Aula 7: Identificação de um modelo de séries temporais

Apresentação

Nesta aula apresentaremos o procedimento de identificação de um modelo de séries temporais, baseado nas estimativas das funções de autocorrelação e de autocorrelação parcial obtidas a partir das observações da série.

A partir da comparação entre essas estimativas e os respectivos padrões teóricos correspondentes a cada modelo - estudados na aula 6 - pode-se identificar as ordens de um modelo preliminar para a série, como etapa inicial da modelagem.

Objetivos

- Estimar a função de autocorrelação (FAC) a partir dos dados;
- Estimar a função de autocorrelação parcial (FACP) a partir dos dados;
- Identificar um modelo inicial para a série, a partir das estimativas da FAC e da FACP.

A metodologia de Box e Jenkins

A metodologia proposta por Box e Jenkins para a análise de uma série temporal consiste basicamente em quatro etapas sequenciais:

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

Identificação

↓

Estimação

↓

Diagnóstico

↓

Diagnóstico

Cada uma dessas etapas será estudada nesta e nas próximas aulas, mas é importante que fique claro que, em alguns casos, pode haver necessidade de um retorno à etapa inicial.

Exemplo

Se na fase de diagnóstico ficar comprovado que o modelo escolhido é inadequado, pode-se voltar à etapa de identificação e reiniciar todo o processo.

A identificação não pretende definir o modelo final para representar a série. Se assim fosse, não haveria necessidade das outras etapas de modelagem. O objetivo é apenas definir os valores de p e q que, inicialmente, sejam adequados para modelar a série.

Inicialmente porque, segundo os autores do método, este diagnóstico costuma ser bastante parcimonioso, ou seja, conduzir a um modelo subespecificado, menor do que o modelo verdadeiro para a série.

A identificação da ordem do modelo será baseada nos seguintes resultados, obtidos na aula 6:

- A FAC de um MA(q) é "truncada" no *lag* q, para qualquer q = 1, 2, 3, ...
- A FACP de um AR(p) é "truncada" no *lag* p, para qualquer p = 1, 2, 3, ...

Assim, por meio da comparação das estimativas da FAC e da FACP com os padrões acima, será possível deduzir se temos um modelo AR, MA ou um ARMA e, a partir da inspeção do *lag* em que ocorre o "truncamento", qual o valor de p e de q que, em princípio, parece apropriado.

O quadro a seguir relembra os padrões de FAC e FACP esperados para os modelos teóricos.

AR(1)	$Y_t = \phi_1 Y_{t-1} + \varepsilon_t$	$\rho_k = \phi_1^k$	Truncada no lag 1
AR(2)	$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + +\varepsilon_t$	$\rho_1 = \frac{\phi_1}{(1-\phi_2)}$ $\rho_2 = \frac{\phi_1^2}{(1-\phi_2)} + \phi_2$ etc.	Truncada no lag 2
MA(1)	$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1}$	$\rho_1 = \frac{-\theta_1}{(1+\theta_1^2)}.$ $\rho_k = 0, k = 2, 3, \ldots$	Infinita, decai exponencialmente ou como uma senoide amortecida
MA(2)	$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$	$\rho_1 = \frac{\theta_1(\theta_2 - 1)}{1+\theta_1^2+\theta_2^2}$ $\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$ $\rho_k = 0, k = 3,4,\ldots.$	Infinita, decai exponencialmente ou como uma senoide amortecida

Se observarmos truncamento em ambas as funções, então devemos estar diante de um modelo ARMA com p e q diferentes de zero. Cabe ressaltar, porém, que neste caso a identificação pode não ser tão evidente, uma vez que, como afirmado ao final da aula 6, a uma estrutura ARMA corresponde uma mistura dos padrões do AR e do MA; portanto, o truncamento estará "disfarçado" por uma estrutura de decaimento senoidal ou exponencial.

Clique no botão acima.

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

FACV e FAC para o modelo MA

Para cada *lag* k, calcula-se a correlação amostral entre Y_t e Y_{t-k}, da seguinte forma:

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T} Y_t - Y Y_{t-k} - Y}{\sum_{t=1}^{T} Y_t - Y^2}, k = 0, \quad 1, \quad 2, \quad ...$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

em que Y é a média amostral da série.

Vamos exercitar?

Considere uma série temporal com as seguintes observações: $Y_1 = 8$, $Y_2 = 4$, $Y_3 = 6$, $Y_4 = 4$ e $Y_5 = 8$. Estime a função de autocorrelação.

Solução:

No *lag* k = 0, sempre: $\hat{p}_0$=1.

No *lag* 1:

$$\hat{p}_1 = \frac{\sum_{t=2}^{5} Y_t - YY_{t-k} - Y}{\sum^{T}{}_{t=1} Y_t - Y^2} =$$

$$\frac{Y_2 - YY_1 - Y + Y_3 - YY_2 - Y + Y_4 - YY_3 - Y + Y_5 - YY_4 - Y}{Y_1 - Y^2 + Y_1 - Y^2 + Y_3 - Y^2 + Y_4 - Y^2 + Y_5 - Y^2} =$$

$$\frac{-22 + 0\text{-}2 + -20 + 2\text{-}2}{2^2 + -2^2 + 0^2 + -2^2 + 2^2} = \frac{-4 + 0 + 0 + -4}{4 + 4 + 0 + 4 + 4} = -\frac{8}{16} = -\frac{1}{2}$$

Nos *lags* 2, 3 e 4:

$$\hat{p}_2 = \frac{\sum_{t=3}^{5} Y_t - Y\,)\,Y_t - k - Y}{\sum^{T}{}_{t=1} Y_t - Y^2} =$$

$$\frac{Y_3 - YY_1 - Y + Y_4 - YY_2 - Y + Y_5 - YY_3 - Y}{Y_1 - Y^2 + Y_2 - Y^2 + Y_3 - Y^2 + Y_4 - Y^2 + Y_5 - Y^2} =$$

$$\frac{6 - 6\,)\;(8 - 6 + 4 - 6\;(4 - 6\,)\; + 8 - 66 - 6}{8 - 6^2 + 4 - 6^2 + 6 - 6^2 + 4 - 6^2 + 8 - 6^2} =$$

$$\frac{02 + -2\text{-}2 + 20}{2^2 + -2^2 + 0^2 + -2^2 + 2^2} = \frac{0 + 4 + 0}{4 + 4 + 0 + 4 + 4} = \frac{4}{16} = \frac{1}{4}.$$

$$\hat{p}_3 = \frac{\sum_{t=4}^{5} Y_t - Y\,)\,Y_t - k - Y}{\sum^{T}{}_{t=1} Y_t - Y^2} =$$

$$\frac{Y_4 - YY_1 - Y + Y_5 - YY_2 - Y}{Y_1 - Y^2 + (Y_2 - Y)^2 + Y_3 - Y^2 + Y_4 - Y^2 + Y_5 - Y^2} =$$

$$\frac{4 - 6\,)\,8 - 6 + 8 - 6\,(4 - 6}{8 - 6^2 + 4 - 6^2 + 6 - 6^2 + 4 - 6^2 + 8 - 6^2} =$$

$$\frac{-22 + 2\text{-}2}{2^2 + -2^2 + 0^2 + -2^2 + 2^2} = \frac{-4 + -4}{4 + 4 + 0 + 4 + 4} = -\frac{8}{16} = -\frac{1}{2}.$$

$$\hat{\rho}_4 = \frac{Y_5 - YY_1 - X}{(Y_1 - Y^2 + Y_2 - Y^2 + Y_3 - Y^2 + Y_4 - Y^2 + (Y_5 - Y^2} =$$

$$\frac{8 - 68 - 6}{8 - 6^2 + 4 - 6^2 + 6 - 6^2 + 4 - 6^2 + 8 - 6^2} =$$

$$\frac{(2)\,(2)}{2^2 + -2^2 + 0^2 + (-2)^2 + 2^2} = \frac{4}{4 + 4 + 0 + 4 + 4} = \frac{4}{16} = \frac{1}{4}.$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Assim, a FAC estimada é:

$$\hat{p}_0 = 1$$

$$\hat{p}_1 = -\frac{1}{2}$$

$$\hat{p}_2 = \frac{1}{4}$$

$$\hat{p}_3 = -\frac{1}{2}$$

$$\hat{\rho}_4 = \frac{1}{4}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Note que só é possível caminhar até o *lag* 4, uma vez que a série possui apenas cinco observações. Perde-se também k observações no cálculo da estimativa para cada *lag*. Tecnicamente, esse número k de observações perdidas é chamado graus de liberdade.

É indesejável perder graus de liberdade, mas o que importa é que este número seja pequeno em relação ao número de observações da série. Séries de interesse prático terão tamanho bem maior do que o exemplo, e isso não será problema.

Por exemplo, em uma série com 300 observações, se quisermos estimar a FAC até o *lag* 30 (que costuma ser mais do que suficiente) teremos uma perda máxima de 30 graus de liberdade, restando ainda 300-30 = 270 graus de liberdade para o cômputo de $\hat{p}_{30}$.

Estimação da FACP

Usa-se o fato de que, no AR(p), $\Phi_{pp} = \phi$. Assim, $\hat{\Phi}_{pp} = \hat{\varphi}_p$.

O procedimento para estimar a FACP consiste em ajustar modelos AR de diferentes ordens, de p = 1 até a ordem k desejada, e, para cada k, tomar, como estimativa de Φ_k, a estimativa do coeficiente Φ_k do termo Y_{t-k} do AR(k).

Por exemplo, inicialmente estima-se um AR(1). Digamos que o resultado tenha sido:

$$Y_t = 0.4 + 0.6Yt_{-1}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Neste caso, $\hat{\Phi}_{11} = \hat{\varphi}_1 = 0.6.$

Em seguida, estima-se um AR(2). Suponha que o resultado tenha sido:

$$Y_t = 0.2 \quad + \quad 0.5Yt_{-1} + 0.3Yt_{-2}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Daí, $\hat{\Phi}_{22} = \widehat{\varphi}_2 = 0.3.$

Estima-se então um AR(3), com resultado:

$$Y_t = 0.3 \quad + \quad 0.4Yt_1 + 0.5Yt_2 + 0.1Yt_3$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Temos assim, $\hat{\Phi}_{33} = \widehat{\varphi}_3 = 0.1.$

E assim por diante, até o número máximo de *lags* para os quais se deseja estimar a FACP. É importante observar que não basta estimar um AR(k) e considerar seus coeficientes. Para cada *lag* de interesse, é necessário estimar um AR(k) e considerar apenas o seu último coeficiente como estimativa de $\hat{\Phi}_{kk}$.

Desta forma, se quisermos, por exemplo, a FACP para 30 *lags*, é necessário estimar 30 modelos AR. Para nossa sorte, o R executa rapidamente esta tarefa para nós. O comando é **ggtsdisplay(serie)**.

Identificação das ordens p e q de um modelo ARMA

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

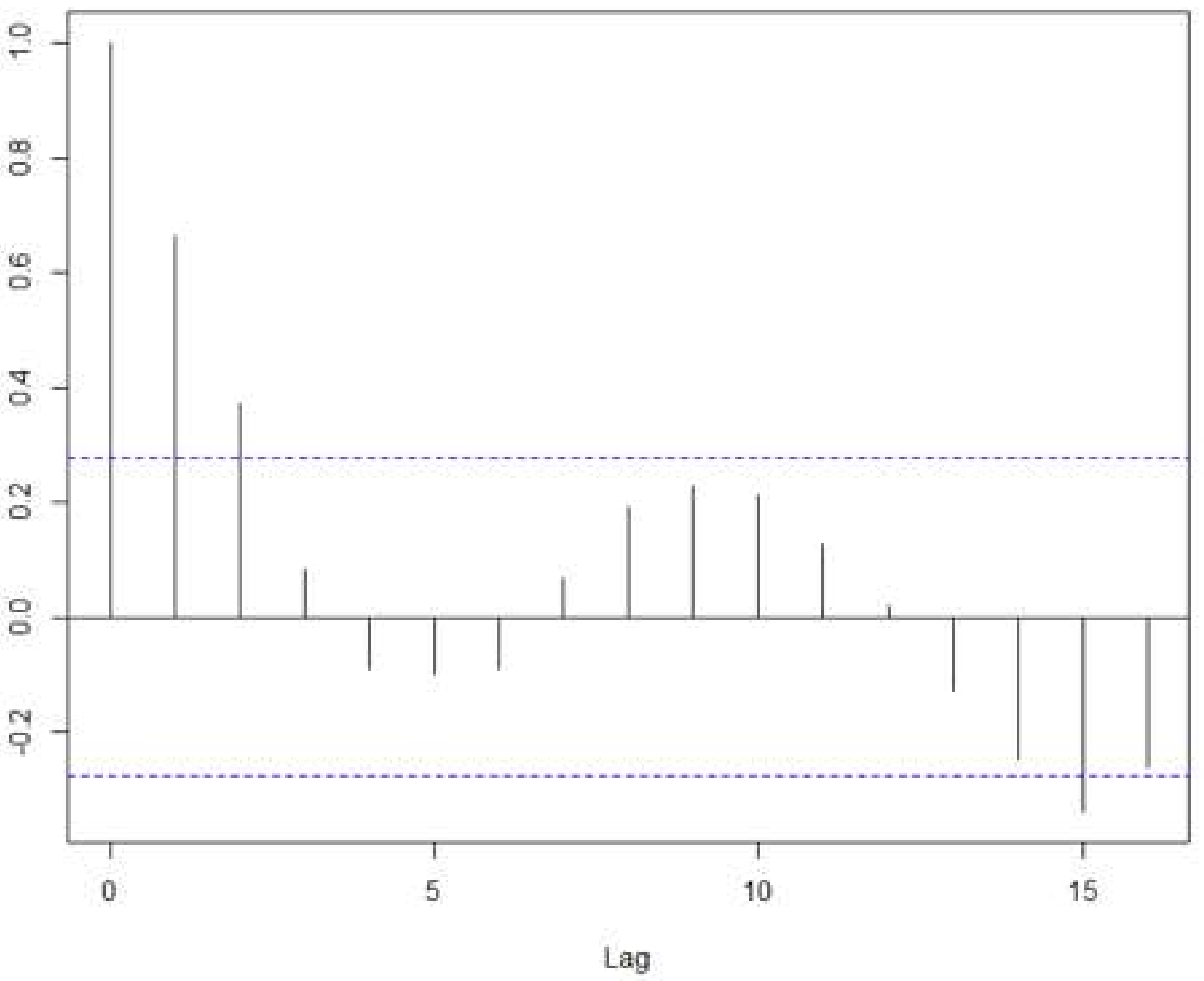
A FAC de um MA(q) é truncada em q. A FACP de um AR(p) é truncada em p. A FAC e FACP de um ARMA(p,q) apresentam uma mistura dos comportamentos acima. Desta forma, o procedimento sugerido por Box e Jenkins para a identificação das ordens p e q de um modelo de séries temporais é:

A partir da série que se quer analisar, estima-se a FAC e obtém-se o gráfico da FAC e da FACP amostral.

- Se a FAC for truncada no *lag* q, então identificamos uma componente MA(q) na série.
- Se a FACP for truncada no *lag* p, então identificamos uma componente AR(p) na série.

FAC e FACP de um ARMA(p,q) apresentam uma mistura dos padrões acima.

Veja um exemplo: seja a seguinte FAC estimada para uma série temporal.



Esta FAC apresenta decaimento exponencial. As linhas pontilhadas em azul representam os limites do intervalo de confiança de 95%. Ao identificar um decaimento exponencial, estamos considerando apenas o que ocorre fora desses limites.

A explicação vem da teoria de testes de hipóteses em modelos estatísticos, em que um parâmetro é dito não significante (ao nível usual 0.05) se a sua estimativa cai dentro do intervalo de confiança (de 95%) para o respectivo parâmetro. Desta forma, o que ocorre dentro dos limites de confiança é considerado estatisticamente não significante (ou seja, para fins práticos, tais valores são iguais a zero).

Se os valores da FAC dentro do intervalo de confiança fossem considerados, identificaríamos um decaimento de acordo com uma senoide amortecida, e não exponencial (o que, no entanto, não afetaria a ordem do modelo identificado).

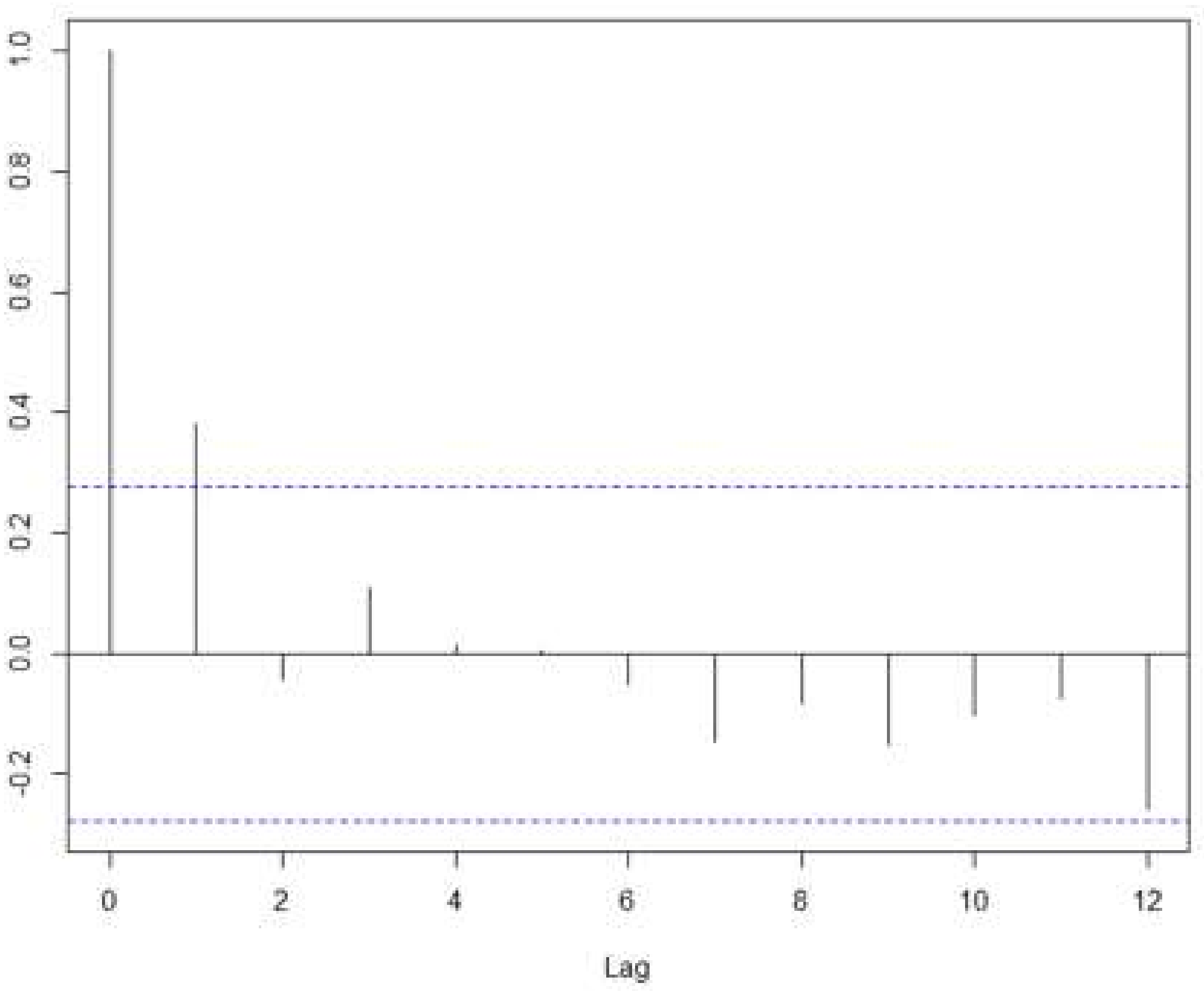
A fórmula para obter o intervalo de confiança para cada ρ_j na FAC é apresentada a seguir:

$$IC_{1000(1-\alpha)\%}\left(\rho_j\right) = \hat{\rho}_j \mp z_{\frac{\alpha}{2}} \sqrt{\hat{V}\hat{\rho}_j},$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

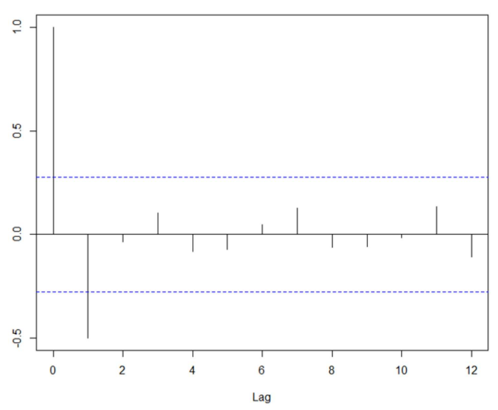
sendo $\hat{V}\hat{\rho} \cong \frac{1}{T}$, contém o zero, para cada j = 1, 2, ... (esta fórmula para a variância é uma aproximação, denominada fórmula de Bartlett).

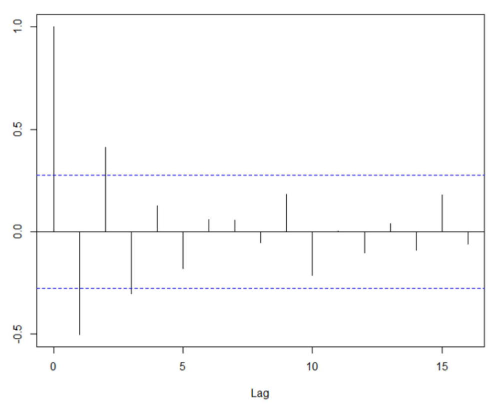
Suponha agora que a FACP seja:



Como a FACP é truncada no *lag* 1, identifica-se um AR(1). Cuidado para não se confundir com os *lags*. A função plotada pelo R começa do *lag* zero, em que trivialmente assume valor 1. O valor no *lag* 1 é, portanto, dado pelo segundo palitinho, e não pelo primeiro, na figura acima.

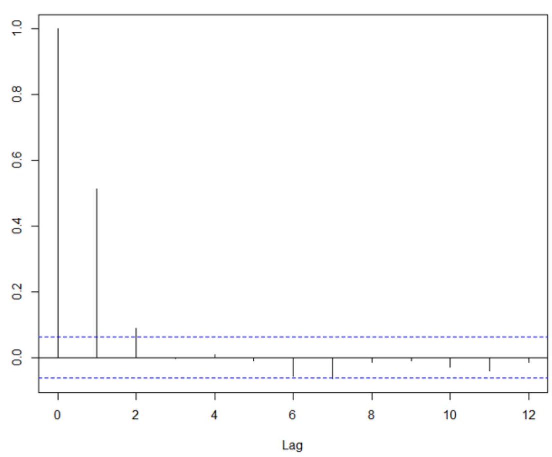
Sejam as seguintes FAC e FACP estimadas para uma série temporal.

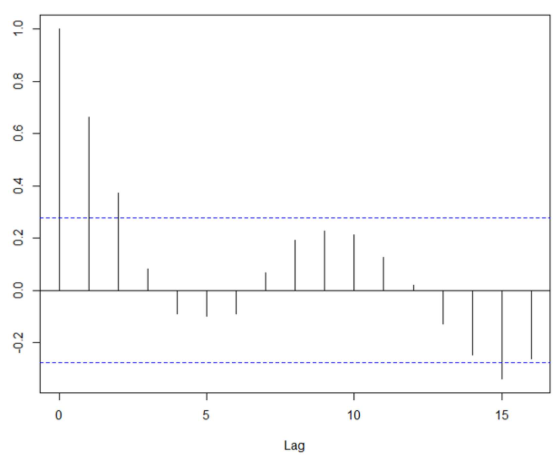




Como a FAC é truncada no *lag* 1 e a FACP apresenta decaimento segundo uma senoide amortecida, identifica-se um MA(1).

Sejam as seguintes FAC e FACPs estimadas para uma série temporal.





Aqui temos uma dúvida entre o truncamento no *lag* 1 ou 2, que nos levaria, respectivamente, a um modelo MA(1) ou MA(2). O que fazer neste caso?

Segundo Box e Jenkins, o procedimento de identificação costuma ser excessivamente parcimonioso, ou seja, tende a subidentificar as ordens p e q do modelo. Será necessário utilizar um procedimento complementar chamado teste de sobrefixação, a ser apresentado na aula 8, que consiste basicamente em estimar sequencialmente modelos de ordens maiores.

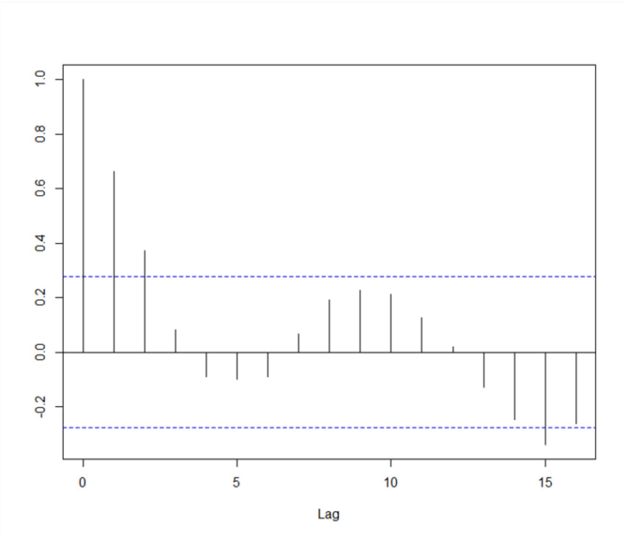
Desta forma, é mais adequado utilizar de parcimônia no momento de escolher o modelo identificado, que será apenas um candidato inicial, escolhendo-se assim o MA(1).

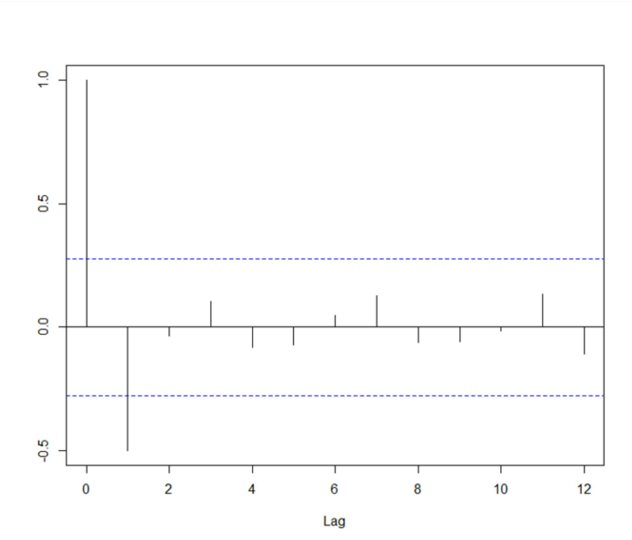
Note que, se escolhêssemos o MA(2), não daríamos chance ao MA(1) no processo de sobreidentificação. Do contrário, escolhendo o MA(1) como modelo inicial, um dos modelos sobrefixados a serem testados será justamente o MA(2) (o outro será o ARMA(1,1)).

Os testes de sobrefixação serão detalhados na aula 8, mas é importante ter em mente desde já que a etapa de identificação não é definitiva e que deve ser complementada por esses testes. Antes disso, porém, falaremos sobre a estimação de parâmetros, no início da aula 8.

Atividades

1. As funções de autocorrelação e de autocorrelação parcial a seguir

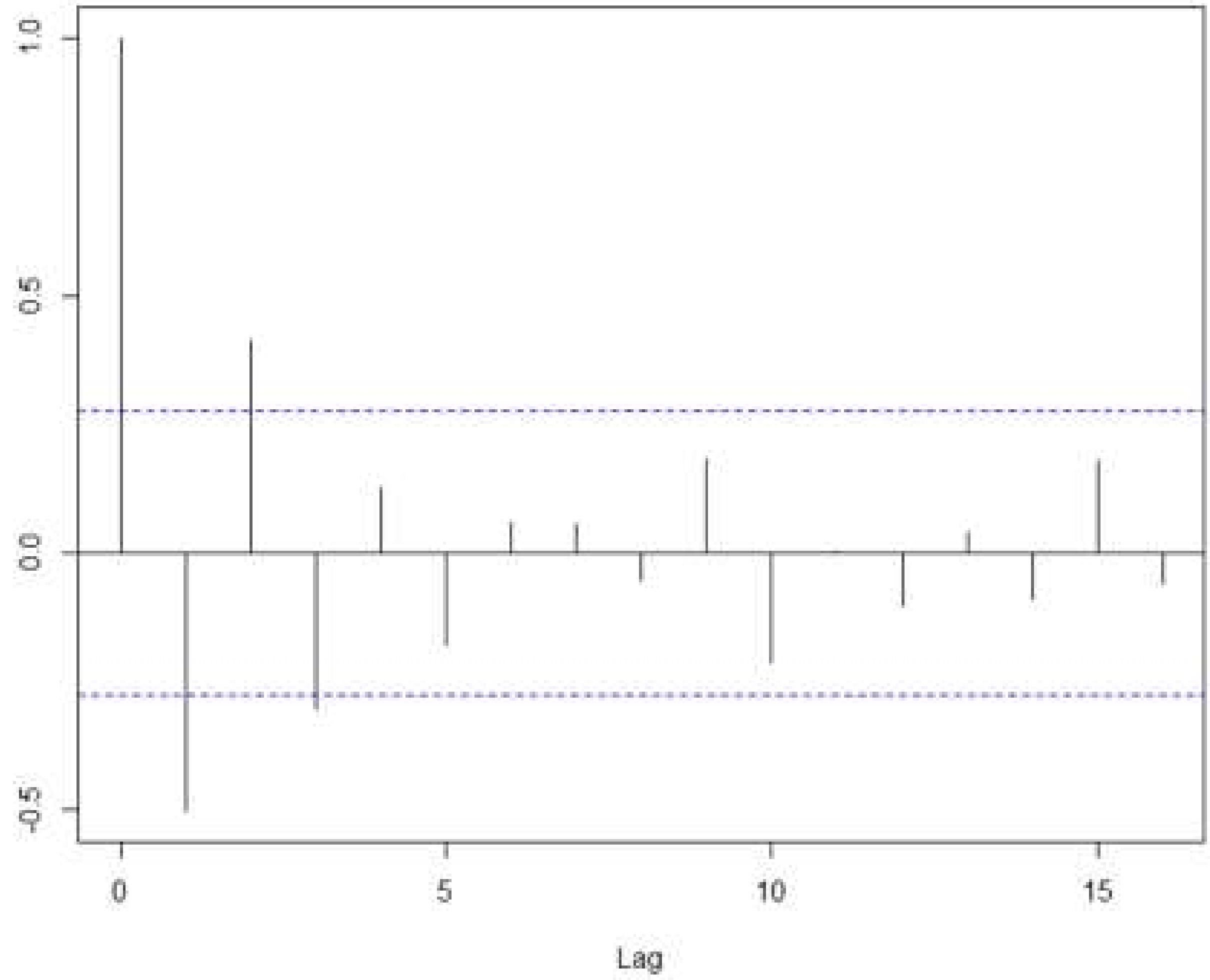




conduzem à identificação de um modelo:

a) AR(1)
b) MA(1)
c) AR(2)
d) MA(2)
e) ARMA(1,1).

2. A função de autocorrelação estimada apresentada a seguir :



conduz à identificação de um modelo:

a) AR(1)
b) MA(1)
c) AR(4)
d) MA(4)
e) AR, cuja ordem não é possível determinar.

3. Considere o seguinte modelo AR(1) estimado: $Y_t = 0.6Y_{t-1} + \varepsilon_t$.

Sobre este modelo, considere as seguintes afirmativas:

I. A FACP estimada no lag 1 é igual a 0.6
II. A FACP estimada no lag 2 é igual a 0.36
III. A FACP estimada no lag 1 não é possível de determinar com a informação dada
IV. A FACP estimada no lag 2 não é possível de determinar com a informação dada

São verdadeiras apenas as seguintes afirmativas:

a) Somente a III
b) I e II
c) I e IV
d) II e III
e) III e IV

Notas

Título modal [1]

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

Título modal [1]

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

Referências

BOX, G.E.P.; JENKINS, G.M. **Time series analysis:** Forecasting and control. San Francisco: Holden-Day, Inc., 1976.

MORETTIN, P.A.; TOLOI, C.M.C. **Análise de Séries Temporais**. São Paulo: Edgard Blucher, 2004.

Próxima aula

- Métodos para a estimação dos parâmetros de um modelo de séries temporais;
- Testes de sobrefixação;
- Etapa de diagnóstico de um modelo.

Explore mais

- Gere modelos de séries temporais artificialmente (usando, por exemplo, o Excel), e estime a FAC e a FACP por meio do comando ggtsdisplay() no R. Esta função está na library "forecast", portanto você terá que usar o comando "library(forecast)" antes de utilizá-la. Compare com o padrão teórico correspondente aos modelos que você gerou.