International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:3, 2012

The Robust Clustering with Reduction Dimension

Dyah E. Herwindiati

***Abstract*—**A clustering is process to identify a homogeneous groups of object called as cluster. Clustering is one interesting topic on data mining. A group or class behaves similarly characteristics. This paper discusses a robust clustering process for data images with two reduction dimension approaches; i.e. the two dimensional principal component analysis (2DPCA) and principal component analysis (PCA). A standard approach to overcome this problem is dimension reduction, which transforms a high-dimensional data into a lower-dimensional space with limited loss of information. One of the most common forms of dimensionality reduction is the principal components analysis (PCA). The 2DPCA is often called a variant of principal component (PCA), the image matrices were directly treated as 2D matrices; they do not need to be transformed into a vector so that the covariance matrix of image can be constructed directly using the original image matrices. The decomposed classical covariance matrix is very sensitive to outlying observations. The objective of paper is to compare the performance of robust minimizing vector variance (MVV) in the two dimensional projection PCA (2DPCA) and the PCA for clustering on an arbitrary data image when outliers are hiden in the data set. The simulation aspects of robustness and the illustration of clustering images are discussed in the end of paper

***Keywords*—**Breakdown point, Consistency, 2DPCA, PCA, Outlier, Vector Variance

I. Introduction

CLUSTERING is one common technique for statistical data analysis used in many fields. A clustering is process to identify a homogeneous groups of object called as cluster. A group or class behaves similarly characteristics. The clustering algoritms are generally classified into hierachical and non hierarchical algorithms. This paper discusses robust non hierarchical clustering process for data images with reduction dimension. A cluster of image is built from robust distance; which is measured from central location observation.

Reduction dimension has been used widely in many application involving high dimensional data, such as application on image processing. The digital number or value of image pixels have loaded resemble character to the near neighbour pixels. It means that the one of variables can be written as a near linear combination of the other variables, and the dispersion of data is close to singularity problem. A standard approach to overcome this problem is dimension reduction, which transforms a high-dimensional data into a lower-dimensional space with limited loss of information.

The author is lecturer at Tarumanagara University, Jln Let. Jend. S Parman 1, Jakarta 1140, Indonesia. (e-mail: herwindiati@untar.ac.id)).

This paper talks deal with a robust clustering process for data images with two reduction dimension approaches; i.e. the two dimensional principal component analysis (2DPCA) and principal component analysis (PCA). One of the most common forms of dimensionality reduction is the principal components analysis (PCA), see Jolife [5]. A principal component analysis focused on reducing the dimensionality of a data set in order to explain as much information as possible. The first principal component is the combination of variables that explains the greatest amount of variation. The second principal component is defined as the next largest amount of variation and is independent to the first principal component. This step will be continued for the entire principal components corresponding to the eigenvectors of covariance matrix sample. One disadvantage of PCA is the elaborate computation.

Yang et.al [6] proposed the application of two dimensional Principal Component (2DPCA) for reducing of computational time of standard PCA on face recognition. The 2DPCA is often called a variant of principal component (PCA). In the 2DPCA, the image matrices were directly treated as 2D matrices; the images do not need to be transformed into a vector so that the covariance matrix of image can be constructed directly using the original image matrices. The 2DPCA has two important benefits over PCA, it is easier to evaluate the covariance matrix and it has the less time for determining the eigenvectors.

The decomposed information variation of classical PCA and 2DPCA becomes pointless if outliers are present in the data. The decomposed classical covariance matrix is very sensitive to outlying observations. The first component consisting of the greatest variation is often pushed toward the anomalous observations.

Robust statistics a convinient modern way of summarising result when outliers are hidden in the data set. Outlier is often difficult to be identified through visual inspection without the analytic tools. There are many different robust estimator of location estimator. In this paper we discuss the robust estimator of minimum vector variance (MVV). The objective of paper is to compare the performance of robust minimizing vector variance (MVV) in the two dimensional projection PCA (2DPCA) and the PCA for clustering of the arbitrary data image. Minimum vector variance (MVV) is the robust measure in an attempt to determine the location estimator and covariance matrix based on a data subset covering approximately an half data which give the minimum vector variance, see Herwindiati et. al [1]. The algorithm of two methods and the clustering cases are comprehensively discussed. The aim of paper is to give the explanations and

comparison of robust minimizing vector variance (MVV) in the two dimensional projection PCA (2DPCA) and the PCA for clustering of the arbitrary data image

II. The Robust Principal Component Analysis

The main idea of principal component analysis (PCA) is to reduce the dimensionality of data set consisting of alarge number of interrelated variable, while retaining as much as possible of variation in the data set, see Jolife [4]. In the image processing, PCA is the statistical teqnique useful to find pattern in data image of high dimension.

Suppose that the random vector $\vec{X}$ of p components has the classical covariance matrix S which is a $p \times p$ symmetric and positive semi definite.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Covariance matrix S has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and eigenvector U such that $U'SU = L$; where L is diagonal matrix.

The principal components are uncorrelated linear combinations $\vec{Y}$ whose variances are as large as possible. The first principal component is given by $\vec{Y}_1 = \vec{U}_1' X$ which has the largest proportion of total variance. Tecnically, a pricipal component can be defined as a linear combination of optimally-weighted observed variable.

The proportion of total variance the k principal component is often explained by the ratio of the eigenvalues $\lambda_k = \sum_{i=i}^{k} \lambda_i$. The determination of k is an important role to the PCA analysis. A larger k gives a better fit in PCA, but a larger k has the larger redundancy of information. The replacement of original variable p to the k principal component must be considered as a goal in optimizing. The decomposed classical covariance matrix S is very sensitive to outlying observations. The k principal component becomes unreliable if outliers are present in the original variable p. The k principal component consisting of the largest proportion of total variance S is often pushed toward the outliers.

Regarding the fact, Huber et al [8] introduced a new method for robust principal component (ROBPCA). ROBPCA is PCA method combining two advantages of both projection pursuit and robust covariance estimation. The robust estimator is computed by the MCD ideas of covariance matrix. Based on our experience in computations, ROBPCA is an effective and efficient method. Herwindiati and Isa [2] proposed the robust principal component minimizing vector variance (MVV) based on the good properties of ROBPCA.

The MVV robust PCA is an impressive method for interpreting the application of PCA, such as the classification and the clustering process. The algorithm of MVV robust PCA is composed with three stages generally listed as follows,

Stage 1. Start with a singular value decomposition of the mean centered data matrix

Stage 2. Estimate the location and covariance matrix using MVV robust approach.

1. Let H_{old} be an arbitrary subset containing $h = \left[\frac{n+k+1}{2} \right]$ data points. Compute the mean vector $\bar{\vec{X}}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old} . Then compute,

   $$d^2_{H_{old}}(i) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}} \right)^t S^{-1}_{H_{old}} \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}} \right)$$

   for all i = 1, 2, … , n
2. Sort these distances in increasing order,
3. Define H_{new} from the order distance H_{new} = $\left\{ \vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \cdots, \vec{X}_{\pi(h)} \right\}$
4. Calculate $\bar{\vec{X}}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$.
5. If $Tr\left(S^2_{H_{new}}\right)$ = 0, repeat steps 1 to 5 and. The process is stopped If $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$,

   Otherwise, the process is continued until the *k*-th iteration if

   $$Tr\left(S^2_1\right) \geq Tr\left(S^2_2\right) \geq Tr\left(S^2_3\right) \geq \cdots \geq Tr\left(S^2_k\right) = Tr\left(S^2_{k+1}\right)$$

Stage 3. Do the clustring images using the MVV robust squared Mahalanobis distance defined as,

$d^2_{MVV}\left(\vec{X}_i, \vec{T}_{MVV}\right) = \left(\vec{X}_i - \vec{T}_{MVV}\right)^t S^{-1}_{MVV}\left(\vec{X}_i - \vec{T}_{MVV}\right)$; for all i = 1, 2, … , *n*. and $\vec{T}_{MVV}$ and S_{MVV} are the location and covariance matrix given by that process.

The Subset *h* in the first stage has the important role in the estimator. Hubert et al [8] suggested to take subset $h = \max\left\{ [\alpha n], \left[(n + k_{\max} + 1)/2 \right] \right\}$, where α is chosen as any real value between 0.5 and 1, $k_{\max}$ as a maximal number of components that will be computed, however Rousseeuw and van Driessen [10] stated that the subset $h = \left[\frac{n+k+1}{2} \right]$ has the high breakdown point estimator.

Breakdown point is the smallest fraction of data which causes the value of estimator to be infinity when the value of all data in the fraction are changed to be infinity, Rousseeuw and Leroy [9]. The good robust estimator must be high breakdown point. The higher breakdown point estimator means the more resistant estimator to agains the contaminant data.

Two subsets; $h_1 = \max\{[\alpha n], [(n+k_{\max}+1)/2]\}$ and $h_2 = \left[\frac{n+k+1}{2}\right]$; are simulated to compare the breakdown points as seen in the Figure 1 and Figure 2. Those figures reveal the fact that the breakdown point of h_2 is higher and more stable than the one of h_2.

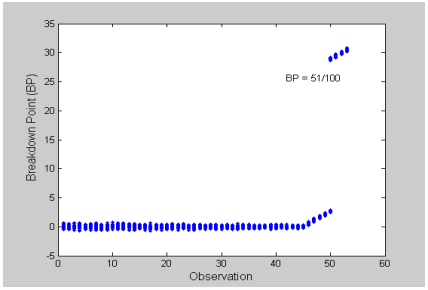


Fig. 1 MVV Breakdown point using $h = \left[\frac{n+k+1}{2}\right]$

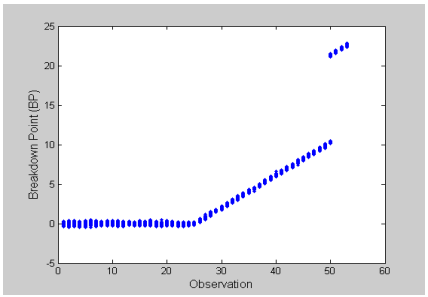
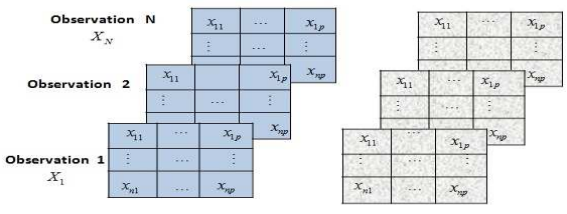


Fig. 2 MVV Breakdown point using $h = 0.75\,n$

III. The Robust Two Dimension Principal Component (Robust 2DPCA)

Two dimensional Principal Component (2DPCA) was proposed by Yang et. al [6]. The method using the projection technique is developed for the gray scale face recognition. Though the 2DPCA is often called as a variant of principal component (PCA), the 2DPCA has two important benefits over PCA. It is easier to evaluate the covariance matrix and it has the less time for determining the eigenvectors. In the 2DPCA, the image matrices are directly treated as 2D matrices; the images need not be transformed into a vector so that the covariance matrix of the image can be constructed directly using the original image matrices.



Class *1*.....Class *k*

Fig. 3 The Illustration of k Clustering data Images

Consider $X_1, X_2, \cdots, X_N$ is a $m \times p$ random image matrix, let $\vec{V}$ is an p dimensional unitary column vector, the idea of 2DPCA is to project X onto $\vec{V}$ by linear transformation

$$\vec{Y} = X\vec{V} \tag{1}$$

Define the image covariance matrix:

$S_M = E\left[(X - EX)^T (X - EX)\right]$ which is a $p \times p$ non negative definite matrix. The covariance matrix of projected feature of sample is defined as

$$S_X = \vec{V}^T E\left[(X - EX)^T (X - EX)\right]\vec{V} = \vec{V}^T S_M \vec{V} \tag{2}$$

Suppose there are *N* image matrices $\{X_i\}$, $i = 1, 2, \cdots, N$ and denote the average image as $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, then S_M can be evaluated by

$$S_X = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^T (X_i - \bar{X}) \tag{3}$$

In line with the PCA algorithm, to have a projection direction of 2DPCA is done by reducing the dimensionality of a data set in order to explain as much information as possible, S_X has the important rule of letting $\vec{V}_{opt}$ as the eigenvector of S_X corresponding to the largest eigenvalue. A set orthonormal projection directions $\vec{V}_1, \vec{V}_2, \cdots, \vec{V}_d$ are the orthonormal eigenvector of S_X corresponding to the d largest eigenvalues, i.e. $\vec{V}_{opt} = \left[\vec{V}_1, \vec{V}_2, \cdots, \vec{V}_d\right]$. Projecting a matrix X onto $\vec{V}_{opt}$ is

$$\vec{Y}_k = X\vec{V}_k, \quad k = 1, 2, \cdots, d \tag{4}$$

In this section author discusses the robust 2DPCA by using the measure of minimizing vector variance (MVV). The MVV robust 2DPCA is primarily a robust approach which describes the variance covariance structure through a linear transformation of the original variables. The technique is a useful device for representing a set of variables by a much smaller set of composite variables that account for much of the variance among the set of original variables. The data reduction based on the classical approach becomes unreliable if outliers are present in the data. The decomposed classical covariance matrix is very sensitive to outlying observations. The first component consisting of the greatest variation is often pushed toward the anomalous observations.

The algorithm of MVV robust 2DPCA has no significant difference with MVV robust PCA except for the criterion projection, Herwindiati [3]. Suppose $X_1, X_2, \cdots, X_N$ is a $m \times p$ random image matrix.

Stage 1. Start with a construction the covariance matrix by using the N original two dimensional (2D) matrices.

Find the orthonormal eigenvectors corresponding to the d largest eigenvalues S_X, $\vec{V}_{opt} = \left[\vec{V}_1, \vec{V}_2, \cdots, \vec{V}_d\right]$.

Projecting a matrix X onto $\vec{V}_{opt}$ is

$\vec{Y}_k = X\vec{V}_k, \quad k = 1, 2, \cdots, d$

Stage 2. Estimate the location and covariance matrix of projected matrix $X_{m\times d}$ by using MVV robust approach.

1. Let H_{old} be an arbitrary subset containing $h = \left[\frac{n+k+1}{2}\right]$ matrix data points. Compute the average matrix as $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}. Then calculate $B_{m\times k} = \left(X - \bar{X}_{H_{old}}\right), \ k = 1, 2, \cdots, d$

2. Compute $d^2_{H_{old}}(i) = \vec{D}^t S^{-1}_{H_{old}} \vec{D}$, for all i = 1, 2, … , N

where $\vec{D}_{1\times d}$ is defined as mean of m rows in each k column $k = 1, 2, \cdots, d$;

3. Sort these distances in increasing order,

4. Define $H_{new} = \left\{\vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \cdots, \vec{X}_{\pi(h)}\right\}$

5. Calculate $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$.

6. If $Tr\left(S^2_{H_{new}}\right) = 0$, repeat steps 1 to 5.

If $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$, the process is stopped.

Otherwise, the process is continued until the r-th iteration if

$Tr\left(S^2_1\right) \geq Tr\left(S^2_2\right) \geq Tr\left(S^2_3\right) \geq \cdots \geq Tr\left(S^2_r\right) = Tr\left(S^2_{r+1}\right)$

Stage 3. Cluster the matrix data based on robust MVV distance

$d^2_{MVV}(i) = \vec{D}^t_{MVV} S^{-1}_{H_{old}} \vec{D}_{MVV}$, for all i = 1, 2, … , N.

IV. The Ilustration of Clustering Images Using MVV Robust PCA and 2DPCA

A. The Illustration 1

A sample set included 97 grass images and 7 wood images are selected for experiment. Two two kinds of images have different color. The extraction of each pixel in the color feature is represented as a point in a 3D RGB color space . Assume that we do not know the caracteristics of sample. Two approaches of MVV robust reduction dimension are used for clustering process.

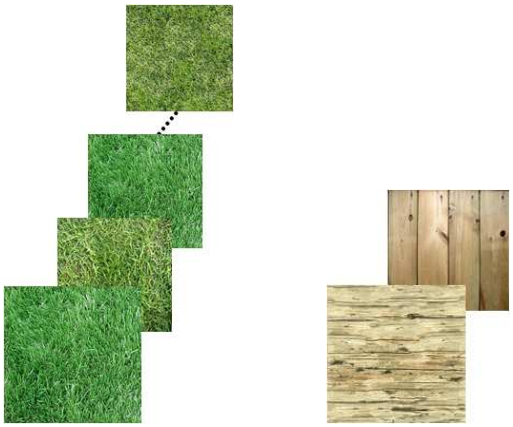


Fig. 4 The Images of Grass and Wood

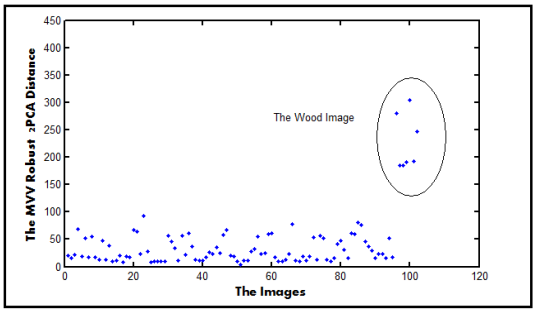


Fig. 5 The Clustering of Grass and wood with MVV Robust 2DPCA

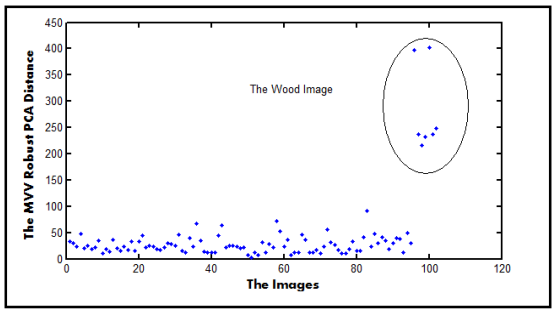


Fig. 6 The Clustering of Grass and wood with MVV Robust PCA

B. The Illustration 2

In this illustration author shows the classification of two kinds of cities; that are cities having the high density and low high of population. The images are captured from satellite. We assume that the dense roof means the dense population. We have 55 images of the high density and 4 images of low density

Fig. 7 The cities having high and low density of population

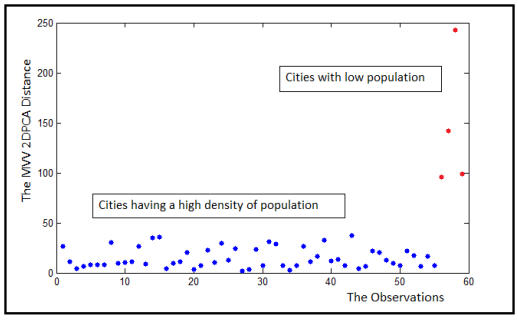


Fig. 8 The Clustering Cities with MVV Robust 2DPCA

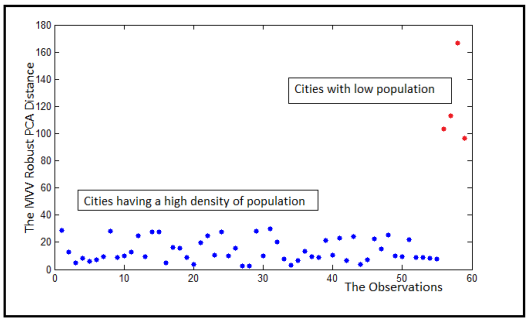


Fig. 9 The Clustering Cities with MVV Robust PCA

Two illustrations tell us that the two robust approaches; i.e. MVV Robust PCA and MVV Robust 2DPCA; have the good performance for clustering process of images. The approaches can separate clearly two classes having different characteristics.

V. The Computation Time Of MVV Robust PCA And 2DPCA

The 2DPCA is often called as a variant of principal component (PCA). To distinguish PCA from 2DPCA, all of the 2D data must be previously transformed into 1D vector before they are processed by PCA approach. The transformation leads to a high dimensional vector space. The 2DPCA has the less time for determining the eigenvectors, the image matrices are directly treated as 2D matrices and the covariance matrix can be constructed directly using the original image matrices, see Yang et al [6].The efficiency or running time of an algorithm is related the length time or the number of steps. In this section authors are going to compare the computational time of MVV robust 2DPCA and MVV robust PCA. To generate n random matrices of 40×40, and defined them as $X_1, X_2, \cdots, X_N$ for experiment. The next step is to calculate the average of computational time of MVV robust PCA and MVV robust 2DPCA for 100 experiment . We repeat the actions using $n=20, 30, 40,\ldots,100$.

The computation time of two methods is presented by Figure 9. The graphic pattern of The MVV robust 2DPCA is more stable than the MVV robust PCA graph. The difference computation time of two methods is more biger for larger of data size.

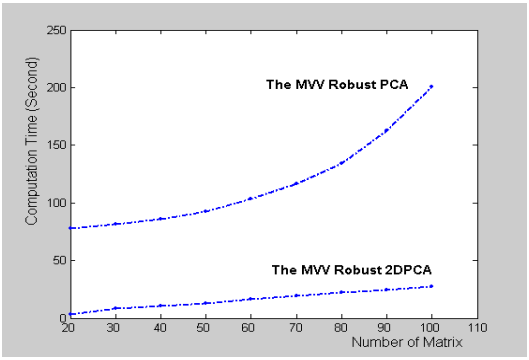


Fig. 10 The Computation Time of MVV robust 2DPCA and MVV robust PCA

VI. The Consistency Estimator

The section discusses the consistency of estimator of the MVV robust PCA and the MVV robust 2DPCA. The Estimation is the process by which sample data are used to indicate the value of an unknown quantity in a population. An estimator is any quantity calculated from the sample data which is used to approximate the unknown parameters. The one desirable property of estimator is the value of an estimator is closed to the value of the true parameter. An estimator for a parameter is consistent if the estimator converges in probability to the true value of the parameter, Kendal and Stuart [11]. Consider an estimator t_n, computed from a sample of n values, will be said to be a consistent estimator if there is some N such that the probability that

$$|t_n - \theta| < \varepsilon \qquad (5)$$

is greater than $(1-\eta)$ for all $n > N$. In the notation of the probability theory,

$$P\{|t_n - \theta| < \varepsilon\} > 1-\eta \qquad n > N \qquad (6)$$

for any positive ε and η however small.

The sample estimator should have a high probability of being close to the population value for large sample size. The formulate (5) means that the distributions of the estimators become more and more concentrated near the true value of the

parameter being estimated, so that the probability of the estimator being arbitrarily close to θ converges to one.The estimator of MVV robust PCA and the MVV robust 2DPCA are consistent estmator. To prove the starement, we do two experiments with 100 replication of an each experiment. For the first experiment, we generate the multivariate normal $N_p(\bar{\mu},\Sigma)$, $p=25$, and $\bar{\mu}=\vec{0}$; $\Sigma=I_2$ and $n=200$. The contaminant data present in a data set beginning 1% and gradually to be higher; i.e. 2%; 3%; 4% and so on till 10 %.

The following figure is the result of simulation experiment for consistency. The figure illustrates that the probability of MVV robust PCA and MVV robust 2DPCA converge to 0.9. Moreover, we see that the line of robust PCA approach; the blue line; is more stable than robust 2DPA approach; the green line. It means that the MVV robust PCA is more consistent against of contaminant.

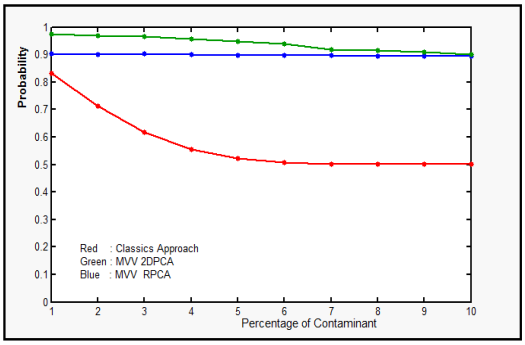


Fig. 11 The Comparison of Concistency of Classics; MVV Robust PCA and MVV Robust 2DPCA estimator

VII. Conclusion

The MVV Robust PCA and MVV Robust 2DPCA can be considered as measure for clustering data images. The good properties of robust can reduce the anomolous images some times coming from human error or different setting instrument when the images are captured. The MVV 2DPCA is more efficient computation than the time of MVV Robust PCA. The simulation experiments of consistent estimator suggestions that the MVV Robust PCA is better performance of clustering than the MVV 2DPCA performance

References

[1] D.E. Herwindiati, M.A. Djauhari, and M. Mashuri, "Robust Multivariate Outlier Labeling', *Journal Communication in Statistics – Simulation And Computation*, Vol. **36**, No 6, pp 1287-1294, April 2007.

[2] D.E. Herwindiati, S.M. Isa, S.M, "The Robust Principal Component Using Minimum Vector Variance", *Electronic Engineering and Computing Technology, SpringerLink*, Volume 60, pp 397-408, 2010

[3] D.E. Herwindiati, "A Robust Two-Dimensional Principal Component Analysis for Classification" *Civil-Comp Proceedings ISSN 1759-3433*, paper No 108, Valencia, September 2010

[4] F.Anguilla and C. Pizzuti, "Outlier Mining and Large High-Dimensional Data Sets", *IEEE Transaction on Knowledge and Data Engineering*, Vol 17, No 2, pp 203-215, 2005

[5] I.T. Jolliffe, I.T. "Principal Component Analysis", Springer Verlag, 1986

[6] J. Yang, D. Zhang, A.F. Frangi and J-yu Yang, "Two-Dimensional PCA: A New Approach to Appearance – Based Face Representation and Recognition", *IEEE Transaction on Pattern Analysis and machine Intelligence*, Vol **26**, No 1, pp 131 -137, 2004

[7] M.A Djauhari,"Improved Monitoring of Multivariate Process Variability", *Journal of Quality Technology*, No 37, pp 32-39, 2005

[8] M. Hubert, P.J. Rousseeuw, K. vanden Branden, "ROBPCA: a New Approach to Robust Principal Component Analysis", *Journal. Technometrics*, 47, pp 64-79, 2003

[9] P.J. Rousseeuw and A.M. Leroy, "Robust Regression and Outlier Detection", John Wiley, New York, 1987

[10] P.J. Rousseeuw and K.van Driessen, "A Fast Algorithm for The Minimum Covariance Determinant Estimator", *Journal. Technometrics*, 41, pp 212-223, 1999

[11] S.M Kendall and A. Stuart, "The Advanced Theory of Statistics", *Charles Griffin & Co Ltd, Vol. 2, Fourth Edition*, London, 1979