




# Robust *versus* traditional methods for outlier detection in the simultaneous equation model

**Anabela Rocha**[1] **Manuela Souto Miranda**[1] **João Branco**[3]

[1]CIDMA
University of Aveiro

[2]CEMAT
University of Lisbon

July 4, 2018




# Summary


# Introduction



We propose a method for outlier detection in the Simultaneous Equations Model (*SEM*) which is based on analysing anomalous departures in the residuals of the dependent variables and in the set of the explanatory variables.

The method adopts a robust procedure according to the following steps:

- Estimation of the coefficients - we used a robust version of the Generalized Method of Moments (*RGMM*), originally suggested in Rocha (2010);
- Introduction of Robust Mahalanobis Distances (*RMD*) - computed with the *MCD* estimators.


# Introduction



The present talk has three main goals:

- To evaluate the behaviour of the proposed method in the contaminated *SEM*
  - with a simulation study.
- To compare the results with those that would be obtained using a similar non robust technique (with the most popular estimator for the *SEM*)
  - Three Stages Least Squares (*3SLS*);
  - Traditional Mahalanobis Distances (*MD*).
- To present the results of the application of the proposed method to real data
  - we analyse a case study with Portuguese economic data.


# Applications of the method



**In the examples we follow general procedures:**

- We confirmed the identifiability of the equations in the model.
- We estimated the parameters by *RGMM* and by *3SLS* and calculated the respective residuals.
- We calculated the *RMD* and *MD* of the residuals and of the observations of the explanatory variables.
- We obtained the cut-off points as the 97.5% quantiles of the Chi-square distributions with the appropriate degrees of freedom.
- We performed the detection of outliers using the robust method and the traditional aproach.
- All the computations were performed with R version 3.3.3.


# Simulation study



We performed a simulation study based on a model proposed in Judge *et al.* (1988), to detect outliers in the *SEM*:

$$\begin{cases} -y_1 + y_2\gamma_{21} + y_3\gamma_{31} + x_1\beta_{11} + e_1 = 0 \\ y_1\gamma_{12} - y_2 + x_1\beta_{12} + x_2\beta_{22} + x_3\beta_{32} + x_4\beta_{42} + e_2 = 0 \\ y_2\gamma_{23} - y_3 + x_1\beta_{13} + x_2\beta_{23} + x_5\beta_{53} + e_3 = 0 \end{cases}$$

where $x_1$ represents a constant term in the model.

We generate observations, with and without contamination, to evaluate the performance of the proposed outlier detection method.

Simulation scenarios were taken considering:
(i) parameters and exogenous variables as in Judge *et al.* (1988);
(ii) the choices of the errors distributions and contaminations as in Hubert *et al.* (2017) for the *SUR* model.


# Simulation scenarios



Thus we considered:

- Normal 3D distribution without contamination and with different levels of contamination (5%, 10% and 30%):
  - contamination was introduced in one variable, repeating for each explanatory variable, either endogenous or exogenous (except the constant term);
  - contamination in two variables was also considered.

- Contamination was introduced by replacing the first observations of each explanatory variable by values generated by a Uniform distribution, with support far from the range of the observed values in that variable, with other variables remaining unchanged.

- There were simulated 100 replications of each scenario with sample sizes of 30 and 100. Therefore the study included 7500 samples from the model.


# Outliers identification



The system has 3 equations and all of them include an error term and a constant term ($x_1$, unitary), which should be ignored in calculating the degrees of freedom.
So, we have:

- $M = 3$ equations;
- $k = 4$ explanatory variables ($x_2$, $x_3$, $x_4$ and $x_5$);
- $rank(\mathbf{\Gamma}) = 3$, with $\mathbf{\Gamma} = \begin{bmatrix} -1 & \nu_{21} & 0 \\ \nu_{12} & -1 & 0 \\ 0 & \nu_{23} & -1 \end{bmatrix}$;
- finally, $\nu = rank(\mathbf{\Gamma}) + k = 3 + 4 = 7$.

**Table of cut-off values**

| | | | |
|---|---|---|---|
| System | Residuals | $\nu = 3$ | $\sqrt{\chi^2_{3;0.975}} = 3.06$ |
| | Explanatory var. | $\nu = 3 + 4 = 7$ | $\sqrt{\chi^2_{7;0.975}} = 4.00$ |


# Evaluation of the method



The performance of the proposed method was evaluated like in Filzmoser *et al.* (2008), calculating proportions of true and of false outliers, because some authors argue that the use of the Chi-squared distribution can increase the number of detected outliers.
Thus, we computed **for the system**:

- the proportion of **false negatives** (*FN*), corresponding to the points that were generated as outliers but were not identified as so,
- the proportion of **false positives** (*FP*), relative to observations that were not generated as outliers but were classified as so by the method.


# Results



**Proportion of false positives (*FP*); $n = 100$, cont. 30%.**

| | | $y_1$ | $y_2$ | $y_3$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|---|---|
| **Robust** | Explanat. var. | 0.3 | 0.5 | 0.6 | 0.3 | 0.3 | 0.5 | 0.4 |
| | Residuals | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.2 |
| **Traditional** | Explanat. var. | 0.3 | 0.3 | 0.2 | 0.3 | 0.3 | 0.3 | 0.3 |
| | Residuals | 0.7 | 0.0 | 0.5 | 0.9 | 1.1 | 0.9 | 1.0 |

- The proposed method was excellent in identifying outliers, with extremely low values of *FP* in detecting both type of outliers (smaller than 1%).
- This seems to contradict the suspicion that the use of the Chi-square distribution tends to highlight more points than it should, at least in the simulated *SEM*.


# Results



## Proportion of false negatives (*FN*); $n = 100$, cont. 30%.

| | | $y_1$ | $y_2$ | $y_3$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|---|---|
| **Robust** | Explanat. var. | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Residuals | 0.0 | 0.0 | 0.0 | 0.3 | 0.4 | 1.1 | 18.6 |
| **Traditional** | Explanat. var. | 98.6 | 99.2 | 98.7 | 98.9 | 99.3 | 99.3 | 98.7 |
| | Residuals | 96.4 | 98.1 | 96.1 | 96.9 | 97.3 | 96.9 | 96.5 |

- The low values of *FN* obtained by the robust method show its excellent performance in detecting both types of outliers.
- The use of classical non robust estimators presents a very high proportion of *FN* for both types of outliers.


# Results



For the others scenarios:

- With the **robust method**, the *FP* and *FN* proportions remained very low for the various simulation scenarios, regardless of the size or the percentage of contamination considered.

- With the **traditional method**, the proportions of *FP* remained very low for the various simulation scenarios. However, the *FN* proportions remained very high and and they increased with higher contamination. It seems that the sample size did not have evident influence.


# Results



- As the proportion of *FN* reflects the observations that should be identified as outliers and were not, and since the non-robust methods present a very high value of *FN*, we conclude that they do not detect many outliers, which can have very dangerous consequences in the statistical analysis. Thus, it is not advisable to use the classical estimators.

- The **robust method has better performance than the traditional method** in detecting both types of outliers, since it leads to slightly lower proportions of *FP* and incomparably lower proportions of *FN*.


# Case study



We considered a macroeconomic equilibrium model to describe data of economic variables. Values were observed in Portugal and the model is presented in the structural form.

$$\begin{cases} -y_1 + y_3\gamma_{31} + x_3\beta_{31} + + x_5\beta_{51} + e_1 = 0 \\ y_1\gamma_{12} - y_2 + x_1\beta_{12} + x_4\beta_{42} + x_5\beta_{52} + e_2 = 0 \\ y_1 - y_2 - y_3 + x_1 + x_2 + x_4 = 0 \end{cases}.$$

Equation 1 - explains the private consumption;
Equation 2 - explains the imports;
Equation 3 - it is an accounting equality.


# Case study



- **Data** - taken from Pordata (https://www.pordata.pt/), in July 2017, corresponding to the period from 1996 to 2016 (n = 21).

- **Endogenous variables** - private consumption ($y_1$), imports ($y_2$) and gross domestic product (GDP) ($y_3$).
- **Exogenous variables** - investment ($x_1$), exports ($x_2$), tax revenues ($x_3$), public consumption ($x_4$) and constant term ($x_5$).


# Case study



Equation 3 - Accounting equality that reflects an equilibrium condition

$$y_1 - y_2 - y_3 + x_1 + x_2 + x_4 = 0$$
$$\Leftrightarrow$$
$$y_2 + y_3 = y_1 + x_1 + x_2 + x_4$$
$$\Leftrightarrow$$

***import. + GDP = priv. cons. + invest. + export. + pub. cons.***

This condition is not always observed in practice due to administrative restrictions and to market fluctuations that are reflected in the economic variables.
We considered a corrected GDP, denoted by GDP *, which verifies the equation 3:
GDP*= $y_3 - D$, with $D = y_3 - (y_1 - y_2 + x_1 + x_2 + x_4)$.

In the following procedures, we used $y_3 =$ GDP*.


# Outliers identification



- The system has 3 equations, but the last one is an accounting identity (without random error), which does not enter in the computation of the degrees of freedom. In the two other equations there are constant terms ($X_5$, unitary), which should be ignored in those calculations.

- Considering the remaining equations of the **system**:
  - $M = 2$ equations.
  - $k = 4$ exogenous variables ($x_1$, $x_2$, $x_3$ e $x_4$).
  - $rank(\mathbf{\Gamma}) = 2$,

$$\mathbf{\Gamma} = \begin{bmatrix} -1 & 0 & \nu_{31} \\ \nu_{12} & -1 & 0 \end{bmatrix}$$

so, $\nu = rank(\mathbf{\Gamma}) + k = 2 + 4 = 6$.


# Outliers identification



- Considering each equation:
  - Equation 1: $\nu = 1 + k_1 = 1 + 1 = 2$.
  - Equation 2: $\nu = 1 + k_2 = 1 + 2 = 3$.

## Table of cut-off values

| | | | |
|---|---|---|---|
| System | Residuals | $\nu = 2$ | $\sqrt{\chi^2_{2;0.975}} = 2.72$ |
| | Explanatory var. | $\nu = 2 + 4 = 6$ | $\sqrt{\chi^2_{6;0.975}} = 3.80$ |
| *Per* equation | Residuals | $\nu = 1$ | $\sqrt{\chi^2_{1;0.975}} = 2.24$ |
| equation 1 | Explanatory var. | $\nu = 1 + 1 = 2$ | $\sqrt{\chi^2_{2;0.975}} = 2.72$ |
| equation 2 | Explanatory var. | $\nu = 1 + 2 = 3$ | $\sqrt{\chi^2_{3;0.975}} = 3.06$ |

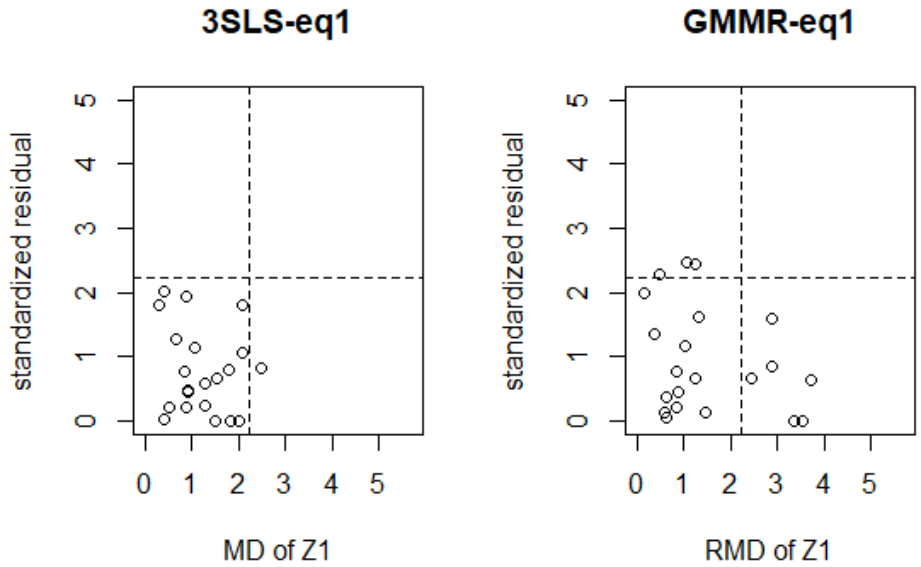

# Results



## Equation 1 - Private consumption

$$y_1 = y_3\gamma_{31} + x_3\beta_{31} + x_5\beta_{51} + e_1$$

Explanatory variables - GDP ($y_3$) and tax revenues ($x_3$).




**Outliers detected by the robust method:**
**Residuals** - 1996, 2002, **2008**.
**Explanatory variables** - ***2009***, **2010, 2013, 2014, 2015, 2016**.

**The years of Portugal's economic crisis were marked at bold.**

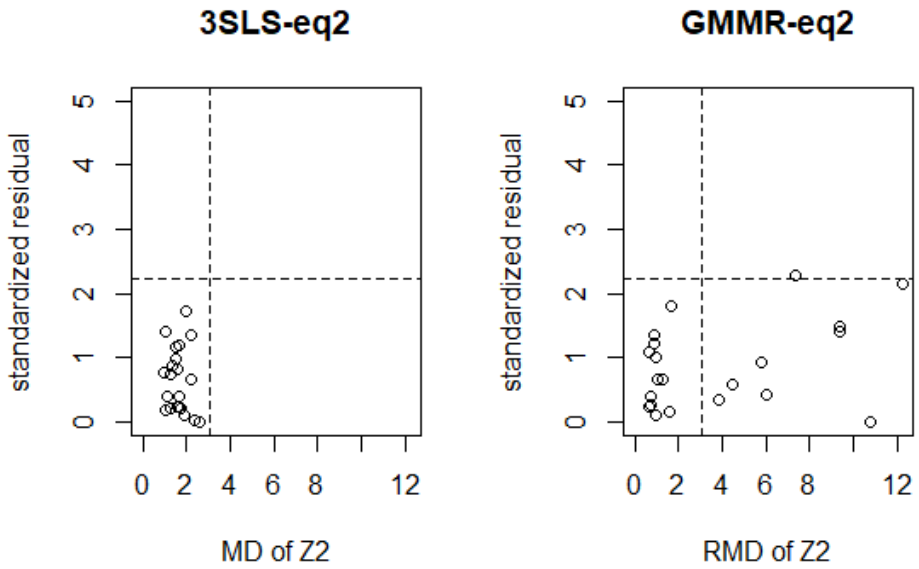

# Results



## Equation 2 - Imports

$$y_2 = y_1\gamma_{12} + x_1\beta_{12} + x_4\beta_{42} + x_5\beta_{52} + e_2$$

Explanatory variables - private consumption ($y_1$), investment ($x_1$), public consumption ($x_4$).



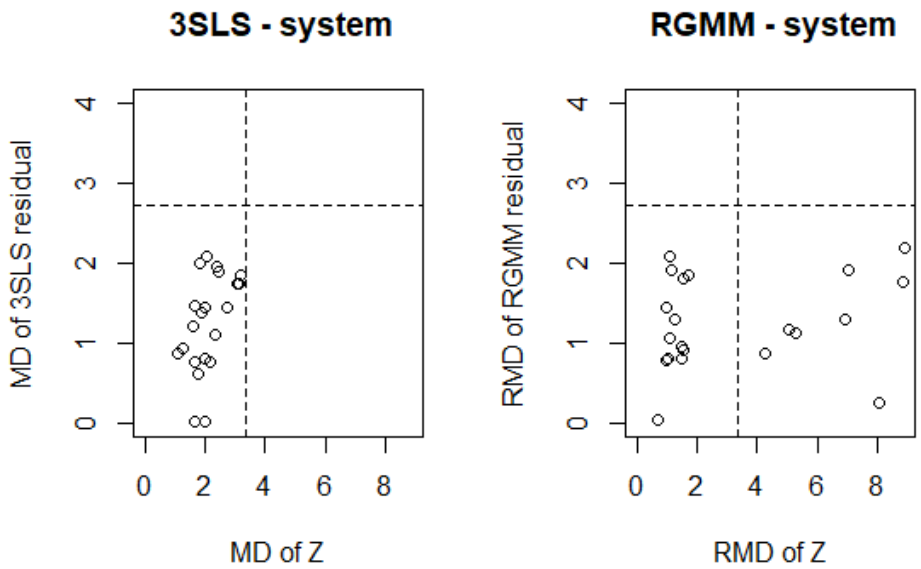

**Outliers detected by the robust method:**
**Explanatory variables and residuals - 2013.**
**Explanatory variables - 2007, 2008, 2010, 2011, 2012, 2014, 2015, 2016.**

# Results



## System

Explanatory variables - private consumption ($y_1$), GDP ($y_3$), investment ($x_1$), tax revenues ($x_3$) and public consumption ($x_4$).




**Outliers detected by the robust method:**
**Explanatory variables - 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016.**


# Final comments



- Economic crisis in Europe began in 2007-2008.
- Crisis effects in Portugal were felt since 2008 (with Troika intervention during 2011-2014), with great impact on the Portuguese economy.
- Years of crisis had a great impact on a number of economic variables, including private consumption and imports (**endogenous variables**), as well as investment, exports, tax revenues and public consumption (**exogenous exogenous**).
- The proposed **Robust method** identifies outliers in several years of the crisis.


# Final comments



- **Non robust method** (using *3SLS* estimator and *MD*) only highlights one of those years. Particularly, it detects an outlier on the explanatory variables of equation 1 for the year 2009, which is also identified in the robust method.
- **Robust method** (with *RGMM* estimator and *RMD*) identified several years as outliers (either *per* equation or for the system). Those years reflect the period of economic crisis.

**Robust method performed better in detecting multivariate outliers for the national macroeconomic variables.**


# Acknowledgements



**Thank you for your attention.**

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.


# References



[1] Filzmoser, P., Maronna, R. and Werner, M., Outlier identification in high dimensions, Computational Statistics & Data Analysis, **52**, 1694–1711 (2008).

[2] Hubert, M., Verdonk, T. and Yorulmaz, O., Fast robust *SUR* with economical and actuarial applications, Statistical Analysis and Data Mining **10**(2), DOI: 10.1002/sam.11313, 77–88 (2017).

[3] Judge, G., Griffiths, W., Lutkepohl, Hill, R. and Lee, T., Introduction to the theory and practice of econometrics, 2nd Edition. John Wiley & Sons, New York (1988).

[4] Rocha, A., Estimação robusta em modelos lineares de equações simultâneas, PhD Thesis, Universidade de Aveiro (2010).

[5] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org (2017).

[6] Zellner, A. and Theil, H., Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations, Econometrica, **30**(1), 54–78 (1962).