Avaliação da acurácia de algoritmo automático de identificação de modelos ARMA(1,1), MA(1), AR(1) e ruído brando.

INTRODUÇÃO:

‘’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’Notação’’’’’’’’’’’’’’

Uma série temporal é qualquer conjunto de observações feitas sequencialmente ao longo do tempo, dentre alguns exemplos, podemos citar os preços diários de uma ação na Bolsa de Valores, as observações mensais de venda de uma loja esportiva e o consumo mensal de energia elétrica em uma cidade. Por ter aplicabilidade em diversas áreas, como finanças, marketing, meteorologia, entre outras, o estudo de séries temporais é muito relevante.

Muitos modelos controlados por leis probabilísticas foram propostos para descrever séries temporais, dentre estes, podemos citar a importante classe de modelos tipo ARMA, que foi abordada por Box e Jenkins (1970). Não serão discutidos aqui os detalhes e as particularidades da classe de modelos ARMA(p,q), para maior esclarecimento veja Morettin (2006), mas um aspecto relevante para este estudo é que, para esta classe de modelos, se é necessário realizar um processo de especificação das ordens dos parâmetros das partes auto regressiva e de médias móveis para que o modelo seja estimado adequadamente à série temporal de interesse. Este processo deve definir os valores de “p” e “q” com base no comportamento da série temporal.

A identificação das ordens do modelo não é uma tarefa simples e direta, ela depende da análise crítica de uma pessoa que tenha certa familiaridade com as técnicas de identificação e teoria relacionada à essa classe de modelos. Mesmo a identificação das ordens de um modelo ARMA sendo altamente dependente da análise feita por um ser humano, existem algoritmos automáticos que tentam realizar esta tarefa com base nos dados observados da série.

A função “auto.arima” disponível no software estatístico R através da utilização do pacote “forecast”, dentre outras tarefas, identifica as ordens apropriadas de um modelo ARMA com base nas observações de uma série temporal fornecida como parâmetro da função. A utilização da função “auto.arima” faria com que o processo de identificação de um modelo ARMA fosse muito mais simples e rápido, porém não se sabe qual o nível de acurácia deste algoritmo na identificação das ordens de um modelo ARMA. Uma maneira de avaliar essa acurácia seria simular séries segundo modelos ARMA, aplica-las na função “auto.arima” e comparar as ordens do modelo sugerido pela função com as ordens reais do modelo simulado. A acurácia pode ser estimada pela proporção de ordens sugeridas que concordam com as ordens simuladas.

OBJETIVO:

Este trabalho visa avaliar a acurácia na identificação automática de modelos ARMA(1,1), ARMA(1,0), ARMA(0,1) e ARMA(0,0) utilizando a função “auto.arima” disponível no software estatístico R através da utilização do pacote “forecast”.

METODOLOGIA:

O presente trabalho discute a acurácia descrita apenas para séries simuladas conforme modelos ARMA(1,1), ARMA(1,0), ARMA(0,1) e ARMA(0,0), sendo os três últimos conhecidos como modelos AR(1), MA(1) e ruído branco, respectivamente.

Seguindo a notação utilizada pelo pacote “forecast”, os modelos abordados nesse trabalho propõe as seguintes leis probabilísticas :

1. :
2. :
3. :
4. :

Onde at é ruído branco, zt é a observação da série no instante t e phi e theta são parâmetros a serem definidos no caso de séries simuladas.

Veja descrição mais detalhada dos modelos em Morettin 2006.

Note que (2), (3) e (4) são casos particulares de (1). Quando se considera phi=0, theta = 0 e phi=theta=0 em (1) se obtém (2), (3) e (4) respectivamente.

Utilizando o software R versão 3.4.4 e a função “arima.sim” disponível também no pacote “forecast”, foram simuladas três quantidades de tamanho amostrais (100, 500, 1000) de séries temporais segundo modelos ARMA(1,1), ARMA(1,0), ARMA(0,1) e ARMA(0,0) com diferentes combinações de valores para os parâmetros phi e theta (phi, theta = 0, 0,25, 0,5 e 0,75).

A variação dos parâmetros, ordens dos modelos objetivam evidenciar possíveis diferenças de performance evidente na utilização da função “auto.arima” para diferentes situações. Optou-se por restringir os valores ‘p’ e ‘q’ como valores máximos iguais a 1, por se tratar de modelos mais simples e usuais, e valores de phi e theta variando entre 0, 0,25 , 0,5 e 0,75 por se tratar de valores que geram séries invertíveis, estacionárias e que não proporcionam cancelamento de raízes.

Aplicou-se a função “auto.arima” em todas as séries simuladas, definindo os parâmetros “seasonal” como ‘FALSE’ e “stationary” como ‘TRUE’, de modo a restringir a busca da função por modelos estacionários e não sazonais, mantendo o restante dos parâmetros como o padrão da função. Comparou-se as ordens dos modelos sugeridos pela função com as ordens do modelo simulado, caso estas concordassem computou-se um acerto. Esta situação foi repetida 1000 vezes e então identificou-se a quantidade de acertos para cada série simulada, assim como a proporção de vezes em que ocorreu o acerto.

RESULTADOS:

Abaixo são apresentadas as proporções de acerto para modelos ARMA com todas as combinações de parâmetros phi e theta descritos anteriormente, para séries simuladas com 100, 500 e 1000 observações nas tabelas 1, 2 e 3 respectivamente. Nestas também são apresentadas as médias de proporções de acerto considerando todos os modelos para séries separadas pelos diferentes tamanhos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n = 100 | MA | | | |
| AR | 0 | .25 | .5 | .75 |
| 0 | 0,861 | 0,724 | 0,78 | 0,371 |
| .25 | 0,333 | 0,274 | 0,5 | 0,044 |
| .5 | 0,689 | 0,655 | 0,754 | 0,185 |
| .75 | 0,769 | 0,656 | 0,732 | 0,293 |
| Média Geral: | | | | 0,53875 |
|  |  |  |  |  |
| n = 500 | MA | | | |
| AR | 0 | .25 | .5 | .75 |
| 0 | 0,716 | 0,622 | 0,519 | 0,545 |
| .25 | 0,466 | 0,604 | 0,557 | 0,335 |
| .5 | 0,627 | 0,637 | 0,575 | 0,553 |
| .75 | 0,671 | 0,625 | 0,564 | 0,613 |
| Média Geral: | | | | 0,576813 |
|  |  |  |  |  |
| n = 1000 | Theta (parte MA) | | | |
| AR | 0 | .25 | .5 | .75 |
| 0 | 0,687 | 0,584 | 0,544 | 0,591 |
| .25 | 0,577 | 0,613 | 0,545 | 0,55 |
| .5 | 0,631 | 0,617 | 0,592 | 0,646 |
| .75 | 0,668 | 0,603 | 0,566 | 0,677 |
| Média Geral: | | | | 0,605688 |

Considerando todas as simulações feitas, a proporção de acerto foi de aproximadamente 0,57.

Discussões dos Resultados:

Observando os resultados obtidos, nota-se maior variabilidade nos valores das proporções de acerto para séries simuladas dos modelos propostos com tamanho 100 do que quando comparadas às de tamanho 500, assim como maior variabilidade destas de tamanho 500 em relação às de tamanho 1000.

Comparando-se as médias gerais de proporções de acertos com os tamanhos das séries simuladas, nota-se que existe um crescimento no valor destas proporções quando se têm um número maior de observações da série a ser analisada, ou seja, existem indícios de que o algoritmo tem performance melhor conforme tem mais observações para analisar.

Por meio da tabela 1, nota-se proporções de acertos com grandes desvios abaixo da média geral para séries de tamanho 100 nos modelos ARMA(0,1) com valor de theta igual a 0,75 e modelos ARMA(1,1) com pares de parâmetros phi e theta iguais à (0.25, 0,25),(0,25, 0,5), (0,25, 0,75), (0,5, 0,75) e (0,75, 0,75). Em especial para o modelo ARMA(1,1) com pares de parâmetros (0,25, 0,75), observou-se uma acurácia de apenas 0,044.

Já para séries de tamanho 500, destacam-se por apresentar proporções com desvios grandes abaixo da média geral de proporções de acertos para este tamanho de séries, os modelos ARMA(1,1) com pares de parâmetros phi e theta iguais a (0.25, 0,25) e (0,25, 0,75).

Para séries de tamanho 1000 nenhum modelo parece ter se destacado dos demais por apresentar grande diferença na acurácia na identificação das ordens dos modelos por meio da utilização da função “auto.arima”.

CONCLUSÃO:

Com exceção dos modelos com parâmetro de médias móveis igual a 0,75 (theta=0.75) para séries de tamanho 100, onde o algoritmo da função auto.arima apresentou baixa acurácia em relação aos demais na identificação das ordens do modelo, não se identificou claro destaque de valores para acurácia em algum modelo e/ou situação específica.

Contudo, de um modo geral a identificação das ordens dos modelos apresentados aqui por meio da função “auto.arima” não parece apresentar uma performance aceitável, uma vez que mesmo para séries simuladas com 1000 observações, que é uma quantidade que as vezes é difícil de se obter em séries reais, a função auto.arima obteve, em termos gerais, proporções de acerto por volta de apenas 60 %.

É difícil mensurar a acurácia de uma análise das ordens de um modelo ARMA adequado à uma séries temporal por um humano, já que é dependente da habilidade deste, porém considerando que a análise fosse feita por um indivíduo treinado para tal fim, acredito que a acurácia na identificação das ordens dos modelos seria, apesar de mais demorada, cabalmente maior do que a acurácia apresentada pela função auto.arima. Deste modo, o presente trabalho reforça a importância das análises criteriosas de um ser humano treinado para a identificação do modelo adequado para descrever o comportamento de uma série temporal.

Referências: