned}$$

Mô hình SARIMA$(p, d, q) \times (P, D, Q)_s$ được áp dụng cho những chuỗi thời gian không dừng và có những dấu hiệu của thành phần mùa với chu kì s.

Có thể nhận thấy $\phi(B)\Phi(B^s)$ là một đa thức bậc $p + Ps$ và $\theta(B)\Theta(B^s)$ là một đa thức bậc $q + Qs$. Như vậy ta có thể coi chuỗi $\{Y_t\}$ là một quá trình ARMA với bậc tự hồi quy là $p + Ps$, bậc trung bình trượt là $q + Qs$. Việc xét các tính chất dừng, nhân quả, khả nghịch hay ước lượng tham số cho mô hình có thể áp dụng trực tiếp từ phần mô hình ARMA.

Chương 2

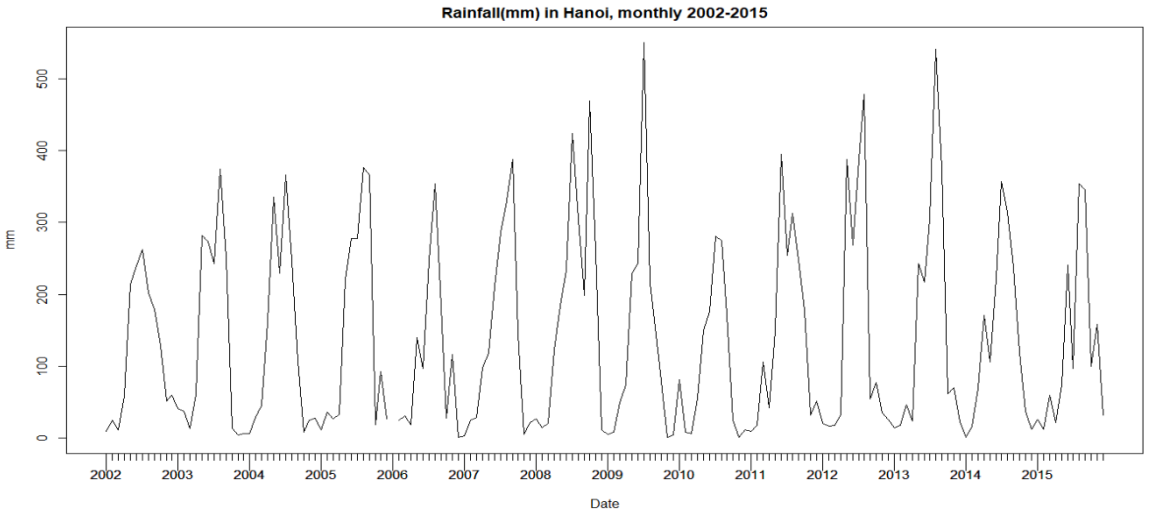
Xử lý chuỗi thời gian có yếu tố mùa sử dụng mô hình SARIMA

Trong phần này, ta sẽ tiến hành ứng dụng mô hình SARIMA để phân tích, xử lý chuỗi thời gian có yếu tố mùa trong thực tế. Bộ dữ liệu là lượng mưa các tháng trong năm đo tại trạm quan trắc Hà Nội, từ tháng 1 năm 2002 đến tháng 12 năm 2015 (Nguồn: Tổng cục Thống Kê [13]).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2002	9.0	24.0	11.0	59.0	214.0	240.0	262.0	202.0	179.0	128.0	51.0	60.0
2003	41.0	37.0	13.0	61.0	282.0	274.0	243.0	375.0	251.0	13.0	4.0	6.0
2004	6.0	29.0	45.0	161.0	335.0	229.0	366.0	247.0	107.0	8.0	24.0	28.0
2005	11.0	36.0	27.0	33.0	221.0	278.0	278.0	377.0	366.0	18.0	92.0	27.0
2006	NA	25.0	31.0	18.0	140.0	97.0	247.0	354.0	183.0	28.0	116.0	1.0
2007	3.0	25.0	29.0	98.0	118.0	211.0	286.0	330.0	388.0	145.0	5.0	21.0
2008	27.0	14.0	20.0	122.0	184.0	234.0	424.0	305.0	199.0	469.0	259.0	11.0
2009	4.9	8.0	49.1	74.3	229.0	242.4	550.5	215.7	154.6	78.8	1.2	3.6
2010	80.9	8.1	5.8	55.6	149.7	175.4	280.4	274.4	171.8	24.9	0.6	11.6
2011	9.3	17.5	105.8	42.0	149.0	395.5	254.4	313.2	247.6	177.6	31.8	51.5
2012	20.3	16.5	16.9	31.8	387.7	268.9	388.3	478.1	54.7	77.5	34.8	25.7
2013	13.8	17.7	46.1	23.3	242.5	216.7	305.9	541.4	374.3	61.2	69.6	22.2
2014	0.7	16.1	68.6	170.4	106.1	221.7	357.3	314.7	237.3	119.4	36.5	11.8
2015	25.6	12.5	59.4	21.6	74.2	241.1	96.8	354.2	345.4	99.7	158.0	31.5

Hình 2.1: Lượng mưa hàng tháng (mm) tại Hà Nội, từ 1/2002 đến 12/2015.

Dưới đây là biểu đồ chuỗi thời gian mô tả dữ liệu trên (xem hình 2.2):

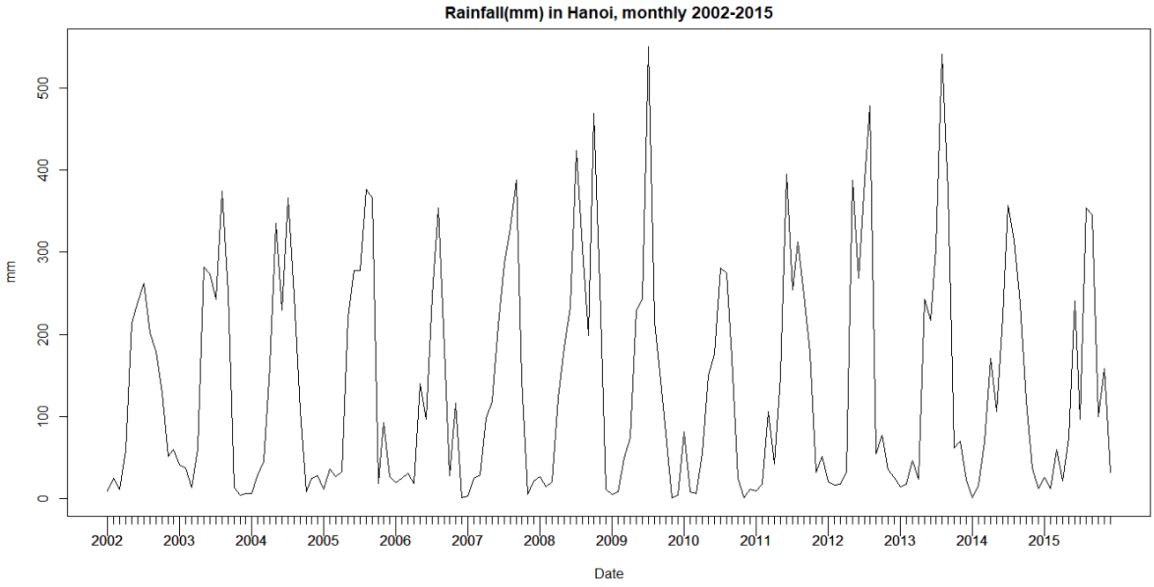


Hình 2.2: Biểu đồ mô tả số liệu lượng mưa hàng tháng tại Hà Nội từ 1/2002 đến 12/2015.

2.1 Tiền xử lý số liệu

Trước khi đi vào chi tiết phân tích, xử lý số liệu, ta cần phải tiền xử lý, làm sạch bộ dữ liệu. Từ bảng số liệu cũng như biểu đồ mô tả dữ liệu ở trên, ta có thể nhận thấy số liệu ở tháng 1 năm 2006 bị thiếu. Do chỉ có một số liệu bị thiếu nên ta sẽ xử lý bằng phương pháp đơn giản là cho giá trị của 1/2006 bằng trung bình cộng của tất cả các tháng 1 của các năm còn lại.

Dưới đây là biểu đồ chuỗi thời gian sau khi đã lấp đầy chỗ thiếu (xem hình 2.3):

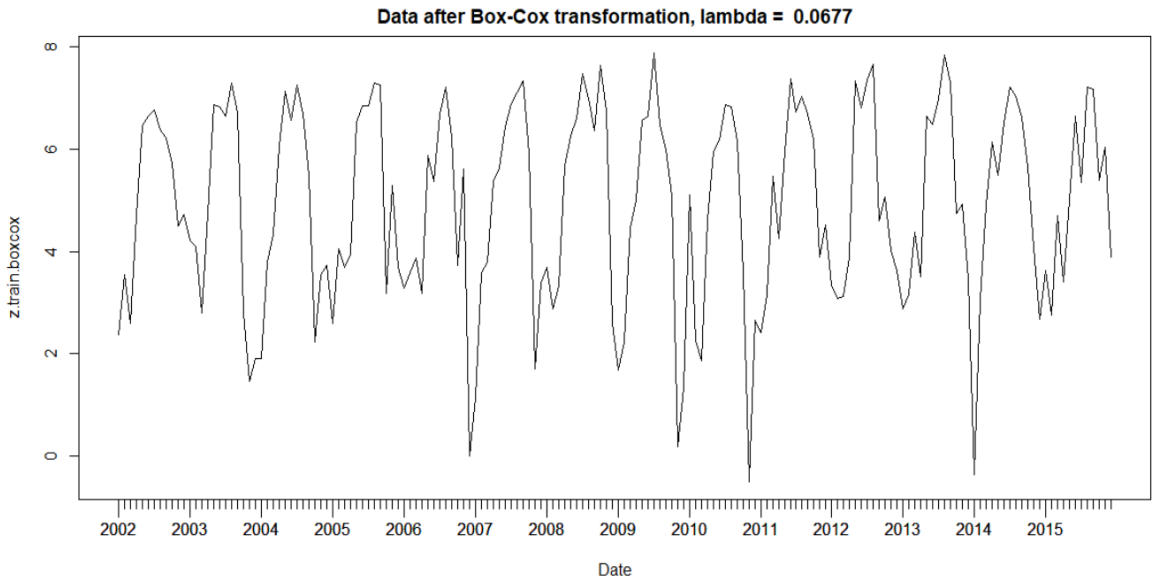


Hình 2.3: Biểu đồ lượng mưa hàng tháng tại Hà Nội từ 1/2002 đến 12/2015 (sau khi lấp đầy dữ liệu thiếu).

Tiếp theo, để giảm sự biến động của phương sai theo thời gian, ta sử dụng biến đổi Box-Cox [3]

$$f_\lambda(X_t) = \begin{cases} \ln(X_t), & \text{với } \lambda = 0. \\ \dfrac{X_t^\lambda - 1}{\lambda}, & \text{với } \lambda > 0. \end{cases} \tag{2.1}$$

Chuỗi thời gian sau khi biến đổi Box-Cox với $\lambda = 0.0677$ (hệ số λ được chọn tự động nhờ phương pháp do Victor M.Guerrero đề xuất [10]) (xem hình 2.4):



Hình 2.4: Chuỗi thời gian sau khi biến đổi Box-Cox.

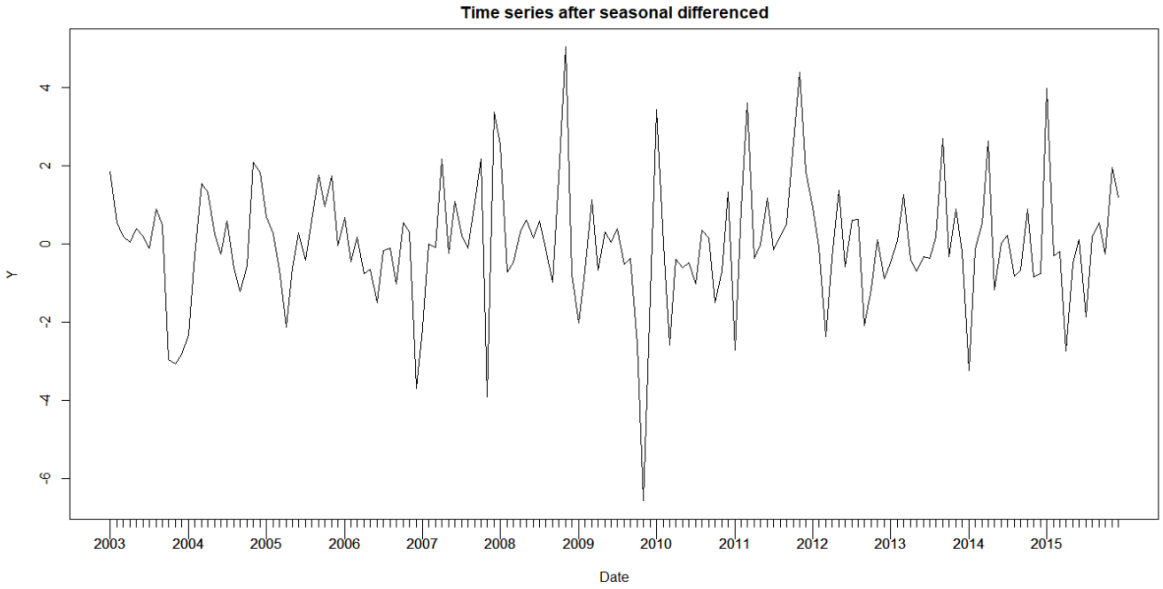
2.2 Dừng hóa chuỗi thời gian

Dựa vào biểu đồ số liệu trên, có thể nhận thấy rõ tính chu kì (năm) của lượng mưa theo từng tháng: Lượng mưa tăng cao đạt đỉnh điểm vào các tháng 6,7,8 và giảm xuống thấp nhất vào tầm tháng 11 đến tháng 2 hàng năm. Do vậy, mô hình ARIMA kết hợp yếu tố mùa (SARIMA) với chu kì bằng 12 tháng (1 năm) khá phù hợp để mô hình hóa chuỗi thời gian này.

Trước tiên, ta lấy sai phân trễ 12 để khử thành phần mùa của chuỗi thời gian

$$Y_t = \nabla_{12} X_t = X_t - X_{t-12} \tag{2.2}$$

Hình 2.5 là biểu đồ của Y_t là chuỗi sau khi đã lấy sai phân trễ 12:

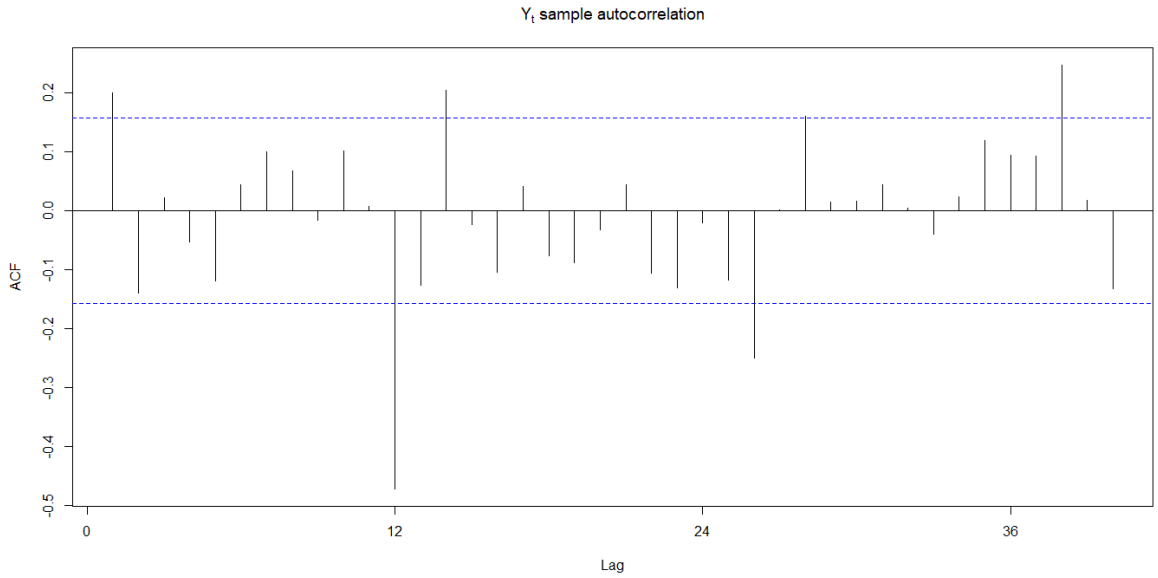


Hình 2.5: Biểu đồ chuỗi thời gian $Y_t = \nabla_{12} X_t = X_t - X_{t-12}$.

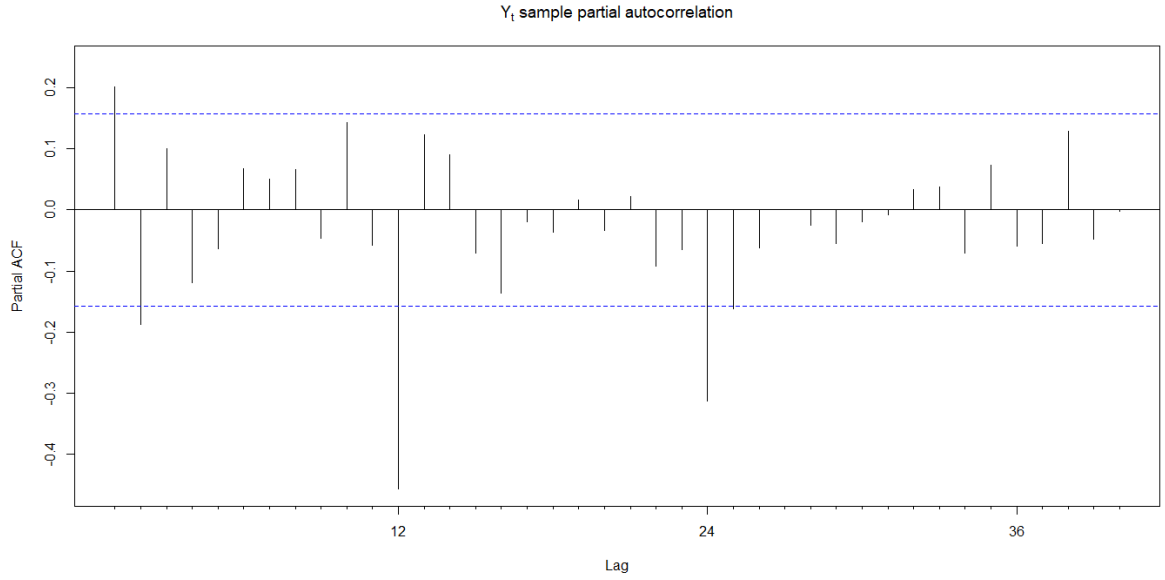
Từ biểu đồ chuỗi Y_t, ta khó có thể nhận ra một xu thế (trend) rõ ràng nào cả. Hơn nữa, tiến hành phép thử Augmented Dickey-Fuller [11] kiểm định giả thuyết H_0 : $\phi(z) = 0$ có nghiệm nằm trên đường tròn đơn vị và đối thuyết H_1 : chuỗi thời gian là chuỗi dừng ta thu được p-value nhỏ hơn 0.05. Như vậy ta có thể coi chuỗi $Y_t = \nabla_{12} X_t$ đã thỏa mãn tính dừng.

2.3 Chọn bậc p, q, P, Q cho mô hình

Hình 2.6 và 2.7 mô tả đồ thị hàm tự tương quan mẫu $r(h)$ và tự tương quan riêng mẫu $\hat{\alpha}(h)$:



Hình 2.6: Tự tương quan mẫu $r(h)$ của chuỗi $Y_t = \nabla_{12} X_t$.



Hình 2.7: Tự tương quan riêng mẫu $\hat{\alpha}(h)$ của chuỗi $Y_t = \nabla_{12} X_t$.

Từ hình 2.6 và 2.7, có thể nhận thấy các giá trị $r(12), \hat{\alpha}(12), \hat{\alpha}(24)$ có độ lớn khá đáng kể (nằm ngoài khoảng $\pm\dfrac{1.96}{\sqrt{n}}$ với n là số lượng dữ liệu quan sát của Y_t [6, tr. 215-216]). Như vậy, các mô hình với bậc $P \leq 2, Q \leq 1$ sẽ phù hợp đối với bộ dữ liệu này. Đồng thời, các giá trị $r(1), \hat{\alpha}(1), \hat{\alpha}(2)$ và một số giá trị $r, \hat{\alpha}$ khác gần kề các trễ là bội của 12 cũng có độ lớn đáng kể, do đó ta cũng nên đưa thêm bậc p, q vào mô hình. Lần

lượt thử các mô hình của X_t với $p = 0, 1, 2; d = 0; q = 0, 1; P = 0, 1, 2; Q = 0, 1; D = 1$ và tính giá trị của tiêu chuẩn AICc (corrected Akaike Information Criterion)

$$AICc = -2 \ln L + \frac{2(m+1)n}{n-m-2} \tag{2.3}$$

với n là số quan sát của chuỗi $\{Y_t\}$ và m là số các hệ số AR và MA khác 0 trong mô hình [7, tr. 171-174]. Mô hình sẽ được ưu tiên chọn nếu mô hình đó có giá trị AICc nhỏ nhất. Kết quả thu được là mô hình SARIMA $(1, 0, 1) \times (0, 1, 1)_{12}$ với giá trị AICc là 504.4043.

2.4 Ước lượng các tham số

Mô hình SARIMA $(1, 0, 1) \times (0, 1, 1)_{12}$

$$(1 - \phi_1 B)\nabla_{12} X_t = (1 - \phi_1 B) Y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t.$$

Sử dụng phương pháp ước lượng hợp lý nhất (maximum likelihood), ta thu được ước lượng các tham số:

$$\begin{aligned} \hat{\phi}_1 &\approx -0.4818, \\ \hat{\theta}_1 &\approx 0.6781, \\ \hat{\Theta}_1 &\approx -0.9996, \\ \hat{\sigma}^2 &\approx 1.172, \end{aligned}$$

có thể nhận thấy hệ số $\hat{\Theta}_1$ rất gần với -1, dẫn tới nghiệm của phương trình $(1+\theta_1 z)(1+\Theta_1 z^{12}) = 0$ nằm ngay sát đường tròn đơn vị. Ta sẽ loại bỏ không dùng những mô hình như thế này. Quay lại bước chọn bậc mô hình ở trên và thử những mô hình khác có giá trị AICc nhỏ, đồng thời nghiệm của phương trình $\phi(z)\Phi(z^{12}) = 0$ và $\theta(z)\Theta(z^{12}) = 0$ đều nằm ngoài và không quá sát hình tròn đơn vị để đảm bảo tính nhân quả và tính khả nghịch. Kết quả thu được là mô hình SARIMA$(1, 0, 1) \times (2, 1, 0)_{12}$ với giá trị AICc là 521.7218

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})\nabla_{12} X_t = (1 - \phi_1 B)(1 - \Phi_1 B^{12} - \Phi_2 B^{24}) Y_t = (1 + \theta_1 B)\epsilon_t. \tag{2.4}$$

Các tham số ước lượng được bằng phương pháp ước lượng hợp lý nhất như sau:

$$\begin{aligned}\hat{\phi}_1 &\approx -0.4033,\\ \hat{\theta}_1 &\approx 0.6249,\\ \hat{\Phi}_1 &\approx -0.6679,\\ \hat{\Phi}_2 &\approx -0.3216,\\ \hat{\sigma}^2 &\approx 1.531.\end{aligned}$$

2.5 Kiểm tra tính phù hợp của mô hình

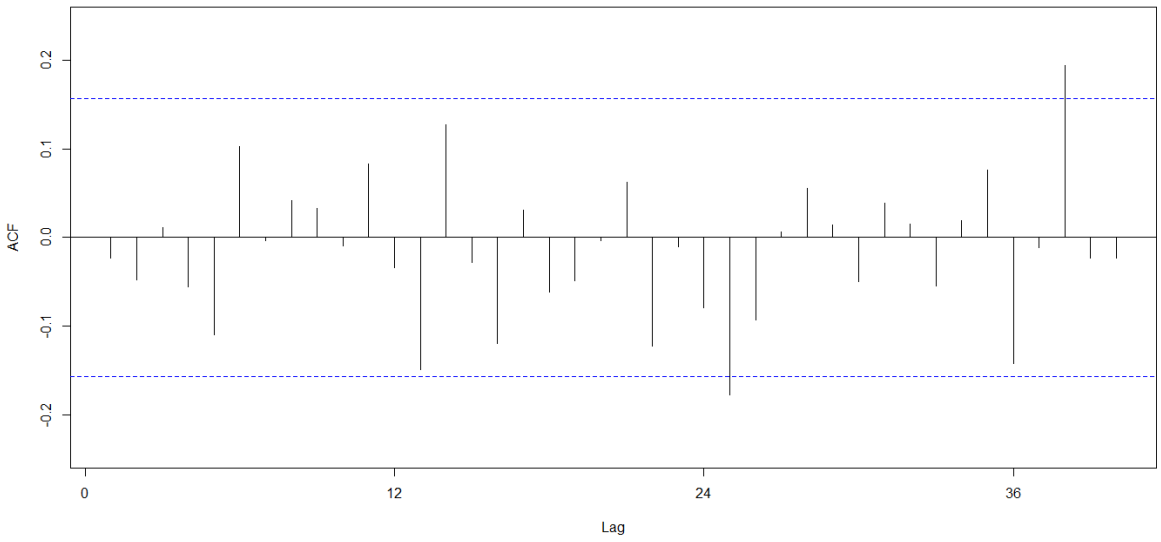
Từ phương trình (2.4)

$$\begin{aligned}&(1-\phi_1 B)(1-\Phi_1 B^{12}-\Phi_2 B^{24})Y_t = (1+\theta_1 B)\epsilon_t\\ \Leftrightarrow&(1-\phi_1 B-\Phi_1 B^{12}+\phi_1\Phi_1 B^{13}-\Phi_2 B^{24}+\phi_1\Phi_2 B^{25})Y_t = (1+\theta_1 B)\epsilon_t\\ \Leftrightarrow&Y_t-\phi_1 Y_{t-1}-\Phi_1 Y_{t-12}+\phi_1\Phi_1 Y_{t-13}-\Phi_2 Y_{t-24}+\phi_1\Phi_2 Y_{t-25} = \epsilon_t+\theta_1\epsilon_{t-1}\\ \Leftrightarrow&Y_t = \phi_1 Y_{t-1}+\Phi_1 Y_{t-12}-\phi_1\Phi_1 Y_{t-13}+\Phi_2 Y_{t-24}-\phi_1\Phi_2 Y_{t-25}+\theta_1\epsilon_{t-1}+\epsilon_t.\end{aligned}$$

Đặt

$$\hat{Y}_t = \hat{\phi}_1 Y_{t-1}+\hat{\Phi}_1 Y_{t-12}-\hat{\phi}_1\hat{\Phi}_1 Y_{t-13}+\hat{\Phi}_2 Y_{t-24}-\hat{\phi}_1\hat{\Phi}_2 Y_{t-25}+\hat{\theta}_1\hat{\epsilon}_{t-1} \tag{2.5}$$

thì phần dư (residuals) $\hat{\epsilon}_t = Y_t - \hat{Y}_t$. Nếu mô hình đúng đắn, chuỗi phần dư $\{\hat{\epsilon}_t\}$ sẽ xấp xỉ là một ồn trắng (thông thường có thể coi tuân theo phân phối chuẩn). Dưới đây là tự tương quan mẫu của chuỗi phần dư (xem hình 2.8)



Hình 2.8: Tự tương quan mẫu của chuỗi phần dư.

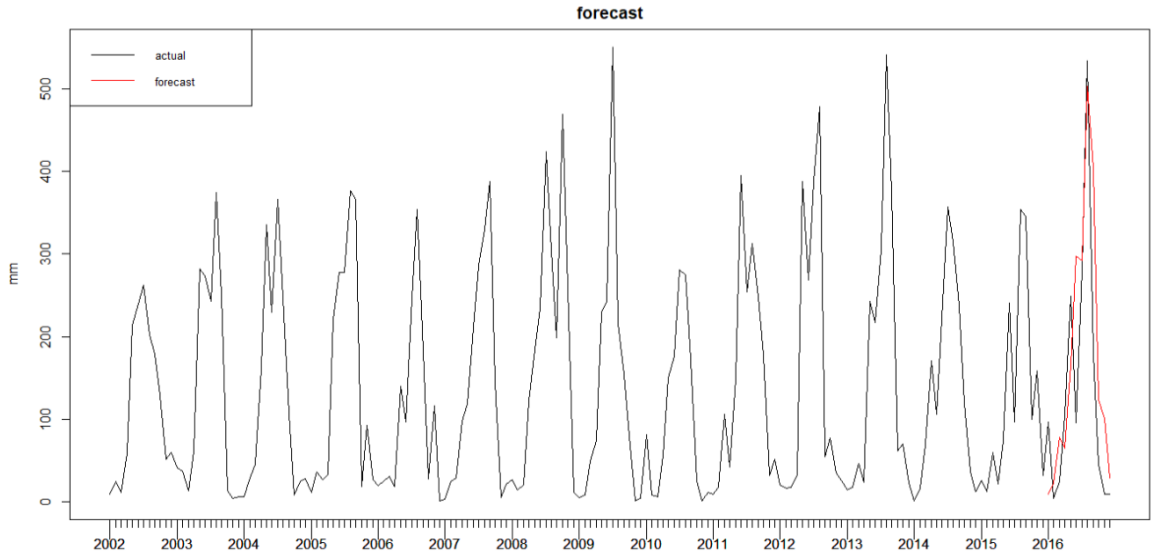
Ta thấy chỉ có 2 giá trị nằm ngoài khoảng $\pm\frac{1.96}{\sqrt{n}}$ trong 40 trễ đầu tiên ($\leq 5\%$). Hơn nữa, tiến hành phép kiểm định Ljung-Box [5] kiểm tra tính không tương quan của chuỗi phần dư ta thu được kết quả

```
> Box.test(residuals, lag = 24, type = "Ljung-Box", fitdf = 4)

        Box-Ljung test

data:  residuals
X-squared = 21.781, df = 20, p-value = 0.3525
```

Giá trị p-value thu được là $0.3525 > 0.05$ nên ta không có cơ sở bác bỏ giả thuyết gốc. Vậy ta có thể coi chuỗi phần dư là ồn trắng và mô hình đưa ra là phù hợp.

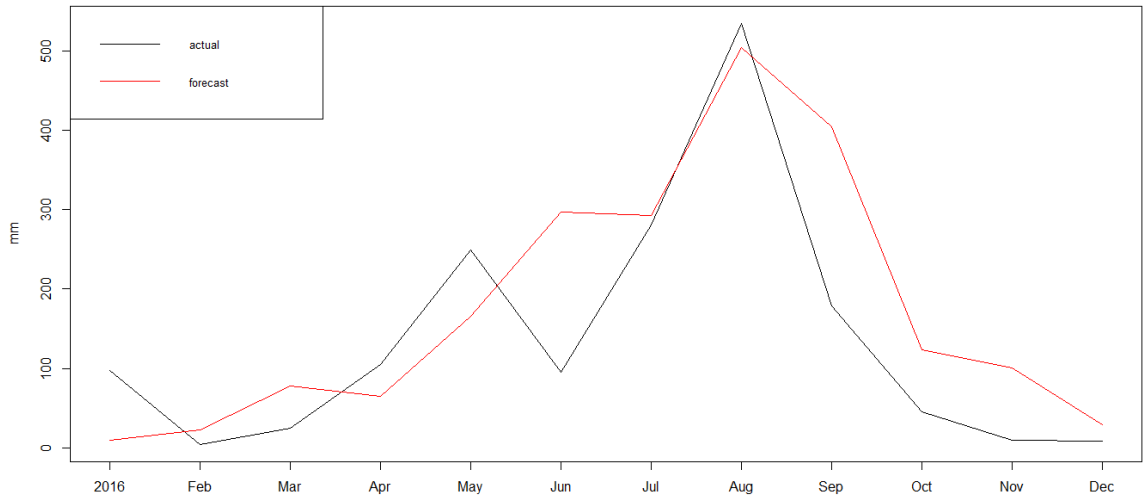
2.6 Đưa ra dự báo

Dự báo lượng mưa từng tháng tại Hà Nội trong năm tiếp theo 2016 dựa vào mô hình SARIMA$(1,0,1)\times(2,1,0)_{12}$ (2.4) được mô tả trong hình 2.9 dưới đây. Đường màu đỏ là giá trị dự báo, còn đường màu đen là giá trị thực tế.



Hình 2.9: Dự báo chuỗi thời gian trong năm tiếp theo.

So sánh lượng mưa thực tế và kết quả dự báo:



Hình 2.10: Biểu đồ minh họa.

Tháng / Giá trị(mm)	1	2	3	4	5	6	7	8	9	10	11	12
Thực tế	97	4	25	105	249	95	280	535	179	45	9	9
Dự báo	9	23	78	65	166	297	292	504	405	123	101	29

Bảng 2.1: Bảng số liệu.

Có thể nhận thấy mô hình SARIMA dự báo tương đối chính xác lượng mưa tăng cao vào mùa mưa (tháng 8) và lượng mưa thấp tại các tháng mùa khô (tháng 12,1,2). Tuy nhiên, tại một số tháng lượng mưa thực tế đột ngột biến đổi (tháng 6, tháng 9) thì kết quả dự báo của mô hình lại chưa phản ánh được điều này.

Kết luận

Trong đồ án này, em đã trình bày những khái niệm cơ bản của chuỗi thời gian và một số mô hình chuỗi thời gian như AR, MA, ARMA, ARIMA và SARIMA. Em cũng phân tích, trình bày phương pháp để xác định bậc của mô hình dựa vào hàm tự tương quan mẫu, tự tương quan riêng mẫu, một số phương pháp để ước lượng tham số cho mô hình như ước lượng moment, ước lượng hợp lý nhất. Cuối cùng là ứng dụng mô hình SARIMA để xử lý chuỗi thời gian lượng mưa hàng tháng tại trạm quan trắc Hà Nội có tính mùa. Kết quả dự báo tương đối chính xác, cho thấy việc áp dụng mô hình SARIMA đối với chuỗi thời gian này là hợp lý. Tuy nhiên, có thể nhận thấy hạn chế của mô hình đã xây dựng là chưa phản ánh tốt lượng mưa tại một số tháng thay đổi một cách bất thường so với mọi năm (tháng 1, tháng 6, tháng 9, tháng 11, trong đó sai số lượng mưa dự báo so với thực tế lên tới hơn 200mm ở tháng 6 và tháng 9). Hướng mở rộng nghiên cứu, phát triển: kết hợp những nhân tố khác ảnh hưởng tới lượng mưa như nhiệt độ khí quyển, áp suất không khí, độ ẩm, v.v. để phân tích, dự báo lượng mưa.

Tài liệu tham khảo

[1] Hoàng Tụy, *Hàm thực và Giải tích Hàm, in lần thứ năm*, NXB Đại học Quốc gia Hà Nội, 2005.

[2] Nguyễn Hồ Quỳnh, *Chuỗi thời gian: Phân tích và nhận dạng*, NXB Khoa học và Kĩ thuật, 2004.

[3] George E.P.Box, David R.Cox, "An analysis of transformations", *Journal of the Royal Statistical Society*, Series B, Vol 26 (Issue 2), 1964, pp. 211–252.

[4] George E.P.Box, Gwilym M.Jenkins, *Time Series Analysis: forecasting and control*, Holden-Day, 1970.

[5] Greta M.Ljung, George E.P.Box, "On a measure of lack of fit in time series models", *Biometrika*, Vol 65(Issue 2), 1978, pp. 297–303.

[6] Peter J.Brockwell, Richard A.Davis, *Time series: Theory and Methods*, Springer, 1987.

[7] Peter J.Brockwell, Richard A.Davis, *Introduction to Time Series and Forcasting, Second Edition*, Springer, 2002.

[8] Robert H.Shumway, David S.Stoffer, *Times Series Analysis and Its Applications With R Examples, Third Edition*, Springer, 2011.

[9] Rob J.Hyndman, Yeasmin Khandakar, "Automatic time series forecasting: the forecast package for R", *Journal of Statistical Software*, Vol 26(Issue 3), 2008, pp. 1–22. `http://www.jstatsoft.org/article/view/v027i03`

[10] Victor M.Guerrero, "Time-series analysis supported by power transformations", *Journal of Forecasting*, Vol 12(Issue 1), 1993, pp. 37–48.

[11] Wayne A.Fuller, *Introduction to Statistical Time Series, second ed.*, New York: John Wiley and Sons, 1996.

[12] Trang chủ dự án R:
`https://www.r-project.org/`

[13] Tổng cục thống kê:
`https://www.gso.gov.vn/`