Sains Malaysiana 54(8)(2025): 2087-2097
http://doi.org/10.17576/jsm-2025-5408-17

Improved Robust Principal Component Analysis based on Minimum Regularized Covariance Determinant for the Detection of High Leverage Points in High Dimensional Data

(Penambahbaikan Analisis Komponen Utama berdasarkan Penentu Kovarian Teratur Minimum bagi Mengecam Titik Tuasan Tinggi untuk Data Dimensi Tinggi)

HABSHAH MIDI[1,2,*], JAAZ SUHAIZA[1,3], MOHD ASLAM[1,2], HANI SYAHIDA[2] & EMI AMIELDA[3]

[1]*Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Department of Mathematics & Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[3]*Faculty of Computing & Multimedia, Universiti Poly-Tech Malaysia, 56100 Cheras, Kuala Lumpur, Malaysia*

Received: 22 April 2024/Accepted: 13 March 2025

ABSTRACT

This paper presents an extension work of robust principal component analysis (ROBPCA) denoted as IRPCA, to improve the accuracy of the detection of high leverage points (HLPs) in high dimensional data (HDD). The IRPCA employs the Principal Component Analysis (PCA) to reduce the dimension of the data set and subsequently a robust location and scatter estimates of the PC scores are obtained based on the Minimum Regularized Covariance Determinant (MRCD). Instead of using robust score distance to detect HLPs as in ROBPCA; in the proposed IRPCA, we have considered using Robust Mahalanobis distance (RMD). The performance of the IRPCA is compared to the ROBPCA and the Minimum Regularized Covariance Determinant and PCA-based method (MRCD-PCA) for the identification of HLPs in HDD. The results signify that all the three methods are very successful in the detection of HLPs with no masking effect. Nonetheless, the ROBPCA suffers from serious swamping problems for less than 30% of HLPs. The proposed IRPCA and the MRCD-PCA have similar performance, having very small swamping effect. However, the MRCD-PCA algorithm is quite cumbersome and required longer computational running time. The attractive feature of the IRPCA is that it provides a simpler algorithm and it is very fast.

Keywords: High Leverage Point; minimum regularized covariance determinant; principal component analysis; robust mahalanobis distance

ABSTRAK

Kertas ini membentangkan kerja lanjutan bagi Analisis Komponen Utama Teguh (ROBPCA) ditandakan dengan IRPCA, untuk meningkatkan ketepatan pengecaman titik tuasan tinggi (HLPs) dalam data dimensi tinggi (HDD). IRPCA menggunakan Analisis Komponen Utama (PCA) bagi menurunkan dimensi set data dan seterusnya penganggar lokasi dan skala skor PC dikira berdasarkan Penentu Kovarian Teratur Minimum (MRCD). Dengan tidak menggunakan jarak skor teguh untuk pengecaman HLPs seperti ROBPCA; dalam kaedah IRPCA yang dicadangkan, kami telah mempertimbangkan penggunaan Jarak Mahalanobis Teguh (RMD). Prestasi IRPCA yang dicadang dibandingkan dengan kaedah ROBPCA dan kaedah Penentu Kovarian Teratur Minimum dan PCA (MRCD-PCA) bagi mengecam HLPs dalam HDD. Keputusan menunjukkan ketiga-tiga kaedah sangat berjaya dalam pengesanan HLPs tanpa kesan penyorokan. Walau bagaimanapun, ROBPCA mengalami masalah kesan limpahan yang serius apabila terdapat HLPs kurang daripada 30%. Prestasi IRPCA yang dicadangkan dan MRCD-PCA ada lah sama; mempunyai kesan limpahan yang sangat kecil. Namun begitu, algoritma MRCD-PCA agak rumit dan memerlukan masa yang panjang. Sifat menarik bagi IRPCA ialah ia memberi algoritma yang mudah dan masa pengiraan yang singkat.

Kata kunci: Analisis komponen utama; jarak Mahalanobis teguh; penentu kovarian teratur minimum; titik tuasan baik

INTRODUCTION

High dimensional data refers to the situations where the number of covariates or predictors is much larger than the number of data points (i.e., $p >> n$). To provide an example, in gene analysis, a single individual may have measurements for millions of genes (Boulesteix & Strimmer 2007), whereas in image analysis, there are thousands of high-resolution pixel images with a limited number of samples (Chiang 2016). Other prominent and critical areas of high dimensional data are usually found in image analysis,

microarray analysis, document classification, astronomy, and atmospheric science. Dealing with this kind of data sets involves new challenging issue since it is difficult to analyze high dimensional data, due to the high correlation between variables and the risk of model overfitting. The application of conventional statistical approaches to high dimensional data tends to be ineffective, can cause serious misleading result and difficult interpretation on the pattern of the data particularly in the presence of outliers.

Outliers in regression problem can be classified into three categories namely the vertical outliers, residual outliers, and high leverage points. Vertical outliers are observations that are outlying in the *Y*-space; residual outliers are data with noticeably large residuals (Habshah, Norazan & Imon 2009; Siti Zahariah & Habshah 2023). In contrast, high leverage points (HLPs) are those observations that are outlying in the *X*-space. While a great deal of study has been done on residual and vertical outlier detection, comparatively there has been less emphasis on addressing high leverage points (HLPs). Huber (1973) and Rana, Midi and Imon (2009) stated that the presence of HLPs may cause apparent non-normality. HLPs in a dataset might cause severe effects on the parameter estimates and would give invalid results to the regression model and become more serious in high dimensional data (Midi et al. 2021; Rashid et al. 2021). Accurate detection of HLPs is of paramount importance in statistical analysis, as an incorrect identification of such points will substantially disrupt the standard error of estimates and give rise to a multicollinearity problem, masking and swamping of outliers, overfitting or underfitting of a model which will lead to insignificant prediction (Chiang 2016; Siti Zahariah, Habshah & Mohd Shafie 2022). Masking refers to outliers misidentified as inliers and swamping, on the other hand, is a phenomenon of incorrectly labelling normal observations as outliers (Rashid et al. 2021). This is the reason why the detection of outliers or HLPs is essential before making any kind of inferences.

Many papers are available in the literatures for the identification of HLPs in linear model and low dimensional data, to name a few (Lim & Midi 2016; Rousseeuw & Driessen 1999). Nonetheless, not many papers are devoted to the detection of HLPs in high dimensional data. This is primarily due to computational burden that one has to face when analyzing a huge number of variables. Robust Mahalanobis distance (RMD) is a very popular diagnostic tool used for the identification of HLPs (Hubert, Rousseeuw & Verdonck 2012). The formulation of the RMD is based on robust location and robust covariance matrix. Minimum covariance determinant (MCD) is an example of highly robust estimators of multivariate location and scatter (Rousseeuw 1985). It is very resistant to outlying observations that makes the MCD highly effective for outlier detection. Nevertheless, most of the robust covariance matrix is only applicable for low dimensional data because it is not invertible in high dimension cases. As a solution to this problem, Boudt et al. (2018) developed a minimum regularized covariance determinant (MRCD) to overcome the curse of dimensionality issue. Afterwards, the robust Mahalanobis distance which is based on the MRCD (RMD-MRCD) is put forward. However, according to Siti Zahariah and Habshah (2023), the RMD-MRCD method indicates a decrease in its performance as the number of independent variables (p) increases. To remedy this problem, Siti Zahariah and Habshah (2023) proposed robust Mahalanobis distance (RMD) based on the combined methods of the minimum regularized covariance determinant and the principal component analysis (MRCD-PCA). It is developed by incorporating the Principal Component Analysis (PCA) method in the MRCD algorithm. The MRCD-PCA consist of two stages whereby in the first stage, the PCA reduces the dimension of data set and generates a fitted $\hat{X}$ matrix in the original dimension p. Subsequently, the fitted $\hat{X}$ matrix will be shrunk to yield an invertible covariance matrix for HDD. The MRCD was then performed on these fitted $\hat{X}$ to determine the robust mean and robust covariance of HDD. In the second stage, the robust Mahalanobis distance based on MRCD-PCA estimators is constructed for the identification of HLPs in HDD. The MRCD-PCA is very successful in the detection of HLPs with small swamping effect. The only shortcoming of this method is that its algorithm is quite cumbersome and takes longer computational running times.

Hubert, Rousseeuw and Vanden Branden (2005) developed Robust Principal Component Analysis (ROBPCA) which is the combination of projection pursuit and robust covariance estimate, i.e., MCD. ROBPCA is one of the popular methods for the detection of HLPs in high dimensional data. PCA transforms high dimensional data into the low dimensional data set and yields k-dimensional subspace. The MCD is then applied to this low dimensional data set to compute robust location and scatter estimates based on the k-dimensional subspace. The robust score distance (SD) or orthogonal distance (OD) is then computed to identify outliers. The ROBPCA is very successful in the identification of HLPs; however, it suffers from serious swamping effects for less than 30% of HLPs (Siti Zahariah & Habshah 2023). Another shortcoming of this method is that it uses Chi-Squared distribution as the cut-off point for SD based on the assumption that the k-dimensional variables follow a multivariate normal distribution. Nonetheless, in a real situation, there is no guarantee that data would come from a multivariate normal distribution. This cut-off point is inappropriate when the assumption of normality is not met.

In this paper, an attempt is made to compromise between the MRCD-PCA and the ROBPCA. We expect that our proposed method provides the best results in term of having 100% correct detection of HLPs with no masking effect and negligible swamping effect with the least computational running times compared to the MRCD-PCA and ROBPCA.

PRINCIPAL COMPONENT ANALYSIS (PCA)

Hotelling (1933) developed Principal Component Analysis (PCA) as a data reduction method for multidimensional data set. The main goal of PCA is to summarize a data without losing too much information by searching fewer linear combinations of variables. The PCA seeks to explain the variance–covariance structure of a multivariate data set (Hubert, Rousseeuw & Vanden Branden 2005). The standard setting for PCA as an exploratory data analysis method involves a dataset with observations on p numerical variables, for each of n observations. These data values define p n-dimensional vectors, $x_1, \ldots, x_p$ or, equivalently, an $n \times p$ data matrix X. The main idea of PCA is to find a linear combination of the columns of data matrix X known as principal components (PCs), that exhibits the maximum variance. The PCs are arranged so that most of the variation found in all the original variables is retained in the first few (Jolliffe, 1986). The factorization of a data matrix X is given by $X = TP'$, where T is the score matrix and P is an orthogonal matrix known as the loading matrix (Jolliffe 1986). Important PCs are given by the first k columns of P, where $k \leq p$. Then, PCs are obtained with $T^{(k)} = XP^{(k)}$. The proportion of the variance explained by PCs is found by $\left(\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i\right) \times 100$ where λ_i is eigenvalues of variance–covariance matrix.

ROBUST PRINCIPAL COMPONENT ANALYSIS (ROBPCA)

Robust Principal Component Analysis (ROBPCA) introduced by Hubert, Rousseeuw and Vanden Branden (2005), is a statistical technique designed to effectively manage data that contains outliers. It is an extension of the classical Principal Component Analysis (PCA), which is widely used for dimensionality reduction and data compression. The goal of robust PCA method is to obtain principal components that are least affected by outliers. To address the issue of outliers in high-dimensional data, Hubert, Rousseeuw and Vanden Branden (2005) combined the concept of projection pursuit with robust covariance estimation in ROBPCA. The procedure of applying projection pursuit in ROBPCA is to reduce the dimension of high dimensional data into low dimensional dataset. Next, within this low dimensional space, a robust Minimum Covariance Determinant (MCD) estimator is then applied to compute the robust location and scatter estimates to replace the classical covariance matrix. In the PCA space, the principal component scores will be calculated using the robust estimators obtained from MCD.

The robust score distance (SD) and orthogonal distance (OD) are the two distances that are employed in the ROBPCA approach to identify outliers in PCA. Score distance measures how far each observation is from the centroid of the data cloud in the principal component space whereas the orthogonal distance (OD) measures the distance between an observation, x_i and its projection, $\hat{x}_i = \hat{\mu} + p_{p,k}\, l_i$ in the k-dimensional PCA subspace.

The score distance is given as,

$$SD_i = \sqrt{\sum_{j=i}^{k} \frac{(t_i)_j^2}{l_j}} \quad (1)$$

where j is the number of robust principal components; $(t_i)_j$ of the PC score; and l_j is the robust eigenvalues corresponding to the PCs. k denotes the maximum number of PCs.
The OD is defined as,

$$OD_i = \left\| x_i - \mu - P_{p,k}\, t_i \right\| \quad (2)$$

The cut-off value for SD is $\sqrt{\chi^2_{k,0.975}}$ when $k > 1$ and $\pm\sqrt{\chi^2_{1,0.975}}$ when $k = 1$ which are approximately χ^2_k distribution with the assumption that scores are normally distributed. The cut-off value for OD is $(\hat{\mu}_{mcd} + \hat{\sigma}_{mcd} z_{0.975})^{3/2}$, where $z_{0.975}$ equal to 97.5% quantile of the Gaussian distribution.

ROBUST MAHALANOBIS DISTANCE (RMD)

Mahalanobis Distance (Mahalanobis 1936) is widely employed in multivariate analysis to measure the gap between two points with multiple variables (Varmuza & Filzmoser 2009). It is also used for the detection of HLPs. Nevertheless, it is not very successful in identifying of HLPs since it is based on classical mean vector and classical variance covariance matrix of X which is easily affected by outliers. As a remedy to this problem, an alternative approach is to find robust location and scatter estimates that are resistant to HLPs or outlying observations. Rousseeuw (1985) proposed robust Mahalanobis distance (RMD) for the identification of HLPs and it is defined as

$$RMD_i = \sqrt{(x_i - T(x))[C(x)]^{-1}(x_i - T(x))^T}, \quad i = 1,2,3,\ldots,n \quad (3)$$

where robust estimated mean, $T(x)$ and robust covariance matrix, $C(x)$ are employed in (3).

Rousseeuw (1985) suggested using Chi-Squared distribution as the cut-off point for RMD based on the assumption that the k-dimensional variables follow a multivariate normal distribution. However, there is no guarantee that data would come from a multivariate normal distribution. Hence, as per Dhhan, Rana and Midi (2015), Habshah, Norazan and Imon (2009), Midi et al. (2023), and Rashid et al. (2022), since the distribution of RMD is intractable, the following confident bound type of cut-off point is utilised:

$$\text{Cut-off point} = median\,(RMD_i) + 3 * MAD\,(RMD_i) \quad (4)$$

where median absolute deviation (MAD) is defined as

$$MAD\ (RMD_i) = median\ (abs\ (RMD_i - median\ (RMD_i)))/0.6745. \quad (5)$$

for $i = 1, 2, 3\ldots, n$.

Any observations that exceed the cut-off point are declared as HLPs. Many robust methods are available to be used such as the Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD), Index Set Equality (ISE), Reweighted Fast Consistent and High Breakdown (RFCH) estimators (Midi et al. 2020). Nevertheless, all these methods can only be applied to low dimensional data and not applicable to HDD because the covariance matrix of X is not invertible.

MINIMUM REGULARIZED COVARIANCE DETERMINANT (MRCD)

It is noted that the requirement to determine the MCD estimators is that, for any h-subset the number of parameters p must satisfy $p < h$; or else the covariance matrix will be singular. Hence, the scatter matrix of the MCD and other estimators such as the MVE are not invertible. In this situation, the RMD as stated in (3) cannot be used for high dimensional data for the detection of HLPs unless we can find robust covariance matrix that is invertible. To rectify this problem, Boudt et al. (2018) modified the MCD algorithm so that it is invertible and called it Minimum Regularized Covariance Determinant (MRCD). The fundamental objective of MRCD is to substitute a regularized covariance estimate to the MCD subset-based covariance. H-subset of MRCD that minimizes the determinant of K(H) is as follows,

$$H_{mrcd} = \arg\min_{H \in h_h} \left(\det\ K(H)\right)^{1/p} \quad (6)$$

where $K(H)$ represents a regularized covariance matrix in MRCD. It can be written as.

$$K(H) = \rho T + (1-\rho) c_\alpha\ S(H) \quad (7)$$

where T is a predetermined, symmetric and positive definite target matrix, in other words, assume $T = I$, $S(H)$is a sample covariance estimates based on subset H and $\rho=(0,1]$ is regularization intensity parameter. The value of ρ is set such that $K(H)$ is well-conditioned such that $\frac{\lambda_{\max}}{\lambda_{\min}} \leq 1000$. The eigenvalue of MRCD covariance is equal to $\rho+(1-\rho)\lambda$ and the regularization is employed when needed. Then, C-step of Boudt et al. (2018) of MCD is applied until the estimated MRCD covariance converges (Due to space constraint, the detailed steps are not reported here, one can refer to Siti Zahariah and Habshah (2023),

Upon convergence, the following estimates are obtained;
Location estimates of MRCD, $M_{MRCD} = V_x + D_x m_u (H_{MRCD})$
Scatter estimates of MRCD,
$K_{MRCD} = D_x Q \Lambda^{\frac{1}{2}} [\rho I + (1-\rho) S_w (H_{MRCD})] \Lambda^{\frac{1}{2}} Q' D_x$
where $m_u (H_{MRCD})$ is a location estimate based on subset H_{MRCD}. $\chi^2_{p,0.9}$ is used as the cutoff point.

THE MINIMUM REGULARIZED COVARIANCE DETERMINANT BASED ON PRINCIPAL COMPONENT ANALYSIS (MRCD-PCA) FOR THE DETECTION OF HIGH LEVERAGE POINTS

Siti Zahariah and Habshah (2023) proposed robust Mahalanobis distance (RMD) based on the combined methods of the minimum regularized covariance determinant and the principal component analysis (MRCD-PCA). At the outset, the PCA is applied to the original data with the main aim of reducing the high-dimensional data to a low-dimensional data set. Subsequently, a new data set is reconstructed based on selected k principal components by mapping it back to the original high dimensional space. The MRCD is then performed to this newly fitted data to determine the robust mean and robust covariance of high-dimensional data. To identify HLPs, they calculated the Robust Mahalanobis Distance (RMD) for each observation.

$$RMD_i\ (MRCD-PCA) = \sqrt{(x_i - \hat{\mu}_{mrcd-pca})^T\ \Sigma^{-1}_{mrcd-pca}\ (x - \hat{\mu}_{mrcd-pca})} \quad (8)$$

where $\mu_{mrcd-pca}$ and $\Sigma^{-1}_{mrcd-pca}$ are the robust location and parameter estimates of MRCD-PCA, respectively. Finally, robust cut-off point is used to identify high leverage points.

The MRCD-PCA technique can be summarized as follows:

Step 1 : Construct centered data matrix X by subtracting median of each column x_j from each observation x_{ij};

$$x_{ij} - median\ (x_j) \quad (9)$$

Step 2 : By using the PCA method, the dimension of the centered data matrix will be reduced. The number of principal components k is chosen based on the Scree plot or Cumulative variance in which the first k loadings $\geq$ 80%(Cao 2006). The $n \times k$ matrix of PCA projections (scores) can be written as $Z = XV$ where V is the $p \times k$ matrix (eigenvector matrix).

Step 3 : The original data is reconstructed based on these k principal components, and map it back to p dimensions as follows;

$$\hat{X} = Z V^T \quad (10)$$

Step 4 : The minimum regularized covariance determinant is then performed on the fitted data, $\hat{X}$ to determine the robust mean and robust covariance estimator for high dimensional data.

Step 5 : Finally, compute the Robust Mahalanobis distance for each observation based on the estimated mean and the estimated covariance matrix of MRCD-PCA.

Step 6 : The following cut-off point is employed;

$$median\,(RMD_{mrcd-pca}) + 3MAD(RMD_{mrcd-pca}) \quad (11)$$

An observations that exceed the cut-off point are declared as HLPs.

THE PROPOSED METHOD FOR THE DETECTION OF HLPS: IMPROVED ROBUST PRINCIPAL COMPONENT ANALYSIS (IRPCA)

As already discussed in the introduction section, ROBPCA can correctly identify outliers. However, it suffers from a serious swamping effect especially for high dimensional data. Hence, we propose to improve the ROBPCA so that the swamping effect can be reduced. The proposed method that we call IRPCA combines the idea of principal component analysis (PCA) and Minimum Regularized Covariance Determinant (MRCD) whereby it is simple to implement and takes less computation running times. This involved transforming a high-dimensional space into a lower-dimensional subspace by using PCA and subsequently conducting our work within this newly established principal component subspace. Then, the Minimum Regularized Covariance Determinant (MRCD) is applied to this newly derived low dimensional subspace to obtain the location and scatter matrix. Instead of using robust score distance (SD) or orthogonal distance (OD) to identify outliers in HDD as suggested by Hubert, Rousseeuw and Vanden Branden (2005), we propose using Robust Mahalanobis distance (RMD) to detect HLPs and suggest a confident bound type of cut-off point.

The IRPCA technique can be summarized as follows:

Step 1 : Center the data by subtracting the median of each column x_j from each observation x_{ij}

$$x_{ij} - median\,(x_j) \quad (12)$$

Step 2 : Apply Principal Component Analysis (PCA) to the centered data to reduce from the original p variables into k dimensional subspace where $k << p$. The number of dimensions k retained is based on the Scree plot or Cumulative Variance.

Step 3 : Project the data points on the k-dimensional subspace and obtain the principal component score where the score are the entries of $n \times k$ matrix

$$Tn,k = (Xn,p - 1n\,\hat{\mu}')Pp,k \quad (13)$$

Step 4 : where $P_{p,k}$ consists of the first k columns of $P_{p,p}$

Estimate the robust scatter matrix of the principal component score within k-dimensional subspace using the Minimum Regularized Covariance Determinant (MRCD) estimator. The robust location and scatter estimates are indicated as $\hat{\mu}_{IRPCA}$ and $\sum_{IRPCA}$, respectively.

Step 5 : Calculate Robust Mahalanobis Distance (RMD) for each observation of the HDD based on the robust location and scatter estimates obtained from Step (4). The RMD of the proposed method is given by

$$RMD_i(IRPCA) = \sqrt{(x_i - \hat{\mu}_{IRPCA})^T \Sigma_{IRPCA}^{-1} (x_i - \hat{\mu}_{IRPCA})} \quad (14)$$

Step 6 : Calculate the cut-off point to identify HLPs. Since the distribution of RMD_i $(IRPCA)$ is intractable, as per Habshah, Norazan and Imon (2009), Rashid et al. (2022) and Siti Zahariah and Habshah (2023), the confident bound type of cutoff point for RMD_i $(IRPCA)$ is employed as follows,

$$median\,(RMD_{IRPCA}) + 3MAD\,(RMD_{IRPCA}) \quad (15)$$

Any observations such that its RMD_i $(IRPCA)$ exceeds the cut-off point are declared as HLPs.

SIMULATION STUDY

A simulation study similar to that of Boudt et al. (2018) and Siti Zahariah and Habshah (2023) was conducted to assess the performance of our proposed IRPCA method. We generated two sample sizes of $n = 50$ and $n = 100$ from a p-variate normal distribution with four different sizes of $p = 100$, 200, 300, and 500 for each sample size. In this simulation study, we compare our proposed method of IRPCA with MRCD-PCA and ROBPCA. Since the PCA, MRCD-PCA and ROBPCA estimators are location and scale equivariant, as per Agostinelli et al. (2015), we make general assumption that the mean is 0, and that the diagonal element of variance are all equal to 1. Since our proposed method of IRPCA applied the algorithm of MRCD, we account for the lack of affine equivariance of our proposed estimator with a similar manner to that of Agostinelli et

al. (2015), by generating different correlation structures. To ensure that the generated correlation matrix is within a tolerance interval around 100, a condition number is fixed at 100. The dataset was set up as to contain both clean and contaminated observations. The clean observation was generated from $x_i \sim N_p(0, I)$ for $i = 1,2,3,..., n - m$. For the contaminated data sets, we follow Maronna and Zamar (2002) by randomly replace $\lfloor \varepsilon n \rfloor$ observations with outliers along the direction of the eigenvector of $\sum$ with the least eigenvalue, since this is the direction where the contamination is most difficult to detect. For contamination model, we generated $x_i \sim N_p(y_0, \delta^2 I)$ for $i > n - m$, where $y_0 = k\, a_0$ and a_0 is the eigenvector to the smallest eigenvalue of Σ. k is denoted as the distance between the outliers and the mean of the good data. In our simulation study, we choose a medium-sized outlier contamination and hence k is set at 50. Various contaminated fractions are considered, i.e., 5%, 10%, 20% and 30%. While Maronna and Zamar (2002) approach primarily focuses on estimating the robust location and scale estimates in the presence of outliers within the dataset, however, they do not address high leverage points (HLPs) which could do more damage to the statistical analysis. In contrast, we aimed to compare our proposed method with MRCD-PCA and ROBPCA in terms of percentage of HLP detected, the swamping and masking effect, and the computation running time. Tables 1-2 present the percentage of high leverage points (HLP) detected and the percentage of masking and swamping for $n = 50$ and $n = 100$. It can be observed from Tables 1-2 that the results of the current study, i.e., IRPCA and the previous study, i.e., MRCD-PCA and ROBPCA, successfully identify all HLPs with zero masking effect regardless of the outlier's contamination percentage, sample size and number of variables. The results also signify that the swamping effects of the ROBPCA, MRCD-PCA, and IRPCA are slightly decreasing as the percentage of HLPs and sample size increases. However, ROBPCA suffers from serious swamping effects for less than 30% HLPs. It is interesting to see that at 30% of HLPs, the swamping effect of ROBPCA tends to be very small and its values are fairly closed to the MRCD-PCA and IRPCA. On the other hand, the swamping effects of both the MRCD-PCA and our proposed IRPCA methods are relatively very small and negligible compared to the ROBPCA irrespective of the percentage of HLPs, number of predictor variables and sample size. The percentage of swamping effects of the MRCD-PCA is very small and reasonably closed to the IRPCA. Nonetheless, we will illustrate later that the MRCD-PCA suffers from longer computation running times, which is undesirable.

We have seen that the results of the simulation study indicate that the performances of our proposed IRPCA and MRCD-PCA are equally good and the ROBPCA performs poorly for less than 30% outliers. We further investigated the properties of the three methods by considering their computational running times. It is important to note that, based on our experience, it took more than an hour to run the simulation for MRCD-PCA for just one dimension, one sample size, and one level of contamination for a large p, i.e., $p = 3{,}000$. This is why we do not include results for $p >> 500$. Table 3 exhibits the running times for the three methods at various level of contaminations, dimensions and sample size. The results of Table 3 show that as the number of predictor variables and sample size increases, the running times for all methods tend to increase. It is interesting to see that the computation running time of the current study (IRPCA) is the shortest compared to the previous study (MRCD-PCA and ROBPCA). It should be noted that the running time for MRCD-PCA is much higher than the IRPCA and ROBPCA. This is due to the fact that the MRCD-PCA approach is more difficult to compute; it requires larger computer storage because it uses PCA as a tool to reduce from high dimensional to low dimensional data, and then mapped again to the original number of dimensions, p. Then, robust location and scatter estimates are obtained from this high dimensions data. Hence, the calculation of robust distance for all these observations requires more time before the HLPs can be discovered. Essentially, the IRPCA is preferred over the MRCD-PCA and ROBPCA methods by virtue of its good performance in terms of having 100% detection rate, no masking effect and very small swamping effect. Moreover, this method is computationally easy and takes the least computational running times.

REAL EXAMPLE 1

The octane data described in Esbensen et al. (1994), was used to further evaluate the performance of our proposed IRPCA compared to the MRCD-PCA and ROBPCA methods. This high dimensional data consists of near-infrared absorbance spectra over $p = 226$ wavelengths of $n = 39$ gasoline samples with certain octane numbers. According to Hubert, Rousseeuw and Vanden Branden (2005), observations 25, 26, 36, 37, 38, 39 are outliers which contain added alcohol. Since the number and position of outliers of this data are exactly known, it has been used by many researchers (Hubert, Rousseeuw & Vanden Branden 2005; Rashid et al. 2022; Siti Zahariah & Habshah 2023) to detect HLPs in high dimensional data. Hence, we applied the IRCPA, MRCD-PCA, and ROBPCA methods to this dataset to illustrate the merit of our proposed IRPCA method. Based on the scree plot of the classical PCA and ROBPCA of Hubert, Rousseeuw and Vanden Branden (2005), only two principal components are retained. The MRCD algorithm is applied to these two principal components score to determine the robust location and robust scatter estimators. To identify the HLPs of the data set, we calculate the robust Mahalanobis distance on the score of the two principal components. Our proposed method has successfully identified observations 25, 26, 36, 37, 38, and 39 as HLPs. Similarly, the MRCD-PCA and the ROBPCA also able to detect the six outliers. The computing

TABLE 1. Percentage of correct detection of HLPs, masking and swamping by MRCD-PCA, IRPCA, and ROBPCA, $n = 50$

Contamination (%)	p	% of correct detection			% of masking			% of swamping		
		MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA
5 (3 outliers)	100	100	100	100	0	0	0	0.812	0.916	7.372
	200	100	100	100	0	0	0	0.900	0.912	7.848
	300	100	100	100	0	0	0	0.976	1.068	7.776
	500	100	100	100	0	0	0	1.200	1.020	8.428
10 (5 outliers)	100	100	100	100	0	0	0	0.464	0.636	5.784
	200	100	100	100	0	0	0	0.436	0.608	6.684
	300	100	100	100	0	0	0	0.472	0.652	7.228
	500	100	100	100	0	0	0	0.340	0.736	7.608
20 (10 outliers)	100	100	100	100	0	0	0	0.136	0.164	2.292
	200	100	100	100	0	0	0	0.176	0.180	3.508
	300	100	100	100	0	0	0	0.164	0.232	4.504
	500	100	100	100	0	0	0	0.168	0.248	5.832
30 (15 outliers)	100	100	100	100	0	0	0	0.020	0.024	0.056
	200	100	100	100	0	0	0	0.044	0.020	0.172
	300	100	100	100	0	0	0	0.032	0.028	0.064
	500	100	100	100	0	0	0	0.036	0.024	0.256

TABLE 2. Percentage of correct detection of HLPs, masking and swamping by MRCD-PCA, IRPCA, and ROBPCA, $n = 100$

Contamination (%)	p	% of correct detection			% of masking			% of swamping		
		MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA
5 (5 outliers)	100	100	100	100	0	0	0	0.522	0.528	5.866
	200	100	100	100	0	0	0	0.496	0.544	6.618
	300	99.92	100	100	0.08	0	0	0.642	0.550	7.046
	500	99.2	100	100	0.8	0	0	0.620	0.564	7.326
10 (10 outliers)	100	100	100	100	0	0	0	0.122	0.276	4.378
	200	100	100	100	0	0	0	0.182	0.250	5.090
	300	100	100	100	0	0	0	0.160	0.182	5.636
	500	100	100	100	0	0	0	0.156	0.160	6.290
20 (20 outliers)	100	100	100	100	0	0	0	0.040	0.044	1.470
	200	100	100	100	0	0	0	0.028	0.038	2.238
	300	100	100	100	0	0	0	0.018	0.028	3.042
	500	100	100	100	0	0	0	0.028	0.034	4.550
30 (30 outliers)	100	100	100	100	0	0	0	0.004	0	0.014
	200	100	100	100	0	0	0	0	0	0.004
	300	100	100	100	0	0	0	0.004	0	0.024
	500	100	100	100	0	0	0	0.002	0.002	0.030

TABLE 3. Running time for simulated data (in seconds), $n = 50$ and $n = 100$ (in parentheses)

Contamination (%)	p	Running time (in seconds)		
		MRCD-PCA	IRPCA	ROBPCA
5	100	1.23140 (1.43447)	0.06425 (0.06204)	0.089023 (0.08887)
	200	4.23890 (4.73505)	0.18848 (0.16347)	0.21462 (0.18677)
	300	10.31759 (16.08993)	0.42762 (0.42292)	0.51335 (0.44525)
	500	31.09337 (49.34740)	1.65628 (1.61324)	1.60271 (1.86286)
10	100	1.0386 (1.70448)	0.0772 (0.07008)	0.1073 (0.08988)
	200	4.1551 (5.40792)	0.2022 (0.15996)	0.2249 (0.22236)
	300	8.5156 (14.28192)	0.3968 (0.4062)	0.4958 (0.52776)
	500	28.5395 (38.26656)	1.5013 (1.56276)	1.6135 (1.53156)
20	100	1.0403 (1.64472)	0.0658 (0.08916)	0.1120 (0.09996)
	200	4.1246 (6.31812)	0.1686 (0.17592)	0.1700 (0.1908)
	300	9.0696 (14.27124)	0.4470 (0.41952)	0.3943 (0.429)
	500	26.3369 (38.9274)	1.8042 (1.58388)	1.8442 (1.90452)
30	100	1.0104 (1.28568)	0.0631 (0.08628)	0.0808 (0.09432)
	200	3.2407 (5.87064)	0.1691 (0.17976)	0.1634 (0.20064)
	300	8.6417 (15.5622)	0.4580 (0.41892)	0.5107 (0.4254)
	500	27.3748 (40.48008)	1.5426 (1.57056)	1.9333 (1.57224)

time for IRPCA, MRCD-PCA, and ROBPCA are 3.593 s, 13.940 s, and 3.090 s, respectively. It should be noted that the computational running time for the MRCD-PCA is much higher than the IRPCA and ROBPCA for this data set. According to Rousseeuw and Van Zomeren (1990), to avoid the curse of dimensionality, it is recommended to set $n > 5k$; therefore, following Hubert, Rousseeuw and Vanden Branden (2005), we considered $k = 7$. While Hubert, Rousseeuw and Vanden Branden (2005) verified that ROBPCA can still detected all the six outliers with $k = 7$, however they did not address the issue of swamping effect in their research study. Both IRPCA and MRCD-PCA yielded the same results as $k = 2$. On the other hand, the ROBPCA can detect the six outliers, but also flagged clean observations 3 and 7 as outliers. The results of octane data were consistent with the results of the simulation study, where our proposed IRPCA successfully detect HLPs with no masking effect, negligible swamping effect with the least computational running time. On the other hand, the shortcomings of ROBPCA and MRCD-PCA are that it suffers from serious swamping effect and has longer computational running times, respectively.

REAL EXAMPLE 2

Our second example is the EPXMA spectra data taken from Lemberge et al. (2000). This data set consists of $p = 750$ wavelengths collected over 180 archaeological glass samples. The chemical analysis was conducted using a Jeol JSM 6300 scanning electron microscope equipped with an energy-dispersive Si(Li) X-Ray detection system (SEM-EDX). This data has a sparse structure and is high dimensional. Since the exact position and number of outliers of this data is known, many researchers used this data to validate their diagnostic methods of identification of HLPs (for instance, Hubert, Rousseeuw & Vanden Branden 2005; Siti Zahariah & Habshah 2023). As for the glass spectra data, the scree plot suggested to keep four principal components which explained about 99.5% of the variance of the entire dataset. It can be observed from Figure 1 that our proposed IRPCA method, clearly separates the HLPs into 2 major groups above the Cut-off line. Our method detected observations 143 – 180 as one set of HLPs and observations (58- 63, 74, and 76) as the second group of HLPs. Identical results are obtained for MRCD-PCA. However, the computational time for IRPCA is much faster than the MRCD-PCA where IRPCA took 4.551 s while MRCD-PCA took 160.290 s to pinpoint HLPs in the glass dataset. Conversely, the results of ROBPCA in Figure 2 show that apart from identifying HLPs from the 2 major groups, it also treats observation 57 and 75 as HLPs due to their clustering with the HLPs. According to Hubert et al. (2015), the algorithm of ROBPCA generates robust but non-sparse loadings, hence, it cannot handle sparse data. Moreover, as already discussed earlier, the ROBPCA suffers from swamping effect due to using MCD in its algorithm. However, the outcome of the glass spectra dataset indicates that our proposed IRPCA method, can successfully detected HLPs under this type of data with the least computational running times.

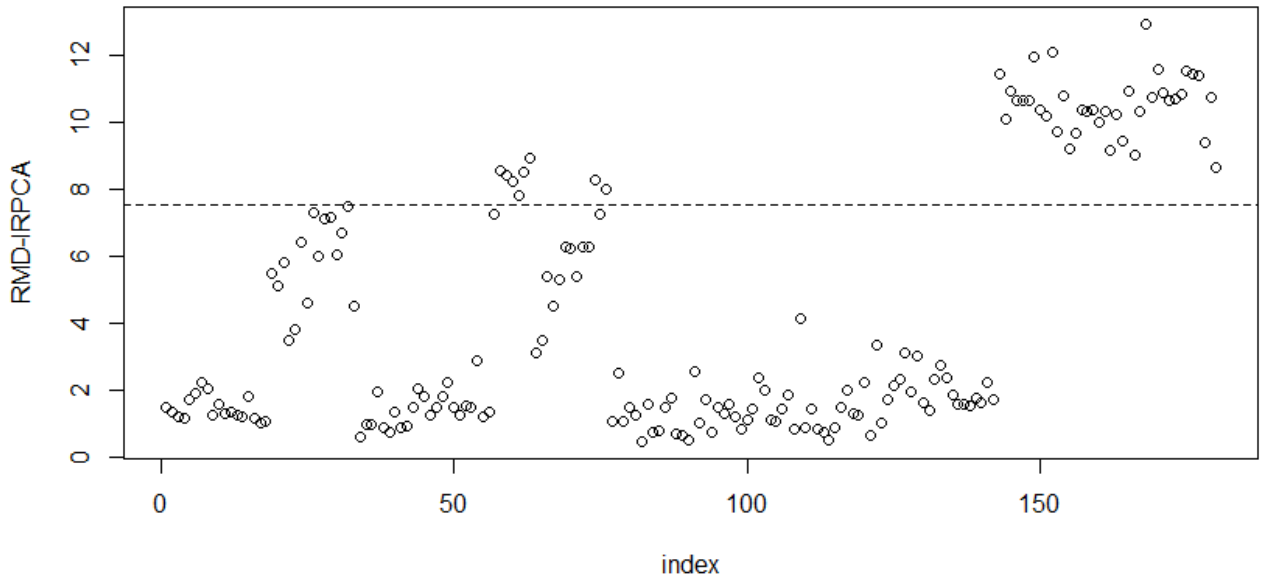

FIGURE 1. Index plot of Glass Spectra data set based on RMD-IRPCA

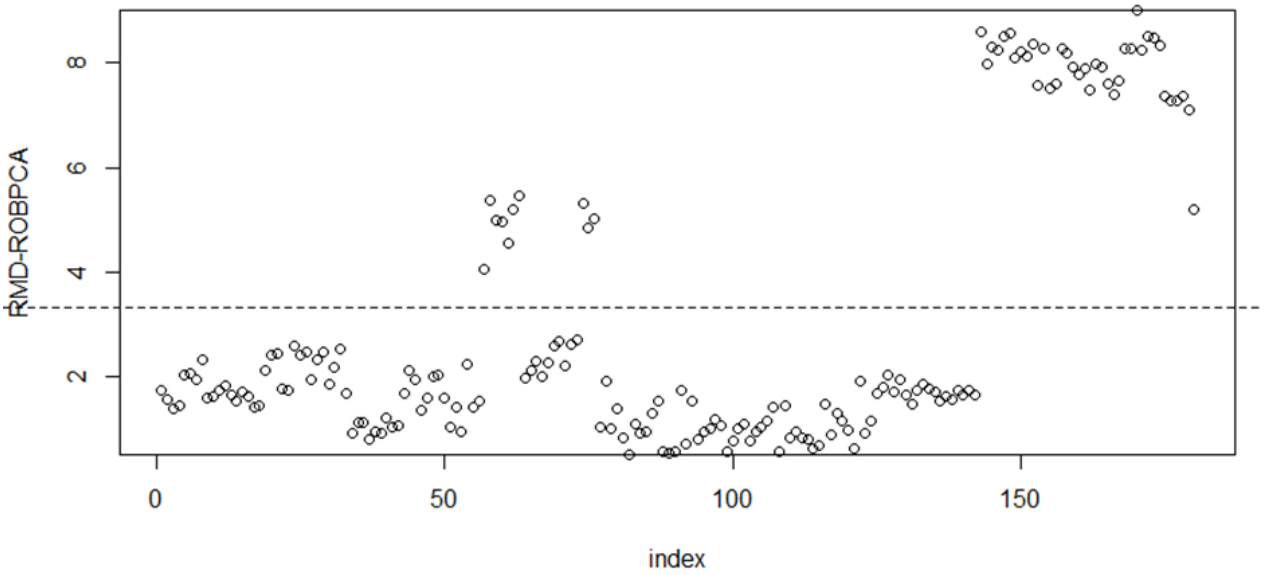

FIGURE 2. Index plot of Glass Spectra data set based on ROBPCA

REAL EXAMPLE 3

Fish oil data taken from Killen et al. (2019) is our third example to provide a preliminary validation of the limitation of the MRCD-PCA in term of having very long computation running times for huge dimensions, i.e., p=3471 and n = 126. These results serve as an indicator of the expected performance of the method for higher-dimensional settings, even though they are not directly derived from the full-scale simulation study due to overly long computation time. For this data, the IRPCA, MRCD-PCA, and ROBPCA detected 9, 10, and 14 HLPs, respectively. ROBPCA shows a swamping effect. It is interesting to note that the computation times for IRPCA are 4 s, for ROBPCA are 7.68 s, and for MRCD-PCA are 2.5 h. The lengthy processing time of MRCD-PCA for detecting HLPs makes this method unattractive, even though it can detect the correct number of HLPs, just like IRPCA.

CONCLUSION

This article provides another procedure of detecting HLPs in HDD that we call IRPCA. The proposed IRPCA methods and two existing methods namely the ROBPCA and MRCD-PCA are very successful in identifying HLPs. However, the empirical study shows that the ROBPCA suffers from

severe swamping effect for less than 30% HLPs. The IRPCA and MRCD-PCA methods are indistinguishable in terms of correct detection of outliers, no masking effects and having very small swamping effects. Nonetheless, the MRCD-PCA algorithm is not straight forward, was somewhat computationally cumbersome and, it takes very long computational running times. On the other hand, the IRPCA algorithm is quite simple and its computational running time is much faster than the MRCD-PCA. The results seem to suggest that the IRPCA may provide the most attractive diagnostic method for identifying HLPs in HDD.

It is worth mentioning that in this study, we sought to validate the proposed IRCPA against the previous methods (MRCD-PCA and ROBPCA) across various values of p, with a particular focus on larger dimensions, such as p=3000, to demonstrate its scalability through a simulation study. However, due to the significant computational demands associated with running simulations for such large values of p for MRCD-PCA, we were unable to complete the full simulation study for p=3000 unless a high-performance computer was used. Running simulations at this scale requires substantial computational resources and time, which limited our ability to conduct these experiments within the scope of such a large dataset.

REFERENCES

Agostinelli, C., Leung, A., Yohai, V.J. & Zamar, R.H. 2015. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24(3): 441-461. https://doi.org/10.1007/s11749-015-0450-6

Boudt, K., Rousseeuw, P.J., Vanduffel, S. & Verdonck, T. 2018. The minimum regularized covariance determinant estimator. *Statistics and Computing* 30: 113-128. https://doi.org/10.1007/s11222-019-09869-x

Boulesteix, A.L. & Strimmer, K. 2007. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* 8(1): 32-44. https://doi.org/10.1093/bib/bbl016

Cao, L. 2006. *Singular Value Decomposition Applied to Digital Image Processing.* Division of Computing Studies, Arizona State University. pp. 1-15. http://www.lokminglui.com/CaoSVDintro.pdf

Chiang, J-T. 2016. The masking and swamping effects using the planted mean-shift outliers models. *International Journal of Contemporary Mathematical Sciences* 2(7): 297-307. https://doi.org/10.12988/ijcms.2007.07024

Dhhan, W., Rana, S. & Midi, H. 2015. Non-sparse ε-insensitive support vector regression for outlier detection. *J. Appl. Stat.* 42: 1723-1739.

Esbensen, K.H., Sch¨onkopf, S., Midtgaard, T. & Guyof, D. 1994. *Multivariate Analysis in Practice.* Camo, Trondheim.

Habshah, M., Norazan, M.R. & Imon, A.H.M.R. 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics* 36(5): 507-520. https://doi.org/10.1080/02664760802553463

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6): 417-441. https://doi.org/10.1037/h0071325

Huber, P.J. 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* 1(5): 799-821.

Hubert, M., Rousseeuw, P.J. & Verdonck, T. 2012. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics* 21(3): 618-637. https://doi.org/10.1080/10618600.2012.672100

Hubert, M., Rousseeuw, P.J. & Vanden Branden, K. 2005. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 47(1): 64-79. https://doi.org/10.1198/004017004000000563

Hubert, M., Reynkens, T., Schmitt, E. & Verdonck, T. 2015. Sparse PCA for high-dimensional data with outliers. *Technometrics* 58(4): 424-434. https://doi.org/10.1080/00401706.2015.1093962

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer Series in Statistics. Berlin: Springer.

Killeen, D.P., Card, A., Gordon, K.C. & Perry, N.B. 2019. First use of handheld Raman spectroscopy to analyze omega-3 fatty acids in intact fish oil capsules. *Applied Spectroscopy* 74(3): 365-371.

Lemberge, P., De Raedt, I., Janssens, K.H., Wei, F. & Van Espen, P.J. 2000. Quantitative analysis of 16-17th century archaeological glass vessels using PLS regression of EPXMA and μ-XRF data. *Journal of Chemometrics* 14(5-6): 751-763. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<751

Lim, H.A. & Midi, H. 2016. Diagnostic robust generalized potential based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics* 31: 859-877.

Midi, H., Hendi, T.H., Uraibi, H., Arasan, J. & Ismaeel, S.S. 2023. An efficient method of identification of influential observations in multiple linear regression and its application to real data. *Sains Malaysiana* 52(12): 3879-3892.

Midi, H., Ismaeel, S.S., Arasan, J. & Mohammad, A.M. 2021. Simple and fast generalized-M (GM) estimator and its application to real data. *Sains Malaysiana* 50(3): 859-867.

Midi, M., Talib, H., Jayanthi, A. & Uraibi, H.S. 2020. Fast and robust diagnostic technique for the detection of high leverage points. *Journal of Science and Technology* 28(4): 1203-1220.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 2(1): 49-55.

Maronna, R.A. & Zamar, R.H. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44(4): 307-317. https://doi.org/10.1198/004017002188618509

Rana, M.S., Midi, H. & Imon, A.H.M.R. 2009. A robust rescaled moment test for normality in regression. *Journal of Mathematics and Statistics* 5(1): 54-62.

Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2021. An efficient estimation and classification methods for high dimensional data using robust iteratively reweighted SIMPLS algorithm based on Nu-support vector regression. *IEEE Access* 9: 45955-45967.

Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2022. Detection of outliers in high-dimensional data using Nu-support vector regression. *Journal of Applied Statistics* 49(10): 2550-2569.

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* 8: 37.

Rousseeuw, P. & Driessen, K. 1999. A fast algorithm for the minimum covariance. *Technometrics* 41(3): 212-223.

Rousseeuw, P.J. & Van Zomeren, B.C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85: 633-651.

Siti Zahariah & Habshah Midi. 2023. Minimum regularized covariance determinant and principal component analysis - based method for the identification of high leverage points in high dimensional sparse data. *Journal of Applied Statistics* 50(13): 2817-2835.

Siti Zahariah, Habshah Midi & Mohd Shafie Mustafa. 2022. An improvised SIMPLS estimator based on MRCD-PCA weighting function and its application to real data. *Symmetry* 13(11): 2211.

Varmuza, K. & Filzmoser, P. 2009. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton: CRC Press. doi:10.1201/9781420059496

*Corresponding author; email: habshah@upm.edu.my