Identificação de *outliers* em modelos *k-Factor* Gegenbauer, resultados de simulação a partir do algoritmo SODA

Ian Danilevicz [1] [3]

Cleber Bisognin[2] [3]

Resumo: Neste trabalho seguimos aprimorando a identificação de *outliers* em processos com longa dependência. O modelo que estamos investigando, são os modelos da classe *k-Factor* Gegenbauer e o método de identificação é o SODA. Nesse algoritmo de identificação é atribuída uma estatística de teste para classificar cada uma das posições da série temporal como outlier aditivo, outliers inovador, ou não outlier. Para avaliar a utilidade dessa estatística Γ procedemos com estudos de simulação em que vários arranjos de modelos *k-Factor* Gegenbauer foram propostos e avaliamos a dispersão dessa estatística em cada um dos casos.

Palavras-chave: *processos estocásticos, longa dependência, Gegenbauer, Outliers*.

1 Introdução

No presente momento, a identificação de outliers ocupa uma posição de grande preocupação entre analistas e estatísticos, dada a dificuldade de encontrá-los e os grandes desequilíbrios que eles acarretam em modelos não robustos. No contexto de séries temporais essa questão se agrava, pois estamos lidando com dados correlacionados e as ferramentas para identificá-los deve levar em conta essa estrutura. O algoritmo SODA é desenhado de forma a levar em conta um modelo de série temporal e a partir desse filtro identificar os possíveis outliers por uma estatística Γ_i para as i observações em uma série. Dessa forma, valores altos da Γ_i indicam maior chance de estarmos em uma posição i que seja outlier.

A função SODA foi originalmente desenhada para processos da classe ARMA. No entanto, redesenhamos essa função para os modelos $k - Factor$ Gegenbauer, um tipo particular de processo estocástico sobre os quais estamos trabalhando a algum tempo. Neste resumo discutimos o comportamento dessa estatística Γ para diversos casos simulados, pois nos interessa saber o que é um valor alto ou baixo, para termos uma ferramenta acurada na identificação de outliers. Além disso, a presente estatística tem duas subdivisões a Γ_{AO} e a Γ_{IO}, valores adaptados para outliers aditivos e inovadores, respectivamente. Um

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *iandanilevicz@google.com*

[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *cleberbisognin@google.com*

[3]Agradecimento ao CNPq pela bolsa de Iniciação Científica

Outlier Aditivo é uma observação anômala que aparece pontualmente na série, já um Outlier Inovador instaura um subperíodo de observações anômalas na série causando uma mudança estrutural na média ou mesmo na variância do processo estocástico.

2 Resultados de Simulação

Nesta seção apresentamos os resultados de simulação para a identificação de outliers pelo método SODA. Todas as simulações tomam como base modelos $k-Factor$-GARMA (0,**u**,λ,0) com **u**={-0.7,0.5,0.8} e λ={0.1,0.2,0.3}, no entanto as contaminações não são sempre as mesmas. Nossa primeira figura 1(a) é apenas um exemplo de uma série contaminada por sete outliers do tipo AO e ω, magnitude de contaminação, igual a 5. Ao lado, 1(b) temos a mesma estrutura de contaminação para um caso multivariado, ou seja, mil séries. Escolhemos ω=5 por entendermos que esse é um outlier de tamanho médio.

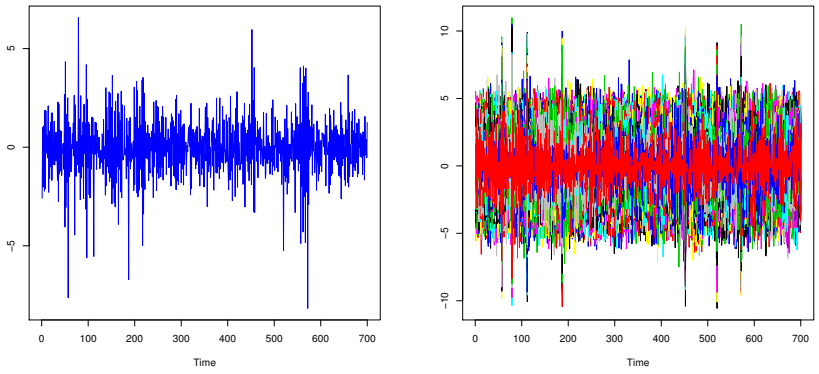

Figura 1: Séries Simuladas que seguem o modelo K-FACTOR-GARMA (0,**u**,λ,0) com **u**={-0.7,0.5,0.8} e λ={0.1,0.2,0.3}, com n=700 e 7 observações contaminadas por outliers AO de magnitude ω=5: (a) Modelo Univariado (b) Modelo Multivariado, ou seja, mil séries

Para testar a eficiência do algoritmo SODA geramos mil repetições de cada tipo de combinação de parâmetros, isto é, mil séries para cada ω = {0.5, 1, 3, 5}, para cada número de observações contaminadas, quais sejam 7, 14 ou 28, e, finalmente para cada tipo de contaminação, quais sejam aditivos e inovadores. No entanto, para poupar espaço apresentamos apenas os casos de ω extremos, ou seja, ω={0.5,5}. Como estamos procedendo com uma contaminação paramétrica, sabemos quais as posições contaminadas e por que tipo de outlier, em cada grupo de séries. Portanto, podemos separar as observações conforme a sua classe e avaliar os valores da estatística Γ.

Apresentamos uma tabela descritiva 2, média e desvio padrão, da estatística Γ do SODA que nos ajuda a classificar corretamente as observações de uma série como outliers ou não. Além disso po-

Tabela 1: Descritivas de Γ para 1000 séries de tamanho 700 com 7, 14 e 28 observações contaminadas por outliers tipo AO ou IO de magnitude ω=5

Série com Outliers Aditivos									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	3.369	2.873	0.465	3.369	2.866	0.477	3.360	2.847	0.495
Desv Pad	0.507	0.640	0.441	0.524	0.642	0.460	0.586	0.721	0.508
Série com Outliers Inovadores									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	1.622	1.448	0.249	1.636	1.445	0.261	1.635	1.441	0.283
Desv Pad	0.254	0.320	0.243	0.259	0.330	0.263	0.268	0.367	0.297

demos especificar esta estatística para outliers do tipo I e II. Para processos da classe ARMA, o valor de referência é 2, sendo valores superiores a dois fortes candidatos a serem corretamente classificados como outliers, e, se $\Gamma_{AO} > \Gamma_{IO}$ a observação deve ser do tipo I e, analogamente, se o contrário. Para os modelos K-FACTOR-GARMA ainda não temos esse valor de referência, nem temos certeza se o comportamento Γ_{AO} e Γ_{IO} é o mesmo.

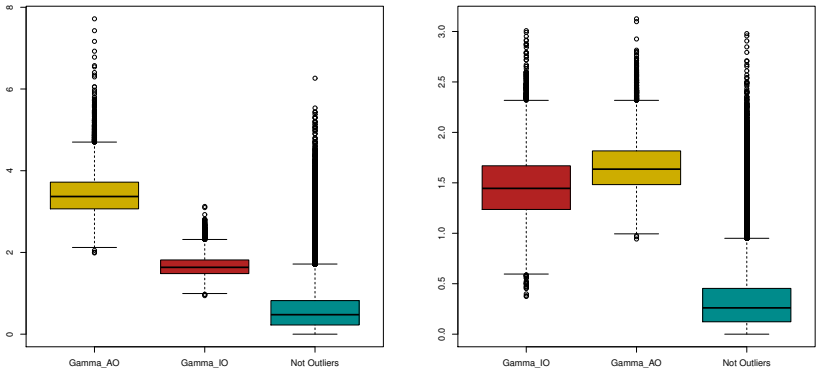

Figura 2: Box-Plot de Γ para 1000 séries de n=700 com 14 observações contaminadas por outliers de magnitude ω=5: (a) Séries contaminadas por AO, (b) Séries contaminadas por IO

Para ilustrar os resultados da tabela 1, apresentamos dois gráficos 2(a) e 2(b) para séries contaminadas por outliers do tipo I e II, respectivamente. Escolhemos apresentar somente o caso com um número intermediário de contaminações, ou seja, 14 observações contaminadas. Enfatizamos que o número de observações contaminadas pouco afeta a distribuição da estatística Γ, ou seja, os demais casos são semelhante. O mesmo não vale para quando mudamos os valores de magnitude, ω, dos outliers, como mostraremos mais adiante.

Pelo gráfico 2(a), fica claro que se a contaminação é do tipo I, podemos separar as observações

contaminadas das não contaminadas, além disso Γ_{AO} apresenta valores superiores a 3, Γ_{IO} valores intermediários entre 1.5 e 2, enquanto as observações não contaminadas valores abaixo de 1.5, com raras exceções. Já para o caso das séries contaminadas por outliers do tipo II 2(a), a separação entre observações contaminadas e não contaminadas funciona bem, no entanto a classificação quanto ao tipo de outlier não é tão simples, pois Γ_{AO} e Γ_{IO} ocupam regiões semelhantes. Sendo que a classe AO apresenta valores inclusive superiores em média, uma contradição, portanto, apenas ficamos seguros de estarmos diante de verdadeiros outliers do tipo I se Γ_{AO} for bem maior do que Γ_{IO}.

Tabela 2: Descritivas de Γ para 1000 séries de tamanho 700 com 7, 14 e 28 observações contaminadas por outliers tipo AO ou IO de magnitude ω=0.5

Série com Outliers Aditivos									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	0.624	0.590	0.378	0.615	0.592	0.380	0.613	0.593	0.379
Desv Pad	0.416	0.385	0.339	0.420	0.387	0.339	0.421	0.387	0.339
Série com Outliers Inovadores									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	0.302	0.279	0.173	0.296	0.282	0.173	0.297	0.282	0.173
Desv Pad	0.180	0.172	0.155	0.183	0.174	0.155	0.184	0.174	0.155

Os gráficos nos ajudam a perceber que as distribuições de todas as estatísticas Γ são assimétricas positivas, ou seja, com grande presença de valores extremos na cauda direita. Por esse motivo escolhemos trabalhar com a mediana ao invés da média em nossas tabelas descritivas, pois essa medida é mais robusta, ou seja, menos influenciável por valores extremos.

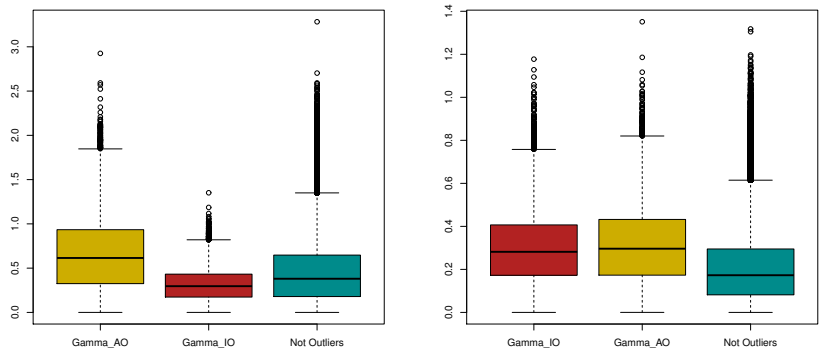

Figura 3: Box-Plot de Γ para 1000 séries de n=700 com 14 observações contaminadas por outliers de magnitude ω=0.5: (a) Séries contaminadas por AO, (b) Séries contaminadas por IO

Já a tabela e gráfico sobre os dados em que a contaminação é da ordem de ω=0.5 temos menos otimismo em identificar os outliers. Pois para o caso de contaminação do tipo AO, todos os tipos de Γ

flutuam na mesma região, ou seja, entre 0.2 e 1.0, sendo difícil a separação dos mesmos. A situação não melhora para o caso IO, em que Γ praticamente tem distribuições idênticas entre 0.15 e 0.4. Temos dessa forma, uma indicação dos limites do nosso trabalho, ou seja, outliers muito pequenos não são bem identificados pelo método proposto, qual seja o SODA especialmente desenhado para processos $k - Factor$ Gegenbauer.

3 Conclusões Parciais

Até este momento, sugerimos que outliers de magninitude media como ω=5 não são dificies de serem identificados pelo algoritmo SODA devidamente ajustado para o respectivo modelo $k - Factor$ Gegenbauer. No entanto, se a magnitude do outlier é muito pequena, como ω=0.5 temos maiores dificuldade de identificar esse outlier pois a estatística Γ das observações anômalas e não anômalas é muito semelhante. Porém, acreditamos que outliers assim pequenos não acarretam muitos problemas de alteração de parâmetros da série temporal, claro está que precisamos ainda investigar essa suposição.

Referências

[1] Fox, R.; Taqqu, M.S. "Large-Sample Properties of Estimates for Strongly Stationary Gaussian Time Series". *Annals of Statistics*, Vol. **14** (2), p.517-532, 1986.

[2] Woodward, W.A., Q.C. Cheng e H.L. Gray (1998). "A k-Factor GARMA Long-Memory Model". *Journal of Time Series Analysis*, Vol. **19**(4), pp. 485-504.