Measures of correlation and computer algebra

Halkos, George and Tsilika, Kyriaki

Department of Economics, University of Thessaly

March 2016

Measures of correlation and computer algebra

George E. Halkos and Kyriaki D. Tsilika
Laboratory of Operations Research,
Department of Economics, University of Thessaly,

Abstract
Our contribution in this work is to set the directions for specialized econometric computations in a free computer algebra system, Xcas. We focus on the programming of a routine dedicated to correlation criteria for multiple regression models. We program several operations for detecting and evaluating collinearity by applying the diagnostic techniques of linear regression analysis. Xcas could constitute a supplemental tool in a collinear data study. Its use is proposed complementary to established econometric software or as substitute software.

Keywords: Multicollinearity; correlation criteria; computational econometrics; CAS software.

JEL Classification: C13; C18; C63; C88.

1. Introduction

Algebraic calculations are widely and strongly involved in econometric analysis. The close connection of econometric analysis and matrix algebra is a scientific fact (see indicatively Frawley, 1985; Schipp & Krämer, 2009). Hence the programming environment of a computer algebra system (CAS) is more than appropriate to estimate metrics, perform methodologies or strategies, test conditions and criteria, identify rules. A number of computer algebra system (CAS) approaches in econometrics has already been proposed (Merckens, 1991; Merckens & Bekker, 1993; Bekker et al., 1994; Hutton & Hutton, 1995; Amman et al., 1996; Belsley, 1999; Kendrick & Amman, 1999; Stroeker & Kaashoek, 1999; Li & Racine, 2008; Bollen & Bauldry, 2010; Halkos & Tsilika, 2015). Novel computational trends in econometrics even propose Python, a general purpose language. Python – with the right set of add-ons – is comparable to domain-specific languages such as R and MATLAB (Sheppard, 2014).

In this paper we set the computational framework for a complete study in correlation analysis with CAS software. It is worthwhile to note that Belsley (1999) was one of the first who proposed a CAS environment for doing econometrics, having proposed a dominant and highly sophisticated survey in collinearity diagnostics (Belsley et.al. 1980; Belsley 1991a; Belsley 1991b). Here, we choose to use the program editor of free CAS software, Xcas[1], to propose computational codes for a number of indicators and formulations related to classic and alternative correlation criteria.

All our programmed functions work in a black box mode, with the user only having to insert simple input (i.e. the sample data) and getting the desired result

[1]The selected software, Xcas, is a computer algebra system accessible to all users interested, free of any charges, available at http://www-fourier.ujf-grenoble.fr/~parisse/giac.html. Xcas is compatible with Mac OSX, Windows (except possibly for Vista) and Linux/Ubuntu.

immediately, in just one entry. Our routine is simple to use, estimates correlation metrics that are not included in the standard output of one (of the widely used) software package and runs in an environment appropriate for more tabular displays and algebraic manipulations associated with the collinearity diagnostics.

Finally, all our computational codes constitute the basis for an automatic testing for collinearity, with a (simple) procedural programming approach. The Xcas collinearity test interprets the eigenanalysis of the correlation matrix and the variance decomposition proportions following the already mentioned above Belsley's diagnostic methodology.

2. Existing Computational Approaches for Correlation Criteria

A number of procedures are used to indicate the presence of collinearity using traditional statistical and econometric packages (SPSS, MINITAB, SAS, STATA, S-Plus):

- A very high R^2 in a multiple regression equation with few significant t statistics may be an indicator of multicollinearity.
- Construction of a correlation matrix among the explanatory variables. Relatively high simple correlations between one or more pairs of explanatory variables may indicate multicollinearity. Correlation values (off-diagonal elements) of at least .7 are sometimes interpreted as indicating a multicollinearity problem. Every package calculates the correlation for every possible pair, and displays the correlation matrix. It also displays p-values for the hypothesis test of the correlation coefficient being zero.
- Estimation of partial correlation coefficients.

- Auxiliary regressions: one way of finding out which X variable is related to other X variables is to regress each X_i on the remaining X variables and compute the corresponding R^2 (Gujarati, 2003, chapter 10).
- Estimation of eigenvalues, condition number and variance decomposition proportions (first presented in Belsley et al., 1980; Belsley, 1991a; Belsley, 1991b).
- Estimation of tolerance (TOL) and variance inflation factor (VIF).

Menu choices in SPSS and MINITAB, the REG procedure in SAS, option Collin in STATA, generate diagnostic results for collinearity evaluating the above metrics. There are also dedicated packages in other software. The download https://github.com/brian-lau/colldiag programmed by Brian Lau (19 Oct 2014, updated 17 Aug 2015) provides a Matlab code for determining the degree and nature of collinearity in a regression matrix (variance decomposition proportions, condition index, VIF, tableplot). The package "perturb" of the statistical software R (Hendrickx, 2010) evaluates collinearity by adding random noise to selected variables. In this R package, collinearity tests are performed by calculation of condition numbers and variance decomposition proportions. Friendly & Kwan (2009) proposed a visual approach for collinearity diagnostics (specifically for condition indices and variance proportions) in SAS and R, by creating table plots and biplots.

3. The Necessary Theoretical Basis

In mathematical modeling, in a n-parameter multiple linear regression

$$y = b_0 + b_1 x_1 + ... + b_n x_n \quad (1)$$

it is essential to ensure first that the variables $(y, x_1, ..., x_n)$ are linearly dependent and it is also necessary that the variables $x_1, ..., x_n$ do not already constitute a linearly

dependent set. Algebraically, a group of n variables is a linearly independent set if there exist n constants $a_1,...,a_n$ different from zero, such that $a_1x_{1i}+...+a_nx_{ni}=0$ for all $i=1,...,N$, where N is the number of observations of each variable.

A criterion for rank: The necessary and sufficient condition for a linear relationship between $(n+1)$ variables to be determined by a given data set is that the group of observations corresponding to the statistical data has a rank of n.

Let m_{ij} be the moments

$$m_{ij}=\sum_t (x_i^{(t)}-\bar{x}_i)(x_j^{(t)}-\bar{x}_j) \tag{2}$$

with t yielding all N observations and $\bar{x}_i$ being the mean of x_i. The determinant of moments

$$M=\begin{vmatrix} m_{11} & m_{12} & ... & m_{1n} \\ m_{11} & m_{22} & ... & m_{2n} \\ ... & ... & ... & ... \\ m_{n1} & m_{n2} & ... & m_{nn} \end{vmatrix} \tag{3}$$

plays a key role in the detection of linear dependency, since the rank of the determinant of moments M equals to the rank of the variables.

Let R be the correlation determinant

$$R=\begin{vmatrix} r_{11} & r_{12} & ... & r_{1n} \\ r_{11} & r_{22} & ... & r_{2n} \\ ... & ... & ... & ... \\ r_{n1} & r_{n2} & ... & r_{nn} \end{vmatrix} \tag{4}$$

The correlation determinant R is always between zero and one. r_{ij} are the correlation coefficients of the variables defined as

$$r_{ij}=\frac{m_{ij}}{\sqrt{m_{ii}m_{jj}}} \text{ or } r_{ij}=\frac{\sum_t (x_i^{(t)}-\bar{x}_i)(x_j^{(t)}-\bar{x}_j)}{\sqrt{\sum_t (x_i^{(t)}-\bar{x}_i)^2 \sum_t (x_j^{(t)}-\bar{x}_j)^2}} \tag{5}$$

A distinction between different correlation coefficients (Spearman's rank correlations and Pearson's correlations) is made in Shirokikh et al. (2013). More measures of dependence of a pair of random variables are examined in Bautin et al. (2013), as the probability measure of similarity and the sign correlation.

Correlation criterion 1: The determinant *R* equals to one stemming from the necessary and sufficient condition for non-correlation. In these lines, for the $x_i, x_j, ...x_k$ variables to be linearly dependent, the determinant *R* has to be equal to zero. Thus, the value of *R* quantifies the degree of scattering in a swarm of observations: if *R* approaches zero, the organization is great and if *R* approaches one there is no organization (Bjerkoholt & Dupont-Kieffer, 2009, lecture 6). Connecting *R* with the multicollinearity phenomenon, Field (2009) claims that when the value of *R* is greater than 0.0001 there is no severe multicollinearity.

Correlation criterion 2: Bjerkoholt & Dupont-Kieffer (2009) in lecture 6 introduce another indicator for the degree of scattering (dispersion) of the data variables. They call it scatter coefficient:

$$r = \sqrt{R} \tag{6}$$

where *R* is the correlation determinant (4). If the scatter coefficient (6) for a group of (*n+1*) variables $x_1, ...x_n, y$ is close to one this implies that *y* is absolutely unrelated to the rest of the system. But if the scatter coefficient is near zero we may expect a linear relation of the form (1). If the scatter coefficient for the set of the *n* explanatory variables $x_1, ...x_n$ is close to one, then it seems reasonable to assume non-correlation and consider a relationship of the form $ay + a_0x_0 + a_1x_1 + ... + a_nx_n = 0$ (Bjerkoholt & Dupont-Kieffer, 2009, lecture 6). Consequently, scatter coefficient is a measure for

testing the correlation among $x_1,...x_n$ and the correlation between series $b_0 + b_1 x_1 + ... + b_n x_n$ and the dependent variable y.

The scatter coefficient (6) performs a quantitative analysis of the degree of the dispersion of the data more sensitive than the determinant R. Evaluating the square root of a number in the interval [0,1], we get a higher range of values within the interval [0,1], as it is obvious from figure 1.

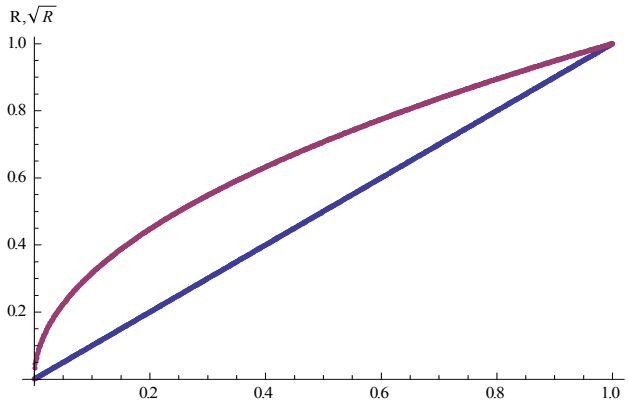


Figure 1: The range of functions R and $\sqrt{R}$

Correlation criterion 3: A different approach uses the eigenanalysis of the correlation matrix and leaves it to the user to decide whether eigenvalues are extreme, indicating that the dimension of the problem could be (or should be) reduced. There are three measures, namely *eigenvalues*, *condition index* and *condition number*. Condition index is an alternative to Variance Inflation Factor (VIF). In case of no collinearity, all eigenvalues would be 1. Eigenvalues smaller or larger than 1 would indicate departures from the ideal situation. "Too" small or large eigenvalues would indicate multicollinearity problems. Eigenvalues are important in multiple regression models. The division of each eigenvalue by the number of discriminator variables used in the analysis calculates the absolute percent which shows the magnitude of between-group

variability explained by each function in relation to the between-group variation (Brown & Wicker 2000).

While the condition index[2] is the ratio between a specific eigenvalue and the maximum of all eigenvalues, the condition number is the root of largest eigenvalue divided by the smallest. That is:

$$\text{Condition Number} = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} \tag{7}$$

A condition index between 100 and 1000 or condition numbers between 10 and 30 would indicate weak to serious collinearity problems. For a complete discussion for the values of condition numbers and indices that indicate dependencies refer to Belsley et al. (1980).

Correlation criterion 4: In case the correlation matrix has a rank of $r < n$, with n being the number of variables, then there will be $n - r$ eigenvalues equal to zero. This may lead us to suspect the existence of multicollinearity.

Correlation criterion 5: The larger the value of VIF_j the more collinear the variable X_j. If the VIF of a variable exceeds 10, the variable is said to be highly nonlinear (Gujarati, 2003, chapter 10; Halkos, 2006, 2011, Chapter 6)

$$VIF_j = \sum_{k=1}^{p} \frac{V_{jk}^2}{\lambda_\kappa} \tag{8}$$

Correlation criterion 6: Variance-decomposition proportions (Fox, 1984, par. 3.1.3)

$$p_{jk} = \frac{V_{jk}^2}{VIF_j \lambda_\kappa} \tag{9}$$

[2] *SAS, STATA, SPSS* and *S-plus* define the condition index as the *square root* of this ratio.

This ratio informs us on how much percentage of the parameter coefficient's variance is associated with each eigenvalue. A usual decisive factor for collinearity relying on high variance decomposition proportions puts the threshold of a variance decomposition proportion greater than 0.50 for two or more variables associated with a high condition index. If a condition index is high (within the interval (100, 1000) or higher) and two or more explanatory variables illustrate high proportions of variation concerning this index, we may infer that these explanatory variables are significantly linear dependent (Belsley, 1991a, b).

4. Programming in Xcas

The codes in Xcas given below create dedicated functions to estimate i) the moments m_{ij} as defined in (2) and ii) the determinant of moments as defined in (3).The user just has to introduce the independent variables (argument *vars*). This is the only input required.

```
m(xn,xk):=sum(((xn[l]-mean(xn))*(xk[l]-mean(xk))),l,0,length(xk)-1)
moments(vars):=makemat((j,k)->m(vars[j],vars[k]),nrows(vars),nrows(vars))
```

The codes in Xcas below generate i) the correlation matrix of data variables (argument *vars*) ii) the scatter coefficient as defined in (6) iii) the variance inflation factors iv) the condition indices as defined in section 3 and the condition number as defined in (7) and v) the proportions of variance for all independent variables as defined in (9) (lying in the rows of a tabular display). The user just has to introduce the independent variables (argument *vars*). This is the only input required.

```
correlation_matrix(vars):=makemat((j,k)->m(vars[j],vars[k])
/sqrt(m(vars[j],vars[j])*m(vars[k],vars[k])),nrows(vars),nrows(vars))
scatter_coefficient(vars):=sqrt(det(correlation_matrix(vars)))
```

```
phi(k,j,vars):=eigenvectors(correlation_matrix(vars))[k,j]^2/(eigenvalues(co
rrelation_matrix(vars))[j])

vif(vars):=seq(sum(phi(l,k,vars),k,0,eigenvalues(correlation_matrix(vars))-
1),l,0,length(eigenvalues(correlation_matrix(gg)))-1)

condition_indices(vars):=seq(max(eigenvalues(correlation_matrix(vars)))/(eig
envalues(correlation_matrix(vars))[j]),j,0,length(eigenvalues(correlation_ma
trix(vars)))-1)

condition_number(vars):=sqrt(max(eigenvalues(correlation_matrix(vars)))/min(
eigenvalues(correlation_matrix(vars))))

vdp(k,j,vars):=eigenvectors(correlation_matrix(vars))[j,k]^2/((eigenvalues(c
orrelation_matrix(vars))[k])*vif(vars)[j])

variance_proportions(vars):=seq(seq(vdp(j,k,vars),j,0,nrows(vars)-
1),k,0,nrows(vars)-1)
```

With the following function, we get a direct answer for the collinearity problem (classifying the cases of weak - moderate to strong - severe collinearity) based on condition numbers. The cutoff values which point that multicollinearity affects estimates are taken from Callaghan & Chen (2008) but can easily be adjusted to the user's preferences.

```
cn_condition(vars):=if(condition_number(vars)<10) "weak collinearity"; else
(if(condition_number(vars)>10 and condition_number(vars)<30) "moderate to
strong collinearity"; else "severe collinearity";);
```

In Xcas programming environment we create a loop that checks the number of variance decomposition proportions and the associated condition indices for all the components (or dimensions according to SPSS). If the number of variance decomposition proportions exceeding 0.50 is greater than two for a component associated with high condition index (say more than 100), the indication "collinearity" is printed (according to correlation criterion 6 of section 3). The indication "collinearity" is printed as many times as the number of components which satisfy the criterion.

```
for k from 0 to nrows(vars)-1 do
 (if((count_sup(0.5,variance_proportions(vars),col)[k]>1)                and
condition_indices(vars)[k]>100) print("collinearity") ; )
end_for;
```

The associated function in Xcas is

```
vdp_test(vars):=for k from 0 to nrows(vars)-1 do
 (if((count_sup(0.5,variance_proportions(vars),col)[k]>1)                    and
condition_indices(vars)[k]>100) print("collinearity") ; )
end_for;
```

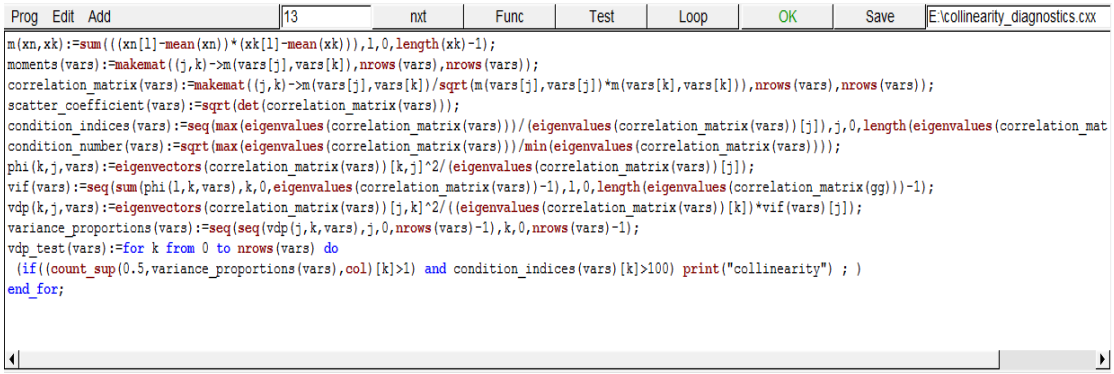


Figure 2: "collinearity_diagnostics" program file in Xcas

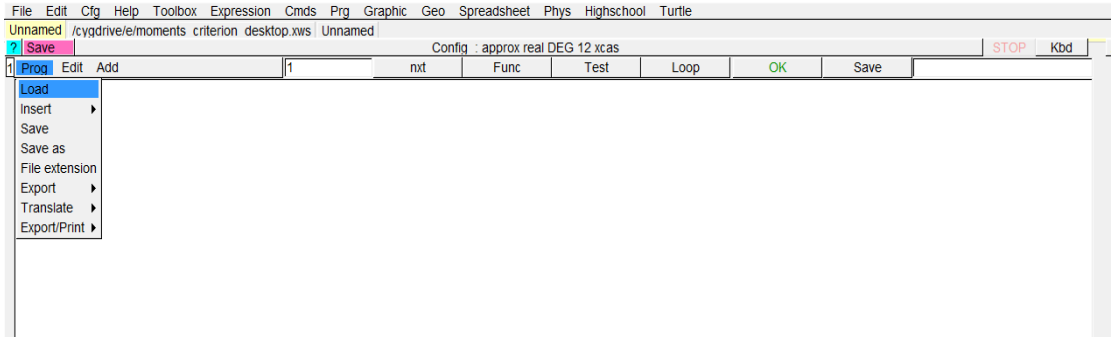


Figure 3: Loading a program file in Xcas

Working in any session in Xcas, by loading collinearity_diagnostics.cxx program file[3] or by writing in a commandline `read("collinearity_diagnostics.cxx")` we can use *moments, correlation_matrix, vif, scatter_coefficient, condition_number, condition_indices, variance_proportions, cn_condition, vdp_test* functions.

5. An illustrative example

We consider the data used by a telephone cable manufacturer to predict sales to a major customer for the period 1968-1983 as presented in Table 1 (the example is taken from Gujarati, 2003, p.290).

[3] All files are available on request.

Here we demonstrate the results from Xcas programmed (and built-in) functions. Similar results are generated in different outputs in SPSS, MINITAB and STATA (but not all together in the same software). Moreover, we generate the scatter coefficient (posed by correlation criterion 2) and the rank of the moments and data matrices. A direct answer for collinearity and its degree is given by *cn_condition* function and *vdp_test* function.

Table 1: Data of the illustrative example

Year	X_2 *GNP*	X_3 housing starts	X_4 unemployment %	X_5 prime rate lag, 6 mos	X_6 customer line gains, %	*Y,* total plastic purchases (MPF)
1968	1051.8	1503.6	3.6	5.8	5.9	5873
1969	1078.8	1486.7	3.5	6.7	4.5	7852
1970	1075.3	1434.8	5.0	8.4	4.2	8189
1971	1107.5	2035.6	6.0	6.2	4.2	7497
1972	1171.1	2360.8	5.6	5.4	4.9	8534
1973	1235.0	2043.9	4.9	5.9	5.0	8688
1974	1217.8	1331.9	5.6	9.4	4.1	7270
1975	1202.3	1160.0	8.5	9.4	3.4	5020
1976	1271.0	1535.0	7.7	7.2	4.2	6035
1977	1332.7	1961.8	7.0	6.6	4.5	7425
1978	1399.2	2009.3	6.0	7.6	3.9	9400
1979	1431.6	1721.9	6.0	10.6	4.4	9350
1980	1480.7	1298.0	7.2	14.9	3.9	6540
1981	1510.3	1100.0	7.6	16.6	3.1	7675
1982	1492.2	1039.0	9.2	17.5	0.6	7419
1983	1535.4	1200.0	8.8	16.0	1.5	7923

Variables X_2, X_3, X_4, X_5, X_6 in Xcas are introduced in matrix notation, having their values lying in rows.

First we define the matrix of the explanatory variables

```
gg:=[[1051.8,1078.8,1075.3,1107.5,1171.1,1235.0,1217.8,1202.3,1271.0,1332.7,
1399.2,1431.6,1480.7,1510.3,1492.2,1535.4],[1503.6,1486.7,1434.8,2035.6,2360
.8,2043.9,1331.9,1160.0,1535.0,1961.8,2009.3,1721.9,1298.0,1100.0,1039.0,120
0.0],[3.6,3.5,5.0,6.0,5.6,4.9,5.6,8.5,7.7,7.0,6.0,6.0,7.2,7.6,9.2,8.8],[5.8,
6.7,8.4,6.2,5.4,5.9,9.4,9.4,7.2,6.6,7.6,10.6,14.9,16.6,17.5,16.0],[5.9,4.5,4
.2,4.2,4.9,5.0,4.1,3.4,4.2,4.5,3.9,4.4,3.9,3.1,0.6,1.5]]
```

From now on in Xcas we may recall the set of X_2, X_3, X_4, X_5, X_6 variables of our example by "gg".

We compute the rank of the variables using built-in Xcas function *rank*:

```
rank(gg)
5
```

Based on this primitive result we can assume that the rank of the observations is equal to five or that the systematic variation of the variables under study has five degrees of freedom.

In order to load the routine for collinearity diagnostics in Xcas, we write:

```
read("collinearity_diagnostics.cxx")
```

Next, we generate the momentsmatrix by the programmed function *moments*:

```
moments(gg)
```

430837.599375,	-355877.635625,	3176.27875,	8693.86375,	-2214.305625
-355877.635625,	2405028.96938,	-4785.50125,	-18988.51625,	4703.279375
3176.27875,	-4785.50125,	44.3575,	76.3475,	-26.95125
8693.86375,	-18988.51625,	76.3475,	267.4575,	-67.97625
-2214.305625,	4703.279375,	-26.95125,	-67.97625,	25.029375

We compute the rank of the moments matrix using built-in Xcas function *rank*:

```
rank(moments(gg))
5
```

We generate the correlation matrix of the five variables, its determinant and its eigenvalues:

```
correlation_matrix(gg))
```

1.0,	-0.349609915876,	0.726571368304,	0.809894856,	-0.674304040551
-0.349609915876,	1.0,	-0.463322880186,	-0.748692502708,	0.606199561778
0.726571368304,	-0.463322880186,	1.0,	0.700944931338,	-0.808854450482
0.809894856,	-0.748692502708,	0.700944931338,	1.0,	-0.830815919353
-0.674304040551,	0.606199561778,	-0.808854450482,	-0.830815919353,	1.0

```
det(correlation_matrix(gg)))
0.00663839557296
```

(this value indicates high degree of organization for the data variables. We cannot detect multicollinearity since the value is greater than the cutoff value of 0.00001)

```
eigenvalues(correlation_matrix(gg))
(0.0396034793407,0.179588118415,0.353410545961,0.710556578822,3.71684127746)
```

(the fact that one eigenvalue is near zero (though >0.01) indicates one near collinear relation)

We compute the variance inflation factors for the five variables:

```
vif(gg)
[6.90516019365,4.3449454702,3.96791652719,14.6830463272,5.42349875894]
```

(the forth value 14.683046>10 indicates collinearity for explanatory variable X_5)

We compute the condition number as defined in (7) and the condition indices:

```
condition_number(gg)
9.68769230702
```
(a value <10 indicates that multicollinearity is not strong)

```
condition_indices(gg)
[93.8513822355,20.6964765278,10.5170638509,5.23088714993,1.0]
```
The sequence of condition indices is presented in accordance with the sequence of the eigenvalues above.

If a value of a condition index exceeds a cutoff value of, say, 100 to 1000, two or more columns of the data matrix have moderate to strong relations (Belsley et al., 1980; Callaghan & Chen, 2008).

Interpreting the value of the condition number as a collinearity diagnostic using the programmed function *cn_condition* we get:

```
cn_condition(gg)
"weak collinearity"
```

```
variance_proportions(gg)
```
0.789747368793,	0.0164981306652,	0.137507682737,	0.0489233567007,	0.00732346110473
0.645096335214,	0.13780700556,	0.00146544443674,	0.206942801592,	0.00868841319737
0.266141043443,	0.442220144046,	0.241456463891,	0.0365199110244,	0.013662437595
0.951491907106,	0.0135453632052,	0.0287934216339,	0.00170158736506,	0.00446772068974
0.362725447838,	0.535084686354,	0.0908705385282,	8.78062052653e-05,	0.0112315210738

The first row contains the variance decomposition proportions (vdps) of the first independent variable X_2, the fifth row contains the vdps of the fifth independent variable X_6, e.t.c. The sequence of vdps in each row is in accordance with the sequence of the eigenvalues and the sequence of condition indices given above. A detailed report for diagnosing collinearity based on variance proportions is made in Belsley et al. (1980) and Callaghan & Chen (2008).

The *count_sup* built-in function below counts the number of the variance proportions strictly greater than 0.5 per column.

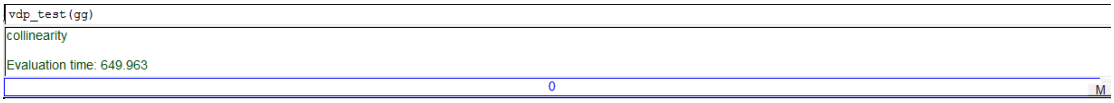
```
count_sup(0.5,variance_proportions(gg),col)
[3,1,0,0,0]
```

The fact that three variance proportions in the first column, associated with variables X_2, X_3, X_5, are greater than 0.5 is not a red flag since they are related with a condition index smaller than 100.

```
vdp_test(gg)
0
```

(the absence of printed output means absence of collinearity in our data)

If we change the cutoff value for the condition indices from 100 to 90 (violating the correlation criterion) in the codes of *vdp_test* function in order to illustrate its performance in this example, we get:



(three variance proportions in the first column of the variance proportion matrix are bigger than 0.5 and the first column is associated with a condition index>90. This fact reflects the "collinearity" indication once).

In order to check the degree of organization of the given data, we generate the scatter coefficient of the explanatory variables X_2, X_3, X_4, X_5, X_6 by the programmed function *scatter_coefficient*:

```
scatter_coefficient(gg)
0.0814763497768
```

(a value close to zero lets us suspect linear dependency for the set of explanatory variables. The scatter coefficient is the first indicator in this example to raise a red flag)

If we consider the set of X_2, X_3, X_4 instead, we calculate the corresponding scatter coefficient to see that its value is considerably higher and closer to one:

```
scatter_coefficient(gg_without_x5_x6)
0.608754651635
```

(a value close to one indicates that X_2, X_3, X_4 do not constitute a linearly dependent set. The corresponding condition number of this set of variables is 2.79940489207)

6. Conclusions

Correlation matrices and moments matrices provide the key links between sample data and best linear unbiased estimation. In this paper we exploited Xcas' built-in matrix functions and Xcas' programming capabilities to examine special topics on correlation analysis. Using simple functional programming techniques with simple input – direct output, we accomplished to evaluate a number of metrics like the correlation matrix of data variables, the scatter coefficient, the variance inflation factors, the condition indices, the condition number and the variance decomposition proportions that help a user check a number of criteria concerning the degree of scattering organization of given data.

Furthermore, with a routine in the Xcas program editor we provided the result of the collinearity study instantly, in a black box mode, avoiding the complex interpretation of a series of indicators. In this way we extract the indications of the degree of multicollinearity.

Xcas seems to be an efficient environment for doing econometrics. Researchers could be inspired of the capabilities of this versatile computing environment and find ways to make further use in econometric methodologies.

Appendix

Relevant output in commercial statistical software

Computing the relevant procedure for the illustrative example in Section 5 in Stata (collin option) we result in the same condition number, the same eigenvalues and the same determinant of correlation matrix:

```
. collin x2 x3 x4 x5 x6, corr
(obs=16)
  Collinearity Diagnostics
                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
        x2      6.91    2.63    0.1448       0.8552
        x3      4.34    2.08    0.2302       0.7698
        x4      3.97    1.99    0.2520       0.7480
        x5     14.68    3.83    0.0681       0.9319
        x6      5.42    2.33    0.1844       0.8156
----------------------------------------------------
  Mean VIF      7.06

                          Cond
        Eigenval          Index
---------------------------------
    1     3.7168          1.0000
    2     0.7106          2.2871
    3     0.3534          3.2430
    4     0.1796          4.5493
    5     0.0396          9.6877
---------------------------------
 Condition Number         9.6877
 Eigenvalues & Cond Index computed from deviation sscp (no intercept)
 Det(correlation matrix)    0.0066
```

Computing the Linear Regression and Factor Analysis procedures in MINITAB and SPSS, the relevant printout for our illustrative example (in Section 5) certifies once again the validity of our results in Xcas:

Data Display

Matrix CORR1

1.00000	-0.34961	0.72657	0.80989	-0.67430
-0.34961	1.00000	-0.46332	-0.74869	0.60620
0.72657	-0.46332	1.00000	0.70094	-0.80885
0.80989	-0.74869	0.70094	1.00000	-0.83082
-0.67430	0.60620	-0.80885	-0.83082	1.00000

Figure 4: The correlation matrix in MINITAB

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	7543,125	156,900		48,076	,000					
	Zscore(x2)	827,669	425,818	,680	1,944	,081	,210	,524	,259	,145	6,905
	Zscore(x3)	946,573	337,777	,778	2,802	,019	,497	,663	,373	,230	4,345
	Zscore(x4)	-1408,608	322,789	-1,157	-4,364	,001	-,263	-,810	-,581	,252	3,968
	Zscore(x5)	50,716	620,934	,042	,082	,937	-,050	,026	,011	,068	14,683
	Zscore(x6)	-1099,789	377,379	-,904	-2,914	,015	,011	-,678	-,388	,184	5,423

a. Dependent Variable: Y

Figure 5: Collinearity Statistics in SPSS (computed considering centered data[4])

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Zscore(x2)	Zscore(x3)	Zscore(x4)	Zscore(x5)	Zscore(x6)
1	1	3,717	1,000	,00	,01	,01	,01	,00	,01
	2	1,000	1,928	1,00	,00	,00	,00	,00	,00
	3	,711	2,287	,00	,05	,21	,04	,00	,00
	4	,353	3,243	,00	,14	,00	,24	,03	,09
	5	,180	4,549	,00	,02	,14	,44	,01	,54
	6	,040	9,688	,00	,79	,65	,27	,95	,36

a. Dependent Variable: Y

Figure 6: Collinearity Diagnostics in SPSS (computed considering centered data)

Correlation Matrix[a]

		x2	x3	x4	x5	x6
Correlation	x2	1,000	-,350	,727	,810	-,674
	x3	-,350	1,000	-,463	-,749	,606
	x4	,727	-,463	1,000	,701	-,809
	x5	,810	-,749	,701	1,000	-,831
	x6	-,674	,606	-,809	-,831	1,000

a. Determinant = ,007

Figure 7: The correlation matrix in SPSS

[4] Running a linear regression on the z-scores

References

Amman, H., Kendrick, D. & Rust, J., (1996). *Handbook of Computational Economics*. Elsevier North-Holland, Amsterdam, The Netherlands.

Bautin, G. A., Kalyagin, V.A., Koldanov, A.P., Koldanov, P.A. & Pardalos, P.M. (2013). Simple measure of similarity for the market graph construction. *Computational Management Science* 10(2), 105-124.

Bekker, P.A., Merckens, A., & Wansbeek, T.J. (1994). *Identification, Equivalent Models, and Computer Algebra*. San Diego, CA: Academic Press. (out of print)

Belsley, D.A. (1991a). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4, 33-50.

Belsley, D.A. (1991b). *Collinearity diagnostics: Collinearity and weak data in regression*. New York: John Wiley & Sons.

Belsley, D.A., Kuh, E. & Welsch, R.H. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Belsley, D.A. (1999). Mathematica as an Environment for Doing Economics and Econometrics. *Computational Economics*, 14(1), 69-87.

Bjerkoholt, O. & Dupont-Kieffer, A. (Eds), (2009). *Problems and Methods of Econometrics. The Poincare Lectures of Ragnar Frisch*, 1993. Routledge, London, New York.

Bollen, K.A. & Bauldry, S. (2010). Model identification and computer algebra. *Sociological Methods and Research*, 39(2), 127-156.

Brown, M.T. & Wicker R.T. (2000). Discriminant analysis, In*:* Tinsley H.E.A. and Brown S.D. (Eds), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, 209-235, Elsevier.

Callaghan, K. & Chen, J. (2008). Revisiting the Collinear Data Problem: An Assessment of Estimator 'Ill-Conditioning' in Linear Regression, *Practical Assessment, Research & Evaluation,* 13(5).

Field, A. (2009). *Discovering Statistics using SPSS for Windows, Third edition*. Sage publications, Los Angeles, London, New Delhi, Singapore, Washington D.C.

Frawley, W.J. (1985). Computer Generation of Symbolic Generalized Inverses and Applications to Physics and Data Analysis. In: R. Pavelle (Ed.) *Applications of Computer Algebra*. Kluwer Academic Publishers, Boston/Dordrecht/Lancaster.

Friendly, M. & Kwan, E. (2009). Where's Waldo? Visualizing Collinearity Diagnostics, *The American Statistician*, 63(1), 56-65.

Fox, J. (1984). *Linear Statistical Models and Related Methods*. John Wiley and Sons, New York.

Gujarati, D.N. (2003). *Basic Econometrics*, 4th ed. Boston: McGraw Hill.

Halkos, G.E. (2006). *Econometrics: Theory and Practice*. Giourdas Publications, Athens.

Halkos, G.E. (2011). *Econometrics: Theory and Practice: Instructions in using Eviews, Minitab, SPSS and Excel.* Gutenberg: Athens.

Halkos, G.E. & Tsilika, K.D. (2015). Programming Identification Criteria in Simultaneous Equation Models. *Computational Economics* 46(1), 157-170. DOI: 10.1007/s10614-014-9444-9.

Hendrickx, J. (2010). perturb: Tools for evaluating collinearity. R package version 2.04. URL http://CRAN.R-project.org/package=perturb

Hutton, J. & Hutton, J. (1995). The Maple computer algebra system: A review. *Journal of Applied Econometrics*, 10(3), 329–337. doi: 10.1002/jae.3950100308

Kendrick, D.A. & Amman, H.M. (1999). Programming languages in economics. *Computational Economics*, 14, 151-181.

Li, J. & Racine, J.S. (2008). Maxima: An open source computer algebra system. *Journal of Applied Econometrics*, 23(4), 515–523. doi: 10.1002/jae.1007

Merckens, A. & Bekker, P.A. (1993). Identification of simultaneous equation models with measurement error: a computerized evaluation. *Statistica Neerlandica*, 47(4), 233–244.

Merckens, A. (1991). *Computer algebra applications in econometrics*. Doctoral Thesis, 164 p. University of Groningen.

Pindyck, R.S. & Rubinfeld, D.L. (1998). *Econometric Models and Economic Forecasts*. 4th Edn., McGrawHill, Boston.

Schipp, B. & Krämer, W. (Eds) (2009). *Statistical Inference, Econometric Analysis and Matrix Algebra: Festschrift in Honor of Götz Trenkler*, Springer, Heidelberg.

Sheppard, K. (2014). *Introduction to Python for Econometrics, Statistics and Data Analysis*. University of Oxford. Tuesday 5th August, 2014. Available at: https://www.kevinsheppard.com/images/0/09/Python_introduction.pdf

Shirokikh, O., Pastukhov, G., Boginski, V. & Butenko, S. (2013). Computational study of the US stock market evolution: a rank correlation-based network model. *Computational Management Science* 10(2), 81-103.

Stroeker, R.J. & Kaashoek, J.F. (1999). *Discovering Mathematics with Maple: An interactive exploration for mathematicians, engineers and econometricians*. Springer Basel AG.