World Applied Sciences Journal 11 (8): 979-987, 2010
ISSN 1818-4952


# Diagnostic-Robust Generalized Potentials for Identifying High Leverage Points in Mediation Analysis

[1,2]*Anwar Fitrianto and* [1,2]*Habshah Midi*

[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia,
Mathematics Building, 43400 Universiti Putra Malaysia, Serdang Selangor, Malaysia
[2]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

**Abstract:** Due to the fact that mediation model involves several linear regression equations, there is concern not only when the data contain observations that are extreme in the response variable but also in the regressor space, namely the leverage points. The Diagnostic Robust Generalized Potentials (DRGP) procedure in multiple linear regression incorporated the Robust Mahanalobis Distance based on the minimum volume ellipsoid and uses Median Absolute Deviation as its cut-off points. In this paper, a slight modification to the DRGP is proposed and we call it ModDRGP. The ModDRGP is applied to the mediation model. The performance of our proposed ModDRGP is evaluated based on Monte Carlo simulation study. The simulation results suggest that ModDRGP has improved the accuracy of the identification of high leverage points when the percentage of high leverage points is medium or high. The method can also be used for the identification of high leverage points in multiple mediation models, as well.

**Key words:** Mediation analysis · Mahalanobis distance · Potentials · Monte Carlo

## INTRODUCTION

Consider the standard $p$ variables multiple linear regression models which can be presented in the matrix notation as $\mathbf{Y} = \mathrm{X}\beta + \varepsilon$, where $\mathbf{Y}$ is an $n\times1$ vector of response (dependent variables), X is an $n \times p$ matrix of predictors (explanatory variables) and $\varepsilon$ is an $n \times 1$ vector of error terms with zero mean and an unknown variance $\sigma^2$. The $\beta$ is $n\times1$ vector of regression coefficients. The predicted value can be written as $\hat{\mathbf{Y}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}'$, where' $\mathrm{X}(\mathrm{XTX})^{-1}\,\mathrm{X}'$ matrix is formally known as a weight matrix or leverage matrix or also hat matrix, denoted by W. Usually the diagonal elements $w_{ii}$ of the weight matrix W are considered as leverage values, which measure influences in the $x$-space. The mean value of leverage $w_{ii}$ for the $n$ points in the sample is $\frac{(p+1)}{n}$, where $p$ is the number of the independent variables and $n$ is the number of observations. Based on the *twice-the-mean-rule* [1], observations are considered unusual when $w_{ii}$ exceeded $\frac{2(p+1)}{n}$. Afterwards, [2] updated the cut-off points for $w_{ii}$ which is considered as large only when it exceeds $\frac{3(p+1)}{n}$. Meanwhile, [3] suggested that the value of $w_{ii}$ ¡ Ü 0.3 of the range of possible values of $w_{ij}$, (0 ! Ü*wii* ! Ü1) appear to be secure, whereas values between 0.2 and 0.5 are risky. Hadi [4] revealed that high leverage points may distort the leverage structure in such a way that the above leverage diagnostics may fail to identify the genuine high-leverage points. He introduced a new type of measure, where the leverage of the $i^{th}$ point is based on a fit to the data with the $i^{th}$ point deleted. Every possible subset of $n$-1 observations is used to form the weight matrix and weight of every deleted observation in turn is generated externally which is known as potentials. Imon [5] defined the $i^{th}$ leverage value as

$$w_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \tag{1}$$

According to Hadi [4], the $i^{th}$ potential can be defined as $p_{ii} = \mathbf{x}_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{x}_i$, where $X_{(i)}$ is the data matrix $X$

**Corresponding Author:** Anwar Fitrianto, Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia Mathematics Building, 43400 Universiti Putra Malaysia,
Serdang Selangor, Malaysia, E-mail: anwarfitrianto@gmail.com.

with the $i^{th}$ row deleted. Meanwhile, $p_{ii}$ is interpreted as the amount of leverage each value $y_i$ has in determining $\hat{y}_i$. Observations corresponding to excessively large potential values are considered as high leverage points. In his article, Hadi [4] also proposed a cut-off point for $p_{ii}$ as: median ($p_{ii}$) + $c$MAD($p_{ii}$), where $\mathrm{MAD}(p_{ii}) = \frac{\text{median}\{|p_{ii} - \text{Median}(p_{ii})|\}}{0.6745}$ and c is a constant usually chosen to be 2 or 3. Subsequently, Imon [5] has shown a simple relationship between $w_{ii}$ and $p_{ii}$ as as $p_{ii} = \frac{w_{ii}}{1 - w_{ii}}$. The leverage value is closely related with the Mahalanobis distance [6, 7] and can be seen as a measure of the distance of the object to the centroid of the data. This mahalanobis distance is a common non-robust multivariate approach that can be presented as diagonal elements of the matrix:

$$MD_i = (x_i - \mu)\,\mathbf{S}^{-1}(x_i - \mu)^T \quad (2)$$

Where $\mu$ is the arithmetic mean vector and S is the covariance matrix. The classical Mahalanobis Square Distance (MSD) is not ideally suited to multivariate outlier detection because it is not resistant to outliers. This is due to the fact that the standard sample location and shape parameters are not robust to outliers and the distributional fit to the distance breaks down when robust measures of location and shape are used in the MSD [8]. Rousseeuw and Leroy [9] recommended to use distance based on robust estimators of multivariate location and scatter ($\mu_R$, $S_R$) to avoid masking effect. A $\chi^2_{p,0.975}$ is used as the cut-off point. Any points which has MSD value larger that the cut-off point is considered as outlier, since for normally distributed data, the MSD is approximately chi square distributed with $p$ degrees of freedom.

General form of Mahalanobis Squared Distance (MSD) can be expressed as

$$MD_i^2 = [x_i - T(X)]^T [C(X)]^{-1} [x_i - T(X)] \quad (3)$$

Where $T(X)$ and $C(X)$ are robust estimations of location and scatter, respectively. Potential multivariate outliers $x_i$ will typically have large MD$i$ values and a comparison with the $\chi^2_p$ distribution can be made. The $T(X)$ can be taken as the center of the minimum volume ellipsoid covering at least half of the $h$ points as suggested by Rousseeuw [10], where $h$ is number of points of the data set $X = \{x_1, x_2, ..., x_n\} \subset \Re^p$ to which the smallest regular ellipsoid could cover. In this case, $h$ can be chosen to be equal to $\left(\frac{n}{2} + 1\right)$. This is called the minimum volume ellipsoid (MVE) estimator. The ellipsoid can be employed as the corresponding covariance estimator. Habshah, *et al*. [11] revealed that the calculation of MVE can be started by drawing a sub sample of ($p$ + 1) different observations, indexed by $J = (i_1, i_2, ... i_{p+1})$. Then they determined the arithmetic mean and the corresponding covariance matrix, given respectively by.

$$C_J = \frac{1}{p} \sum_{i \in J} (x_i - \bar{x})^T (x_i - \bar{x}) \quad \text{and} \quad \bar{x}_J = \frac{1}{p+1} \sum_{i \in J} x_i \quad (4)$$

Where $C_J$ is non-singular. The corresponding ellipsoid should then be inflated or deflated to contain exactly $h$ points, which corresponds to compute $m_J^2 = \text{Med}(x_i - \bar{x}_J) C_J^{-1} (x_i - \bar{x}_J)^T$. The volume of the resulting ellipsoid, corresponding to $m_J^2 C_J$ is proportional to [det(*m*

It is repeated for many $J$ so that the above determinant becomes the minimum and its corresponding values yield $T(X) = \bar{x}_J$ and $C(X) = \left(\chi^2_{p,0.5}\right)^{-1} m_J^2 C_J$, where $\chi^2_{p,0.5}$ is the median of the chi-squared distribution with $p$ degrees of freedom. This correction factor is required to attain the consistency for multivariate normal data.

After obtaining the robust multivariate location and scale estimates given by MVE, we compute the robust Mahalanobis distance $RMD_i = \sqrt{[x_i - T(X)]^T [C(X)]^{-1} [x_i - T(X)]}$. Rousseeuw and Leroy [9] suggested a cut-off point for $RMD_i$ as $\sqrt{\chi^2_{p,0.5}}$. This cut-off value comes from the assumption that the $p$-dimensional variables follow a multivariate normal distribution. Nevertheless, in a real life problem there is no guarantee that data would come from a multivariate normal distribution. Another disadvantage of the usual cut-off point is that it depends only on the dimension of the regressors, but does not take any account of the number of observations. To overcome these shortcomings, Imon [12] suggested a cut-off value for the robust Mahalanobis distances as: *Median*($RMD_i$)+3MAD($RMD_i$).

**Diagnostic-Robust Generalized Potentials (DRGP):** There is evidence that all diagnostic techniques discussed above fail to identify multiple high-leverage points [5]. To overcome the problem, Imon [5] extended the idea of Hadi's potential to a group deletion study. Let us denote a set of points 'remaining' in the analysis by $R$ and a set of points 'deleted' by $D$. Hence, $R$ contains $(n - d)$ points after $d < (n - k)$ points in $D$ are deleted. Without loss of generality, assume that these observations are the last of $d$ rows of $X$ and $Y$ so that the weight matrix $W = X(X^TX)^{-1}X^T$ can be partitioned as

$$\mathbf{W} = \begin{bmatrix} \mathbf{U}_R & \mathbf{V} \\ \mathbf{V}^T & \mathbf{U}_D \end{bmatrix}$$

Where $UR = X_R(X^TX)^{-1}X_R^T$ and $UD = X_D(X^TX)^{-1}X_D^T$ are symmetric matrices of order $(n\text{-}d)$ and $d$ respectively and $V = X_R(X^TX)^{-1}X_D^T$ is a $(n\text{-}d)\times d$ matrix.

Habshah, *et al*. [11] proposed Diagnostic Robust Generalized Potentials (DRGP) for the identification of high leverage point. It was done by computing generalized potentials based on a set $R$ obtained from the robust Mahalanobis distances (RMD) which were obtained from minimum variance ellipsoid (MVE) method suggested by Rouesseuw and Leroy [9]. They applied the cut-off value proposed by [12] to identify whether all elements of the deletion set have potentially high-leverages or not. The set is expected to possess the right deleted points with high $p_{ii}^*$ values due to the fact that the $D$ set is based on RMD.

Based on a group of deleted points indexed by $D$, Habshah, *et al*. [11] defined.

$$w_{ii}^{(-D)} = \mathbf{x}_i^T (\mathbf{X}_R^T \mathbf{X}_R)^{-1} \mathbf{x}_i, \quad i = 1,2,\cdots\cdots,n \tag{5}$$

It should be noted that $w_{ii}^{(-D)}$ is the $i^{th}$ diagonal element of $X(X_R^TX_R)^{-1}X^T$ matrix. When the size of $R$ is $(n\text{-}1)$ and $D = i$, we observe from the equation 1 that $w_{ii}^{(-D)} = x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i = p_{ii}$ which shows that $w_{ii}^{(-D)}$ is a natural extension of $p_{ii}$.

Suppose now that a further point $I$ is removed from the remaining subset $R$ and joins the deletion subset $D$. For any such $I$, it is easy to show that.

$$w_{ii}^{(D+i)} = \mathbf{x}_i^T (\mathbf{X}_R^T \mathbf{X}_R)^{-1} \mathbf{x}_i + \frac{(\mathbf{x}_i^T (\mathbf{X}_R^T \mathbf{X}_R)\ \mathbf{x}_i)^2}{1 - \mathbf{x}_i^T (\mathbf{X}_R^T \mathbf{X}_R)\ \mathbf{x}_i} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \tag{6}$$

This tells us that the potential value of any point $i$, generated externally should be equivalent to the quantity $\frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}}$ when $w_{ii}^{(-D)}$ is generated internally on a reduced space. Using these facts, Habshah, *et al*. [11] used the generalized potentials for all members in a data set which are defined as.

$$p_{ii}^* = \begin{cases} \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}}; & \text{for } i \in R \\ w_{ii}^{(-D)}; & \text{for } i \in D \end{cases} \tag{7}$$

Where $D$ is any arbitrary deleted set of points. There exists no finite upper bound for $p^*_{ii}$'s and it may not be easy to derive a theoretical distribution of them. Habshah, *et al*. [11] considered $p^*_{ii}$ to be large if $p^*_{ii} > \text{Median}(p^*_{ii}) + c\text{MAD}(p^*_{ii})$. The merit of this method is swamping less good leverage as high leverage points.

**New Approach of Diagnostic-robust Generalized Potentials:** The identification of high leverage point is an important area of research in the mediation model as its presence will have an unduly effect on the estimation of the mediation model. This motivates us to apply the DRGP technique to the mediation model with a slight modification. The DRGP used the MAD as the cut-off point in the identification of outlier. We call our proposed approach of DRGP as Modified Diagnostic-Robust Generalized Potentials (ModDRGP). The ModDRGP involves a sophisticated $Q_n$ estimator to which MAD is developed.

A very robust scale estimator is the median absolute deviation about the median, given by $\text{MAD}_n = c\ \text{med}_i|xi - \text{med}_j x_j|$. This estimator is also known as median absolute deviation (MAD) or even median deviation. The MAD is an estimator of scale with a 50% breakdown. It was first promoted by Hampel [13], who attributed it to Gauss. Here $c$ is a constant, a small sample correction factor that can be chosen depending on the sample size to achieve unbiasedness.

Rousseeuw and Croux [14] mentioned that the MAD also has some drawbacks. First, its efficiency at Gaussian distributions is very low; whereas the location median's asymptotic efficiency is still 64%, the MAD is only 37% efficient. Second, the MAD takes a symmetric view on dispersion, because one first estimates a central value (the median) and then attaches equal importance to positive and negative deviations from it. Rousseeuw and Croux [14] then proposed a location-free estimator, $Q_n$, with 50% breakdown point and higher efficiency based on an order statistic of all pairwise distance, which ca be written as: $Q_n = c.\{|x_i - x_j|;_i<_j\}_{(k).}$ The $c$ is a constant factor and $k = \binom{h}{2}$ which is approximately $\binom{n}{2}\Big/4$ and $h$ is a subsample size in the data set ($h \leq n$). The value of $h$ is $\left(\frac{n}{2}+1\right)$ (i.e., roughly half the number of observations). In words, we take $k^{th}$ order statistic of the $\binom{n}{2}$ interpoint distances. We use the value 2.2219 (as in [14]} since this is the value that makes $Q_n$ a consistent estimator for Gaussian data. The $Q_n$ estimator has positive small-sample [17].

Their $Q_n$ estimator is given by the $k^{th}$ order statistic of the $\frac{n(n-1)}{2}$ inter-point distances. It also possesses a breakdown point of 50%, i.e. it can resist up to almost 50% large outliers without becoming extremely biased (50% breakdown point, bounded influence function). Additionally, its Gaussian efficiency is 82% in large samples, which is much higher than that of other robust scale estimators. The $Q_n$ estimate does not depend on symmetry [15]. The $Q_n$ scale estimate is motivated by the Hodges-Lehmann [16] estimate of location of $\hat{\mu} = \text{median}\frac{x_i + x_j}{2}$, for $1< i \leq j < n$.

**The Proposed Methods (MODDRGP):** We propose a diagnostic technique to identify multiple high-leverage points in mediation analysis. To the best of our knowledge, no work is found in the literature on the identification of high leverage points in mediation models. Most of such works are devoted only to multiple linear regression models (as in [4], [5], [11]). Our proposed method is based on the DRGP that has been proposed by Habshah *et al.* [11]. The proposed method employs the $Q_n$ estimator instead of MAD. Croux and Rousseeuw [17] verified that both MAD and $Q_n$ have the same breakdown point that is 50%. Nonetheless, the efficiency of the $Q_n$ is higher (86%) than the MAD (37%). This work inspire us to incorporate the $Q_n$ instead of MAD. By using the $Q_n$ rather than the MAD in the equation of cut-off point in the DRGP, we hope a more powerful scheme that can detect more outliers in mediation analysis which involves several regression equations.

We use the following procedure to identify potential outliers in mediation analysis as follows:

Step 1 : For each $i$ point on $(x_i, m_i)$ pair, calculate the $RMD_i$,

Step 2 : An $i^{th}$ point with ($RMD_i$) exceeds cut-off point of median($RMD_i$)+ 3MAD($RMD_i$) is suspected a high-leverage point and included in the deleted set $D$. The rest of the points are put into the $R$ set,

Step 3 : Based on the above $D$ and $R$ sets, compute the $p^*_{ii}$ using the formula written in the equation 7,

Step 4 : Any deleted point having $p^*_{ii}$ exceeds cut-off point of median($p^*_{ii}$)+ $c$ $Q_n(p^*_{ii})$ is finalized and declared as the high-leverage points, where $c$ =3.

For convenience, we refer the above new method of identifying potential outliers in mediation analysis as ModDRGP1 where the MAD is incorporated in the second step of the ModDRGP1 algorithm. In this paper we also propose another diagnostic method which we call ModDRGP2 whereby a slight modification is made on the second step of ModDRGP1 and keeps other steps the same. The ModDRGP2 modifies the criteria of determining the cut-off point in the step 2 of ModDRGP1. Instead of using median($RMD_i$)+ 3MAD($RMD$), to suspect a high-leverage point, the ModDRGP2 employs the median($RMD_i$)+ 3$Q_n$($RMD_i$).

**Examples**

**Harris and Rosenthal's Data:** This data set which is taken from Harris and Rosenthal [18] consists of 40 subjects that describes teacher expectancies and student achievement. They described potential mediational processes for how expectancies about a person's behavior lead to actual changes in behavior. The dependent variable ($Y$) was the score on a test of basic skills after one semester in the classroom. There are two mediator variables, $M_1$ and $M_2$ to represent social climate and teacher input, but we arbitrary use $M_1$ for

Table 1: Robust mahalanobis distance (RMD), DRGP, ModDRGP1 and ModDRGP2 for Harris and Rosenthal data

| Index | RMD (3.3914) | DRGP (0.0891) | ModDRGP1 (0.0920) | ModDRGP1 (0.0920) |
|---|---|---|---|---|
| 1 | 0.9505 | 0.0511 | 0.0511 | 0.0511 |
| 2 | 1.4991 | 0.0236 | 0.0236 | 0.0236 |
| 3 | 1.3038 | 0.0766 | 0.0766 | 0.0766 |
| 4 | 2.4250 | 0.0613 | 0.0613 | 0.0613 |
| 5 | 1.4226 | 0.0694 | 0.0694 | 0.0694 |
| 6 | 0.8779 | 0.0399 | 0.0399 | 0.0399 |
| 7 | 0.6064 | 0.0324 | 0.0324 | 0.0324 |
| 8 | 0.3178 | 0.0297 | 0.0297 | 0.0297 |
| 9 | 1.6705 | 0.0216 | 0.0216 | 0.0216 |
| 10 | 2.7710 | 0.0803 | 0.0803 | 0.0803 |
| 11 | 1.5762 | 0.0973 | 0.0973 | 0.0973 |
| 12 | 1.0961 | 0.0577 | 0.0577 | 0.0577 |
| 13 | 2.2946 | 0.1006 | 0.1006 | 0.1006 |
| 14 | 0.7960 | 0.0253 | 0.0253 | 0.0253 |
| 15 | 0.5287 | 0.0249 | 0.0249 | 0.0249 |
| 16 | 0.5227 | 0.0307 | 0.0307 | 0.0307 |
| 17 | 2.0602 | 0.0644 | 0.0644 | 0.0644 |
| 18 | 1.7370 | 0.1018 | 0.1018 | 0.1018 |
| 19 | 1.0028 | 0.0353 | 0.0353 | 0.0353 |
| 20 | 0.3232 | 0.0293 | 0.0293 | 0.0293 |
| 21 | 1.9348 | 0.1361 | 0.1361 | 0.1361 |
| 22 | 1.7160 | 0.1104 | 0.1104 | 0.1104 |
| 23 | 1.7151 | 0.0484 | 0.0484 | 0.0484 |
| 24 | 2.1754 | 0.0261 | 0.0261 | 0.0261 |
| 25 | 2.4637 | 0.0349 | 0.0349 | 0.0349 |
| 26 | 0.8243 | 0.0401 | 0.0401 | 0.0401 |
| 27 | 0.8964 | 0.0493 | 0.0493 | 0.0493 |
| 28 | 0.9222 | 0.0479 | 0.0479 | 0.0479 |
| 29 | 0.5063 | 0.0300 | 0.0300 | 0.0300 |
| 30 | 2.0079 | 0.1240 | 0.1240 | 0.1240 |
| 31 | 2.4477 | 0.1849 | 0.1849 | 0.1849 |
| 32 | 0.9370 | 0.0318 | 0.0318 | 0.0318 |
| 33 | 1.0944 | 0.0392 | 0.0392 | 0.0392 |
| 34 | 0.7729 | 0.0335 | 0.0335 | 0.0335 |
| 35 | 0.8624 | 0.0282 | 0.0282 | 0.0282 |
| 36 | 0.3127 | 0.0258 | 0.0258 | 0.0258 |
| 37 | 0.4650 | 0.0265 | 0.0265 | 0.0265 |
| 38 | 2.0069 | 0.0193 | 0.0193 | 0.0193 |
| 39 | 0.5507 | 0.0318 | 0.0318 | 0.0318 |
| 40 | 0.8507 | 0.0326 | 0.0326 | 0.0326 |

this example. It is hypothesized that the general social warmth provided to the student is what leads him or her to achieve more. On the other hand, teacher expectancy may lead to increase student achievement.

With the purpose of evaluating the performance of the proposed ModDRGP1 and ModDRGP2, we present comparisons with the DRGP proposed by Habshah *et al.* [11]. Using RMD, at cut-off value of 3.391, no observations can be identified as high leverage points.

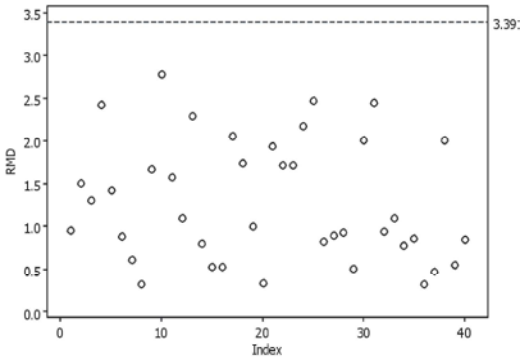


Fig. 1: Index plot of robust mahalanobis distance for Harris and Rosenthal data

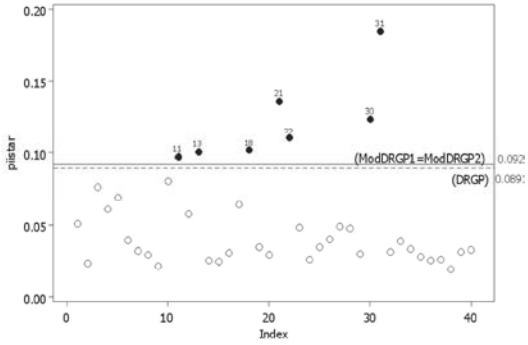


Fig. 2: Index plot of DRGP, ModDRGP1 and ModDRGP2 for Harris and Rosenthal data

Since we could not detect any observation as high leverage point using the well known RMD method, we need to investigate it using more credible method. In this study, we used a newly measure for the identification of high-leverage point which is based on Diagnostic Robust Generalized Potentials (DRGP).

While the numerical result is also presented in Table 1, the graphical result is presented in Figure 2. The plot in the Figure 2 clearly indicates much improvement of the outlier identification when we compared to Figure 1. The ModDRGP1, ModDRGP2 and the previously established DRGP detect same observations as high leverage points. Those observations are observation 11, 13, 18, 21, 22 and 30. If we compare them to the RMD which could detect no observation, all DRGP-based methods perform better in this Harris and Rosenthal data. It shows the merit of the ModDRGP1 and ModDRGP2 which is at least possesses the same as the DRGP method in this data. Moreover, as we can observe in the Table 1, all the DRGP-based method produced the same $p^*_{ii}$ which mean that in this Harris and Rosenthal data, the modification of the cut-off values did not have any effects to the $p^*_{ii}$.

Table 2: Robust mahalanobis distance (RMD), DRGP, ModDRGP1 and ModDRGP2 for Woodworth data

| Index | RMD (2.4877) | DRGP (0.1004) | ModDRGP1 (0.0745) | ModDRGP1 (0.0745) |
|---|---|---|---|---|
| 1 | 1.0315 | 0.0391 | 0.0391 | 0.0391 |
| 2 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 3 | 2.1760 | 0.1218 | 0.1218 | 0.1218 |
| 4 | 2.1281 | 0.1275 | 0.1275 | 0.1275 |
| 5 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 6 | 1.0315 | 0.0391 | 0.0391 | 0.0391 |
| 7 | 1.2111 | 0.0201 | 0.0201 | 0.0201 |
| 8 | 2.0581 | 0.1074 | 0.1074 | 0.1074 |
| 9 | 1.0315 | 0.0391 | 0.0391 | 0.0391 |
| 10 | 2.2214 | 0.1004 | 0.1004 | 0.1004 |
| 11 | 1.5423 | 0.0507 | 0.0507 | 0.0507 |
| 12 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 13 | 2.0581 | 0.1074 | 0.1074 | 0.1074 |
| 14 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 15 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 16 | 1.1002 | 0.0478 | 0.0478 | 0.0478 |
| 17 | 1.0315 | 0.0391 | 0.0391 | 0.0391 |
| 18 | 1.2111 | 0.0201 | 0.0201 | 0.0201 |
| 19 | 1.7788 | 0.0227 | 0.0227 | 0.0227 |
| 20 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 21 | 1.5423 | 0.0507 | 0.0507 | 0.0507 |
| 22 | 1.7788 | 0.0227 | 0.0227 | 0.0227 |
| 23 | 3.3285 | 0.1000 | 0.1000 | 0.1000 |
| 24 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 25 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 26 | 0.2356 | 0.0209 | 0.0209 | 0.0209 |
| 27 | 1.1002 | 0.0478 | 0.0478 | 0.0478 |
| 28 | 1.2111 | 0.0201 | 0.0201 | 0.0201 |
| 29 | 1.7543 | 0.0403 | 0.0403 | 0.0403 |
| 30 | 0.2356 | 0.0209 | 0.0209 | 0.0209 |
| 31 | 1.5423 | 0.0507 | 0.0507 | 0.0507 |
| 32 | 2.1281 | 0.1275 | 0.1275 | 0.1275 |
| 33 | 1.1002 | 0.0478 | 0.0478 | 0.0478 |
| 34 | 1.1002 | 0.0478 | 0.0478 | 0.0478 |
| 35 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 36 | 2.1920 | 0.0423 | 0.0423 | 0.0423 |
| 37 | 1.7227 | 0.0369 | 0.0369 | 0.0369 |
| 38 | 1.2111 | 0.0201 | 0.0201 | 0.0201 |
| 39 | 0.2356 | 0.0209 | 0.0209 | 0.0209 |
| 40 | 2.2034 | 0.0195 | 0.0195 | 0.0195 |
| 41 | 2.2034 | 0.0195 | 0.0195 | 0.0195 |
| 42 | 1.0315 | 0.0391 | 0.0391 | 0.0391 |
| 43 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 44 | 1.4210 | 0.0450 | 0.0450 | 0.0450 |
| 45 | 1.2111 | 0.0201 | 0.0201 | 0.0201 |
| 46 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 47 | 1.0136 | 0.0379 | 0.0379 | 0.0379 |
| 48 | 0.7882 | 0.0217 | 0.0217 | 0.0217 |
| 49 | 1.7227 | 0.0369 | 0.0369 | 0.0369 |
| 50 | 1.1002 | 0.0478 | 0.0478 | 0.0478 |

**Woodworth Data:** This example is a stimulus-organism-response mediation study [19], in which the effect of a stimulus on a response as mediated by the organism. The data for the 50 subjects in this hypothetical study of the effects of room temperature on water consumption is taken from [20], where $X$ is temperature in degrees Fahrenheit, $M$ is a self-report measure of thirst at the end of a 2-hour period and $Y$ is the number of deciliters of water consumed during the last 2 hours of the study.

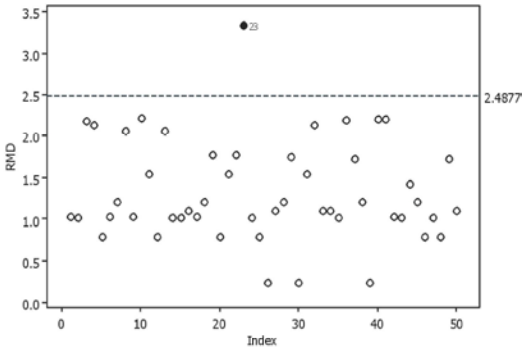


Fig. 3: Index plot of robust mahalanobis distance for Woodworth data

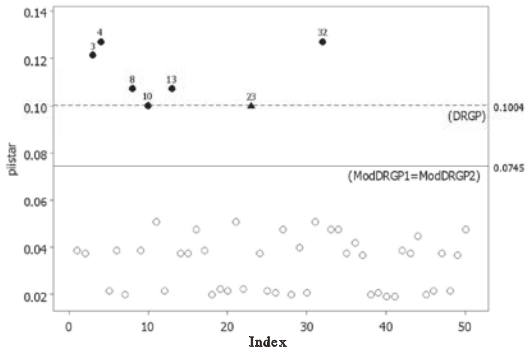


Fig. 4: Index plot of DRGP, ModDRGP1 and ModDRGP2 for Woodworth data

In order to evaluate the merit of the ModDRGP1 and ModDRGP2 methods, we present a comparison of it with the usual DRGP method. Table 2 displays Robust Mahalanobis Distance, $p^*_{ii}$, of each observation of the Woodworth data. Figure 3 shows a scatter plot of the Woodworth based on usual Robust Mahalanobis Distance. We can see from the figure that nothing extraordinary on the observations except only one observation (observation 23) which is located far from the remaining observations. Only this single observation is detected as outlier by the RMD. Six observations specifically observation 3, 4, 8, 10, 13 and 32 can be detected as high leverage points using the DRGP. And finally, when ModDRGP1 and ModDRGP2 are applied, we can detect one more observation to have high leverage, which is observation 23 as can be sought in Figure 4 below. It denotes that in this Woodworth data, our proposed method has an advantage in detecting high leverage points.

Table 3: Identification of multiple high leverage points based on 10,000 simulations

| | | | | Diagnostic Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| % HLP | $n$ | HLP Dist. | HLP | Twice | Thrice | Huber | $p_{ii}$ | DRGP | Mod DRGP1 | Mod DRGP2 |
| 5 | 20 | 5 | 1 | 2.9738 | 1.3834 | 2.9738 | 2.5902 | 1.7032 | 1.6154 | 1.6046 |
| | | 10 | | 2.8325 | 1.3546 | 2.8325 | 2.9841 | 1.7033 | 1.6152 | 1.6045 |
| | 40 | 5 | 2 | 5.8812 | 2.8122 | 2.0736 | 4.6527 | 2.9110 | 2.9166 | 2.9027 |
| | | 10 | | 5.6165 | 2.7622 | 2.0667 | 5.5025 | 2.9110 | 2.9161 | 2.9022 |
| | 80 | 5 | 4 | 11.7001 | 5.6668 | 2.0440 | 8.7304 | 5.2810 | 5.4914 | 5.4825 |
| | | 10 | | 11.2035 | 5.5657 | 4.0000 | 10.5188 | 5.2810 | 5.4903 | 5.4814 |
| | 100 | 5 | 5 | 14.6262 | 7.0838 | 0.3188 | 10.7902 | 6.4896 | 6.8039 | 6.7967 |
| | | 10 | | 14.0203 | 6.9774 | 1.0353 | 13.0620 | 6.4896 | 6.8024 | 6.7952 |
| | 160 | 5 | 8 | 23.2832 | 11.3281 | 0.0000 | 16.9073 | 10.0982 | 10.6791 | 10.6758 |
| | | 10 | | 22.3550 | 11.1456 | 0.0000 | 20.4672 | 10.0982 | 10.6766 | 10.6733 |
| 10 | 20 | 5 | 2 | 3.9200 | 2.3986 | 3.9200 | 3.3190 | 2.5300 | 2.4271 | 2.4235 |
| | | 10 | | 3.8527 | 2.3870 | 3.8527 | 3.5849 | 2.5252 | 2.4224 | 2.4187 |
| | 40 | 5 | 4 | 7.7612 | 4.8607 | 4.0844 | 6.1795 | 4.6233 | 4.5532 | 4.5478 |
| | | 10 | | 7.6449 | 4.8371 | 4.0951 | 6.7580 | 4.6233 | 4.5510 | 4.5457 |
| | 80 | 5 | 8 | 15.4860 | 9.7410 | 0.0031 | 11.8317 | 8.7990 | 8.8115 | 8.8089 |
| | | 10 | | 15.2857 | 9.7067 | 0.0013 | 13.0944 | 8.7990 | 8.8079 | 8.8053 |
| | 100 | 5 | 10 | 19.3369 | 12.1838 | 0.0000 | 14.6690 | 10.9194 | 10.9646 | 10.9630 |
| | | 10 | | 19.1051 | 12.1498 | 0.0000 | 16.2423 | 10.9194 | 10.9610 | 10.9594 |
| | 160 | 5 | 16 | 30.8483 | 19.5065 | 0.0000 | 23.1468 | 17.2536 | 17.3950 | 17.3947 |
| | | 10 | | 30.5060 | 19.4487 | 0.0000 | 25.6479 | 17.2536 | 17.3883 | 17.3880 |
| 15 | 20 | 5 | 3 | 4.9101 | 3.0842 | 4.9101 | 3.8540 | 3.3770 | 3.2807 | 3.2806 |
| | | 10 | | 4.8725 | 3.4086 | 4.8725 | 4.1522 | 3.3678 | 3.2698 | 3.2691 |
| | 40 | 5 | 6 | 9.7406 | 6.2361 | 0.8620 | 7.2565 | 6.3775 | 6.3044 | 6.3033 |
| | | 10 | | 9.6779 | 6.8928 | 0.8892 | 7.9537 | 6.3775 | 6.3011 | 6.3000 |
| | 80 | 5 | 12 | 19.4321 | 12.5014 | 0.0001 | 14.0879 | 12.4377 | 12.3792 | 12.3787 |
| | | 10 | | 19.3606 | 13.8095 | 0.0000 | 15.5384 | 12.4377 | 12.3730 | 12.3725 |
| | 100 | 5 | 15 | 24.2797 | 15.6529 | 0.0000 | 17.5549 | 15.4803 | 15.4249 | 15.4246 |
| | | 10 | | 24.1866 | 17.2763 | 0.0000 | 19.3105 | 15.4803 | 15.4175 | 15.4172 |
| | 160 | 5 | 24 | 38.7446 | 25.0536 | 0.0000 | 28.0973 | 24.6100 | 24.5850 | 24.5850 |
| | | 10 | | 38.6264 | 27.6627 | 0.0000 | 30.6727 | 24.6100 | 24.5747 | 24.5747 |
| 20 | 20 | 5 | 4 | 5.8832 | 1.2463 | 5.8832 | 2.3214 | 4.2500 | 4.1700 | 4.1688 |
| | | 10 | | 5.8764 | 1.2533 | 5.8764 | 2.8390 | 4.2489 | 4.1665 | 4.1662 |
| | 40 | 5 | 8 | 11.7179 | 2.7236 | 0.3666 | 3.4765 | 8.2144 | 8.1524 | 8.1521 |
| | | 10 | | 11.7054 | 2.7663 | 0.3351 | 4.5875 | 8.2144 | 8.1476 | 8.1473 |
| | 80 | 5 | 16 | 23.3686 | 5.5882 | 0.0000 | 5.5618 | 16.1980 | 16.1447 | 16.1447 |
| | | 10 | | 23.3570 | 5.7254 | 0.0000 | 7.6122 | 16.1980 | 16.1390 | 16.1390 |
| | 100 | 5 | 20 | 29.1918 | 7.0286 | 0.0000 | 6.5780 | 20.1976 | 20.1417 | 20.1417 |
| | | 10 | | 29.1848 | 7.1965 | 0.0000 | 8.9722 | 20.1976 | 20.1373 | 20.1373 |
| | 160 | 5 | 32 | 46.6307 | 11.3456 | 0.0000 | 9.6294 | 32.2160 | 32.1601 | 32.1601 |
| | | 10 | | 46.6316 | 11.6489 | 0.0000 | 13.0530 | 32.2160 | 32.1529 | 32.1529 |
| 25 | 20 | 5 | 5 | 5.5154 | 0.8865 | 5.5154 | 1.0927 | 5.1426 | 5.0895 | 5.0897 |
| | | 10 | | 5.9327 | 0.8639 | 5.9327 | 1.2551 | 5.1573 | 5.0984 | 5.0983 |
| | 40 | 5 | 10 | 11.0335 | 1.9649 | 0.2581 | 1.4374 | 10.1012 | 10.0633 | 10.0639 |
| | | 10 | | 11.8272 | 1.9177 | 0.2490 | 1.6867 | 10.1020 | 10.0604 | 10.0604 |
| | 80 | 5 | 20 | 22.1212 | 4.0665 | 0.0000 | 2.1232 | 20.0713 | 20.0433 | 20.0433 |
| | | 10 | | 23.7109 | 3.9814 | 0.0000 | 2.5249 | 20.0713 | 20.0405 | 20.0405 |
| | 100 | 5 | 25 | 27.6527 | 5.1440 | 0.0000 | 2.4572 | 25.0597 | 25.0392 | 25.0392 |
| | | 10 | | 29.6384 | 5.0299 | 0.0000 | 2.8869 | 25.0597 | 25.0362 | 25.0362 |
| | 160 | 5 | 40 | 44.2109 | 8.2731 | 0.0000 | 3.4167 | 40.0463 | 40.0278 | 40.0278 |
| | | 10 | | 47.3548 | 8.1163 | 0.0000 | 4.0230 | 40.0463 | 40.0248 | 40.0248 |
| 30 | 20 | 5 | 6 | 4.2001 | 0.7834 | 4.2001 | 0.8289 | 6.0583 | 6.0270 | 6.0289 |
| | | 10 | | 4.1328 | 0.7690 | 4.1328 | 0.8102 | 6.0890 | 6.0523 | 6.0523 |
| | 40 | 5 | 12 | 8.4153 | 1.6947 | 0.2327 | 1.1274 | 12.0383 | 12.0246 | 12.0246 |
| | | 10 | | 8.2969 | 1.6690 | 0.2381 | 1.0734 | 12.0396 | 12.0231 | 12.0231 |
| | 80 | 5 | 24 | 16.7952 | 3.4955 | 0.0000 | 1.7156 | 24.0211 | 24.0116 | 24.0116 |
| | | 10 | | 16.6170 | 3.4347 | 0.0000 | 1.6074 | 24.0211 | 24.0098 | 24.0098 |
| | 100 | 5 | 30 | 21.0023 | 4.3995 | 0.0000 | 2.0033 | 30.0148 | 30.0067 | 30.0067 |
| | | 10 | | 20.7886 | 4.3249 | 0.0000 | 1.8563 | 30.0148 | 30.0054 | 30.0054 |
| | 160 | 5 | 48 | 33.5222 | 7.1183 | 0.0000 | 2.8888 | 48.0062 | 48.0027 | 48.0027 |
| | | 10 | | 33.2571 | 6.9848 | 0.0000 | 2.6650 | 48.0062 | 48.0021 | 48.0021 |

**Results on Simulated Datasets:** The Monte Carlo simulation study in this section focuses on the detection of high leverage points. The performances of our new proposed cut-off point, ModDRGP1 and ModDRGP2, was done by comparing it to several other common method of identification of high leverage points. They are *twice-the-mean rule* [1], *thrice-the-mean-rule* [2], Huber choice with $w_{ii}$ >0.2 cut-off point [3], Hadi's potentials based on MAD [4] and Habshah's DRGP [11]. We have considered the constant, *c*, equals to 3 in this simulation study which is needed in Hadi's potentials and the Habshah's DRGP.

In this simulation study, we considered five different sample sizes of 20, 40, 80, 100 and 160 to reflect small (20-40), medium (80-100) and large sample sizes (160), respectively. Two variables were generated to reflect the necessary condition in simple mediation model of *x* and *m* which are selected at random from the *U*(0,1) distribution. Several contamination scenarios were done with regard to distance (in $\sigma$ units) and percentage of contaminated observations of the sample sizes. We used HLP distances of $\sigma$ and $\sigma$. Meanwhile, 5%, 10%, 15%, 20%, 25% and 30% observations were used to replace the last observations of the generated data. We consider of low, medium and high percentage of HLP when the percentage of the HLP are 5%-10%, 15%-20% and 25%-30%, respectively. By doing that manner, each generated sample will contain $(1 - \alpha)$% clean observations and the last $\alpha$% observations of both variables are considered as equally high leverage points. We repeated the simulation by 10000 on each combination.

Result of the Monte Carlo simulation discussed is presented in Table 3. After the huge number of simulation, a method which can identify the same or nearly the same number of pre-defined HLP is considered as the good one. The first four columns of the table show the simulation combination of percentage of high leverage points (% HLP), sample size (*n*), HLP distance (HLP Dist.) and the number of high leverage points (No. HLP). The value in columns 5 onwards is the mean or average number of HLP which can be identified by particular method.

Following the columns in the Table 3, we initiate to talk about the Monte Carlo simulation result based on the percentage of high leverage points. When the percentage of HLP is small, diagnostic tools based on $w_{ii}$, such as *twice* or *thrice the mean rule*, look to have better performance than the other methods, but it is disturbed by our some good result from new proposed method and DRGP as well. As a result, at smaller percentage of HLP, we cannot easily conclude which ones perform better. It also can be seen that at the small proportion of HLP, we can easily say that the DRGP has advanced performance than our proposed ModDRGP1 and ModDRGP2, but both *twice-the-mean rule* and *thrice-the mean rule* looks to have slightly better performance than the DRGP-based methods.

Meanwhile, in medium or large percentage of HLP, from the Table 3, we can say that in general, both DRGP and the proposed ModDRGP1 and ModDRGP2 have better performance compared to the other methods. When the percentage of HLP increases, it clearly shows that ModDRGP1 and ModDRGP2 also have outstanding presentation. Our proposed ModDRGP2 is the only method achieving best result amongst the other method under study regardless the sample size as long as the percentage of the high leverage points are relatively medium or high. Especially when we compare the previous DRGP and our new proposed ModDRGP2, at medium or high percentage of HLP, it is clear that this new method enhanced the previous DRGP and ModDRGP1 method.

## CONCLUSION

There were two case studies that comprised the material in this study. In the first case, we presented Harris and Rosenthal data which is well known example of mediation analysis. From the result we found that our newly proposed cut-off point of the DRGP is equally good in detecting of high leverage point. Both methods could detect more observations compared to the Hadi's potentials. The second case study was dealt with the Woodworth data. The data is also very common in discussions of mediation analysis, especially the simple mediation analysis. It was demonstrated that when both Hadi's potentials and the previous DRGP could identify 1 observation and 6 observations respectively, our newly proposed ModDRGP1 and ModDRGP2 method could detect 7 observations. It validated that the ModDRGP1 and ModDRGP2 deserve as an alternative method in detection of high leverage point.

In order to strengthen the analysis, we provided a Monte Carlo simulation to evaluate the performance of our proposed ModDRGP1 and ModDRGP2. The simulation results suggested that by applying our newly proposed method has improved the accuracy of the identification of high leverage point when the percentage of high leverage points is medium or high. Even though the method was studied in simple mediation analysis, but it can be used to identify multivariate high leverage point as well.

## REFERENCES


1. Hoaglin, D.C. and R.E. Welsch, 1978. The Hat matrix in regression and ANOVA. The Amer. Statist., 32: 17-22.
2. Velleman, P.F. and R.E. Welsch, 1981. Efficient computing of regression diagnostics. The Amer. Statist., 35: 234-242.
3. Huber, P.J., 1981. Robust Statistics. New York: Wiley.
4. Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Computational. Statist. and Data Analysis, pp: 14: 1-27.
5. Imon, A.H.M.R., 2002. Identifying Multiple High Leverage Points in Linear Regression, J. Statistical Studies, Special Volume in Honor of Professor Mir Masoom Ali, pp: 207-218.
6. Næs, T., 1989. Leverage and influence measures for principal component regression, Chemom. Intell. Lab. Sys., 5: 155-168.
7. Weisberg, S., 2005. Applied Linear Regresion, 3rd. Edition, New York: John Wiley and Sons.
8. Rousseeuw, P.J. and B.C. Van Zomeren, 1991. Robust distances: simulations and cutoff values, in Directions in Robust Statistics and Diagnostics, part II, eds. Stahel, W. and Weisberg, S., Springer Verlag: New York, vol. 34 of the IMA Volumes in Mathematics and Its Applications, pp: 195-203.
9. Rousseeuw, P.J. and A.M. Leroy, 1987. Robust Regression and Outlier Detection. Wiley, New York.
10. Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Mathematical Statistics and Applications (Vol: B, Edited by W. Grossmann *et al*.), 283-297, Reidel Publishing.
11. Habshah, M .H., M.R. Norazan and A.H.M. Rahmatullah Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression, J. Applied Statistics, 36(5): 507-520.
12. Imon, A.H.M.R., 2008. Identification of multiple high leverage points using robust Mahalanobis distances, J. Stat. Stud. (under review).
13. Hampel, F., 1974. The influence curve and its role in robust estimation, J. Am. Statist. Assoc., 69: 383-393.
14. Rousseeuw, P.J. and C. Croux, 1993. Alternatives to the median absolute deviation: J. Amer. Stat. Assoc., 88(424): 1273-1283.
15. Nunkesser, R., K. Schettlinger and R. Fried, 2007. Aplying the Qn-estimator online, Data Analysis, Machine Learning and Applications. Poceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007, Springer, Berlin, Heidelberg, pp: 277-284.
16. Hodges, J.L. and E.L. Lehmann, 1963. Estimates of location based on rank tests, Ann. Math. Statist, 34: 598-611.
17. Croux, C. and P.J. Rousseeuw, 1992. Time-efficient algorithms for two highly robust estimators of scale. In Dodge, Y. and Whittaker, J.C. (eds), Computational Statistics, 1, 411-428, Heidelberg. Physica-Verlag.
18. Harris, M.J., and R. Rosenthal, 1985. Mediation of interpersonal expectancy effects: 31 meta-analyses. Psychological Bulletin, 97: 363-386.
19. Woodworth, R.S., 1928. Dynamic psychology. In C. Murchison (Ed.), Psychologies of 1925 (pp. 111-126). Worcester, MA: Clark University Press.
20. MacKinnon, D.P., 2008. Introduction to Statistical Mediation Analysis. New York: Taylor and Francis.