



Detecção de *Outliers* em Modelos de Fatores Dinâmicos

Palavras-chave: Detecção de *Outliers* ; Modelos de Fatores Dinâmicos; Séries Temporais.

Autores:

Cauã Pereira Masseu, IMECC – UNICAMP

Prof. Dr Luiz K. Hotta (orientador), IMECC – UNICAMP

Prof. Dr. Pedro Galeano (orientador externo, coautor), UC3M

1 Introdução

A estimativa da matriz de volatilidade condicional de séries de retornos financeiros é muito importante em finanças. Neste sentido, vários modelos estatísticos foram propostos; por exemplo, os modelos da família GARCH multivariados como BEKK, CCC e cDCC. Entretanto, com os avanços tecnológicos e a crescente conectividade global, a obtenção de dados tornou-se cada vez mais fácil e abundante, resultando em séries temporais de alta dimensão, consequentemente, no surgimento do problema de *curse of dimensionality*, que torna a estimação dos modelos inviável devido à elevada quantidade de parâmetros a serem estimados, conforme discutido em De Almeida, Hotta e Ruiz (2018).

Uma abordagem promissora em séries temporais para contornar essa limitação envolve o uso de modelos de fatores dinâmicos, que reduzem a dimensionalidade ao representar a estrutura das séries temporais por meio de combinações lineares de poucos fatores latentes e componentes idiossincráticas; veja, por exemplo, Stock e Watson (2016). Tais modelos preservam a capacidade de capturar a comovimentação entre séries, ao mesmo tempo que mitigam a complexidade computacional.

Por outro lado, a presença de valores extremos ou aberrantes, conhecidos como *outliers*, é frequente em dados de finanças e compromete significativamente a estimação dos modelos, conforme discutido por Carnero, Daniel Peña e Ruiz (2012), consequentemente, podendo prejudicar a qualidade das aplicações, como previsões. Isso ressalta a importância de uma atenção especial a esse tipo de dado, promovendo o desenvolvimento de métodos de estimação mais robustos. Ademais, a detecção de *outliers* fornece informações relevantes sobre a estrutura e possíveis anomalias nos dados.

Para abordar simultaneamente os desafios da dimensionalidade e da presença de *outliers*, Galeano e D. Peña (2024) propõem um método de detecção de *outliers* em modelos aproximados de fatores dinâmicos, que torna a estimação mais robusta por meio da identificação e remoção desses valores atípicos. O método foi desenvolvido e aplicado exclusivamente para *outliers* aditivos, que constituem o tipo mais frequente nas séries temporais. Contudo, outras formas de *outliers* também ocorrem em séries financeiras, como inovativos e de mudança de nível (*level shift*). Dessa forma, torna-se fundamental avaliar o desempenho do método frente a esses diferentes tipos de *outliers*, com o objetivo

de identificar possíveis limitações e apontar ajustes necessários para aprimorá-lo. Nesse contexto, o presente trabalho teve como objetivo avaliar, por meio de simulações controladas, a eficácia do método de Galeano e Peña (2024) na detecção de variados tipos de *outliers* em modelos de fatores dinâmicos, evidenciando que sua eficácia é limitada para *outliers* distintos dos aditivos.

2 Metodologia

Seja $\mathbf{x}_t = (x_{1t}, \dots, x_{mt})'$, para $t = 1, \dots, T$, um vetor m -dimensional estacionário com média zero. O modelo aproximado de fatores dinâmicos assume que $\mathbf{x}_t$ pode ser expressa como uma combinação linear de fatores dinâmicos e de componentes idiossincráticas, ou seja:

$$\mathbf{x}_t = \mathbf{P} \mathbf{f}_t + \boldsymbol{\eta}_t, \quad (1)$$

onde $\mathbf{P}$ é a matriz de carga de dimensão $m \times r$, $\mathbf{f}_t = (f_{1t}, f_{2t}, \dots, f_{rt})'$ representa os r factores dinâmicos e $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{mt})'$ representa as componentes idiossincráticas, modeladas geralmente como um ruído branco gaussiano multivariado. Na literatura, é comum considerar os fatores como modelos autorregressivos de ordem baixa.

Seguindo a Baragona e Battaglia (2007), os valores atípicos em modelos de fatores dinâmicos estacionários se incorporam da seguinte maneira:

$$\mathbf{x}_t^* = \mathbf{P} \mathbf{f}_t + \boldsymbol{\eta}_t + \mathbf{w} I_t^{(o)} = \mathbf{x}_t + \mathbf{w} I_t^{(o)}, \quad (2)$$

onde $\mathbf{w} = (w_1, \dots, w_m)'$ representa a magnitude do valor atípico no instante de tempo $t = o$, determinado pela variável indicadora $I_t^{(o)}$, que assume valor 1 quando $t = o$ e 0 caso contrário.

Na equação (2), os valores atípicos são considerados de maneira conjunta, sem distinguir entre aqueles que afetam os fatores e os que estão presentes nas componentes idiossincráticas. Entretanto, estes valores não podem ser identificados diretamente; veja, Baragona e Battaglia (2007). Para abordar esta limitação, Galeano e D. Peña (2024) propõem distingui-los, separando os atípicos que afetam os fatores daqueles que influenciam nas componentes idiossincráticas, com o objetivo de facilitar sua identificação. No entanto, ambos tipos de valores atípicos ainda podem ocorrer simultaneamente, afetando tanto os fatores como as componentes idiossincráticas.

2.1 O método de Galeano e D. Peña (2024)

O método proposto por Galeano e D. Peña (2024) apresenta um procedimento sistemático para a detecção de *outliers* em séries temporais estacionárias de alta dimensão. Esse procedimento baseia-se nas propriedades dos Modelos de Fatores Dinâmicos (DFM) e pode ser resumido nas quatro etapas descritas a seguir. Dada a série temporal observada $x_1^*, \dots, x_T^*$, de maneira resumida, o processo segue o seguinte esquema:

- 1. Pré-processamento inicial:** realiza-se uma limpeza preliminar da série temporal $x_1^*, \dots, x_T^*$, obtendo-se uma série modificada $\tilde{x}_1, \dots, \tilde{x}_T$. A partir desta série, obtêm-se estimativas robustas para o número de fatores r , para a matriz de cargas fatoriais $\mathbf{P}$ e para a matriz complemento ortogonal $\mathbf{V}$, denotadas por $\hat{r}$, $\hat{\mathbf{P}}$ e $\hat{\mathbf{V}}$, respectivamente.
- 2. Detecção de *outliers* idiossincráticos:** projeta-se a série original $x_1^*, \dots, x_T^*$ com $\hat{\mathbf{V}}^\top$ para detectar e limpar a série de *outliers* idiossincráticos, resultando na série $x_1^\circ, \dots, x_T^\circ$.
- 3. Detecção de *outliers* fatoriais:** projeta-se a série $x_1^\circ, \dots, x_T^\circ$ com $\hat{\mathbf{P}}^\top$ para detectar e limpar a série de *outliers* nos fatores, resultando na série $x_1^\bullet, \dots, x_T^\bullet$.

4. Iteração: utiliza-se a série $x_1^\bullet, \dots, x_T^\bullet$ para reestimar a matriz de cargas fatoriais $\mathbf{P}$ e a matriz complemento ortogonal $\mathbf{V}$, e então repete-se os passos 2, 3 e 4, substituindo a série original $x_1^*, \dots, x_T^*$ pela série $x_1^\bullet, \dots, x_T^\bullet$, até que não sejam mais encontrados *outliers*. Esse passo contribui para a mitigação dos efeitos de *masking* e *swamping* na detecção de *outliers*.

Essas quatro etapas são automáticas e permitem detectar e remover *outliers* nos fatores e nas componentes idiossincráticas da série temporal observada. Como recomendado pelos autores, caso também haja interesse em ajustar um DFM, recomenda-se estimar conjuntamente os parâmetros do modelo e os efeitos dos *outliers*, refinando o modelo ajustado, se necessário.

2.1.1 Outros tipos de *outliers*

Quando os fatores são afetados por *outliers* no instante de tempo o , temos:

$$\mathbf{f}_t^* = \mathbf{f}_t + \Psi \mathbf{w}_k^{(o)} I_t^{(o)}, \quad (3)$$

onde $\mathbf{w}_k^{(o)} = (0, \dots, w_k^{(o)}, \dots, 0)'$ representa a magnitude do *outlier* no k -ésimo fator. Além disso, a matriz Ψ , a qual determina o tipo do *outlier* é dada por:

$$\Psi = \begin{bmatrix} \psi_1(B) & 0 & 0 & \cdots & 0 \\ 0 & \psi_2(B) & 0 & \cdots & 0 \\ 0 & 0 & \psi_3(B) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \psi_r(B) \end{bmatrix}$$

Assim, para cada fator individual:

$$f_{jt}^* = f_{jt} + \psi_j(B) w_j^{(o)} I_t^{(o)}, \quad j = 1, \dots, r \quad (4)$$

O único tipo de *outlier* analisado para o qual o procedimento de Galeano e D. Peña (2024) foi desenhado é o *outlier* **aditivo**. Para incluir *outlier* aditivos nos fatores, na equação (4), assumimos $\psi_j(B) = 1$. De este modo, a matriz Ψ na equação (3) é a matriz identidade de ordem r . Apenas neste tipo de *outlier* os valores atípicos também se introduzem nas componentes idiossincráticas, as quais tem sido especificamente modeladas para permitir sua identificação e tratamento adequado. Para incluir *outlier* consecutivos, basta considerar valores atípicos nos instantes de tempo t_j e $t_j + 1$.

Os *outliers* **inovativos** podem considerar-se como *outlier* aditivos consecutivos cujo tamanho diminui com o tempo. Neste caso, temos que $\psi_j(B) = \frac{\theta_j(B)}{\phi_j(B)}$ na equação (4).

Já os *outliers* de **mudança de nível** podem ser vistos como uma sequência de *outliers* aditivos consecutivos em todas as observações posteriores à ocorrência da mudança de nível, isto é, nos fatores, na equação (4) temos $\psi_j(B) = \frac{1}{1-B}$; veja Chen e Liu (1993) para mais detalhes nos diferentes tipos de *outliers*.

3 Resultados e Discussão

As simulações realizadas tiveram como objetivo avaliar o desempenho do método proposto em diferentes cenários, com diferentes parâmetros, utilizando diversas métricas, igualmente ao adotado em Galeano e D. Peña (2024). Entre essas métricas, destacam-se: a proporção de *outliers* corretamente e erroneamente detectados na etapa de limpeza inicial; a proporção de acertos na estimativa do número

de fatores; e a proporção de *outliers* corretamente e erroneamente identificados tanto na etapa de limpeza dos fatores quanto nas componentes idiossincráticas. Importante destacar que as posições dos *outliers* nos fatores não ocorreram nem no início nem no final da série, seguindo o mesmo critério adotado no artigo original.

Foram considerados três tipos de *outliers*: inovativos, de mudança de nível e dois *outliers* aditivos consecutivos (também conhecidos como *patch*). Os resultados obtidos, como vemos nas Tabelas 1, 2, 3 e 4, indicam uma degradação no desempenho do procedimento diante da presença de outros tipos de *outliers*, na maioria dos cenários e métricas avaliadas. Optou-se por inserir somente os resultados obtidos para *outlier* inovativos por via de ilustração.

Dando continuidade ao estudo, vêm sendo conduzidas análises da acurácia na estimação e previsão do modelo sob a presença de *outliers*, com a integração do método de detecção proposto a técnicas de estimação, como em Doz, Giannone e Reichlin (2011) e Barigozzi e Luciani (2024).

Tabela 1: Porcentagem média da quantidade de fatores corretamente selecionados com *outliers* inovativos.

		m = 50			m = 100			m = 200		
		100	200	400	100	200	400	100	200	400
w	5	100	100	100	100	100	100	99.6	100	100
	7.5	100	100	100	98.8	100	100	94.8	100	100
	10	100	100	100	100	96.4	100	83.2	100	100

Tabela 2: Porcentagem média de detecção de *outliers* na limpeza inicial com *outliers* inovativos.

		m = 50			m = 100			m = 200		
		100	200	400	100	200	400	100	200	400
Reais	5	18.94	14.43	11.17	21.1	18.64	12.25	20.37	16.95	17.07
	7.5	21.08	16.89	11.93	24.11	19.97	13.08	23.74	18.75	16.68
	10	23.71	18.51	12.75	26.5	21.64	14.1	26.84	19.66	17.58
Falsos	5	21.20	14.81	11.32	20.85	18.58	11.74	18.61	15.88	16.60
	7.5	24.87	16.85	11.85	23.13	19.22	12.17	21.43	15.43	16.30
	10	30.05	19.60	12.42	27.40	21.80	12.92	25.15	16.85	16.54

Tabela 3: Porcentagem média de detecção de *outliers* nos fatores com *outliers* inovativos.

		m = 50			m = 100			m = 200		
		100	200	400	100	200	400	100	200	400
Reais	5	86.00	83.60	86.40	89.40	85.40	86.80	70.30	73.00	77.70
	7.5	96.40	98.00	99.60	96.60	97.60	99.40	93.10	97.70	98.80
	10	97.60	99.20	100.0	96.20	96.40	100.0	86.20	99.50	99.80
Falsos	5	0.81	0.43	0.20	0.41	0.24	0.10	0.24	0.10	0.04
	7.5	1.40	0.78	0.41	0.94	0.49	0.25	0.62	0.27	0.13
	10	1.88	1.00	0.57	1.32	0.74	0.40	1.13	0.48	0.25

4 Conclusões

Este projeto proporcionou uma base sólida sobre os fundamentos dos modelos fatoriais dinâmicos e sobre práticas de simulação, abordando um tema ainda pouco explorado na literatura: a presença de

Tabela 4: Porcentagem média de detecção de *outliers* erroneamente identificados nas componentes idiossincráticas com *outliers* inovativos.

$\begin{smallmatrix} T \\ w \end{smallmatrix}$	$m = 50$			$m = 100$			$m = 200$		
	100	200	400	100	200	400	100	200	400
5	0.26	0.02	0.001	0.03	0.24	0.01	0.03	0.01	0.18
7.5	0.28	0.02	0.001	0.08	0.25	0.01	0.12	0.02	0.16
10	0.28	0.02	0.001	0.01	0.25	0.01	0.26	0.05	0.16

outliers em tais modelos.

Os resultados obtidos evidenciaram limitações no método de Galeano e D. Peña (2024), especialmente frente a diferentes tipos de *outliers*, o que reforça a necessidade de aprimoramentos metodológicos, oferecendo uma contribuição relevante e motivações para investigações futuras.

Como desdobramento do trabalho, vêm sendo conduzidas análises adicionais de estimação e previsão em modelos aproximados de fatores dinâmicos, com o objetivo de avaliar os impactos da presença de *outliers* quando não tratados adequadamente.

Agradecimentos: Cauã agradece o auxílio da FAPESP (processos 2024/20677-0 e 2024/08171-3) e Luiz agradece auxílio da FAPESP (processos 2023/01728-0 e 2023/02538-0). Agradecemos ao Prof. Dr. Pedro Galeano por acolher Cauã como estudante temporário durante dois meses.

Referências

- Baragona, Roberto e Francesco Battaglia (out. de 2007). “Outliers in dynamic factor models”. Em: *Electronic Journal of Statistics* 1, pp. 392–432.
- Barigozzi, Matteo e Matteo Luciani (2024). *Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the EM algorithm*. arXiv:1910.03821v5 [math.ST].
- Carnero, M. Angeles, Daniel Peña e Esther Ruiz (2012). “Estimating GARCH volatility in the presence of outliers”. Em: *Economics Letters* 114.1, pp. 86–90.
- Chen, C. e L. M. Liu (1993). “Joint Estimation of Model Parameters and Outlier Effects in Time Series”. Em: *Journal of the American Statistical Association* 88, pp. 284–297.
- De Almeida, Daniel, Luiz K Hotta e Esther Ruiz (2018). “MGARCH models: Trade-off between feasibility and flexibility”. Em: *International Journal of Forecasting* 34.1, pp. 45–63.
- Doz, Catherine, Domenico Giannone e Lucrezia Reichlin (2011). “A two-step estimator for large approximate dynamic factor models based on Kalman filtering”. Em: *Journal of Econometrics* 164.1. Annals Issue on Forecasting, pp. 188–205.
- Galeano, P. e D. Peña (2024). “Detecting outliers in high-dimensional time series by dynamic factor models”. Em: *Recent Advances in Econometrics and Statistics: Festschrift in Honour of Marc Hallin*. Ed. por Matteo Barigozzi, Siegfried HÖrmann e Davy Paindaveine. Springer, pp. 361–383.
- Stock, J.H. e M.W. Watson (2016). “Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics”. Em: ed. por John B. Taylor e Harald Uhlig. Vol. 2. Handbook of Macroeconomics. Elsevier, pp. 415–525.