



## A Comparison between robust methods in canonical correlation by using empirical influence function

Zahraa Khaleel Al-Ameed [1]  
Ministry of Oil, Falastin St., Baghdad

Prof. dr. Lekaa Ali Al-Alway [2]  
University Of Baghdad, Adhamiya, Baghdad, Iraq

[Zahraa.khaleel1101a@coadec.uobaghdad.edu.iq](mailto:Zahraa.khaleel1101a@coadec.uobaghdad.edu.iq)  
[Zahraaeyes84@gmail.com](mailto:Zahraaeyes84@gmail.com)

[Lekaa.ali.1968@gmail.com](mailto:Lekaa.ali.1968@gmail.com)

Received: 29/3/2021

Accepted: 14/4/2021

Published: June/ 2021



This work is licensed under a [Creative Commons Attribution-Non Commercial 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

### Abstract

Canonical correlation analysis is one of the common methods for analyzing data and know the relationship between two sets of variables under study, as it depends on the process of analyzing the variance matrix or the correlation matrix. Researchers resort to the use of many methods to estimate canonical correlation (CC); some are biased for outliers, and others are resistant to those values; in addition, there are standards that check the efficiency of estimation methods.

In our research, we dealt with robust estimation methods that depend on the correlation matrix in the analysis process to obtain a robust canonical correlation coefficient, which is the method of Biweight Midcorrelation coefficient (Bi) and Kendall-tau correlation coefficient (Ke).

From the comparison between these two methods through the empirical influence function with standard scaled and transformed estimator, the results indicated the efficiency and the preference of the (Bi) method. The study also has application with real data followed a multivariate normal distribution with two sets; the first group represents monthly averages for quantities of exported oil from three OPEC countries, namely Saudi Arabia, Iraq, and Kuwait, the other group represents the returns of those quantities for the period from 2015 to 2019, after applied (Bi) method and estimate IF, the strongest influence about CC was at thirty four-months and the lowest was at twenty-seven.

**Keywords:** Canonical Correlation, Outliers, Robust methods, Biweight Midcorrelation Coefficient, Kendall Tau Coefficient, Influence Function.

## 1- Introduction

The canonical correlation coefficient is the generalization of multiple correlations as it consists of two sets of variables, the first are dependent variables  $(Y_1, Y_2, \dots, Y_p)$  and the second is explanatory variables  $(X_1, X_2, \dots, X_q)$  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 saying that the concept of the canonical correlation appeared in the period 1935/1936 by the scientist (Hotelling), and it became clear that the multiple correlations are 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]

In (1992), the scientist (Mario Romanazzi) presented the derivation of the influence function for the square of the correct and multiple correlation coefficients, 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 Eigenvalues and Eigenvectors and the characteristic values, depending on the study of (Hampe 1974) in the early seventies [2]. In (2000), the researcher (Abd-Aljabar Anaam) a comprehensive study of the influence function in the canonical correlation analysis and the study of all its characteristics and all the robust measures derived from it, in the case of the one dimension and the multidimensional also the relationship of this function with the (Jackknife) variables, and then using the M estimator to estimate the parameters using real data for the variables of two groups of students' grades in the subjects. It has been shown. The extent of sensitivity of the influence functions to canonical vectors associated with the canonical correlation of the robust data, in addition, values of the influence function providing the researcher with basic information about the observations that contaminate data [3]. Researcher (Aziz Thaka) introduced In (2012), a study about the relationship between knowledge management processes and administrative corruption by relying on a questionnaire to know the viewpoints for professors from the College of Administration and Economics by comparing classical canonical with robust canonical correlation by using (MSE) scale, the results indicated that the robust canonical correlation is better than the classical correction correlation [4]. While the researchers (Alkenani & Keming) represented at (2013) two types of Estimators divided into two groups (M-estimators) which include (Percentage Bend, Biweight mild correlation, Winsorized, 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 (2019), the researchers (Yipeng, Zhilong, Kui & Xudong) represented fractional theory due to its rapid development and intersection with other sciences and disciplines such as internet networks, gene interaction networks, and urban infrastructure. Researchers used (Kendall-Tau) correlation coefficient as one of

the robust test indicators. The fractional dimension was calculated separately after assigning the ranks to the ability of the networks based on the required indicators and then calculating the correlation coefficients between the ordered sequences, which led to a strong relationship between the fractional dimensions and the strong indicators of the network's complex. [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. The 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 the empirical influence function of scaled and transformed estimators 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.

## 2- 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.

Assuming the study of two sets of variables:

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

$Y_{q \times 1}$  is a vector with dimension  $p \times 1$  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 multivariate normal distribution as each group has the following specifications:

$$E(x) = \mu_x$$

$$E(y) = \mu_y$$

$$\text{Var}(y) = \Sigma_{yy}$$

$$\text{Var}(x) = \Sigma_{xx}$$

And the homogeneity matrix between the two sets known as:

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

$\Sigma_{xx} > 0$ ,  $\Sigma_{yy} > 0$  And assume  $p \leq q$ , so we can define a number of linear combination equal to the number of  $\text{Min}(p, q)$  by using this equation:

$$\begin{aligned} u_i &= \bar{a}_i \underline{x} \\ &= a_{1i}x_1 + a_{2i}x_2 + \dots + a_{pi}x_p \quad i = 1, 2, \dots, n \quad \text{..(1)} \end{aligned}$$

$$\begin{aligned} v_i &= \bar{b}_i \underline{y} \\ &= b_{1i}y_1 + b_{2i}y_2 + \dots + b_{qi}y_q \quad i = 1, 2, \dots, n \quad \text{..(2)} \end{aligned}$$

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  $U_i$  or  $V_i$

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

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

$$\text{Var} (\underline{\hat{a}} \underline{x}) = \underline{\hat{a}} \underline{\Sigma}_{xx} \underline{\hat{a}} = 1 \dots \dots \dots (3)$$

$$\text{Var} (\underline{\hat{b}} \underline{y}) = \underline{\hat{b}} \underline{\Sigma}_{yy} \underline{\hat{b}} = 1 \dots \dots \dots (4)$$

$$\underline{\hat{a}} \underline{\Sigma}_{xx} \underline{\hat{a}} = \underline{\hat{b}} \underline{\Sigma}_{yy} \underline{\hat{b}} = 1 \dots \dots \dots (5)$$

And the cover between linear combination

$$\text{Cov} (\underline{\hat{a}} \underline{x}, \underline{\hat{b}} \underline{y}) = \underline{\hat{a}} \underline{\Sigma}_{xy} \underline{\hat{b}} \dots \dots \dots (6)$$

So the correlation is :

$$\text{Corr}(\underline{\hat{a}} \underline{x}, \underline{\hat{b}} \underline{y}) = \frac{\underline{\hat{a}} \underline{\Sigma}_{xy} \underline{\hat{b}}}{\sqrt{\underline{\hat{a}} \underline{\Sigma}_{xx} \underline{\hat{a}}} \sqrt{\underline{\hat{b}} \underline{\Sigma}_{yy} \underline{\hat{b}}}} \dots (7)$$

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

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

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

$$\rho_{(u_1, v_1)} = \max_{\underline{a}, \underline{b}} (\underline{\hat{a}} \underline{x}, \underline{\hat{b}} \underline{y}) \dots \dots \dots (8)$$

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

The second pair of variables  $(u_1, v_1)$  is 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  $(u_1, v_1)$ , meaning that

$$\text{Cov} (\underline{\hat{a}} \underline{x}, u_1) = 0 \dots \dots \dots (9)$$

$$\text{Cov} (\underline{\hat{b}} \underline{y}, v_1) = 0 \dots \dots \dots (10)$$

$$\text{Var} (\underline{\hat{b}} \underline{y}) = \text{Var} (\underline{\hat{a}} \underline{x}) = 1 \dots \dots (11)$$

Maximizing the correlation between  $\underline{\hat{b}}_2 \underline{y}$  and  $\underline{\hat{a}}_2 \underline{x}$  Is called the second canonical correlation coefficient, and generally, the pair  $(U_j, V_j)$  of the canonical variables are chosen to maximize the heterogeneity of Cov  $(u_1, v_1)$  and the condition that the linear compositions of pair J are perpendicular to the previous pair  $[(j-1)]^\wedge$  of compositions.

Thus, the coefficients of correlation in the significance of the variables and variance are estimated in the relationship

$$r_c = \frac{\hat{u} S_{xy} v}{\sqrt{\hat{u} S_{xx} u} \sqrt{\hat{v} S_{yy} v}} \dots \dots \dots (12)$$

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 variable.

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

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

$$r_c = \frac{\hat{C}R_{xy}D}{\sqrt{\hat{C}R_{xx}\hat{C}}\sqrt{\hat{D}R_{yy}\hat{D}}} \dots\dots\dots (13)$$

Since:

**C&D:** is the canonical variable that is chosen to maximize heterogeneity.

To estimate canonical weight, which maximizes canonical correlation, the function:

$$g = \hat{C}R_{xy}D - \frac{\sqrt{\lambda_1}}{2} CR_{xx}C - \frac{\sqrt{\lambda_2}}{2} \hat{D}R_{yy}\hat{D} \dots\dots\dots (14)$$

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

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

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

$$\frac{\partial g}{\partial d} = \hat{c}R_{xy} - \sqrt{\lambda_2}\hat{d}R_{yy} \dots\dots\dots (16)$$

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

$$\underline{c} = \frac{1}{\sqrt{\lambda_1}} R_{xx}^{-1} R_{xy} \underline{d} \dots\dots\dots (17)$$

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 roots  $\lambda_i$  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:

$$r_c^2 = \sqrt{\lambda} \quad . [7, \text{pp.378}]$$

### 3- Concepts of Robust Statistics

The robust methods of estimation came as an alternative to the traditional classical methods because of their different dealings with data, wherein some cases and through the process of statistical analysis there appears a percentage of observations deviate from the assumptions of the basic model, which is called atypical values which cause negative effects on the results of the analysis. [8, pp.1,6] The symmetric robust of Pearson correlation coefficient has been divided into two parts, the first type provides protection against abnormal values without taking into account the general structure of the data, and the second section takes into consideration the general structure when the extreme values are present and called (O-correlation), and therefore the first type will be used in this research. [10, pp.446]

#### 4- Mahalanobis Distance (MD)

The presence of samples of multivariate on the uncontaminated observation that results due to sampling errors or in the case of data recording is more complicated in the case of one variable. In order to test the presence of these observations, we resort to the use of classic Mahalanobis squared distances:

$$d_i^2 = (\underline{X}i - \bar{X})' S^{-1} (\underline{X}i - \bar{X}) \dots\dots(18)$$

Then eliminate it and apply estimates to the good observations. Also, we can use Robust Mahalanobis squared distances:

$$Rd_i^2 = (\underline{X}i - \bar{\mu}_n)' \Sigma_n^{-1} (\underline{X}i - \bar{\mu}_n) \dots\dots(19)$$

It is calculated for each observation and then compared with the table value of  $\chi^2_{(p,0.05)}$ . If the calculated distance for the observation is greater than the table value, then that observation is considered contaminated. [9, pp. 17]

#### 5- 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  $\psi$  an odd function,  $\mu_x$  &  $\mu_y$  location standard for random variable X, Y straightly and let  $\tau_y$  &  $\tau_x$  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:

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

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

$$\gamma_{xy} = \frac{nk^2 \cdot \tau_x \cdot \tau_y \cdot E(\psi(u) \cdot \psi(v))}{E(\psi(u)) \cdot E(\psi(v))} \dots\dots\dots (20)$$

Since correlation scale  $\rho_b$  calculate as:

$$\rho_b = \frac{\gamma_{xy}}{\sqrt{\gamma_{xx} \cdot \gamma_{yy}}} \quad -1 \leq \rho_b \leq 1 \quad \dots\dots\dots (21)$$

By choosing  $K = 9$  (previous studies have shown David (1985) that choosing this value makes (Bicov) significantly more efficient dispersion is estimated than the variance [11, pp. 7]) and the function, which represents the weight function, which is known as the following relationship:

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

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

$$U_i = \frac{(X_i - med_x)}{9 \cdot MAD_x}, \quad V_i = \frac{(Y_i - med_y)}{9 \cdot MAD_y} \quad [14, pp. 16]$$

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:

$$MAD_x = med_i |x - med_{xi}| = med |x - med_x| \quad [11, pp. 119]$$

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 Midcovariance between X & Y:

$$Bicov(x, y) = \frac{n \sum a_i (X_i - med_x) (1 - U_i^2)^2 b_i (Y_i - med_y) (1 - V_i^2)^2}{[\sum a_i (1 - U_i^2) (1 - 5U_i^2)] [\sum b_i (1 - V_i^2) (1 - 5V_i^2)]} \dots\dots\dots (22)$$

After applying the correlation formula, the estimation Biweight mild correlation:

$$r_{bi} = \frac{bicov(x, y)}{\sqrt{bicov(x, x) \cdot bicov(y, y)}} \dots\dots\dots (23)$$

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:

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

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  $\alpha$ .

[5, pp. 696-697]

**6- Kendall's Tau Correlation (Ke)**

Kendall-tau correlation coefficient is a non-parametric correlation (M-type), which is suitable for dealing with quantitative and qualitative data; it is defined as the difference between the probabilities that the pair of observations is concordant minus the probability that the pair of observations is discordant. [5, pp. 698]

Let the pairs  $(X_i, Y_i), i = 1, 2, \dots, n$ , so the rank for  $(\tau)$  defined as [16, pp. 5]

$$\hat{\tau} = \frac{1}{n(n-1)} \sum_{i \neq j} \text{sign}(X_i - X_j) \text{sgn}(Y_i - Y_j) \dots\dots\dots (24)$$

If there are n of data pairs  $(X_i, Y_i), i = 1, 2, \dots, n$ , we can describe  $(\tau)$  as:

$$\hat{\tau} = \frac{2 \sum_{i < j} S_{ij}}{n(n-1)} \dots\dots\dots (25)$$

Since:

$S_{ij}$  Has the following values through arranging observations rank:

$$S_{ij} = \begin{cases} 1 & \text{if } j \text{ (concordant)} \\ -1 & \text{if } j \text{ (discordant)} \end{cases}$$

If  $|Z| < Z_{1-\alpha/2}$ , the decision is to reject  $H_0$

Since Z can calculate from this formula:

$$Z = \frac{\hat{\tau}}{\sigma_{\tau}} \quad \text{And the value of } \sigma_{\tau}:$$

$$\sigma_{\tau}^2 = \frac{2(2n+5)}{9n(n-1)} \dots\dots\dots (26)$$

So, the value of Z:

$$Z = \frac{6 \sum_{i < j} S_{ij}}{\sqrt{2n(n-1)(2n+5)}} \dots\dots\dots (27)$$

For the comparison between the estimations of the CC coefficients calculated according to the Kendall-tau formula with (Bi) computed by other correlation methods, the Sin ( $\pi / 2$ ) transformation is applied in order to obtain consistent estimates under the conditions of a normal distribution.

[13 pp.153]

Then, we calculate the coefficient for two sets of variables to obtain this matrix:

$$R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$$

### 7- Measurement of Robust Estimators

Robust estimator measures are important tools for describing and knowing the behavior of estimators, including the influence function (IF).

This function provides us with the extent to which the estimate is affected when the basic statistical model malfunctions, where IF value is useful when examining the robustness and strongest of the estimator also the model, as well as calculating covariance and heterogeneity matrix for certain types of estimators.

The concept of (IF) back to the contributions of (Hampel) and his colleagues (1974), where they called the term influence curve and thus developed a new approach to obtaining infused statistics. [14, pp. 40-41]

### 8- Influence Function (IF)

The IF basically consider an analytic tool can use to evaluate the effect of observation on the estimator  $T_n$  at distribution function F by:

$$IF_{T_n, F(x)} = \lim_{\omega \rightarrow 0} \frac{[T_n(F_\omega) - T_n(F)]}{\omega} \dots\dots\dots (28)$$

Since:

F: cumulative distribution function

$T_n(F)$ : function to F called mathematically (functional).

$IF_{T_n, F(x)}$  : Derivative for  $T_n(F)$  which is represent IF.

And:

$$F_\omega = (1 - \omega)F + \omega\delta_x \dots\dots\dots(29)$$

Since:

$\omega$  : Contaminated ratio  $0 < \omega < 1$

$\delta_x$  : Probability scale

$$\delta_{x_0}(x) = \begin{cases} 1 & \text{for } x \geq x_0 \\ 0 & \text{for } x < x_0 = CDF \text{ for a point mass 1 at } x_0 \end{cases}$$

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 Estimators of the influence function, which work the same as the IF: [5, pp.55]

### 9- Unscaled and Untransformed Estimators

Let  $x_1, x_2, \dots, x_{n-1}$  a sample consist of n-1 size, we can define the estimator:

$$\perp_{(x)}^1 = T_n(x_1, x_2, \dots, x_{n-1}, x) \dots\dots\dots (30)$$

Since  $\perp_{(x)}^1$  represent the estimator values when we add a new observation x to the sample, and it called addition – corruption.

Or we assume  $x_1, x_2, \dots, x_n$  a sample with n size, so we can define the estimator:



$$\perp_{(x)}^2 = T_n(x_1, x_2, \dots, x_n, x) \dots\dots\dots (31)$$

Since  $\perp_{(x)}^1$  represent the estimator value after we replacing the observation  $x_n$  with observation  $x$ , and it called replacement – corruption. [15, pp.22]

**10- Unscaled and Transformed**

The following estimator consider translated estimator from  $\perp_{(x)}^2$  &  $\perp_{(x)}^1$

$$\perp_{(x)}^3 = T(x, x_1, x_2, \dots, x_n) - T(x_1, x_2, \dots, x_n) \dots\dots\dots (32)$$

And the estimator

$$\perp_{(x)}^4 = \sum_{i=1}^n [T(x \text{ replacing } x_j) - T(x_1, x_2, \dots, x_n)] \dots\dots\dots (33) [15, pp.23]$$

**11- Biased and Transformed Estimators.**

Empirical influence function defined as depending on the previous estimators with this formula:

$$\begin{aligned} \text{EIF}(x, F_n) &= \text{IF}(x, F_n) \\ &= \lim_{\omega \rightarrow 0} \frac{T(F_n + \omega(\delta_x - F_n)) - T(F_n)}{\omega} \dots\dots\dots (33) \end{aligned}$$

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  $T(F_n)$  represents the original estimator resulting from the original distribution function  $F_n$  of sample size (n).

It is better to estimate the empirical effect function (influence function) in relation to:

$$\begin{aligned} \text{EIF}_e(x, F_n) &= \text{IF}_e(x, F_n) \\ &= \frac{T\left(F_n + \frac{1}{100n}(\delta_x - F_n)\right) - T(F_n)}{\frac{1}{100n}} \dots\dots\dots (34) \end{aligned}$$

Since:

$\frac{1}{100n}$  : represent the ratio that is taken to contaminate data.

From this, the empirical influence function can define as:

$$I_j = \text{IF}_{(x_j)} = \text{EIF}(x_j, F_n) \dots\dots\dots (35)$$

Which can be rounded by choosing different values to  $\omega$  (contamination data) as  $(\frac{1}{n}, \frac{1}{\sqrt{n}}, \frac{1}{n+1}, \frac{1}{n-1})$  and other values without taking the limit for the amount. [11, pp.23, 331]

## 12- Simulation

The 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  $\mu$  and covariance matrix  $\Sigma$  for real data ( Oil Exports and Returns), as well as generating multivariate contaminant normal distribution tracking data by employing mean vectors, covariance matrices, and different contamination ratios, The canonical correlation coefficients were also estimated according to these methods: Biweight Midcorrelation and Kendall- Tau's correlation, then comparison was made between these robust methods based on the experimental effect function standard with the scaled and transformed estimators.

### 13- Steps of Simulation:

a- Generating six variables following the multivariate normal distribution  $N_p(\underline{\mu}, \Sigma)$  which are on the order  $x_1, x_2, x_3, z_1, z_2, z_3$  depending on the mean vector  $\mu$  and the CV matrix  $\Sigma$  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:

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

And that the six variables are distributed into two equal groups, namely the set of variables  $x_1, x_2, x_3$  and the corresponding set of variables  $z_1, z_2, z_3$

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

$$(1 - \omega) N_p(\underline{\mu}, \Sigma) + \omega N_p(\underline{\mu}_j, \Sigma_j), \quad j = 1, 2, 3, \quad \omega \neq 0$$

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

Model II:  $\mu_2 = \mu + I_1, \Sigma_2 = 0.9 * \Sigma$  Compared with Model I, which is uncontaminated data with  $\omega = 0\%$

Since,  $I_1 = 1$ , and use two size samples in generating data,  $n = 30 \& 60$

c- After generating data, we estimate canonical correlation according to two robust methods also estimate Eigenvalues and Eigenvectors.

d- 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 replacing the uncontaminated data with contaminated data.

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

From simulation results, we note the following:

1- Table (1), the maximum value for (EIFST) was at second observation when  $\omega = 0\%$  and (Bi) method gave the least value of the method (Ke), but at the

Model II with  $\omega = 10\%$ , the max. Value for (EIFST) was at twenty-two obs. since the (Bi) method gave the least value of the method (Ke).

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

$\omega = 0\%$						$\omega = 10\%$					
Met	Bi	Ke	Met	Bi	Ke	Met	Bi	Ke	Met	Bi	Ke
Obs.			Obs.			Obs.			Obs.		
1	0.1061	0.1396	16	0.149	0.1831	1	0.2538	0.2937	16	0.2963	0.3362
2	0.1617	0.1952	17	0.113	0.1465	2	0.3151	0.355	17	0.2489	0.2888
3	0.1096	0.1431	18	0.144	0.1784	3	0.2488	0.2887	18	0.2943	0.3342
4	0.1470	0.1805	19	0.113	0.147	4	0.3015	0.3415	19	0.2509	0.2908
5	0.1157	0.1492	20	0.141	0.1745	5	0.2507	0.2906	20	0.3214	0.3613
6	0.1354	0.1689	21	0.112	0.1458	6	0.2995	0.3395	21	0.2505	0.2904
7	0.1064	0.1399	22	0.141	0.1748	7	0.2454	0.2854	22	0.3229	0.3628
8	0.1380	0.1715	23	0.110	0.1438	8	0.3104	0.3503	23	0.2514	0.2913
9	0.1171	0.1506	24	0.147	0.1813	9	0.2463	0.2862	24	0.3137	0.3536
10	0.1437	0.1772	25	0.113	0.1472	10	0.3135	0.3535	25	0.2466	0.2865
11	0.1108	0.1443	26	0.145	0.1788	11	0.2568	0.2967	26	0.3054	0.3453
12	0.1436	0.1771	27	0.112	0.146	12	0.3172	0.3571	27	0.2431	0.283
13	0.1142	0.1477	28	0.148	0.1817	13	0.256	0.2959	28	0.2672	0.3071
14	0.1501	0.1836	29	0.104	0.1382	14	0.3223	0.3623	29	0.2231	0.263
15	0.1168	0.1503	30	0.144	0.1776	15	0.2365	0.2764	30	0.2878	0.3277

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

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

$\omega = 0\%$						$\omega = 10\%$					
Meth	Bi	Ke	Meth	Bi	Ke	Meth	Bi	Ke	Meth	Bi	Ke
Obs.			Obs.			Obs.			Obs.		
1	0.0743	0.1138	31	0.0744	0.114	1	0.1695	0.2129	31	0.1742	0.2176
2	0.1036	0.1431	32	0.1003	0.1398	2	0.2122	0.2556	32	0.214	0.2574
3	0.0823	0.1218	33	0.0774	0.1169	3	0.1722	0.2156	33	0.17	0.2134
4	0.1012	0.1408	34	0.1045	0.144	4	0.2144	0.2578	34	0.214	0.2573
5	0.0773	0.1169	35	0.0792	0.1187	5	0.1717	0.2151	35	0.1691	0.2125
6	0.0979	0.1374	36	0.0992	0.1387	6	0.2181	0.2615	36	0.2237	0.2671
7	0.079	0.1185	37	0.0775	0.117	7	0.17	0.2134	37	0.1709	0.2142
8	0.1004	0.1399	38	0.1016	0.1411	8	0.2283	0.2717	38	0.2195	0.2629
9	0.0782	0.1177	39	0.082	0.1215	9	0.1696	0.213	39	0.1736	0.217
10	0.1012	0.1407	40	0.1088	0.1483	10	0.2112	0.2546	40	0.2153	0.2587
11	0.0777	0.1173	41	0.0808	0.1203	11	0.1742	0.2176	41	0.1646	0.208
12	0.103	0.1426	42	0.0974	0.1369	12	0.2017	0.2451	42	0.2132	0.2566
13	0.0801	0.1196	43	0.0815	0.121	13	0.1774	0.2208	43	0.1676	0.211
14	0.0984	0.1379	44	0.0974	0.1369	14	0.2176	0.261	44	0.2053	0.2487
15	0.0771	0.1166	45	0.0793	0.1189	15	0.1696	0.213	45	0.1642	0.2076
16	0.0961	0.1356	46	0.1003	0.1399	16	0.2084	0.2518	46	0.2079	0.2513
17	0.0775	0.117	47	0.0749	0.1144	17	0.1735	0.2169	47	0.1726	0.216
18	0.0989	0.1384	48	0.1013	0.1409	18	0.2155	0.2589	48	0.2202	0.2636
19	0.0744	0.114	49	0.0775	0.117	19	0.1686	0.212	49	0.1692	0.2126
20	0.0988	0.1383	50	0.1023	0.1418	20	0.2091	0.2525	50	0.2164	0.2598
21	0.0802	0.1197	51	0.0832	0.1227	21	0.173	0.2163	51	0.1687	0.212

22	0.1092	0.1487	52	0.0973	0.1368	22	0.216	0.2594	52	0.2116	0.255
23	0.0709	0.1105	53	0.0763	0.1158	23	0.1732	0.2166	53	0.173	0.2163
24	0.09	0.1295	54	0.0999	0.1394	24	0.2128	0.2562	54	0.2113	0.2547
25	0.0739	0.1135	55	0.0736	0.1132	25	0.1769	0.2203	55	0.1568	0.2002
26	0.0987	0.1382	56	0.0999	0.1394	26	0.2077	0.2511	56	0.2058	0.2492
27	0.0821	0.1217	57	0.0785	0.1181	27	0.173	0.2164	57	0.1623	0.2057
28	0.1084	0.1479	58	0.0999	0.1394	28	0.2065	0.2499	58	0.2116	0.255
29	0.0748	0.1144	59	0.0789	0.1184	29	0.1683	0.2117	59	0.1611	0.2045
30	0.0949	0.1344	60	0.098	0.1376	30	0.2102	0.2535	60	0.2013	0.2447

3- We note from table 3 & 4 that estimated Eigenvalue 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 (Ke).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) & (Ke) methods

Model	$\omega$	n	Bi			Ke		
I	0%	30	0.9131	0.8469	0.5448	0.8468	0.7383	0.4625
		60	0.9160	0.8599	0.5573	0.8406	0.7378	0.4650
II	10%	30	0.9147	0.8562	0.6334	0.8443	0.7483	0.5591
		60	0.9159	0.8697	0.6443	0.8368	0.7462	0.5639

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

Model	$\omega$	n	Bi	Ke
I	0%	30	0.9556	0.9202
		60	0.9571	0.9168
II	10%	30	0.9557	0.9189
		60	0.9532	0.9148

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 sizes n= 30&60 used in simulation experiments.

4- Figures 1, 2, 3&4 represent estimated EIFST for vectors (a) & (b) at uncontaminated data; we note that the values of EIFST increase at n=60 and become the highest at n=30. Also, a method (Bi) has surpassed a method (Ke) 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)

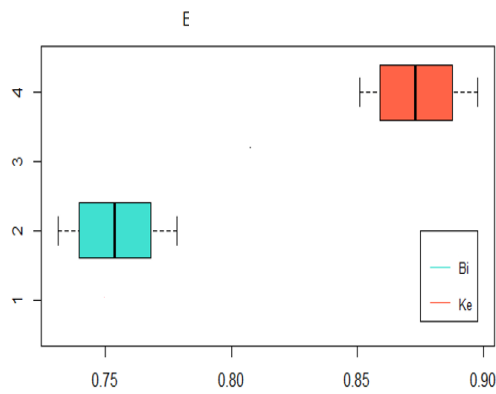


Figure 1: Model I: EIFST for Vector (a) at n=30

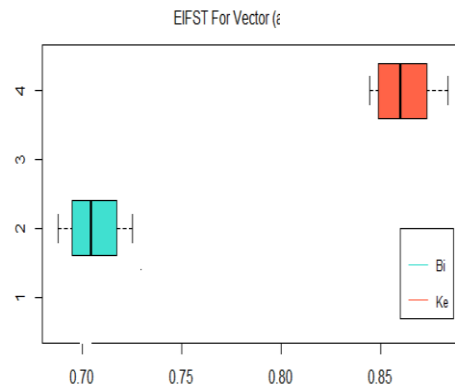


Figure 2: Model I: EIFST for Vector (a) at n=60

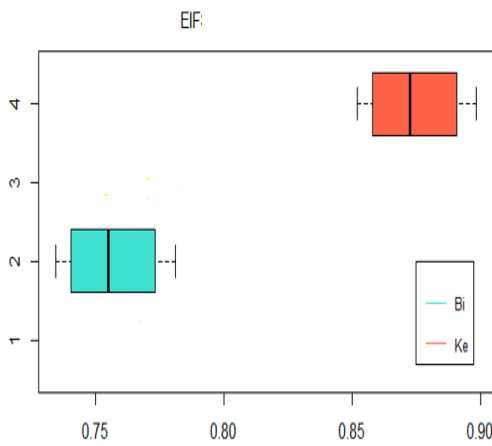


Figure 3: Model I: EIFST for Vector (b) at n=30

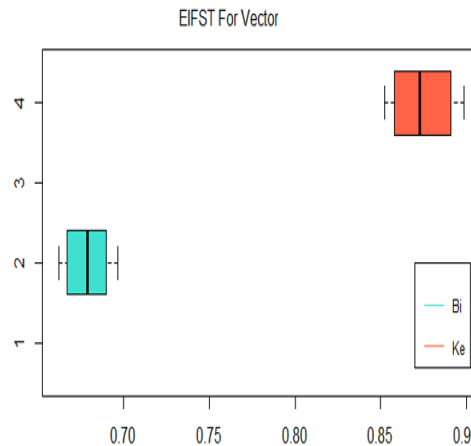


Figure 4: Model I: EIFST for Vector (b) at n=60

5- The figures below show that (Bi) method was better than method (Ke) , also there was a simple difference between vectors (a) & (b) in their values

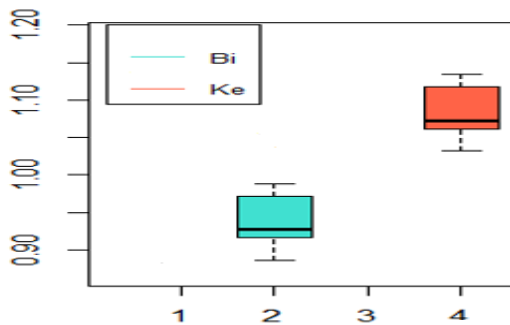


Figure 5: Model II: EIFST for Vector (a)

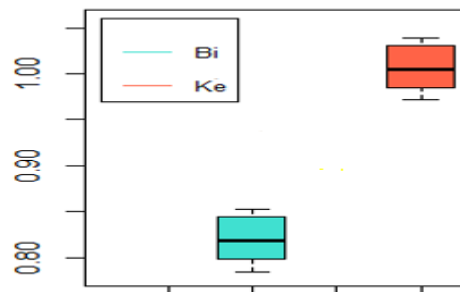


Figure 5: Model II: EIFST for Vector (a) Model II, n=60

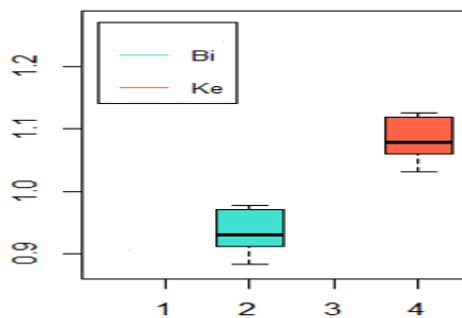


Figure 7: Model II: EIFST for Vector (b) Model II, n=30

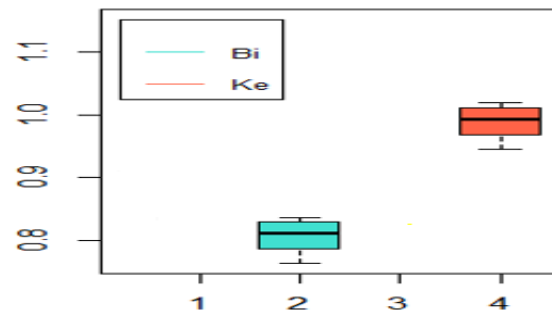


Figure 8: Model II: EIFST for Vector (b) Model II, n=60

### 14- Application

The study based on two sets of variables; the first one includes monthly quantities of oil exported (Million Barrels) for three oil-producing countries within the OPEC (Saudi  $x_1$ , Iraq  $x_2$ , Kuwait  $x_3$  Recorded for a period of sixty months in the years starting January 2015, the second set is  $(z_1, z_2, z_3)$  Represents returns from those quantities ( Million Dollar) for the same countries.

### 15- Determination of Contaminated Data

The Mahalanobis Distance (MD) method was also employed to determine the contaminated observations accurately; the results in Table (6) and below indicate that the contaminated observations were (9, 11, 12, 32, 41, 48, 54, 56, and 60) and that is because the MD values calculated for those observations were greater than the table value  $\chi^2_{(6,0.05)}$  of (12.59).

Table 5: Mahlanobis Distance

obs	MD	obs	MD	obs	MD	obs	MD
1	6.641	16	2.422	31	2.390	46	9.722
2	0.600	17	0.001	32	14.10	47	9.722
3	5.059	18	3.688	33	6.944	48	12.834
4	7.316	19	6.279	34	1.385	49	8.131
5	3.322	20	0.594	35	2.454	50	6.358
6	0.386	21	4.558	36	4.458	51	0.084
7	4.870	22	5.801	37	5.579	52	7.725
8	1.094	23	4.893	38	0.001	53	8.503
9	13.545	24	3.866	39	2.842	54	17.068
10	6.636	25	4.992	40	0.033	55	5.561
11	17.877	26	4.509	41	15.51	56	20.972
12	14.009	27	2.699	42	0.001	57	1.741
13	0.136	28	7.971	43	12.57543	58	0.001
14	10.347	29	0.477	44	5.971	59	1.081
15	4.273	30	3.304	45	6.628	60	12.642

### 16- Estimating Canonical Correlation and Weights Vectors

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

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

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

### 17- Estimation of Influence Function

By finding empirical influence function according to scaled and transformed estimator, it is possible to clarify the effect of the studied data observations on the CC between the variables of the quantities of oil exported (first set) and the corresponding returns (second set).

From Table 6 & 7, it becomes clear that the highest value of the effect function was (0.7188), which is the value that the observation possesses (56), while the lowest value of the effect function is the value that the observation possesses (39) and reached (0.0766), and 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 the observation (34) is the strongest influence in CC estimation, while the lowest value of the influence function was equal to (0.0039) when replacing observation ( 27), and 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 their 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

1- Empirical influence function (EIFST) is an important measure to determine the importance of each observation of the studied data in the estimation process, as well as its role in determining the influence of outliers in the estimation process. The canonical correlation coefficient and weights vectors case of contaminated and uncontaminated data.

2- Robust estimation methods showed convergence in estimated values of the CC and the influence function of the CC coefficient (EIFST).

3- The values of the empirical influence function (EIFST) increase as the sample size decreases.

4- Robust estimation methods are efficient in estimating the CC coefficient in case of data contamination. The values of influence function (EIFST) are close in case of contaminated distribution and in case of uncontaminated data, and the (Bi) method is less. They are affected by contaminated distribution than the (Ke) method.

5- Variables of quantities for exported oil and returns obtained from them for three oil-producing countries within the OPEC organization, Saudi Arabia, Iraq, and Kuwait, follow the contaminated natural distribution, the nature of the relationship between quantities of exported oil and the corresponding returns is strong, while the relationship between the variables of the quantities of oil exported among countries can be described weak. The three are said to be weak, and the same case applies to the nature of the relationship between oil returns of the three countries.

6- Canonical correlation coefficient estimated by (Be) 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 a too strong relationship between the exported quantities of oil and their revenues, and also indicates to an influence of contaminated values when estimating CC.



**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- Romanazzi, M, (1992) "Influence Function in Canonical Correlation Analysis" *Psychometrika*, Vol.57, No.2.
- 3- Abd-Aljabar, Anaam, Abd. , (2000), "Study of the Influence function in Canonical Correlation Analysis" PHD, Economics and Administrative College, University of Baghdad.
- 4- Aziz, Thaka Y., (2012),"Comparison Study Between Canonical Correlation and Robust Canonical Correlation" , *Future Research Magazine*, folder 1, PP.127-147,no.39.
- 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- Yipeng, WU , Chen, Z. , Yao K. , Zhao, X. & Chen Y.( (2019) " On The Correlation Between Fractal Dimension And Robustness Of Complex Networks" *World Scientific Publishing Company*, Vol. 27, No. 4 1950067 (9 pages)
- 7- Al-Ali, Ibrahim M., (2020) "Foundations of multivariate statistical analysis" Syria, Teshreen University – College of Economics, PP 378.
- 8- Farcomeni, A.; Greco, L., (2015)" Robust methods for data reduction", Boca Raton, FL: Chapman & Hall/CRC Press.
- 9- Croux, C. and Dehon, C, (2002), " Analyse Canonique basée sur des estimateurs robustes de La matrice de Covariance ", *Revue de statistique Appliquée* 2, PP. 5-26.
- 10- Wilcox RR. , (2013) "Introduction to Robust Estimation and Hypothesis Testing. "3rd edition, A volume in *Statistical Modeling and Decision Science*
- 11- 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).
- 12- Song, Lin, (2012) "Comparison of co-expression measures: mutual information, correlation, and model based indices". *BMC Bioinformatics*. 13 (328).
- 13- Shevloykov, GL. and Oja, H.,(2016) *Robust Correlation: Theory and Applications* (Wiley Series in Probability and Statistics) 1<sup>st</sup> edition .
- 14- F. Hampel, E. Ronchetti, P. Rousseeuw, W. Stahel, (2011) *Robust statistics: The approach based on influence functions*, Wiley.
- 15- Nasser, M. and Mesbahul, A. Md, (2006) "Estimators of Influence Function" *Communications in Statistics—Theory and Methods*, 35: 21–32.
- 16- Li, G., Peng, H., Zhang, J. and Zhu, L., (2012) "Robust Rank Correlation Based Screening," *the Annals of Statistic*, Vol. 40, No. 3, 1846-1877.
- 17- P. Morris, T., R. White, I. and J. Crowther, M.,(2019) "Using Simulation Studies To Evaluate Statistical Methods " *Statistics in Medicine.*; 38: 2074–2102.

## المقارنة بين الطرق الحصينة في تحليل الارتباط القويم باستخدام دالة التأثير التجريبية

أ.د. لقاء علي محمد العلوي  
جامعة بغداد / كلية الإدارة والاقتصاد ،  
بغداد ، الاعظمية

[Lekaa.ali.1968@gmail.com](mailto:Lekaa.ali.1968@gmail.com)

الباحث/ زهراء خليل حمودي العميد  
وزارة النفط / بغداد ، شارع فلسطين

[Zahraa.khaleel1101a@coadec.uobaghdad.edu.iq](mailto:Zahraa.khaleel1101a@coadec.uobaghdad.edu.iq)  
[Zahraaeyes84@gmail.com](mailto:Zahraaeyes84@gmail.com)

Received:29/3/2021

Accepted: 14/4/2021

Published:

June/ 2021

هذا العمل مرخص تحت اتفاقية المشاع الابداعي نسب المُصنّف - غير تجاري - الترخيص العمومي الدولي 4.0

[Attribution-NonCommercial 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)



## المستخلص

يعد تحليل الارتباط القويم أحد الأساليب الشائعة لتحليل البيانات ومعرفة العلاقة بين مجموعتين من المتغيرات قيد الدراسة ، حيث يعتمد في تحليله على مصفوفة التباين أو مصفوفة الارتباط . يلجأ الباحثون إلى استخدام العديد من الطرق لتقدير الارتباط القويم ، بعضها متحيز للقيم المتطرفة والبعض الآخر مقاوم لتلك القيم ، بالإضافة إلى وجود معايير تكشف عن كفاءة طرق التقدير. في بحثنا هذا، تعاملنا مع طرق تقدير حصينة تعتمد على مصفوفة الارتباط في عملية التحليل للحصول على معامل ارتباط قويم حصين ، وهي طريقتي معامل ارتباط متوسط ثنائي الوزن (BM) ومعامل الارتباط كيندال - تاو (KT).

ومن خلال استخدام المحاكاة والمقارنة بين هاتين الطريقتين باستخدام دالة التأثير التجريبية للمقدرات المرجحة والمنقولة ، أشارت النتائج إلى كفاءة وتفضيل طريقة معامل الارتباط متوسط ثنائي الوزن ، كذلك فإن الدراسة إحتوت على تطبيق لبيانات حقيقية لمجموعتين تتبع متغيراتها التوزيع الطبيعي متعدد المتغيرات ، المجموعة الأولى تمثل المتوسطات الشهرية لكميات النفط المصدر من ثلاث دول من أوبك وهي السعودية والعراق والكويت ، وتمثل المجموعة الأخرى عوائد تلك الكميات للفترة من 2015 إلى 2019 ، بعد تطبيق طريقة (BM) وتقدير دالة التأثير ، كان أعلى تأثير لمعامل الارتباط القويم عند الشهر الرابع والثلاثين وأقلها كان عند السابع والعشرين .

المصطلحات الرئيسية للبحث: الارتباط القويم ، القيم الشاذة ، الطرق الحصينة ، معامل ارتباط متوسط ثنائي الوزن ، معامل ارتباط كندال تاو ، دالة التأثير .