Research Article

Using Scaled And Translated Measure To Compare Between Robust Estimators In Canonical Correlation

Zahraa Khaleel Hammoodi[a] , Lekaa Ali Mohammad[b]
[a] Baghdad University, College of Administration and Economic, Iraq, (zahraaeyes84@gmail.com)
[b] Baghdad University, College of Administration and Economic, Iraq, (lekaa.ali.1968@gmail.com)

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract :Many researches have dealt with analysis of classical canonical correlation based on either covariance (heterogeneity) or correlation matrix where the coefficient of correlation used is Pearson which is biased to the outlier's values, because of it depends on mean in the calculation. In our research we find robust canonical correlation depend on robust methods which is insensitive towards outliers value. Methods are used Percentage bend correlation coefficient (Pe) & Biweight midcorrelation coefficient correlation (Bi) to estimate canonical correlation (CC) instead of Pearson correlation.
The researchers addressed robustness measurement to check the ability of robust methods for contaminated values, we used biased and translated estimator of empirical influence function to make the comparison between robust methods when we use simulation and choose (Bi) method to apply it on real data.

Key Words: Canonical Correlation, Outliers, Percentage Bend Correlation, Biweight Midcorrelation Coefficient, Influence Function.

Introduction

Canonical correlation coefficient is generalization of multiple correlation as it consists of two sets of variables, the first are dependent variables ($Y_1, Y_2, , , , , , , Y_P$) and the second is explanatory variables ($X_1, X_2, , , , , , , X_q$) , and both groups have a common distribution.
Canonical correlation analysis contributes to describe two sets of variables, one of which is auxiliary and the other is the original variables corresponding to the helpful variables.
It is worth to say that the concept of the canonical correlation appeared in the period 1935/1936 by the scientist (Hotelling), and it became clear that the multiple correlation is a special case of the canonical correlation.in (1940) the scientist (Fischer) was the first to use the canonical correlation to analyze harmonic tables with ordered categories. [1]
The most central concept in Hampel's fundamental contribution to robustness theory (Hampel, 1968, 1971, 1974) is the "influence function". He and his co-researchers used heuristics of influence function and developed a new approach to Robust Statistics. [2] .In (1992), the scientist (Mario Romanazzi) presented the derivation of the influence function for the square of the correct and multiple correlation coefficient in addition an explanation and detailed description of three types of sample transformations of the influence function which are (the influence function, the deleted experimental influence function and the sample effect function) as well as finding influence function of the Eigen values and Eigen vectors and the characteristic values, depending on the study of (Hample 1974) in the early seventies[3]. The researchers (Nasser And Alam) introduced in (2006) articles about estimators of influence function included six estimators have the same process as original influence function [4].In (2013) (Alkenani & Keming) represented two types of Estimators divided in to two groups (M-estimators) which includes (Percentage Bend , Biweight midcorrelation, Winsor zed, Kendall , Spearman correlation) to estimate correlation matrix instead of Pearson correlation, the second group (O-estimators) includes (MVE,MCD,FCH,RFCH breakdown and RMVN estimators),the results mentioned the preference for (Biweight) , to estimate correlation matrix and in the second groups the preference was to (FMCD) to estimate heterogeneity matrix[5]. In (2016),(Veenstra , Cooper & Phelps) introduced A study in analyzing the relationship between the returns of different securities because of its fundamental importance in many areas of finance, such as improving the stock market by using the Biweight Midcorrelation (Bicor) (instead of the Pearson correlation coefficient) as it is considered one of the more powerful measures. To find out the relationship between the returns, and the results showed that the (Bicor) method can be used to improve the method of building a financial portfolio based on the chart when dealing with the correlation matrix, thus obtaining better performance [6].
In many phenomena include data that follow a normal distribution, we find some violations of the distribution conditions represented by the presence of outliers, thus the resulting estimates will be inconsistent and inefficient.

Canonical correlation coefficient is one of the most important estimations in describing the nature and strength of the relationship between two sets of variables, which in turn is also affected by the outliers if it is estimated by the classical methods. Here, the concept of our research was launched in order to address this problem by employing some robust methods that can be described as resistance to outlier values.

In our research, we use empirical influence function of scaled and translated version to check the effect of outliers by making a comparison between two robust methods and show the influence function for canonical correlation and weights vectors.

Canonical Correlation Analysis (CCA)

Canonical correlation aims to study the relationship between a set of X explanatory variables and a set of Y response variables. [7]

Assuming the study of two sets of variables:

X_{p*1} is a vector with dimension $p*1$ for the first set

Y_{q*1} is a vector with dimension $p*1$ for the second set

P: is the number of variables in the first group (X) and q: represents the number of variables in the second group (Y). The variables of both groups follow the normal multivariate distribution as each group has the following specifications:

E(y) = μ_y E(x) = μ_x

Var (y) = Σ_{yy} Var (x) = Σ_{xx}

And the homogeneity matrix between the two sets known as:

$$\binom{X}{Y} \sim \text{MVN}\left[\binom{\mu_x}{\mu_y},\ \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}\right]$$

$\Sigma_{xx} > 0$ ، $\Sigma_{yy} > 0$ And assume$p \le q$, so we can define number of linear combination equal to number of $Min_{(p,q)}$ by using this equation:

$$u_i = \bar{a}_i\,\underline{x}$$

$$i = 1,2,\ldots\ldots,n \quad ..(1) \qquad = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p$$

$$v_i = \bar{b}_i\,\underline{y}$$

$$= b_{1i}y_1 + b_{2i}y_2 + \cdots + b_{pi}y_q \qquad i = 1,2,\ldots\ldots,n \quad ..(2)$$

Every linear combination differ in weight values for every variable because of the important variable difference inside the set and its effect on canonical variates Ui or Vi

To calculate the canonical correlation coefficient between two variables: Corr $\left(\frac{x}{\underline{y}}\right)$

And Based on the basis of the variance of each set of variables:

$$\text{Var}\,(\underline{\acute{a}}\,\underline{x}) = \underline{\acute{a}}\,\Sigma_{xx}\,\underline{a} = 1 \ldots\ldots..(3)$$

$$\text{Var}\,(\underline{\acute{b}}\,\underline{y}) = \underline{\acute{b}}\,\Sigma_{yy}\,\underline{b} = 1 \ldots\ldots\ldots(4)$$

$$\underline{\acute{a}}\,\Sigma_{xx}\,\underline{a} = \underline{\acute{b}}\,\Sigma_{yy}\,\underline{b} = 1 \ldots\ldots\ldots(5)$$

And the cov between linear combination

$$\text{Cov}\,(\underline{\acute{a}}\,\underline{x}, \underline{\acute{b}}\,\underline{y}) = \underline{\acute{a}}\,\Sigma_{xy}\,\underline{b} \ldots\ldots\ldots\ldots(6)$$

So the correlation is :

$$\text{Corr}(\underline{\acute{a}}\,\underline{x}, \underline{\acute{b}}\,\underline{y}) = \frac{\underline{\acute{a}}\,\Sigma_{xy}\,\underline{b}}{\sqrt{\underline{\acute{a}}\,\Sigma_{xx}\,\underline{a}}\ \sqrt{\underline{\acute{b}}\,\Sigma_{yy}\,\underline{b}}} \qquad \ldots(7)$$

The main objective of the analysis of the canonical correlation is to explain the structure of the correlation between the X and Y variables through the linear compositions (variables) U and V, so it is necessary to find $\underline{a}$, $\underline{b}$ and their components while maximizing the correlation.

The first pair of variables (u_1, v_1) are chosen in order to maximize the heterogeneity between them, the linear compositions of the husband

$$u_1 = \underline{a_1}\,\underline{x} \quad , \ v_1 = \underline{b}_1\underline{y}$$

And since the variation of the variables of the first pair is equal to the one, the canonical correlation:

$$\rho_{(u1,v1)} = \max_{\underline{a,b}}(\underline{\acute{a}}\,\underline{x}, \underline{\acute{b}}\,\underline{y}) \ldots\ldots\ldots.(8)$$

The resulting correlation represents the coefficient of the canonical correlation of the first pair

The second pair of variables (u1,v1) are selected in order to maximize the heterogeneity of cov (u,v) provided that the linear compositions of the pair are perpendicular to the first pair (u1,v1) meaning that

Cov $(\underline{\acute{a}}\,\underline{x}\,, u1) = 0 \ldots\ldots\ldots..(9)$

Cov $(\,\underline{\acute{b}}\,\underline{y}\,, v1) = 0 \ldots\ldots\ldots..(10)$

= 1…. (11) Var ($\underline{\acute{b}}\,\underline{y}$) = Var ($\underline{\acute{a}}\,\underline{x}$)

Maximizing the correlation between $\underline{\acute{b}_2}\underline{y}$ *and* $\underline{\acute{a}_2}\underline{x}$ is called the second canonical correlation coefficient and generally the pair (Uj,Vj) of the canonical variables is chosen to maximize the heterogeneity of Cov (u1,v1) Thus, the coefficients of correlation in the significance of the variables and variance are estimated in the relationship

$$r_c = \frac{\acute{U}S_{xy}V}{\sqrt{\acute{U}S_{xx}U}\sqrt{\acute{V}S_{yy}V}} \quad \ldots\ldots.. (12)$$

We can calculate the CCA by correlation matrix:

S= DRD

Since:

R: is a correlation matrix for X & Y sets or the homogeneity between them.

D: is a diagonal matrix its component represents the root of variance for every variables.

D=diag$\left(\sqrt{S_{ij}}\right)$

Thus, the canonical correlation by correlation matrix can describe as:

$$r_c = \frac{\acute{C}R_{xy}D}{\sqrt{\acute{C}R_{xx}C}\sqrt{\acute{D}R_{yy}D}} \quad \ldots\ldots.. (13)$$

Since:

C&D: is the canonical variables which is chosen to maximize heterogeneity.

To estimate canonical weight which is maximize canonical correlation, the function:

$$g = \acute{C}R_{xy}D - \frac{\sqrt{\lambda 1}}{2}\acute{C}R_{xx}C - \frac{\sqrt{\lambda 2}}{2}\acute{D}R_{xx}D \ldots\ldots(14)$$

And to $max_{c,d}(g)$ through:

$$\frac{\partial g}{\partial d} = 0\,, \frac{\partial g}{\partial c} = 0$$

$$\frac{\partial g}{\partial c} = R_{xy}\underline{d} - \sqrt{\lambda_1}R_{xx}\underline{c} \ldots\ldots\ldots\ldots\ldots\ldots... (15)$$

$$\frac{\partial g}{\partial d} = \underline{\acute{c}}R_{xy} - \sqrt{\lambda_2}\underline{\acute{d}}R_{yy} \ldots\ldots\ldots\ldots\ldots\ldots... (16)$$

From equation (17) we will find that the weight canonical:

$$\underline{C} = \frac{1}{\sqrt{\lambda_1}}\,R_{xx}^{-1}\,R_{xy}\,\underline{d} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (17)$$

And by compensating C in the second equation we get the relationship:

$$R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy} - \lambda I)\,\underline{d} = \underline{0}$$

It represents the Eigen equations of the$R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$ and the rootsλ_i which not equal to zero achieved by the solution of this equation are equal to q and are called subjective values, and the square coefficient of the coefficient of correlation between each pair of variables is equal to the value of the characteristic root according to the following formula:

$$\boldsymbol{r_c^2} = \sqrt{\boldsymbol{\lambda}}$$

Biweight Midcorrelation Coefficient (Bi)

One of the disadvantages of the Pearson correlation coefficient is that it is easily exposed to the effects of outliers, so a number of alternatives have been relied on from the strong correlation coefficients, including the two-weight mean correlation coefficient.

Let ψ an odd function, μ_x & μ_y location standard for random variable X , Y straightly and let τ_y & τ_x measuring scale for random variable X&Y , If K is a constant magnitude, define the variables in terms of the previous features with the formula: [5] [6]

$$U = \frac{(X-\mu_x)}{K\tau_x}\;,\;\; V = \frac{(Y-\mu_y)}{K\tau_y}$$

So, the heterogeneity scale between X&Y describe as:

$$\gamma_{xy} = \frac{nk^2.\tau_x.\tau_y\,E(\psi(u).\psi(v))}{E(\psi(u)).E(\psi(v))} \quad \ldots\ldots\ldots\ldots\ldots (18)$$

Since correlation scale ρ_b calculate as:

$$\rho_{b=}\frac{\gamma_{xy}}{\sqrt{\gamma_{xx}.\gamma_{yy}}} \qquad -1 \le \rho_b \le 1 \quad \ldots\ldots\ldots\ldots (19)$$

By choosing K = 9 and the function, which represents the biweight function, which is known as the following relationship:

$$\psi(x) = \begin{cases} x(1-x^2) & if\ |x| < 1 \\ 0 & if\ |x| \geq 1 \end{cases}$$

And let med $_x$ & med $_y$,variable median for X&Y straightly calculate from random sample for observation pairs order (X_1,Y_1)،(X_2,Y_2)،،،، (X_n,Y_n) From this results in the definition of the variables:

$$U_i = \frac{(X_i - med_x)}{9.MAD_x}, V_i = \frac{(Y_i - med_y)}{9.MAD_y}$$

We note U_i Proportional to the distance between X_i and the median for X. [6, pp. 4]

Since Median Absolute Deviation(MAD_y& MAD_x) represent:

$$MAD_x = med_i|x - med_{xi}| = med|x - med_x|$$

If we define variables b_i & a_i about their relationship to the variables U_i & V_i

$$a_i = \begin{cases} 1 & -1 \leq U_i \leq 1 \\ 0 & O.W \end{cases}$$

$$b_i = \begin{cases} 1 & -1 \leq V_i \leq 1 \\ 0 & O.W \end{cases}$$

So, we obtain Biweight Midcoveriance between X & Y:

$$Bicov(x,y) = \frac{n\Sigma a_i(X_i - med_x)(1-U_i^2)^2 b_i(Y_i - med_y)(1-V_i^2)^2}{[\Sigma a_i(1-U_i^2)(1-5U_i^2)][\Sigma b_i(1-V_i^2)(1-5V_i^2)]} \quad …….. (20)$$

After apply correlation formula, the estimation Biweight midcorrelation:

$$r_{bi} = \frac{bicov(x,y)}{\sqrt{bicov(x,x).bicov(y,y)}} \quad … … … … (21)$$

To check r_{bi} , we test this assumption

$H_0: \rho_b = 0$

Which is refer that X&Y independent variables, to calculate statistic test:

$$T_b = r_b.\sqrt{\frac{n-2}{1-r_b^2}}$$

And we reject H_0 if

$|T_b| > t_{1-\frac{\alpha}{2}}$

$t_{1-\frac{\alpha}{2}}$ Table value at T distribution with d.f., V=n-2 and error type I equal α.

Percentage Bend Correlation Coefficient(Pe)

Percentage bend correlation consider one of resistance estimators towards outliers, we find correlation value between X & Y.

Let X a random variable with distribution function F and let ψ is non-decreasing odd function, w_x is a constant measure attached with X, then M measure which is related with ψ is ϕ_x and achieve: [8] [9]

$$\int \psi\left(\frac{X-\phi_x}{w_x}\right) = 0$$

If $\psi(x) = x$ & $\phi_x = M$, then the mean represent one of ϕ_x , called(M-estimator), determine from:

$$\Sigma\, \psi\left(\frac{x_i - \hat{\phi}_x}{\hat{w}_x}\right) = 0$$

SinceX_1,X_2,….X_n is random sample & $\hat{w}_x$ is an estimator to w_x , the variance measure called (Midvariance)

$$\gamma_x^2 = \frac{k^2 w^2 E(\psi^2(u))}{[E(\psi(\acute{u}))]^2} ……(22)$$

Since: U= $\frac{(X-\phi_x)}{Kw_x}$ & k: is a constant.

Let Y is another variable, then variance measure between X&Y described as :

$$\gamma_{xy} = \frac{K^2 w_x w_y E(\psi(u).\psi(v))}{E(\psi(\acute{u}))E(\psi(\acute{v}))} ………(23)$$

Since: V= $\frac{(Y-\phi_y)}{Kw_y}$, then

So, correlation coefficient ρ_{pb} described as:

$$\rho_{pb} = \frac{E(\psi(u).\psi(v))}{[E(\psi^2(u)).E(\psi^2(v))]^{1/2}} …..(24)$$

And to test correlation according to null hypothesis H_0

$H_0: \rho_{pb} = 0$
Which is mentioned that X&Y independent, we calculate:

$$T_{pb} = r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}}$$

Then we reject H_0 if :
$|T_{pb}| > t_{1-\alpha}$
We compare calculated value for test with table value for t distribution with degree of freedom (n-2) and (α).

Influence Function (IF)

The IF basically consider analytic tool, can use it to evaluate the effect of observation on estimator T_n at distribution function F by: [10]

$$\mathrm{IF}_{T_n,F(x)} = \lim_{\omega\to 0}\frac{[T_n(F_\omega)-T_n(F)]}{\omega} \quad \text{………} \ (25)$$

Since:

$$F_\omega = (1-\omega)F + \omega\delta_x \text{……….}(26)$$

Since:
ω : Contaminated ratio $0 < \omega < 1$
δ_x : Probability scale
The denominator is a constant amount and the numerator contains the basic information about the IF effect function. Therefore, it became necessary to go into some detail on the Estimator of the influence function, which are work the same as the IF :

Biased and Translated Estimators.

Empirical influence function defined as depending on the (unscaled and untranslated & unscaled and translated estimators) [4] with this formula:

$$\mathrm{EIF}(x, F_n) = \mathrm{IF}(x, F_n)$$
$$= U_{\omega\to 0}\frac{T(F_n+\omega(\delta_x - F_n)) - T(F_n)}{\omega} \quad \text{…………} \quad (27)$$

Since:
F_n: distribution function
$(\delta_x - F_n)$: the difference between contaminated observation distribution and original observation distribution
Therefore, the magnitude $T(F_n + \omega(\delta_x - F_n))$ is obtained through an estimator (T) with two distributions, most of which follow the normal distribution (the original distribution), but contain few observations that follow the contaminated distribution (resulting from the addition or substitution of a contaminated observation).
The expression T (Fn) represents the original estimator resulting from the original distribution function Fn of sample size (n).
It is better to estimate the empirical effect function (influence function) in relation to:

$$\mathrm{EIF}_e(x, F_n) = \mathrm{IF}_e(x, F_n)$$
$$= \frac{T\left(F_n + \frac{1}{100\,n}(\delta_x - F_n)\right) - T(F_n)}{\frac{1}{100\,n}} \quad \text{…………} \ (28)$$

Since:
$\frac{1}{100\,n}$: represent the ratio which is taken to contaminate data.
From this, the empirical influence function can defined as:

$$I_j = \mathrm{IF}_{(xj)} = \mathrm{EIF}(x_j, F_n) \quad \text{…………..} \ (29)$$

Which can be rounded by choosing different values to $\boldsymbol{\omega}$ (contamination data) as ($\frac{1}{n}$, $\frac{1}{\sqrt{n}}$, $\frac{1}{n+1}$, $\frac{1}{n-1}$) and other values without take the limit for the amount. [4]

Simulation

Simulation method is an important tool and computer experiments that include creating data by taking random samples and generating data in several ways to prove and evaluate the success and efficiency of methods also models used in statistical research. Simulation studies are used to obtain experimental results about the performance of the statistical methods that are used in the analysis. Statistician for the research under study [17, pp.2047]

Simulation experiments included generating multivariate normal distribution data with different sample sizes based on means vector μ and covariance matrix **Σ** for real data (Oil Exports and Returns) , as well as generating multivariate contaminant normal distribution tracking data by employing mean vectors, co-variance matrices and different contamination ratios, The canonical correlation coefficients were also estimated according to these methods : Percentage bend correlation coefficient & Biweight Midcorrelation cosfficient , then make a comparison between these robust methods based on the empirical influence function standard with the scaled and transformed estimators.

Steps of Simulation:

Generating six variables following the multivariate normal distribution $N_p\left(\underline{\mu}, \Sigma\right)$which are on the order$x_1$,$x_2, x_3, z_1, z_2, z_3$ depending on the mean vector μ and the CV matrix ∑ of the real data after converting it to the standard form. For the non-conformity of the units of measure for those data, a vector means and a matrix of variance and covariance mentioned below were obtained:

$$\underline{\mu} = \underline{0}\,, \quad \Sigma = \begin{matrix} x_1 \\ x_2 \\ x_3 \\ z_1 \\ z_2 \\ z_3 \end{matrix} \begin{pmatrix} 1 & -0.49 & -0.14 & 0.9 & -0.51 & 0.05 \\ -0.49 & 1 & -0.05 & -0.51 & 0.96 & -0.17 \\ -0.14 & -0.05 & 1 & 0.004 & -0.04 & 0.83 \\ 0.9 & -0.51 & 0.004 & 1 & -0.45 & 0.28 \\ -0.51 & 0.96 & -0.04 & -0.45 & 1 & -0.11 \\ 0.05 & -0.17 & 0.83 & 0.28 & -0.11 & 1 \end{pmatrix}$$

And that the six variables are distributed into two equal groups, namely the set of variablesx_1,x_2, x_3 and the corresponding set of variables z_1, z_2, z_3

Generating contaminated data with $\omega = 10\%$, depending on this formula

$$(\mathbf{1} - \boldsymbol{\omega})\, \boldsymbol{N_p}\left(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right) + \boldsymbol{\omega}\, \boldsymbol{N_p}\left(\underline{\boldsymbol{\mu_j}}, \boldsymbol{\Sigma_j}\right)\,,\ \ \boldsymbol{j} = \mathbf{1}, \mathbf{2}, \mathbf{3}\,,\ \ \boldsymbol{\omega} \neq \mathbf{0}$$

Therefore, the data will be obtained according to the following Model:

Model II: $\underline{\mu_1} = \underline{\mu},\ \Sigma_1 = 1.5 * \Sigma$ Compared with Model I which is uncontaminated data with $\omega = 0\%$

We use two size samples in generating data , n= 30&60

After generating data, we estimate canonical correlation according two robust methods also estimate Eigen values and Eigen vectors.

Estimate empirical influence function for scaled and transformed (EIFST) estimators to canonical correlation and estimate (EIFST) for weighted canonical for both methods before and after replace the uncontaminated data with contaminated data.

Make a comparison between canonical correlation coefficient and estimated weighted canonical before and after outlier values, since the comparison mechanism based on maximum and minimum (IF) for robust methods.

After apply simulation, we note the following:

Table (1), the maximum value for (EIFST) was at second observation when $\boldsymbol{\omega = 0}\%$ and (Bi) method gave the least value of method (Pe), but at the Model II with $\boldsymbol{\omega = 10}\%$,the max.value for (EIFST) was at twenty eight obs. , since (Bi) method gave the least value of method (Pe).

Table 1: estimated EIFST for canonical correlation (CC) at $\boldsymbol{\omega = 0}\%$ & $\boldsymbol{10}\%$ when n= 30

	$\boldsymbol{\omega = 0}\%$						$\boldsymbol{\omega = 10}\%$				
Meth Obs.	Bi	Pe	Meth Obs.	Bi	Pe	Meth Obs.	Bi	Pe	Met Obs	Bi	Pe
1	0.1061	0.1067	16	0.149	0.1502	1	0.2468	0.2479	16	0.3178	0.3189
2	0.1617	0.1633	17	0.113	0.1135	2	0.2971	0.2981	17	0.2477	0.2487
3	0.1096	0.1102	18	0.144	0.1455	3	0.2514	0.2525	18	0.3056	0.3066
4	0.1470	0.1476	19	0.113	0.1141	4	0.3193	0.3204	19	0.2424	0.2434
5	0.1157	0.1162	20	0.141	0.1416	5	0.2486	0.2496	20	0.3163	0.3173
6	0.1354	0.1360	21	0.112	0.1129	6	0.2903	0.2914	21	0.2427	0.2438
7	0.1064	0.1069	22	0.141	0.1418	7	0.2303	0.2314	22	0.3115	0.3126
8	0.1380	0.1386	23	0.110	0.1109	8	0.3188	0.3199	23	0.2371	0.2382
9	0.1171	0.1177	24	0.147	0.1484	9	0.2404	0.2415	24	0.3029	0.304
10	0.1437	0.1443	25	0.113	0.1142	10	0.3155	0.3165	25	0.2468	0.2479
11	0.1108	0.1114	26	0.145	0.1458	11	0.2447	0.2458	26	0.3111	0.3122
12	0.1436	0.1442	27	0.112	0.1130	12	0.3008	0.3018	27	0.2405	0.2415
13	0.1142	0.1148	28	0.148	0.1488	13	0.2374	0.2385	28	0.3373	0.3393

14	0.1501	0.1507	29	0.104	0.1053	14	0.3257	0.3267	29	0.2481	0.2492
15	0.1168	0.1173	30	0.144	0.1446	15	0.2477	0.2487	30	0.3354	0.3365

Table (2), the maximum value for (EIFST) was at twenty two observation when $\omega = 0\%$ and (Bi) method gave the least value of method (Pe), but at the Model II with $\omega = 10\%$,the max. value for (EIFST) was at sixty obs. , since (Bi) method gave the least value of method (Pe).

Table 1: estimated EIFST for canonical correlation (CC) at $\omega = 0\%$ & 10% when n= 60

	$\omega = 0\%$						$\omega = 10\%$				
Meth Obs.	Bi	Pe	Meth Obs.	Bi	Pe	Meth Obs.	Bi	Pe	Meth Obs.	Bi	Pe
1	0.0743	0.0776	31	0.0744	0.0778	1	0.1738	0.1779	31	0.1635	0.1676
2	0.1036	0.107	32	0.1003	0.1036	2	0.2086	0.2128	32	0.2205	0.2246
3	0.0823	0.0857	33	0.0774	0.0808	3	0.168	0.1721	33	0.1672	0.1713
4	0.1012	0.1046	34	0.1045	0.1078	4	0.2075	0.2117	34	0.2145	0.2187
5	0.0773	0.0807	35	0.0792	0.0825	5	0.1671	0.1712	35	0.1586	0.1627
6	0.0979	0.1012	36	0.0992	0.1025	6	0.2145	0.2186	36	0.2147	0.2189
7	0.079	0.0823	37	0.0775	0.0808	7	0.1613	0.1654	37	0.1674	0.1715
8	0.1004	0.1037	38	0.1016	0.1049	8	0.2081	0.2122	38	0.2121	0.2162
9	0.0782	0.0815	39	0.082	0.0853	9	0.1637	0.1678	39	0.1695	0.1736
10	0.1012	0.1045	40	0.1088	0.1121	10	0.2165	0.2206	40	0.2153	0.2194
11	0.0777	0.0811	41	0.0808	0.0842	11	0.1644	0.1685	41	0.1633	0.1674
12	0.103	0.1064	42	0.0974	0.1007	12	0.2129	0.217	42	0.2083	0.2125
13	0.0801	0.0834	43	0.0815	0.0848	13	0.1689	0.173	43	0.1658	0.17
14	0.0984	0.1017	44	0.0974	0.1007	14	0.2141	0.2182	44	0.2076	0.2118
15	0.0771	0.0804	45	0.0793	0.0827	15	0.1628	0.1669	45	0.164	0.1681
16	0.0961	0.0995	46	0.1003	0.1037	16	0.2122	0.2163	46	0.2126	0.2167
17	0.0775	0.0808	47	0.0749	0.0782	17	0.1632	0.1673	47	0.165	0.1691
18	0.0989	0.1022	48	0.1013	0.1047	18	0.2109	0.215	48	0.2148	0.219
19	0.0744	0.0778	49	0.0775	0.0808	19	0.1613	0.1654	49	0.1648	0.1689
20	0.0988	0.1022	50	0.1023	0.1056	20	0.2151	0.2192	50	0.218	0.2221
21	0.0802	0.0835	51	0.0832	0.0865	21	0.1661	0.1702	51	0.1679	0.172
22	0.1092	0.1125	52	0.0973	0.1006	22	0.2204	0.2245	52	0.2114	0.2155
23	0.0709	0.0743	53	0.0763	0.0796	23	0.1648	0.1689	53	0.1677	0.1718
24	0.09	0.0934	54	0.0999	0.1032	24	0.2106	0.2147	54	0.2062	0.2103
25	0.0739	0.0773	55	0.0736	0.077	25	0.1614	0.1656	55	0.1874	0.1915
26	0.0987	0.1021	56	0.0999	0.1032	26	0.2115	0.2156	56	0.2445	0.2486
27	0.0821	0.0855	57	0.0785	0.0819	27	0.1638	0.1679	57	0.1887	0.1928
28	0.1084	0.1118	58	0.0999	0.1032	28	0.2143	0.2184	58	0.2387	0.2429
29	0.0748	0.0782	59	0.0789	0.0822	29	0.1701	0.1742	59	0.1891	0.1932
30	0.0949	0.0983	60	0.098	0.1014	30	0.214	0.2182	60	0.2530	0.2571

We note from table 3 & 4 that estimated Eigen value and CC are so closed in their values and unstable with respect to sample sizes and the largest values for Eigen and CC that is estimated by (Bi) followed by (Pe).also we note that the differences are not clear except in the case of uncontaminated data, as it is less than its values in the case of contaminated data.

Table 3: Eigen values for (Bi) & (Pe) methods

Model	ω	n	Bi	Pe

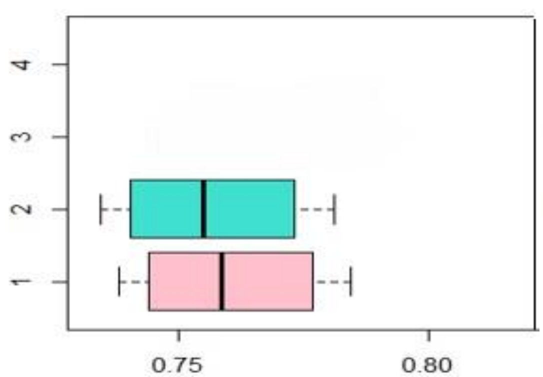
I	0%	30	0.9131	0.8469	0.5448	0.9130	0.8500	0.5525
		60	0.9160	0.8599	0.5573	0.9091	0.8512	0.5533
II	10%	30	0.9167	0.8493	0.5492	0.9160	0.8522	0.5545
		60	0.9170	0.8596	0.5585	0.9102	0.8512	0.5542

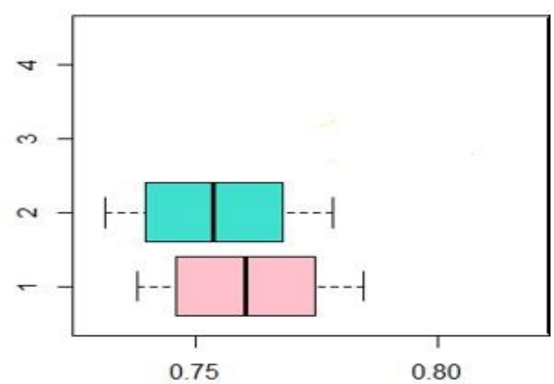
Table 4: CC for (B) & (P) methods

Model	ω	n	Bi	Pe
I	0%	30	0.9556	0.9550
		60	0.9571	0.9534
II	10%	30	0.9574	0.9571
		60	0.9576	0.9540

The box diagram was also used to analyze the effect of observations in estimating the weights vectors corresponding to the coefficient CC of contaminated and uncontaminated data. The (IF) of weights vectors (a) and (b) were estimated for two models and two estimation methods, contamination ratios, and different sample size n= 30&60 used in simulation experiments.

Figures 1, 2, 3&4 show estimated EIFST for (a) & (b) vectors, when uncontaminated data, we note that the values of EIFST increase at n=60 and became the highest at n=30. Also, a method (Bi) has surpassed a method (Pe) based on the lowest values of the (IF),noting that the values of (IF) for vector (b) are slightly higher than the values of the (IF) for vector (a)





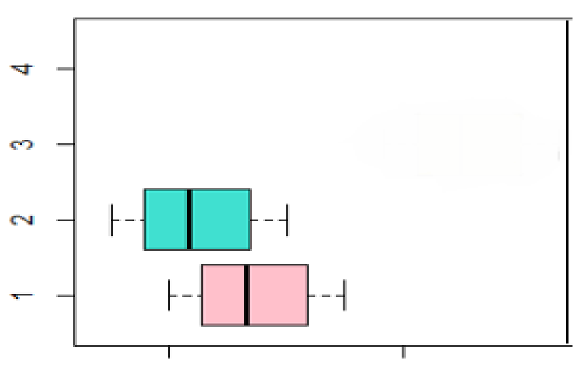


Figure 3: Model I: EIFST for vector (a), n=60

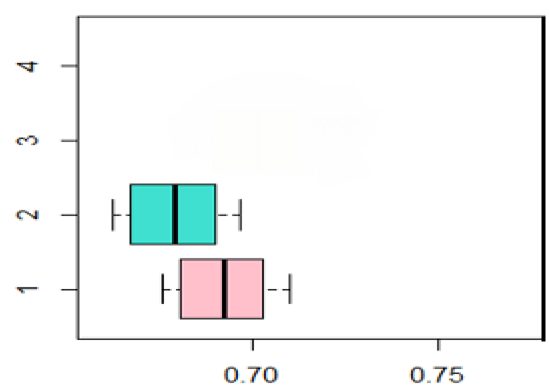


Figure 4: Model I: EIFST for vector (b), n=60

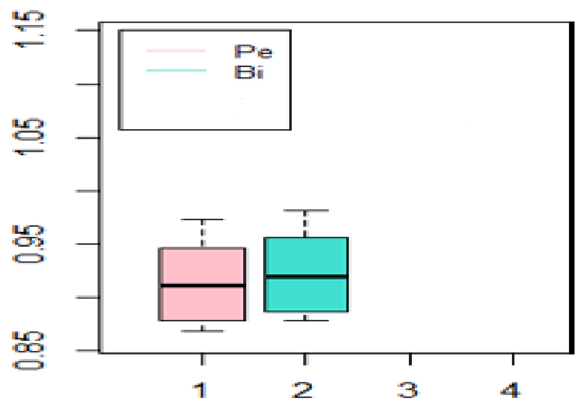


Figure 5: Model II: EIFST for vector (a), n=30

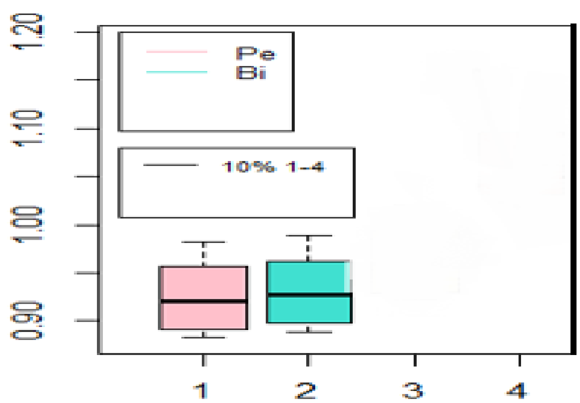


Figure 6: Model II: EIFST for vector (b), n=30

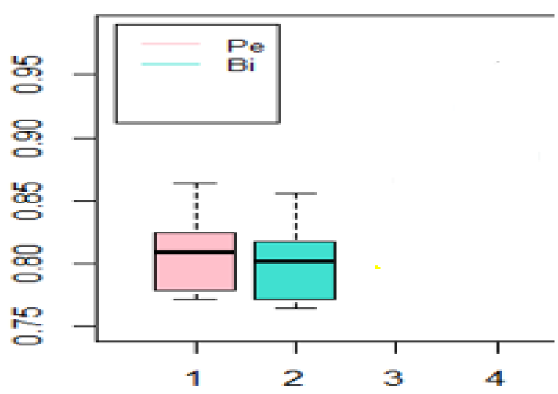


Figure 7: Model II: EIFST for vector (a), n=60

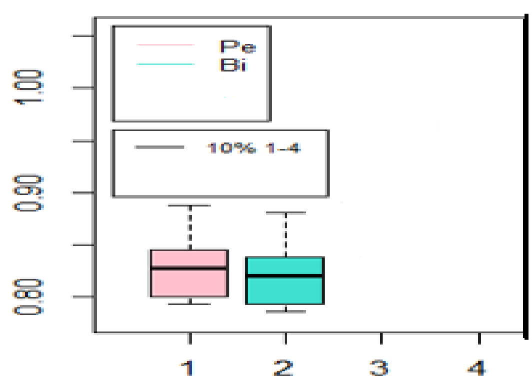


Figure 8: Model II: EIFST for vector (b), n=60

The figures 5, 6, 7 & 8 above show that (Bi) method was better than method (Pe) , also there was a simple difference between vectors (a) & (b) in their values

Case Study

Our study based on real data consist of two variables groups, first one includes monthly quantities of oil exported for three oil-producing countries within OPEC (Saudi x_1 , Iraq x_2, Kuwait x_3) Recorded for a period of sixty months in the years starting at January 2015 , the second set are(z_1 ،z_2 ،z_3) represents returns for those quantities .
Estimating Canonical Correlation Eigen Value

Table below shows that the result for CC estimated by (Bi) method was (0.9501) at Contaminated data and (0.9755) for uncontaminated data, also there were a differences between weights vectors $\hat{a}$ & $\hat{b}$ at two cases.

Table 6: Eigen's and weights Vectors for CC by using (Bi) method for contaminated and uncontaminated data.

	Contaminated data			Uncontaminated data		
Eigenvalues	0.9028	0.8185	0.7602	0.9517	0.8302	0.5909
$\hat{\mathbf{a}}$	-0.0988	0.7082	-0.6710	0.5146	-0.5012	0.6802
$\hat{\mathbf{b}}$	-0.9926	0.1473	0.1848	-0.9482	0.0831	-0.5083

14- Estimation of Influence Function

After finding empirical influence function according to scaled and transformed estimator, it is possible to explain the influence of the studied data observations on the CC between the variables of two sets.
Table 7&8 below, show that the highest value of the influence function was (0.7188), which is return to observation no. (56), while the lowest value of the influence function was the value return to observation no. (39) and reached (0.0766), the highest value of the influence function estimator for CC By using (Bi) method after replacing the contaminated observations, it reached (0.4027) when replacing the observation (34), meaning that observation no. (34) is highest influence in CC estimation, while the lowest value of the influence function was (0.0039) when replacing observation (27), this means that the influence of observation (27) is very poor on the estimated values of CC, as well,the values of the estimated influence function in the case of contaminated data are greater than values if the contaminated observations are excluded and replaced with uncontaminated values.

Table 7: IF of CC for contaminated data

Obs	EIFST	obs	EIFST	obs	EIFST	obs	EIFST
1	0.133	16	0.0148	31	0.0196	46	0.2043
2	0.0797	17	0.0072	32	0.0164	47	0.049
3	0.0254	18	0.0124	33	0.0011	48	0.0665
4	0.0126	19	0.0618	34	0.4027	49	0.0105
5	0.0623	20	0.0798	35	0.0032	50	0.0964
6	0.0168	21	0.0048	36	0.3808	51	0.0055
7	0.0768	22	0.3808	37	0.0053	52	0.0092
8	0.0191	23	0.0004	38	0.1862	53	0.0623
9	0.2166	24	0.1043	39	0.0309	54	0.0012
10	0.0151	25	0.0042	40	0.0352	55	0.0102
11	0.0623	26	0.0389	41	0.0201	56	0.0115
12	0.0301	27	0.0039	42	0.1655	57	0.0213
13	0.0124	28	0.0221	43	0.029	58	0.017
14	0.0123	29	0.0044	44	0.0993	59	0.0077
15	0.0623	30	0.3808	45	0.0623	60	0.3808

Table 8: IF of CC after replace contaminated observations

obs	EIFST	obs	EIFST	obs	EIFST	obs	EIFST
1	0.1807	16	0.0826	31	0.1723	46	0.3616
2	0.0956	17	0.0777	32	0.0985	47	0.1874
3	0.118	18	0.0766	33	0.1036	48	0.2136
4	0.3042	19	0.0784	34	0.1493	49	0.2825
5	0.0875	20	0.1394	35	0.0776	50	0.1443
6	0.0853	21	0.0783	36	0.0884	51	0.0796
7	0.0904	22	0.079	37	0.0937	52	0.3893
8	0.1026	23	0.1044	38	0.1837	53	0.1723
9	0.0774	24	0.2706	39	0.0766	54	0.094
10	0.1211	25	0.077	40	0.1376	55	0.135
11	0.0924	26	0.0862	41	0.0769	56	0.7188

12	0.3856	27	0.113	42	0.1902	57	0.1113
13	0.2218	28	0.2238	43	0.0766	58	0.0766
14	0.0812	29	0.1178	44	0.1896	59	0.0888
15	0.1007	30	0.0925	45	0.1542	60	0.1019

Conclusions

Empirical influence function (EIFST) is an important standard to clarify the effect of each observation for data that we studied, as well as its determining the influence of outliers in estimation of canonical correlation coefficient and weights vectors in case of contaminated and uncontaminated data.
(EIFST) values increase as the sample size decreases.
Robust estimation methods showed a high convergence at CC estimation and of CC coefficient (EIFST).
Robust methods are efficient in estimating CC coefficient in case of data contamination. The values of (EIFST) are close in case of contaminated distribution and uncontaminated data, (Bi) method are less affected by contaminated distribution than (Pe) method.
Variables of quantities for exported oil and returns obtained from them for three oil-producing countries within OPEC organization, Saudi, Iraq and Kuwait, follow the contaminated natural distribution, the nature of the relationship between quantities of exported oil and the corresponding returns is strong,
CC estimated by (Bi) method between the quantities of exported oil and the oil returns of the three countries reached (0.9501) before replacing the contaminated observations, while CC estimated in the same way after replacing the contaminated observations reached (0.9755), and this indicates to strong relationship between two sets.

References

1. Al-Rawi, Ziad R, (2017) “Methods of multivariate statistical analysis” Hashemite Kingdom of Jordan ,Arab Institute for Training and Statistical Research, PP(8-7).
2. Maronna, R., Martin, R., Yohai, V., and Salibián-B. M., (2019) Robust Statistics: Theory and Methods (with R), Second Edition
3. Romanazzi, M, (1992) "Influence Function in Canonical Correlation Analysis" Pychometrika, Vol.57, No.2.
4. Nasser, M. and Mesbahul, A. Md, (2006) "Estimators of Influence Function" Communications in Statistics—Theory and Methods, 35: 21–32.
5. Ali Alkenani & Keming Yu, (2013) "A comparative study for robust canonical correlation methods "Journal of Statistical Computation and Simulation, 83:4, 692-720.
6. Veenstra, P. , Cooper, C. & Phelps, S. ,(2016)" The use of Biweight Mid Correlation to improve graph based portfolio construction " Computer Science and Electronic Engineering Conference, CEEC 2016 - Conference Proceedings (pp. 101-106).
7. Al-Ali, Ibrahim M., (2020) “Foundations of multivariate statistical analysis" Syria, Teshreen University – College of Economics, PP 378.
8. Rand R. Wilcox (1994),"The percentage bend correlation coefficient" Psychometrika , Vol. 59; Iss. 4.
9. Wilcox RR. , (2013) "Introduction to Robust Estimation and Hypothesis Testing. "3rd edition, A volume in Statistical Modeling and Decision Science
10. F. Hampel, E. Ronchetti, P. Rousseeuw, W. Stahel, (2011) Robust statistics: The approach based on influence functions, Wiley