Imagem: bit.ly/3lb6a15

ERHA7016 – Hidrologia Estocástica

Geração de Séries Sintéticas (pt. 2)

estimação e validação

Daniel Detzel
detzel@ufpr.br

Agenda

Geração de séries sintéticas

- modelos ARIMA: estimação
- modelos ARIMA: validação

Geração de séries sintéticas | modelos ARIMA

Construção do modelo ARIMA (relembrando...)

1. Identificação do modelo

Determinar quais componentes farão parte da formulação

2. Estimação dos parâmetros

Obter os valores dos parâmetros envolvidos

3. Validação do modelo

Avaliar, teoricamente, se o ajuste é satisfatório

GERAÇÃO DE SÉRIES SINTÉTICAS

modelos ARIMA: estimação

Geração de séries sintéticas | modelos ARIMA: estimação

Objetivo: determinar os valores de φ_p $(p = 1,2,...)$ e θ_q $(q = 1,2,...)$

Métodos a serem discutidos:

Momentos

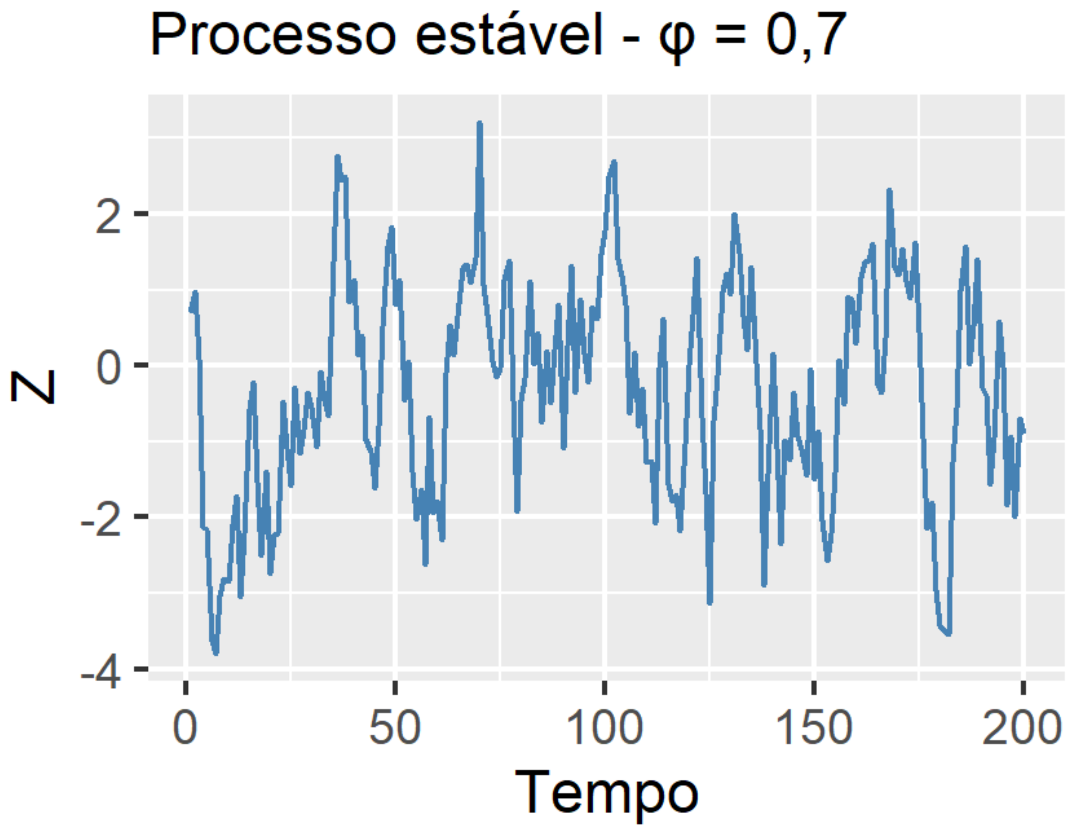
Máxima Verossimilhança

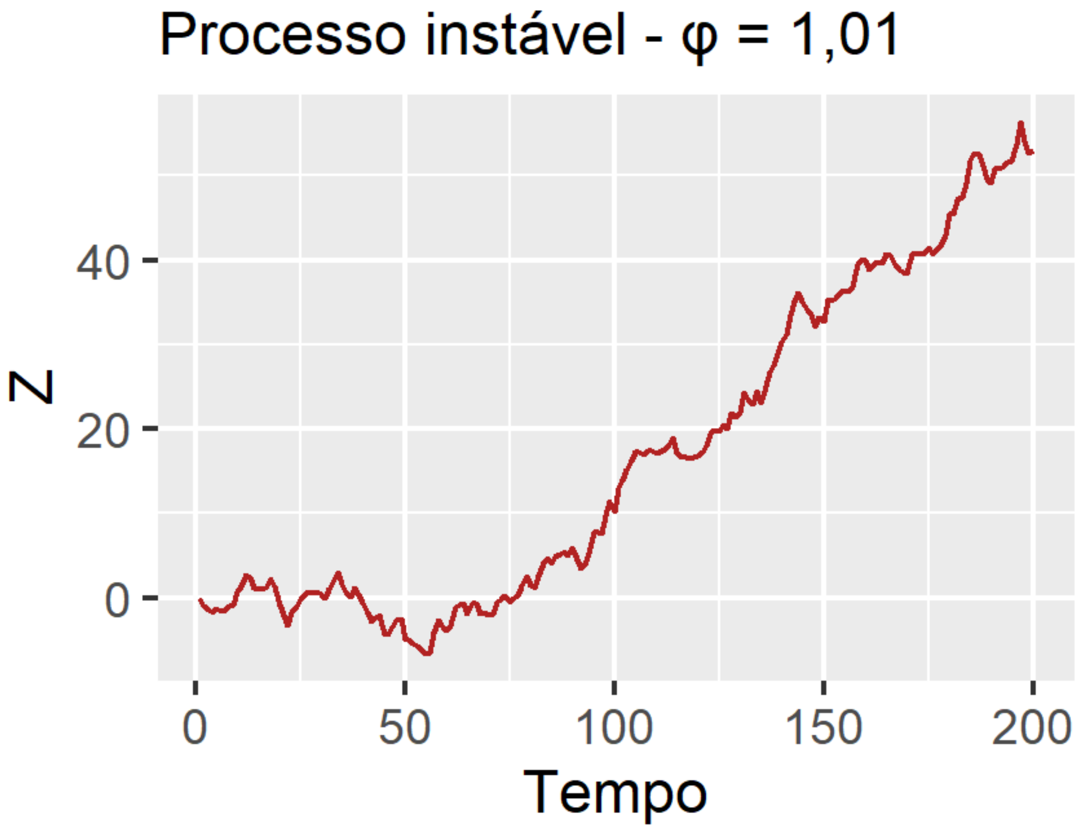
Os parâmetros devem obedecer a certas restrições para que produzam modelos estáveis

um modelo estável produz valores que flutuam no entorno de um nível médio sem “explodir”

Geração de séries sintéticas | modelos ARIMA: estimação

Representação da estabilidade e instabilidade do modelo AR(1)





Geração de séries sintéticas | modelos ARIMA: estimação

As restrições aos parâmetros são denominadas condições de estacionariedade e invertibilidade

estacionariedade: aplicada à componente AR
invertibilidade: aplicada à componente MA

A compreensão dessas condições requer o entendimento da definição do polinômio característico

Geração de séries sintéticas | modelos ARIMA: estimação

Condição de estacionariedade:

Seja um modelo AR(p):

$$z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \cdots + \varphi_p z_{t-p} + a_t$$

Ele pode ser reescrito na forma:

$$z_t - \varphi_1 z_{t-1} - \varphi_2 z_{t-2} - \cdots - \varphi_p z_{t-p} = a_t$$

Geração de séries sintéticas | modelos ARIMA: estimação

Usando o operador defasagem B (lembrando: $B^k z_t = z_{t-k}$):

$$B^0 z_t - \varphi_1 B^1 z_t - \varphi_2 B^2 z_t - \cdots - \varphi_p B^p z_t = a_t$$

Colocando z_t em evidência:

$$\underbrace{\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)}_{\text{polinômio característico } \Phi(B)} z_t = a_t$$

Geração de séries sintéticas | modelos ARIMA: estimação

A condição de estacionariedade é verificada por meio deste polinômio característico $\Phi(B)$

Para tanto, os coeficientes φ_i $(i = 1,2, \dots, p)$ são inseridos no polinômio e as raízes B_i são determinadas

nota: as raízes podem ser complexas

Um modelo é estacionário quando a combinação de coeficientes φ_i $(i = 1,2, \dots, p)$ produz raízes $|B_i| > 1$

diz-se nesse caso que as raízes caem fora do círculo unitário
(círculo formado pelas porções reais e imaginária das raízes B_i)

Representação do círculo unitário e das raízes de $\Phi(B)$ dos modelos AR(1) estável e instável mostrados anteriormente

obtenção do polinômio:

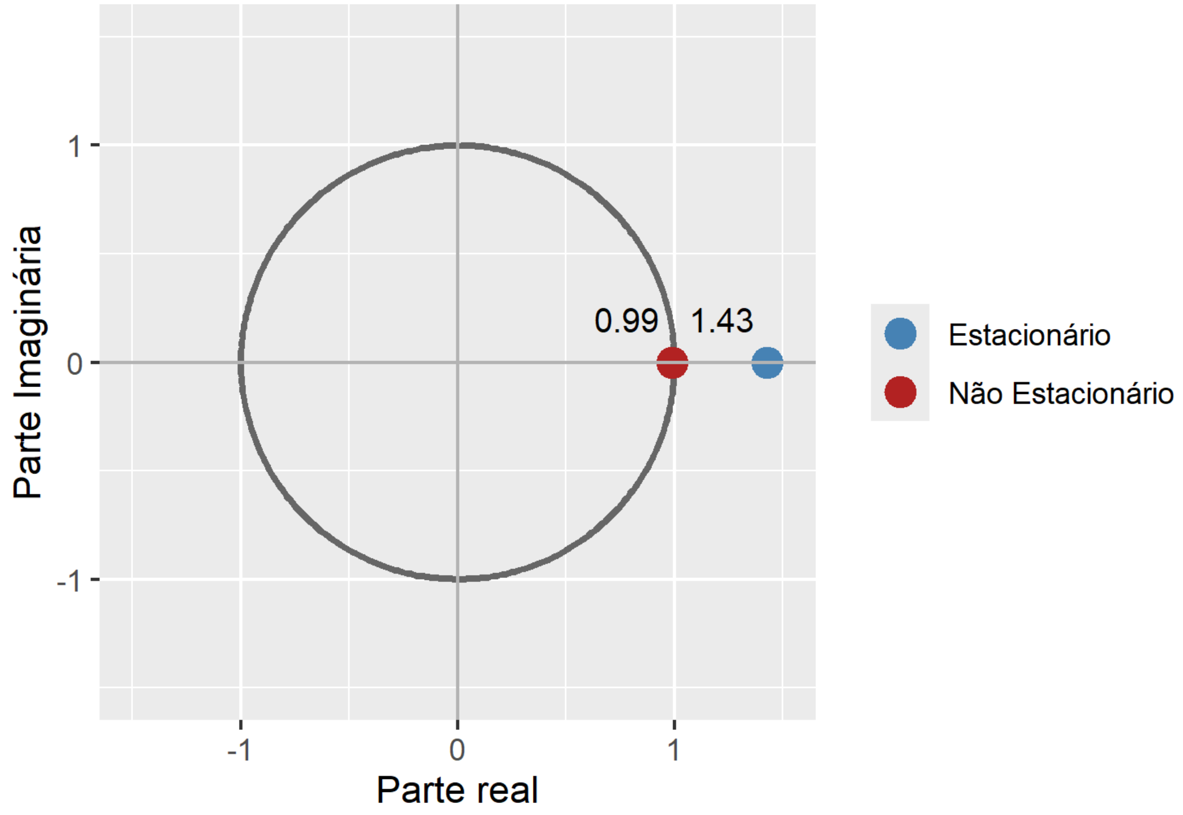
$$1 - \varphi_1 B = 0$$

AR(1) estacionário:

$$\varphi_1 = 0{,}7 \rightarrow B = 1{,}43$$

AR(1) não estacionário:

$$\varphi_1 = 1{,}01 \rightarrow B = 0{,}99$$



Condição de invertibilidade:

Seja um modelo MA(q):

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

Ele pode ser reescrito com operadores B:

$$z_t = \underbrace{\left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p\right)}_{\text{polinômio característico } \Theta(B)} a_t$$

Geração de séries sintéticas | modelos ARIMA: estimação

A condição de invertibilidade é verificada por meio deste polinômio característico $\Theta(B)$

Para tanto, os coeficientes θ_i $(i = 1,2, \dots, q)$ são inseridos no polinômio e as raízes B_i são determinadas

nota: as raízes podem ser complexas

Um modelo é invertível quando a combinação de coeficientes θ_i produz raízes $|B_i| > 1$

Geração de séries sintéticas | modelos ARIMA: estimação

Em resumo:

Modelos puramente AR:

Estacionariedade atendida quando as raízes de $\Phi(B)$ caem fora do círculo unitário

Modelos puramente MA:

Invertibilidade atendida quando as raízes de $\theta(B)$ caem fora do círculo unitário

Modelos ARMA:

Ambas as condições devem ser atendidas

Geração de séries sintéticas | modelos ARIMA: estimação

[Exemplo] No ajuste de um modelo Box & Jenkins, chegou-se à seguinte equação:

$$z_t = 0{,}6 \cdot z_{t-1} + 0{,}2 \cdot z_{t-2} + a_t + 0{,}1 \cdot a_{t-1} - 0{,}8 \cdot a_{t-2}$$

Pede-se:

a) Identificar o modelo.

b) Verificar se ele atende às condições de estacionariedade e invertibilidade.

Geração de séries sintéticas | modelos ARIMA: estimação

Método dos Momentos (MoM):

Recomendado para modelos puramente AR

produz estimadores enviesados quando da presença de componentes MA

Para modelos AR, considera-se a função de autocorrelação teórica:

$$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2} + \cdots + \varphi_p \rho_{k-p}, k > 0$$

onde

ρ_k autocorrelação teórica com defasagem k

φ_i parâmetro AR de ordem i $(i = 1,2, \ldots, p)$

Geração de séries sintéticas | modelos ARIMA: estimação

A expansão para múltiplas defasagens k $(1,2,\ldots,p)$ resulta nas equações de Yule-Walker:

$$
\begin{gathered}
\rho_1 = \varphi_1 + \varphi_2\rho_1 + \cdots + \varphi_p\rho_{p-1} \\
\rho_2 = \varphi_1\rho_1 + \varphi_2 + \cdots + \varphi_p\rho_{p-2} \\
\vdots \\
\rho_p = \varphi_1\rho_{p-1} + \varphi_2\rho_{p-2} + \cdots + \varphi_p
\end{gathered}
$$

O MoM consiste em substituir os momentos teóricos (do modelo) pelos momentos empíricos (da série)

> dado um modelo AR(p), substituem-se as correlações teóricas ρ_p por seus estimadores amostrais r_p

Geração de séries sintéticas | modelos ARIMA: estimação

Lembrando: os estimadores amostrais r_k são retirados da FAC (ou correlograma)

[Exemplo] Obter os estimadores para os modelos AR de primeira e segunda ordens.

1ª ordem: Modelo AR(1) → $p = 1$

Das equações de Yule-Walker: $\rho_1 = \varphi_1$

Substituindo pelo estimador amostral: $r_1 = \varphi_1$

$$\boxed{\therefore \hat{\varphi}_1 = r_1}$$

Geração de séries sintéticas | modelos ARIMA: estimação

Este é um resultado particular muito importante!

interpretação: o coeficiente $\hat{\varphi}_1$ de um modelo AR(1) é o próprio coeficiente de correlação (de Pearson) de primeira ordem r_1

Desdobramento: todo modelo AR(1) será sempre estacionário (estável), pois:

$$1 - \varphi_1 B = 0$$

$$1 - r_1 B = 0$$

$$B = \frac{1}{r_1}$$

Como $-1 < r_1 < 1$ por definição, tem-se que $|B| > 1$*

* Em teoria, $|r_1|$ pode assumir exatamente 1. Nesse caso, $|B| = 1$ e o modelo é não estacionário. Porém, casos de correlação perfeita não acontecem no contexto de séries de fenômenos naturais.

Geração de séries sintéticas | modelos ARIMA: estimação

2ª ordem: Modelo AR(2) → $p = 2$

Das equações de Yule-Walker:

$$\begin{cases} \rho_1 = \varphi_1 + \varphi_2 \rho_1 \\ \rho_2 = \varphi_1 \rho_1 + \varphi_2 \end{cases}$$

Substituindo pelos estimadores amostrais r_1 e r_2 e resolvendo o sistema, chega-se a:

$$\boxed{\begin{aligned} \hat{\varphi}_1 &= \frac{r_1(1 - r_2)}{1 - r_1^2} \\ \hat{\varphi}_2 &= \frac{r_2 - r_1^2}{1 - r_1^2} \end{aligned}}$$

Geração de séries sintéticas | modelos ARIMA: estimação

Método da Máxima Verossimilhança (MLE):

Recomendado para modelos puramente MA ou mistos ARMA

produz estimadores equivalentes às equações de Yule-Walker para modelos AR ajustados a séries com tamanho suficiente ($n > 500$) e sob normalidade

Princípio: busca pelo melhor conjunto de parâmetros que associa os resultados do modelo às observações

As equações a seguir assumem que os dados são normalmente distribuídos

Geração de séries sintéticas | modelos ARIMA: estimação

A função de verossimilhança para um processo ARMA é dada por:

$$L = \frac{1}{(2\pi\hat{\sigma}_a)^{n/2}} \cdot exp\left[-\frac{1}{2\hat{\sigma}_a} \cdot \sum_{t=1}^{n} \hat{a}_t(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})^2\right]$$

onde

$\widehat{\boldsymbol{\varphi}}$ conjunto dos parâmetros AR estimados com base na amostra

$\widehat{\boldsymbol{\theta}}$ conjunto dos parâmetros MA estimados com base na amostra

$\hat{a}_t$ série de resíduos obtidos do ajuste, com variância $\hat{\sigma}_a$

O método busca o conjunto de parâmetros que maximiza a função L

Geração de séries sintéticas | modelos ARIMA: estimação

Matematicamente, é conveniente aplicar o logaritmo, chegando-se à função de log-verossimilhança:

$$\ln L = -n \cdot \ln \hat{\sigma}_a - \frac{S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})}{2\hat{\sigma}_a}$$

onde

$S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ função soma dos quadrados dos resíduos:

$$S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} a_i^2$$

Geração de séries sintéticas | modelos ARIMA: estimação

Qualquer conjunto de parâmetros que maximize a função $S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$, também maximiza a função $\ln L$

O MLE é computacionalmente intenso pois a estimativa de $S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ requer cálculos recursivos para toda série e para cada combinação de parâmetros (recursividade: quando a estimativa do valor atual depende dos valores anteriores)

Lembra-se que as combinações devem respeitar as condições de estacionariedade e invertibilidade

Geração de séries sintéticas | modelos ARIMA: estimação

[Exemplo] Obter a função $S(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ para um modelo ARMA(1,1).

Modelo:

$$z_t = \varphi_1 z_{t-1} + a_t - \theta_1 a_{t-1}$$

Escreve-se a equação em função de a_t:

$$a_t = z_t - \varphi_1 z_{t-1} + \theta_1 a_{t-1}$$

Para os cálculos, a série z_t assume os valores observados

Geração de séries sintéticas | modelos ARIMA: estimação

Variam-se os índices de tempo para $t = p + q, \dots, n$

$t = 2$: $a_2 = z_2 - \varphi_1 z_1 + \theta_1 a_1{}^*$

$t = 3$: $a_3 = z_3 - \varphi_1 z_2 + \theta_1 a_2 = z_3 - \varphi_1 z_2 + \theta_1 (z_2 - \varphi_1 z_1 + \theta_1 a_1)$

$t = 4$: $a_4 = z_4 - \varphi_1 z_3 + \theta_1 a_3$

$= z_4 - \varphi_1 z_3 + \theta_1 [z_3 - \varphi_1 z_2 + \theta_1 (z_2 - \varphi_1 z_1 + \theta_1 a_1)]$

$\vdots$

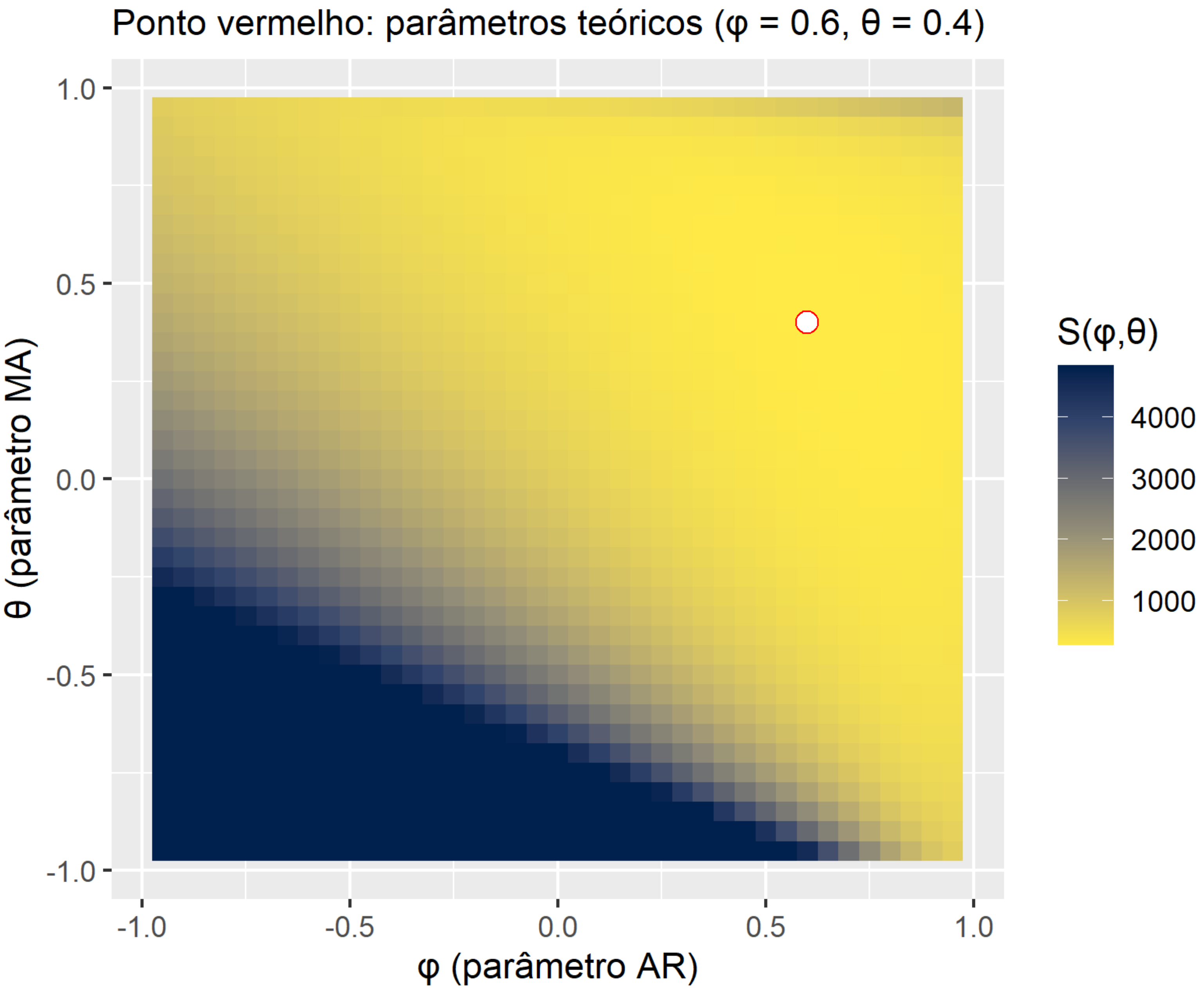
$t = n$: $a_n = z_n - \varphi_1 z_{n-1} + \theta_1 a_{n-1}$

* conveniona-se $a_1 = E[a_t] = 0$

Geração de séries sintéticas | modelos ARIMA: estimação

A cada novo valor para $\hat{\varphi}$ ou $\hat{\theta}$ a função $S(\hat{\varphi}, \hat{\theta})$ é atualizada

Resultado da combinação de parâmetros para um modelo teórico conhecido



Geração de Séries Sintéticas

modelos ARIMA: validação

Geração de séries sintéticas | modelos ARIMA: validação

Última etapa antes do uso do modelo

Uma vez que os parâmetros foram estimados, a série de resíduos a_t resultantes do ajuste é calculada

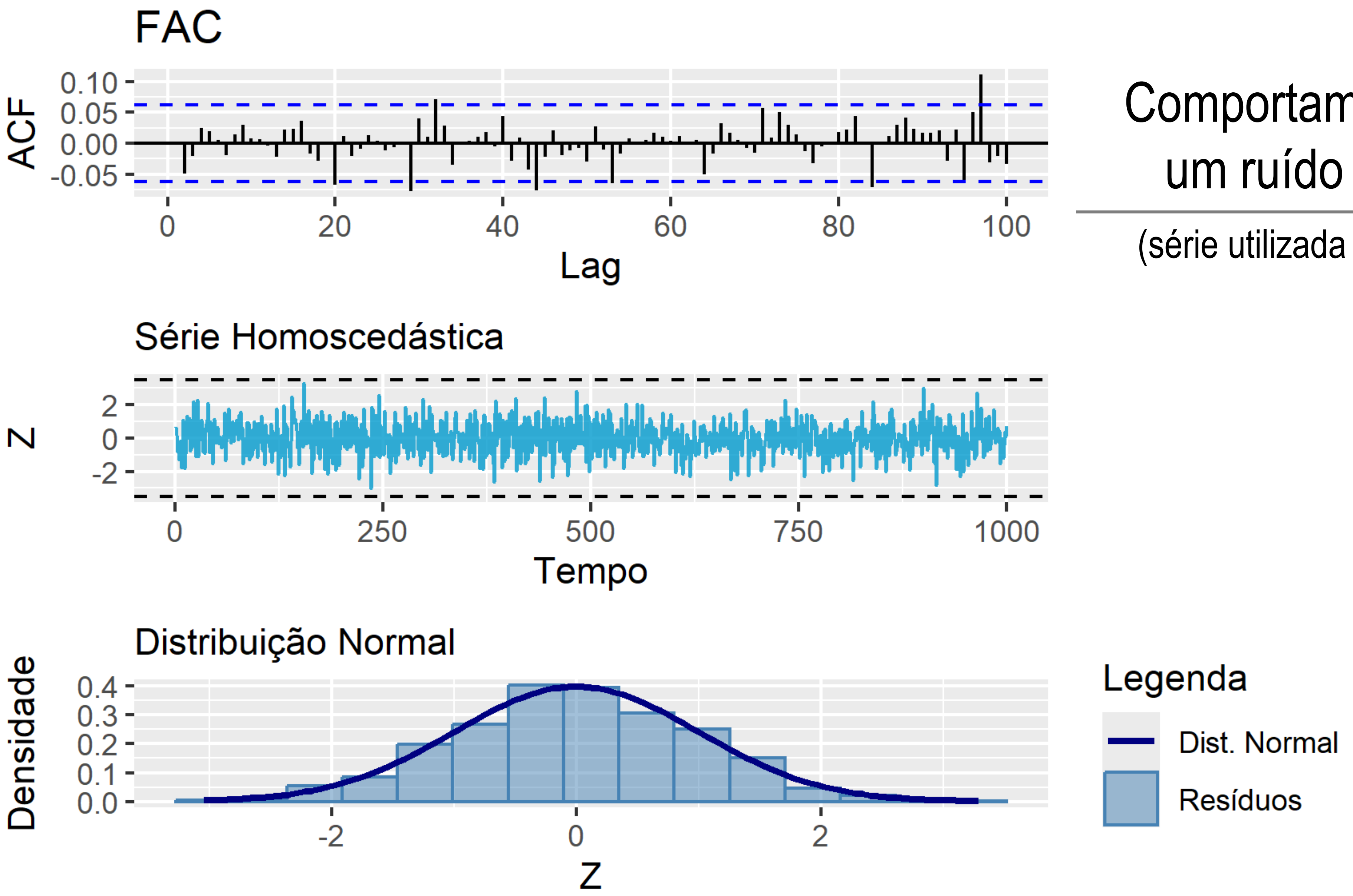
caso a estimação tenha sido feita via MLE, a série a_t já estará calculada

Os resíduos do modelo funcionam como indicadores da qualidade do ajuste

um modelo bem ajustado produz resíduos com características estatísticas próximas a um ruído branco gaussiano

interpretação: as características da série z_t foram "filtradas" pelo, modelo, sobrando apenas os resíduos a_t

Geração de séries sintéticas | modelos ARIMA: validação



Comportamento esperado de um ruído branco gaussiano

(série utilizada para o próximo exemplo)

Geração de séries sintéticas | modelos ARIMA: validação

<u>Independência:</u>

Procedimentos usuais

inspeção visual da FAC dos resíduos

testes estatísticos

Teste estatístico recomendado: Ljung-Box *[Box & Pierce, 1970]*

H_0: Os resíduos não são correlacionados

H_1: Os resíduos são correlacionados

Box, G. E. P.; Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models". Journal of the American Statistical Association. 65 (332): 1509–1526. doi:10.1080/01621459.1970.10481180

Geração de séries sintéticas | modelos ARIMA: validação

O procedimento se utiliza da própria função de autocorrelação para testar a significância até um *lag* pré-determinado k

A estatística do teste é:

$$LB = n(n-2)\sum_{k=1}^{h}\frac{r_k^2}{n-k}$$

onde

n	tamanho da série
r_k^2	autocorrelação de lag k
h	máxima defasagem a ser testada

Geração de séries sintéticas | modelos ARIMA: validação

A estatística LB tem distribuição qui-quadrado, com $h - p - q$ graus de liberdade:

$$LB \sim \chi^2_{h-p-q}$$

Acerca do número de *lags* a ser considerado

atribuído pelo analista de acordo com a série em análise

[Box et al., 1994] recomendam $h \cong 20$

Geração de séries sintéticas | modelos ARIMA: validação

Homoscedasticidade:

Procedimento usual

testes estatísticos

Teste estatístico recomendado: Breusch-Pagan *[Breusch & Pagan, 1978]*

H_0: Homoscedasticidade presente (a variância dos resíduos é constante)
H_1: Heteroscedasticidade presente (a variância dos resíduos não é constante)

Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". *Econometrica*. **47** (5): 1287–1294. doi:10.2307/1911963.

Geração de séries sintéticas | modelos ARIMA: validação

O teste se utiliza (novamente) da série de resíduos ao quadrado a_t^2 como indicador local da variância

ou seja, a variância dos erros para cada observação z_t

para isso, ele assume que os resíduos são normalmente distribuídos

O procedimento envolve ajustar uma regressão linear à série a_t^2:

$$a_t^2 = \alpha t + \beta + \varepsilon_t$$

onde

α e β parâmetros da regressão

ε_t resíduos da regressão (não confundir com o resíduo a_t!!)

Geração de séries sintéticas | modelos ARIMA: validação

A estatística do teste é dada por:

$$BP = n \cdot R^2$$

onde

n tamanho da série a_t^2

R^2 coeficiente de determinação da regressão

A estatística BP tem distribuição qui-quadrado, com um grau de liberdade:

$$BP \sim \chi_1^2$$

Geração de séries sintéticas | modelos ARIMA: validação

A lógica do teste é verificar se podemos explicar a variância dos resíduos com uma regressão em função do tempo

- se o ajuste for bom (medido pelo R^2), a resposta é afirmativa e os resíduos são heteroscedástico

Geração de séries sintéticas | modelos ARIMA: validação

<u>Normalidade:</u>

Procedimento usual

análise visual via histogramas/densidades (ver aula 1)

testes estatísticos (ver aula 2)

Nota: a condição de normalidade dos resíduos é a mais difícil de ser atingida. Portanto, admite-se certo relaxamento nessa condição

pode-se verificar a aproximação dos resíduos da distribuição normal, via análises visuais

Geração de séries sintéticas | modelos ARIMA: validação

[Exemplo] Verificar as condições de independência, homoscedasticidade e normalidade da série de resíduos do início do capítulo

Independência (Ljung-Box)

```
Box.test(ruidoBranco, lag = 20, type = "Ljung-Box")
X-squared = 14.134, df = 20, p-value = 0.8236
```

Homoscedasticidade (Breusch-Pagan)

```
bptest(residuo ~ tempo)
BP = 1.481, df = 1, p-value = 0.2236
```

Normalidade (Shapiro-Wilk)

```
shapiro.test(ruidoBranco)
W = 0.99906, p-value = 0.9001
```

Resumo

Estimação: calcular os valores dos parâmetros do modelo identificado na etapa anterior

observar as condições de estacionariedade e invertibilidade para que o modelo seja estável

Base dos métodos de estimação

modelos puramente AR: Yule-Walker (MoM)

modelos puramente MA ou ARMA: soma dos quadrados dos resíduos (MLE)

Validação: verificação teórica do modelo, com base nos resíduos

comportamento de ruídos brancos gaussianos

independentes, homoscedásticos e (aprox.) normais

ERHA7016 – Hidrologia Estocástica

Daniel Detzel
detzel@ufpr.br