Deteção de *outliers* no modelo de equações simultâneas usando o estimador GMM robusto

Anabela Rocha
ISCA e CIDMA, Universidade de Aveiro, *anabela.rocha@ua.pt*

Manuela Souto de Miranda
DMat e CIDMA, Universidade de Aveiro, *manuela.souto@ua.pt*

João Branco
IST e CEMAT, Universidade de Lisboa, *jbranco@math.ist.utl.pt*

Palavras–chave: SEM (*Simultaneous Equation Model*), SUR (*Seemingly Unrelated Regressions*), robustez, *outliers*, GMM (*Generalized Method of Moments*).

Resumo: O modelo SEM é uma generalização do modelo de regressão multivariado que assume dependência entre equações. Esta característica do SEM cria dificuldades adicionais às que já existem na deteção de *outliers* em modelos multivariados. Neste trabalho, propõe-se um novo método para detetar *outliers* em SEM. A proposta baseia-se numa versão robusta do estimador GMM e adapta ao SEM uma metodologia que foi recentemente utilizada para o modelo SUR, uma vez que este modelo também pressupõe dependência entre equações. As técnicas aplicadas mostraram-se adequadas para a deteção de *outliers*; o desempenho deste método foi comparado com o dos métodos convencionais, com base num estudo de simulação e num conjunto de dados reais. Os resultados mostraram vantagens na utilização da metodologia robusta que aqui se propõe, o que resulta numa mais valia do uso destes modelos na resolução de uma grande variedade de problemas que surgem na prática.

1 Introdução

Os modelos SEM e SUR são frequentemente usados em Econometria e generalizam o modelo de regressão multivariado. O SEM apresenta algumas caraterísticas específicas exigindo processos de estimação mais elaborados do que os que se usam no modelo de regressão ou no SUR.
De entre os estimadores tradicionais para o SEM destacam-se o estimador $3SLS$ (*Three Stages Least Squares*), que é o mais popular, e o estimador GMM. Estes estimadores apresentam boas propriedades, mas não são robustos, sendo muito sensíveis a desvios em relação ao modelo especificado ou à presença de *outliers*. Uma versão robusta para o estimador GMM foi apresentada em Rocha [8].
No presente trabalho foram consideradas sugestões de estimação robusta desenvolvidas para o modelo SUR por Bilodeau e Duchesne [2] e em Hubert *et al* [3]. Estudou-se o desempenho do estimador GMM robusto com base num estudo de simulação, no qual se mantiveram os cenários e os critérios contemplados em Hubert *et al* [3] para o modelo SUR. Este estudo evidenciou a vantagem da estimação robusta quando se verificam desvios dos pressupostos assumidos para o modelo, quer ao nível da localização, quer ao nível da dispersão. Por outro lado, estudou-se um conjunto de dados reais com o objetivo de proceder à deteção de *outliers* univariados e multivariados, adaptando ao SEM os procedimentos robustos usados em Bilodeau e Duchesne [2] e em Hubert *et al.* [3] para o modelo SUR. Este estudo mostrou vantagem nesta metodologia robusta para a deteção de observações atípicas, tanto a nível univariado como multivariado.

Todos os cálculos foram realizados com o programa *R-3.2.1.*

2 Modelo de equações simultâneas

O SEM é caraterizado por um sistema de equações interdependentes que inclui variáveis endógenas e variáveis exógenas. O SEM gene-

raliza o modelo de regressão multivariado, no sentido em que admite erros correlacionados com regressores e erros heterocedásticos.

Exemplo 2.1 *Um exemplo clássico de SEM é o Modelo Keynesiano simples, definido por:*

$$\begin{cases} y_t &= c_t + x_t \\ c_t &= \beta + \gamma y_t + \varepsilon_t \ , \end{cases}$$

onde, para um momento t, c_t representa o consumo (variável endógena), y_t representa o rendimento (variável endógena), x_t representa o investimento (variável exógena), ε_t é o erro aleatório, γ e β são os parâmetros estruturais.
Como se pode observar na 1^aequação, o rendimento depende do consumo, mas o consumo também é influenciado pelo rendimento, de acordo com a 2^aequação, mostrando a interdependência existente entre as equações do modelo.

Uma forma muito usada para escrever o SEM é a forma estrutural:

$$\mathbf{Y\Gamma + XB + E = 0},$$

onde $\mathbf{Y}$ e $\mathbf{X}$ são as matrizes de observações das variáveis endógenas e exógenas, respetivamente, $\mathbf{E}$ é a matriz dos erros aleatórios, $\mathbf{\Gamma}$ e $\mathbf{B}$ são as matrizes dos parâmetros estruturais.
Outra representação do SEM que é conveniente para a estimação dos parâmetros é dada pela equação:

$$\mathbf{y = Z\delta + e}, \tag{1}$$

onde $\mathbf{Z} = diag\left[\begin{array}{ccc} \mathbf{Z}_1 & \cdots & \mathbf{Z}_M \end{array}\right]$, com $\mathbf{Z}_i = \left[\begin{array}{cc} \mathbf{Y}_i & \mathbf{X}_i \end{array}\right]$.
Note-se que entre as variáveis $\mathbf{Z}_i$, que são as variáveis explicativas do SEM, há variáveis endógenas que são correlacionadas com os erros, fazendo com que a estimação por GLS (*Generalized Least Squares*) conduza a um estimador não consistente. Este problema pode ser resolvido utilizando variáveis instrumentais e aplicando a seguir a estimação por GLS. Este processo é designado por estimador $3SLS$.

O SEM escrito na forma (1) é semelhante em termos formais ao modelo SUR. No entanto importa distinguir os dois tipos de modelos: enquanto que no SEM há variáveis endógenas entre as variáveis explicativas em (1), no SUR tal não acontece e a correlação entre equações é devida a fatores externos ao modelo, que se refletem apenas na correlação não nula entre erros de diferentes equações.

Exemplo 2.2 *Um exemplo de SUR, publicado em Judge et al. [4], refere-se a duas empresas americanas do mesmo ramo (General Electric e Westinghouse), onde cada equação traduz a relação entre o investimento bruto anual dessa empresa (Y_1 e Y_2) e as ações emitidas (X_1) e o capital social (X_2) da empresa. O modelo é constituído por um sistema de duas equações da forma:*

$$\begin{cases} Y_{1t} &= \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + u_{1t} \\ Y_{2t} &= \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t} \end{cases},$$

A presença de fatores que influenciam ambas as empresas vai provocar a existência de correlação entre os erros das duas equações. Ao contrário do que acontecia no SEM, as variáveis explicativas do SUR não são correlacionadas com os erros, pelo que a estimação por GLS permite obter um estimador consistente, ao contrário do que acontece na estimação do SEM.

3 Estimação do SEM

De entre os estimadores tradicionais do SEM, destacam-se o estimador $3SLS$ e o estimador GMM. Estes estimadores têm boas propriedades sob um conjunto de pressupostos do modelo, nomeadamente no modelo normal, mas podem sofrer grandes perturbações quando há desvios em relação ao modelo e, em particular, na presença de observações atípicas na amostra.
A estimação robusta surge como uma alternativa conveniente pois é pouco sensível a ligeiros afastamentos dos pressupostos assumidos

para o modelo. De entre os principais trabalhos sobre estimação robusta em *SEM*, destacam-se as propostas de Amemiya [1], Maronna e Yohai [6], Krishnakumar e Ronchetti [5] e Rocha [8].
No seguimento, vai usar-se a versão robusta do estimador GMM proposta em Rocha [8], a qual será designada por estimador $GMMR$. Resumidamente, o algoritmo que permite obter esse estimador, consiste no seguinte procedimento:
P.1. Obter estimativas iniciais dos resíduos, aplicando regressão robusta por equação com base no estimador LTS (*Least Trimmed Squares*), proposto em Rousseeuw [9].
P.2. Estimar a matriz de covariâncias dos erros, usando o estimador OGK (*Orthogonalized Gnanadesikan-Kettenring*), publicado em Maronna e Zamar [7], aplicado aos resíduos obtidos no passo P.1.
P.3. Resolver o problema de minimização de uma função de Huber com resíduos ponderados pelas estimativas das covariâncias obtidas no passo P.2.
Neste trabalho foram adaptados ao SEM os procedimentos robustos sugeridos em Rousseeuw e Van Zomeren [10]. No estudo de simulação desenvolveu-se uma adaptação ao SEM dos cenários e critérios de avaliação do desempenho de estimadores, também usados em Hubert *et al.* [3] para o modelo SUR.
Para a deteção de *outliers* univariados e multivariados procedeu-se à adaptação ao SEM dos princípios usados em Bilodeau e Duchesne [2] e em Hubert *et al.* [3], os quais foram originalmente propostos por Rousseeuw e Van Zomeren [10] para estimadores LMS (*Least Median of Squares*) e MVE (*Minimum Volume Ellipsoid*).

4 Estudo de simulação

Para estudar o desempenho do estimador $GMMR$, efetuou-se um estudo de simulação, gerando as observações de acordo com um SEM particular, já trabalhado por outros autores.
O SEM considerado foi proposto por Judge *et al.* [4], com forma

estrutural definida pelo sistema:

$$\begin{cases} -\mathbf{Y}_1 + \mathbf{Y}_2\gamma_{21} + \mathbf{Y}_3\gamma_{31} + \mathbf{X}_1\beta_{11} + \mathbf{e}_1 = \mathbf{0} \\ \mathbf{Y}_1\gamma_{12} - \mathbf{Y}_2 + \mathbf{X}_1\beta_{12} + \mathbf{X}_2\beta_{22} + \mathbf{X}_3\beta_{32} + \mathbf{X}_4\beta_{42} + \mathbf{e}_2 = \mathbf{0} \\ \mathbf{Y}_2\gamma_{23} - \mathbf{Y}_3 + \mathbf{X}_1\beta_{13} + \mathbf{X}_2\beta_{23} + \mathbf{X}_5\beta_{53} + \mathbf{e}_3 = \mathbf{0} \end{cases}.$$

Na simulação, mantiveram-se os valores dos parâmetros e das variáveis exógenas tal como em Judge *et al.* [4]. Para comparar o desempenho dos estimadores em diferentes condições, simularam-se amostras adaptando ao SEM os cenários usados por Hubert *et al.* [3]: consideraram-se várias distribuições dos erros, nomeadamente, distribuição Normal 3D, com percentagens de contaminação 0, 5, 10 e 30%. Contaminaram-se os valores da variável $\mathbf{Y}_2$, por esta variável ser explicativa nas primeira e terceira equações e por ser variável dependente na segunda equação.
Para cada distribuição anteriormente referida, geraram-se 100 amostras de dimensões 30 e 100, calcularam-se as estimativas dos parâmetros e os resíduos a partir dos estimadores $GMMR$ e $3SLS$.
Com o objetivo de avaliar o desempenho dos estimadores, utilizaram-se os indicadores usados por Hubert *et al.* [3] para o modelo SUR, com base em N amostras:

$$\textbf{Viés} : \left\| 1/N \sum_{k=1}^{N} \hat{\delta}^{(k)} - \delta \right\|, \tag{2}$$

$$\textbf{Erro Quadrático Médio (EQM)} : 1/N \sum_{k=1}^{N} \left\| \hat{\delta}^{(k)} - \delta \right\|^2. \tag{3}$$

Na Tabela 1 encontram-se os resultados relativos ao viés dos estimadores $GMMR$ e $3SLS$, no caso da dimensão amostral n=30 e para diferentes graus de contaminação, de acordo com (2). Os valores mostram que o estimador $GMMR$ tem melhor desempenho nos cenários de contaminação.
Na Tabela 2 encontram-se os resultados relativos ao erro quadrático médio obtido para os estimadores $GMMR$ e $3SLS$, no caso da dimensão amostral n=30 e para os mesmos cenários de contaminação, de acordo com (3). Os valores mostram que, tal como aconteceu em relação ao viés, também relativamente ao critério do erro quadrático médio, os melhores resultados são encontrados com o estimador $GMMR$, desde que a distribuição esteja contaminada. Os resultados obtidos para a dimensão amostral n=100 conduzem às mesmas conclusões, pelo que não são aqui apresentados; ainda

Viés	*3SLS*	*GMMR*
Normal	0.789	2.638
Normal-ct5	80.326	29.071
Normal-ct30	101.744	8.748

Tabela 1: Valores do viés dos estimadores $3SLS$ e $GMMR$, para amostras de dimensão n=30 e diferentes graus de contaminação.

EQM	*3SLS*	*GMMR*
Normal	142.979	208.191
Normal-ct5	6 481.576	6 037.276
Normal-ct30	10 419.45	6 432.463

Tabela 2: Valores do EQM para os estimadores $GMMR$ e $3SLS$, para dimensão amostral n=30 e diferentes graus de contaminação.

assim, é de notar que o estimador $GMMR$ apresentou menor variabilidade. Por motivos idênticos, os resultados para a contaminação 10% não são relatados nas tabelas 1 e 2, uma vez que conduzem a conclusões análogas às dos restantes graus de contaminação.
Em face dos resultados e para as situações simuladas, podemos concluir que o estimador $3SLS$ apenas produz melhores resultados no modelo Normal sem contaminação. Desde que exista contaminação, e para qualquer dos graus considerados, o estimador $GMMR$ mostra-se superior.

5 Deteção de *outliers* no SEM

Como já referimos, a deteção de *outliers* é uma tarefa difícil neste tipo de modelos, não só por estarem presentes as dificuldades conhecidas com observações multivariadas, mas também porque a dependência entre equações mascara ainda mais as observações realmente atípicas.

Motivados pela necessidade de dispor de um meio de diagnóstico de deteção de *outliers* em SEM, e na ausência de outras propostas na bibliografia sobre o assunto, decidiu-se seguir de perto a metodologia aplicada por outros autores para o modelo SUR, nomeadamente em Bilodeau e Duchesne [2] e em Hubert *et al.* [3]. Esses autores sugerem que se investigue, separadamente, a deteção de *outliers* univariados e multivariados.
A deteção de *outliers* univariados baseia-se numa representação gráfica, para cada equação. Propomos que os valores dos resíduos obtidos para cada estimador sejam representados contra os valores da distância de Mahalanobis robusta das observações das variáveis explicativas de cada equação. Os limites a considerar para o eixo dos resíduos são as retas horizontais definidas pelos valores $+2.5$ e -2.5, subjacentes à hipótese de que os erros têm distribuição Normal; no eixo horizontal, onde se registam as distâncias de Mahalanobis, sugere-se a reta vertical definida pelo valor da raiz quadrada de um quantil elevado da distribuição qui-quadrado com $k_i - 1$ graus de liberdade, onde k_i é o número de variáveis explicativas da i-ésima equação, incluindo o termo constante. Este tipo de representação gráfica de resíduos permite simultaneamente avaliar a qualidade do ajustamento (através do eixo dos resíduos) e identificar pontos de alavanca (através do eixo da distância de Mahalanobis).
Importa também e sobretudo detetar *outliers* multivariados no modelo. Para a determinação de *outliers* multivariados propõe-se um outro tipo de gráfico, representando nas ordenadas as distâncias de Mahalanobis (clássicas ou robustas) dos resíduos multivariados do ajustamento robusto e nas abcissas a sequência (ou os índices) das observações. Relativamente aos limites para detetar *outliers* multivariados, os princípios foram os já referidos para o caso univariado, isto é, no eixo das ordenadas usar a reta horizontal definida pelo valor da raiz quadrada de um quantil elevado da distribuição qui-quadrado com $k-1$ graus de liberdade, onde k é o número de variáveis explicativas do modelo, incluindo o termo constante.
Para ilustrar o método proposto, apresenta-se um exemplo de um SEM com dados reais, já trabalhado por outros autores, permitindo deste modo a comparação de resultados.

Exemplo 5.1 *Em Maronna e Yohai [6] é modelado por um SEM um conjunto de dados reais da economia da Argentina, relativos ao período entre 1956 e 1984, com a seguinte forma estrutural:*

$$\begin{cases} -\mathbf{Y}_1 & +\mathbf{Y}_3\gamma_{31} & & & +\mathbf{X}_5\beta_{51}+\mathbf{e}_1 & = \mathbf{0} \\ \mathbf{Y}_1\gamma_{12}-\mathbf{Y}_2 & & +\mathbf{X}_1\beta_{12} & & +\mathbf{X}_4\beta_{42}+\mathbf{X}_5\beta_{52}+\mathbf{e}_2 & = \mathbf{0} \\ \mathbf{Y}_1 & -\mathbf{Y}_2-\mathbf{Y}_3 & +\mathbf{X}_1 & +\mathbf{X}_2-\mathbf{X}_3+\mathbf{X}_4 & & = \mathbf{0} \end{cases},$$

onde as variáveis exógenas são X_1, o investimento bruto, X_2, o volume de exportações, X_3, os custos de impostos, X_4, o consumo público e X_5, a variável unitária; as variáveis endógenas são Y_1, o consumo privado, Y_2, o volume de importações e Y_3, o rendimento. Os dados correspondem a registos anuais das respetivas variáveis do sistema anterior.

Como se referiu, procede-se separadamente à deteção de *outliers* univariados e multivariados. Para identificação das observações atípicas, optou-se pelo quantil 0.975 da distribuição qui-quadrado, para a distância de Mahalanobis, e marcaram-se os anos no eixo das abcissas.

5.1 Análise do Exemplo 5.1 no caso univariado:

Observando a Figura 1, não se deteta a existência de *outliers*, nem de grandes valores de resíduos para a equação 1, uma vez que em nenhum dos dois gráficos aparecem pontos para além das retas limites. Logo, seja na perspetiva clássica com o estimador $3SLS$ e a distância de Mahalanobis convencional, seja do ponto de vista robusto com o estimador $GMMR$ e a distância de Mahalanobis robusta, não há indícios de observações discordantes na 1ª equação.

Relativamente à 2ª equação, a que se refere a Figura 2, as conclusões são diferentes. De facto, ambas as imagens mostram que não há resíduos a destacar (seja por $3SLS$, seja por $GMMR$), uma vez que não existem pontos para além dos limites considerados no eixo vertical (dos resíduos). Ao analisar o limite relativo à distância de Mahalanobis (no eixo horizontal), verifica-se que na imagem da direita da Figura 2 são destacados diversos pontos, não realçados na imagem da esquerda. Isto significa que, com a metodologia robusta baseada na estimação por $GMMR$ e na distância de Mahalanobis robusta, foi possível detetar *outliers* que passavam despercebidos na metodologia clássica, baseada em estimação por $3SLS$ e na distância de Mahalanobis clássica.

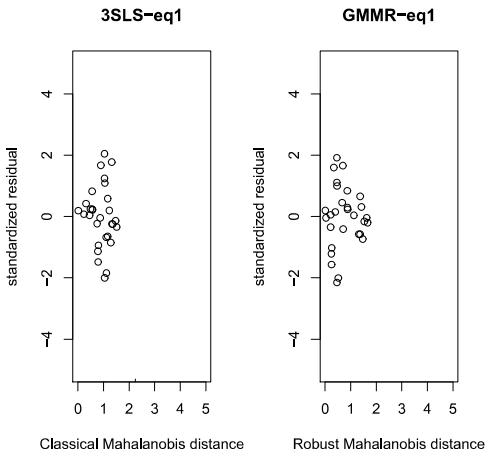

Figura 1: Deteção gráfica de *outliers* da 1ªequação: resíduos com os estimadores $3SLS$ e $GMMR$, contra a distância de Mahalanobis clássica e robusta.

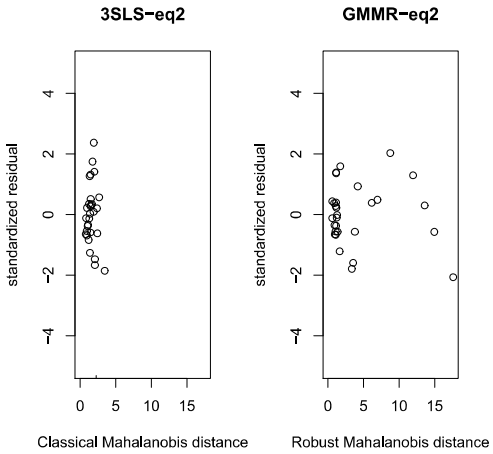

Figura 2: Deteção gráfica de *outliers* da 2ªequação: resíduos com os estimadores $3SLS$ e $GMMR$, contra a distância de Mahalanobis clássica e robusta.

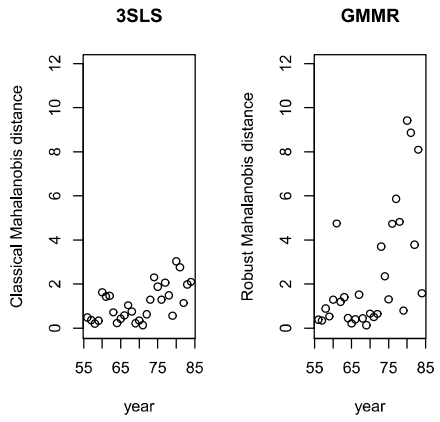

Figura 3: Deteção gráfica de *outliers* do sistema: distância de Mahalanobis clássica e robusta dos resíduos multivariados, com o estimador $GMMR$, contra os anos.

5.2 Análise do Exemplo 5.1 no caso multivariado:

Na Figura 3, observando o limite relativo à distância de Mahalanobis (no eixo vertical), destacam-se diversos pontos na imagem da direita, os quais não aparecem na imagem da esquerda. Isto traduz que a metodologia robusta, que combina a estimação por $GMMR$ com a distância de Mahalanobis robusta, permitiu detetar *outliers* multivariados que não eram notados com a metodologia clássica.

6 Comentários finais

Realizou-se um estudo de simulação que evidenciou a vantagem da estimação robusta ($GMMR$), quando se verificam desvios dos pressupostos assumidos para o modelo, quer ao nível da localização, quer ao nível da dispersão. Estudou-se a deteção de *outliers* univariados e multivariados no SEM procedendo à adaptação de metodologias propostas anteriormente para outros modelos. Os novos procedimentos para a deteção de *outliers* mostraram-se mais eficazes. Os métodos robustos que se propõem neste trabalho mostraram-se preferíveis na deteção de observações atípicas no modelo SEM, quer na perspetiva univariada, quer na multivariada.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação Portuguesa para a Ciência e Tecnologia (FCT-Fundação para a Ciência e a Tecnologia), por meio do CIDMA - Centro de Investigação e Desenvolvimento em Matemática e Aplicações, dentro do projeto UID / MAT / 04106/2013.

Referências

[1] Amemiya, T. (1982). Two stage least absolute deviation estimators, *Econometrica*,50, 689–711.

[2] Bilodeau, M. and Duchesne, P. (2000). Robust estimation of the SUR model. *The Canadian Journal of Statistics*, Vol. 28, 2, 277–288.

[3] Hubert, M., Verdonk, T. and Yorulmaz, O. (2014). Fast robust SUR with applications to the multivariate chain ladder method. *ROBUST@Leuven, Publications, Technical reports.*

[4] Judge, G., Griffiths, W., Lutkepohl, Hill, R. and Lee, T. (1988). *Introduction to the theory and practice of econometrics, second edition*, John Wiley & Sons, New York.

[5] Krishnakumar, J. e Ronchetti, E. (1997). Robust estimators for simultaneous equations models, *Journal of Econometrics*, 78, 295–314.

[6] Maronna, R. e Yohai, V. (1997). Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference*, 57, 233–244.

[7] Maronna, R. e Zamar, R. (2002). Robust estimates of location and disersion for high-dimensional datasets, *Technometrics*, 44, 307–317.

[8] Rocha, A. (2010). *Estimação robusta em Modelos Lineares de Equações Simultâneas, Tese de Doutoramento*, Universidade de Aveiro.

[9] Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.

[10] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, 85, 633–639.