Classificação e agrupamento de séries temporais correlacionadas

Douglas Roberto Mesquita Azevedo

Orientador: Marcio Valk

IM - UFRGS - Brasil

Resumo

No presente trabalho é utilizado o método baseado em quase U-estatísticas proposto por [1] para classificação e agrupamento de séries temporais na presença de correlação entre as séries.

O interesse está em estender os resultados já obtidos para séries não correlacionadas e testar várias métricas como núcleo das U-estatísticas.

Com os resultados obtidos através de inúmeras simulações, foi possível perceber a robustez do método mesmo quando as séries possuem dependência.

Introdução

Atualmente, a criação de bancos de dados e os avanços na área da informática nos permitem o armazenamento e o processamento de um grande volume de dados, neste sentido se faz necessário desenvolver técnicas que reduzam a quantidade de dados sem redução significativa de informação.

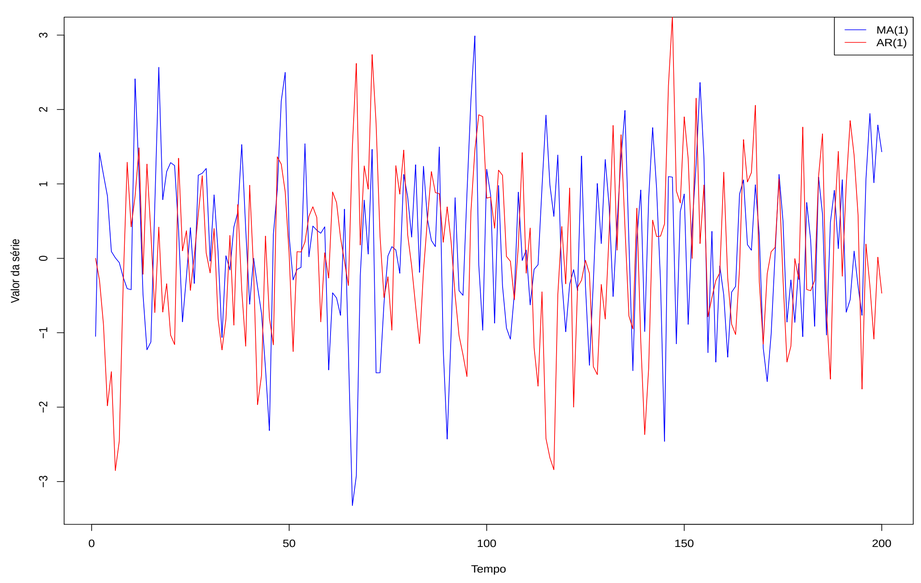
Nesta direção, [1] e [2], adaptando os resultados de [3] para séries temporais, propõe métodos de classificação e agrupamento baseados em quase U-estatísticas, cujas propriedades assintóticas são obtidas sob a hipótese de independência entre as séries.

O que propomos aqui é a generalização destes resultados assintóticos para U-estatísticas baseadas séries correlacionadas.

Com estes resultados devidamente comprovados, podemos então classificar e agrupar séries em grupos com comportamento semelhante.

Estas duas séries são iguais ou diferentes?

Séries temporais simuladas.



200 valores de séries AR(1) e MA(1) com parâmetros iguais (0.5) e com correlação = 0.5.

Resultados preliminares obtidos mostram que a técnica para classificação e agrupamento de séries temporais que está sendo proposta é bastante geral, aplicável a uma ampla classe de séries.

Métodos

O desenvolvimento deste projeto depende inicialmente da compreensão da teoria de U-estatísticas.

U-estatísticas formam uma classe muito ampla de funcionais especialmente importantes na teoria de estimação. Elas surgem naturalmente na construção de estimadores não-viesados de mínima variância.

Nelas, estão incluídas muitas estatísticas usuais, como por exemplo, os estimadores da média e variância.

As técnicas mais conhecidas para agrupar e classificar séries temporais se baseiam em métricas, que são utilizadas para medir a distância ou dissimilaridade entre séries temporais, no presente trabalho algumas métricas serão utilizadas como núcleo de U-estatísticas.

Para detectar dissimilaridade entre grupos de séries temporais, foram utilizadas três métricas:

- Periodograma Normalizado, o qual será chamado de LP
- Logaritmo do Periodograma Normalizado, o qual será chamado de LNP
- Autocorrelação Serial, o qual será chamado de ACF

Utilizamos a estatística Bn proposta por [1] para identificação ou não de dissimilaridade entre grupos de séries a partir de simulações de diferentes processos e parametrizações. São comparados também grupos de séries de diferentes tamanhos e com alguns níveis de dependência entre as séries.

Resultados

Serão apresentados dois resultados, um para classificação de séries e outro para agrupamento de séries.

Para cada combinação de n (número de elementos na série) e ρ (correlação entre as séries) foi aplicado o teste 25 vezes e em cada uma destas vezes foram utilizados 1000 bootstraps.

Todas as análises foram implementadas no software R Versão 3.0.1.

Classificação de séries temporais

P-valores do teste de homogeneidade. AR(1)x MA(1).

$\phi = \theta = 0.4$						$\phi = \theta = 0.4$					
N = 4						N=10					
n		100	300	500	1000	n		100	300	500	1000
ρ	Distância					ρ	Distância				
0	DLNP	0.425	0.325	0.367	0.401	0	DLNP	0.300	0.564	0.224	0.384
0	DNP	0.614	0.282	0.468	0.399	0	DNP	0.248	0.538	0.212	0.360
0	DACF	0.324	0.213	0.368	0.407	0	DACF	0.078	0.368	0.115	0.061
0.3	DLNP	0.160	0.193	0.355	0.520	0.3	DLNP	0.210	0.166	0.204	0.292
0.3	DNP	0.101	0.488	0.364	0.512	0.3	DNP	0.246	0.125	0.170	0.291
0.3	DACF	0.222	0.413	0.343	0.195	0.3	DACF	0.012	0.020	0.086	0.036
0.5	DLNP	0.178	0.230	0.206	0.432	0.5	DLNP	0.003	0.011	0.022	0.063
0.5	DNP	0.141	0.173	0.245	0.427	0.5	DNP	0.014	0.004	0.021	0.086
0.5	DACF	0.110	0.238	0.203	0.356	0.5	DACF	0.682	0.000	0.009	0.038
0.8	DLNP	0.026	0.061	0.109	0.184	0.8	DLNP	0.000	0.000	0.000	0.006
0.8	DNP	0.018	0.056	0.113	0.193	0.8	DNP	0.000	0.000	0.000	0.009
0.8	DACF	0.060	0.066	0.028	0.063	0.8	DACF	0.000	0.000	0.000	0.000
0.99	DLNP	0.006	0.007	0.006	0.007	0.99	DLNP	0.000	0.000	0.000	0.000
0.99	DNP	0.006	0.006	0.006	0.007	0.99	DNP	0.000	0.000	0.000	0.000
0.99	DACF	0.004	0.005	0.006	0.005	0.99	DACF	0.000	0.000	0.000	0.000
$\phi = \theta = 0.9$						$\phi = \theta = 0.9$					
0	DLNP	0.017	0.006	0.007	0.007	0	DLNP	0.013	0.000	0.000	0.000
0	DNP	0.075	0.026	0.022	0.024	0	DNP	0.000	0.000	0.000	0.000
0	DACF	0.010	0.008	0.006	0.008	0	DACF	0.000	0.000	0.000	0.000
0.3	DLNP	0.012	0.004	0.006	0.007	0.3	DLNP	0.000	0.000	0.000	0.000
0.3	DNP	0.035	0.017	0.019	0.022	0.3	DNP	0.000	0.000	0.000	0.000
0.3	DACF	0.007	0.015	0.005	0.007	0.3	DACF	0.000	0.000	0.000	0.000
0.5	DLNP	0.008	0.003	0.009	0.007	0.5	DLNP	0.000	0.000	0.000	0.000
0.5	DNP	0.014	0.023	0.019	0.011	0.5	DNP	0.000	0.000	0.000	0.000
0.5	DACF	0.012	0.005	0.011	0.006	0.5	DACF	0.000	0.000	0.000	0.000
0.8	DLNP	0.005	0.007	0.005	0.007	0.8	DLNP	0.000	0.000	0.000	0.000
0.8	DNP	0.008	0.008	0.008	0.008	0.8	DNP	0.000	0.000	0.000	0.000
0.8	DACF	0.007	0.004	0.004	0.004	0.8	DACF	0.000	0.000	0.000	0.000
0.99	DLNP	0.003	0.007	0.006	0.004	0.99	DLNP	0.000	0.000	0.000	0.000
0.99	DNP	0.003	0.005	0.006	0.007	0.99	DNP	0.000	0.000	0.000	0.000
0.99	DACF	0.003	0.005	0.003	0.003	0.99	DACF	0.000	0.000	0.000	0.000

Foram utilizadas 4 séries em cada grupo.

Agrupamento de séries temporais

Taxa de acerto no agrupamento de séries temporais.

AR(1)xMA(1), ϕ= 0.5 e θ = 0.8														
ρ = 0.5								ρ = 0.8						
Método	Métrica	Min	Q1	Med	Média	Q3	Max	Var	Min	Q1	Med	Média	Q3	Max	Var
Bn	LNP	0.88	1.00	1.00	1.00	1.00	1.00	0.00002	0.75	1.00	1.00	1.00	1.00	1.00	0.00012
Bn	LP	0.50	1.00	1.00	0.96	1.00	1.00	0.01745	0.50	0.50	0.50	0.50	0.50	0.50	0.00000
Bn	ACF	0.75	1.00	1.00	1.00	1.00	1.00	0.00008	0.75	1.00	1.00	0.99	1.00	1.00	0.00170
Clusters	LNP	0.63	0.75	1.00	0.91	1.00	1.00	0.01864	0.75	0.75	0.88	0.88	1.00	1.00	0.01564
Clusters	LP	0.50	1.00	1.00	0.93	1.00	1.00	0.02351	0.50	0.50	0.50	0.50	0.50	0.50	0.00000
Clusters	ACF	0.63	0.75	1.00	0.92	1.00	1.00	0.02021	0.63	0.75	1.00	0.93	1.00	1.00	0.01289

Foram utilizadas 4 séries em cada grupo.

Bibliografia

- Valk, M. (2011). O Uso de Quase U-Estatísticas para Séries Temporais Uni e Multivariadas. Tese de Doutorado. Universidade Estadual de Campinas.
- Valk, M. and A. Pinheiro (2012). Time-series clustering via quasi U-statistics. *Journal of Time Series Analysis*, vol.33(4), 608–619.
- Pinheiro A., P.K. Sen and H.P. Pinheiro (2009). Decomposability of high-dimensional diversity measures: Quasi *U*-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, vol. 100(8), 1645-1656.

douglas-mesquita@hotmail.com