Verossimilhança Composta: Um Estudo Através de Simulações

Palavras-Chave: Verossimilhança aproximada, comparação de estimadores, valores aberrantes

Leonardo Mazzamboni Colussi [IMECC] - [UNICAMP]

Prof. Dr. Luiz Koodi Hotta (orientador) [IMECC] - [UNICAMP]

INTRODUÇÃO:

O método da verossimilhança composta é uma alternativa aos estimadores de máxima verossimilhança (EMV) quando o modelo é complexo e/ou sua dimensão é muito alta. Dessa forma, uma alternativa é a verossimilhança composta, que é uma verossimilhança aproximada, calculada através de combinações de termos de dimensões menores, reduzindo a complexidade e os custos computacionais.

DEFINIÇÕES:

Seja $\boldsymbol{y} = (y_1, \dots, y_m)^T$ o vetor de uma única observação amostral de uma variável aleatória de dimensão m, cuja densidade é $f(\boldsymbol{y}; \boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é um vetor de parâmetros de dimensão p, tal que $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Ainda, seja $\{A_1, \dots, A_K\}$ o conjunto de eventos condicionais e marginais, que é uma partição do espaço amostral, associados às verossimilhanças $\mathcal{L}_K \propto f(\boldsymbol{y} \in A_K; \theta)$ (Varin e Vidoni, 2011; Varin et al., 2011).

A verossimilhança composta $\left(\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y})\right)$ é dada pelo produto das verossimilhanças de um conjunto de observações de uma amostra elevada aos seus respectivos pesos $(w_i, w_i \geq 0)$, ou seja:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{K} \mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{y})^{w_i}.$$

Sendo assim, a log-verossimilhança composta ($\mathrm{l}_c(\boldsymbol{\theta}; \boldsymbol{y})$) será:

$$l_c(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{K} w_i \, \mathrm{l}_i(\boldsymbol{\theta}; \boldsymbol{y}),$$

em que $\mathrm{l}_i(\boldsymbol{\theta}; \boldsymbol{y}) = log[\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{y})]$.

A verossimilhança composta pode ser dividida em dois grandes grupos: condicionais e marginais.

Condicionais:

No contexto de estudos longitudinais (Molenberghs e Verbeke, 2005) e bioinformática (Mardia et al., 2008), obteve-se a verossimilhança composta por agrupamento de densidades condicionais de pares:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{r=1}^{m} \prod_{s=1}^{m} f(y_r | y_s; \theta).$$

Alternativamente, agrupando densidades condicionais completas:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{r=1}^{m} f\left(y_r \middle| y_{(-r)}; \boldsymbol{\theta}\right),$$

em que $y_{(-r)}$ representa o vetor de todas as observações exceto y_r .

Marginais:

A verossimilhança marginal composta mais simples é a usualmente conhecida, a qual assumimos independência da amostra (Chandler e Bate, 2007):

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{m} f(y_i; \boldsymbol{\theta}).$$

No entanto, quando os parâmetros de interesses são relacionados por meio de dependência, pode-se modelar blocos de observações, a exemplo da verossimilhança de pares:

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \boldsymbol{\theta}),$$

obtida de um total de $\binom{m}{2}$ pares de verossimilhanças. Ainda, em alguns casos, há a possibilidade de selecionar aleatoriamente tais pares, de forma que não seja necessário a utilização de todos para a obtenção dos estimadores de máxima verossimilhança composta de pares, diminuindo ainda mais o custo computacional.

METODOLOGIA E OBJETIVO:

O método foi implementado dentro do contexto de Séries Temporais Financeiras, com modelos que envolvem volatilidade, em particular o modelo cDCC (corrected dynamic conditional correlation) (Aielli, 2013) que pode ser considerado como uma das extensões multivariada dos modelos GARCH (Generalized Autoregressive Conditional Heteroskedasticity) de volatilidade univariada (Bollerslev, Engle e Wooldridge, 1988). Além das propriedades teóricas do modelo, foi realizado um estudo de simulações

de Monte Carlo para se estimar os parâmetros α e β deste modelo por meio da verossimilhança composta de pares (Pakel et al., 2021; Cont, 2000).

Nas simulações foram consideradas séries temporais de tamanho amostral e dimensões iguais a $T = 500$, $L \in \{3, 5, 10, 50, 100\}$, respectivamente, e, para alta dimensão, $T = 1000$ e $L = 750$. A matriz de correlação não condicional do modelo cDCC foi tomada como em (Pakel et al., 2021). Ainda, contaminou-as com percentuais diferentes de *outliers* ($\delta \in \{0\%, 25\%, 50\%\}$), localizados na observação central, nas duas centrais e aleatoriamente das séries, os quais foram obtidos acrescentando 5 vezes o desvio-padrão condicional da observação em que o *outlier* foi acrescentado, com o *outlier* tendo o mesmo sinal que a observação.

Além do estimador de máxima verossimilhança (verossimilhança clássica), foram estudadas as seguintes formas da verossimilhança composta por pares: $i)$ todos os possíveis pares ($MCLE$); $ii)$ apenas uma permutação de pares contíguos ($MSCLE$); $iii)$ N permutações de pares das séries temporais, $N \in \{1, 10, 30, 50, 100\}$, com o estimador dado pela média das N estimativas ($MSCLEp$); $iv)$ N permutações de pares das séries temporais, $N \in \{1, 10, 30, 50, 100\}$, com a verossimilhança dada pela soma das verossimilhanças das N permutações ($MSCLE_S$). Os quatro estimadores baseados no método da verossimilhança composta foram comparados entre si e com o estimador de máxima verossimilhança em termos de vício, variância, erro quadrático médio e custo computacional.

RESULTADOS E CONCLUSÕES:

Dos resultados, notou-se que a superfície da log-verossimilhança clássica tende a ficar plana para altas dimensões, não se obtendo estimativas confiáveis. Por sua vez, esse problema não ocorreu considerando todos os possíveis pares das séries temporais, no entanto, o custo computacional ainda foi muito alto. Logo, a alternativa mais adequada neste cenário foi considerar pares contíguos, em que o custo computacional foi significativamente mais baixo e as estimativas foram factíveis.

Além disso, o número N de permutações das séries não influenciou na média das estimativas, apenas no desvio-padrão, de modo que, quanto maior o número de permutações realizadas, menor foi o valor para esta estimativa.

A partir de certo número de permutações o erro quadrático médio é dominado pelo vício. Como o vício não depende do número de permutações, a partir de certo ponto, o aumento do número de permutações não é compensado pela diminuição do erro quadrático médio. Por exemplo, na Figura 1 é colocado o EQM das estimativas dos parâmetros α e β no método $MSCLEp$, considerando um valor do vício e variâncias obtidos através de simulações e considerando independência entre os estimadores das diferentes permutações, suposições aproximadamente válidas.

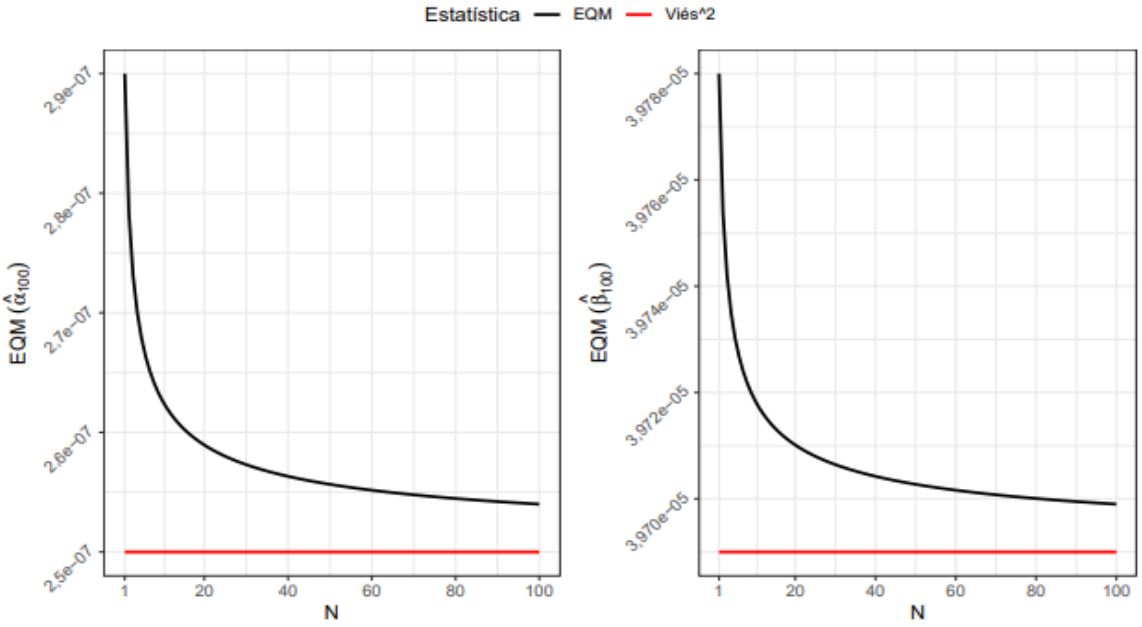

Figura 1 – Evolução do EQM de $\hat{\alpha}_N$ e $\hat{\beta}_N$ conforme o aumento do número (N) de permutações simuladas, utilizando o método $MSCLEp$; $N \in [1{,}100]$.

Ademais, os efeitos dos *outliers* foram diferentes para as estimativas de cada um dos dois parâmetros: para α, quanto maior foi o grau de contaminação, mais subestimadas as estimativas ficavam, enquanto para β, as estimativas ficavam superestimadas (Figura 2).

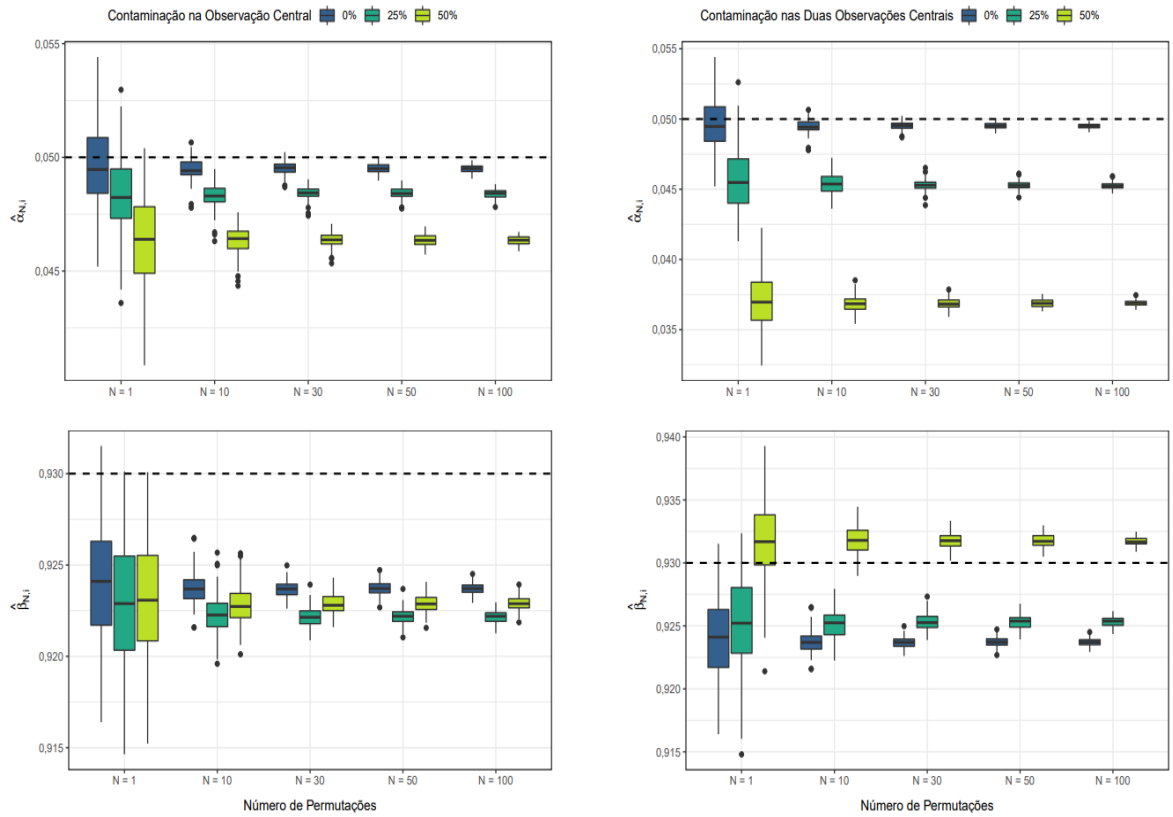

Figura 2 - Boxplots dos valores das estimativas de $\hat{\alpha}_{N,i}$ e $\hat{\beta}_{N,i}$ referentes às 100 replicações de N permutações, pelo método $MSCLEp$; porcentagem de contaminação de outliers em $\delta \in \{0\%, 25\%, 50\%\}$, dispostos na observação central (à esquerda) e nas duas centrais (à direita); $T = 500$ e $L = 100$; verdadeiro valor do parâmetro (linha tracejada): $\alpha = 0{,}05$ e $\beta = 0{,}93$.

Como trabalhos futuros, pode-se considerar alternativas de diminuir o vício e estimadores robustos a *outliers*.

AGRADECIMENTOS:

Os autores agradecem o apoio da FAPESP: Leonardo (Processo Nº 2021/01516-7, bolsa de iniciação científica) e Luiz (Processo Nº 2018/04654-9, projeto temático).

BIBLIOGRAFIA

AIELLI, G. P. Dynamic condicional correlation: on properties and estimation. **Journal of Busines & Economic Statistics**, v. 31 (n.3), p. 282-299, 2013.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. **Journal of Busines & Economic Statistics**, v. 31 (n.3), p. 307-327, 1986.

CHANDLER, R. E. e BATE, S. Inference for clustered data using the independence loglikelihood. **Biometrika,** v. 94, p. 167-183, 2007.

CONT, R. Empirical properties of asset returns: stylized facts and statistical issues. **Quantitative Finance**, p. 223–236, 2000.

MARDIA, K. V., HUGHES, G., TAYLOR, C. C. e SINGH, H. A multivariate von Mises distribution with applications to bioinformatics. **Canadian Journal of Statistics**, v. 36, p. 99-109, 2008.

MOLENBERGHS, G. e VERBEKE, G. **Models for Discrete Longitudinal Data**. Nova Iorque, Springer-Verlag, 2005.

PAKEL, C., SHEPARD S., SHEPARD K. e Engle F. R. Fitting vast dimensional time-varying covariance models. **Journal of Busines & Economic Statistics**, v. 39, n. 3, p. 652–668, 2021.

VARIN, C. e P. VIDONI. Pairwise Likelihood Inference for General State Space Models. **Econometric Reviews**, v. 28 (1-3), p. 170-185, 2011.

VARIN, C., N. Reid e D. Firth. An overview of composite likelihood methods. **Statistica Sinica**, p. 5–42, 2011.