RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2011.37.355

ScienceAsia 37 (2011): **355–359**

New statistics for detection of outliers using the last few principal components

Rungrawee Amnarttrakul*, **Ampai Thongteeraparp**

Statistics Department, Kasetsart University, Bangkok

*Corresponding author, e-mail: g4984008@ku.ac.th

Received 11 Jul 2011
Accepted 10 Nov 2011

ABSTRACT: Two test statistics are proposed for detecting outliers that are not extreme on any of the original variables in multivariate data. The test statistics are derived via principal component analysis and some normal distribution properties. Moreover, the last few principal components or minor principal components from principal component analysis have an important role in these simple test statistics. Finally, the tests are applied to a simulated data and the real data of a financial institution in Thailand as examples. Moreover, a comparative study was carried out using data on milk from Daudin, Duby, and Trecourt [*Statistics*, **19**, 241].

KEYWORDS: minor principal components, multivariate data, principal component analysis

INTRODUCTION

Many researchers have proposed definitions for an outlier with no universally accepted definition. For example, according to Grubbs[1], "an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". For Barnett and Lewis[2], an outlier is "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". However, in this study, the definition is slightly different; multivariate outliers are those that deviate from the usual correlation structure in the p-dimensional space defined by the variables. A major problem in detecting outliers in multivariate data is that an observation that is not extreme in any of the original variables can still be an outlier, because it does not conform with the correlation structure of the remainder of the data[3]. Suppose that the weights and heights are collected from a sample of healthy children of ages between 5 and 15 years old; then an observation with weight and height of 20 kg and 175 cm, respectively, is not extreme on either the weight or height variables individually, as 20 kg is a plausible weight for the youngest children and 175 cm is a plausible height for the older children. Nevertheless, the combination of weight and height is virtually impossible, and will be a clear outlier because it combines a small weight with a large height. Thus this violates the general pattern of a positive correlation between the two variables. This type of outlier is problematic to detect in multivariate data.

There are several proposed ideas for detecting outliers in multivariate data. The traditional method for detection of outliers is known as the Mahalanobis distance. A large distance may indicate that the corresponding observation is an outlier, but two problems occur in practice: masking and swamping[4]. For other methods, see, for example, Refs. 4–7.

Several methods for detecting outliers in multivariate data have been proposed. Among these, principal component analysis is an interesting approach for detecting such outliers which are rather difficult and uncommon. But our approach uses the last few principal components to find new test statistics that can be used to detect this kind of outlier in multivariate data.

In the following section, the concept of principal components analysis is given. Then the test statistics using the last few principal components are derived. The example of one simulation and two applications of these test statistics and conclusions are also included in the next section.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis or PCA is one of the key tools in multivariate statistical analysis and is often used to reduce the dimension of data for easy exploration and further analysis, such as regression analysis, clustering and discriminant analysis. It is concerned with explaining the variance-covariance structure of a set of variables through a few new variables. All principal components are particular linear combinations of the p random variables with

three important properties which are:

1. The principal components are uncorrelated.
2. The first principal component has the highest variance, the second principal component has the second highest variance, and so on.
3. The total variation in all the principal components combined is equal to the total variation in the original variables.

The principal components are computed from an eigenanalysis of the covariance matrix or the correlation matrix, but results from the covariance matrix and the correlation matrix are usually not the same. If the variables are measured on scales with widely different ranges or if the units of measurement are not commensurate, it is better to perform PCA on the correlation matrix.

The observations that are outliers with respect to the first few principal components or the major principal components usually correspond to outliers on one or more of the original variables. On the other hand, the last few principal components or the minor principal components represent linear functions of the original variables with the minimal variance. The minor principal components are sensitive to the observations that are inconsistent with the correlation structure of the data, but are not outliers with respect to the original variables[8].

THE NEW TEST STATISTICS

PCA and normal distribution properties have been applied to the new test statistics. It is assumed that the data come from a multivariate normal distribution.

Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_p \end{bmatrix}'$ be a random sample from multivariate normal distribution with mean vector $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_p \end{bmatrix}'$ and variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \\ \vdots & & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}.$$

This p-dimensional normal density is denoted by $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

In matrix notation, the standardized vector $\mathbf{z}$ is

$$\mathbf{z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{V}^{1/2}$ is the diagonal standard deviation matrix, given by

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}.$$

Clearly, $E(\mathbf{z}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{z}) = \boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is the correlation matrix of the standardized vector. That is to say, $\mathbf{z}$ is distributed as $N_p(\mathbf{0}, \boldsymbol{\rho})$, or $\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\rho})$.

Principal components depend solely on the variance-covariance matrix or the correlation matrix of the random vector. Their calculation does not require a multivariate normal assumption. On the other hand, principal components derived from multivariate normal populations are useful for inference.

As $\boldsymbol{\rho}$ is the variance-covariance matrix associated with the random vector $\mathbf{z}$, and $\boldsymbol{\rho}$ has the eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{\alpha}_1), (\lambda_2, \boldsymbol{\alpha}_2), \dots, (\lambda_p, \boldsymbol{\alpha}_p)$, where $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_p \geqslant 0$, then the ith principal component is given by

$$y_i = \boldsymbol{\alpha}_i'\mathbf{z} = \boldsymbol{\alpha}_{i1}z_1 + \boldsymbol{\alpha}_{i2}z_2 + \dots + \boldsymbol{\alpha}_{ip}z_p \qquad i = 1, 2, \dots, p$$

with these choices:

$$\begin{aligned} \mathrm{Var}(y_i) &= \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_i = \lambda_i \qquad i = 1, 2, \dots, p \\ \mathrm{Cov}(y_i, y_k) &= \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_k = 0 \qquad i \neq k. \end{aligned}$$

The eigenvectors of $\boldsymbol{\rho}$ are orthogonal if all the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ are distinct. In practice, the eigenvalues of the correlation matrix are all different, and thus none of the terms is zero[3].

Therefore, for any two eigenvectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_k$, $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_k = 0,\ i \neq k$.

Since $\boldsymbol{\rho}\boldsymbol{\alpha}_k = \lambda_k\boldsymbol{\alpha}_k$, premultiplication by $\boldsymbol{\alpha}_i'$ gives $\mathrm{Cov}(y_i, y_k) = \boldsymbol{\alpha}_i'\boldsymbol{\rho}\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_i'\lambda_k\boldsymbol{\alpha}_k = 0,\ i \neq k$.

In matrix notation,

$$\mathbf{y} = \boldsymbol{\alpha}\mathbf{z},$$

where $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\rho})$.

Consider

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & & \\ \vdots & & \ddots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pp} \end{bmatrix}.$$

$\boldsymbol{\alpha}$ is a constant $p \times p$ matrix of rank p, the p linear combinations in $\boldsymbol{\alpha}\mathbf{z}$ have a multivariate normal distribution.

From properties of multivariate normal random variables, if $\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\rho})$, then $\boldsymbol{\alpha}\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\alpha}\boldsymbol{\rho}\boldsymbol{\alpha}')$ or $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\alpha}\boldsymbol{\rho}\boldsymbol{\alpha}')$[9].

But,

$$\begin{aligned} \mathrm{Var}(\mathbf{y}) &= \lambda_i \qquad i = 1, 2, \dots, p \\ \mathrm{Cov}(y_i, y_k) &= 0 \qquad i \neq k. \end{aligned}$$

Substituting,

$$\boldsymbol{\alpha\rho\alpha}' = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

The variance-covariance matrix of $\mathbf{y}$ is defined as

$$\boldsymbol{\Sigma}_y = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

If y_i and y_k are jointly multivariate normal and if $\mathrm{Cov}(y_i, y_k) = 0$, then y_i and y_k are independent [10].

Any subset of the y's in $\mathbf{y}$ has a multivariate normal distribution, with mean vector consisting of the corresponding submatrix of $\boldsymbol{\Sigma}_y$.

From normality of marginal distribution: if $\mathbf{y} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_y)$, each y_i in $\mathbf{y}$ has the univariate normal distribution, then $y_i \sim N(0, \lambda_i),\ i = 1, 2, \ldots, p$.

The last two principal components are y_{p-1} and y_p, respectively.

A possible method for detecting outliers is to combine information from the last two principal components in order to form a new test statistic.

Since

$$\begin{aligned} y_i &\sim N(0, \lambda_i) \quad i = 1, 2, \ldots, p, \quad \text{then} \\ y_{p-1} &\sim N(0, \lambda_{p-1}) \quad \text{and} \\ y_p &\sim N(0, \lambda_p). \end{aligned}$$

From the distribution of the sum of two subvectors: if y_{p-1} and y_p are the same size and independent, then $y_{p-1} + y_p$ is $N(0, \lambda_{p-1} + \lambda_p)$.

Consequently,

$$R^2_{2z} = \frac{(y_{p-1} + y_p)^2}{\lambda_{p-1} + \lambda_p} \sim \chi^2_{(1)}. \qquad (1)$$

Similarly, the test statistic using the last three principal components can be shown as

$$R^2_{3z} = \frac{(y_{p-2} + y_{p-1} + y_p)^2}{\lambda_{p-2} + \lambda_{p-1} + \lambda_p} \sim \chi^2_{(1)}. \qquad (2)$$

In conclusion, R^2_{2z} and R^2_{3z} are the test statistics using minor principal components or the last few principal components in detecting multivariate outliers that are observations which are not extreme on any of the original variables, but may be outliers as they do not conform with the correlation structure of the remainder of the data when the data are multivariate normal distribution.

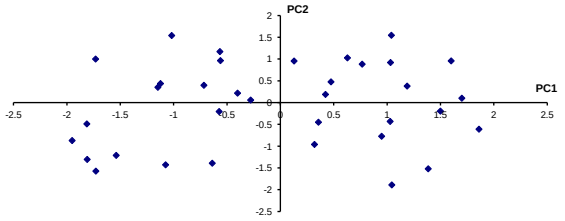

Fig. 1 Plot of the observations with respect to the first two principal components of the simulated data.

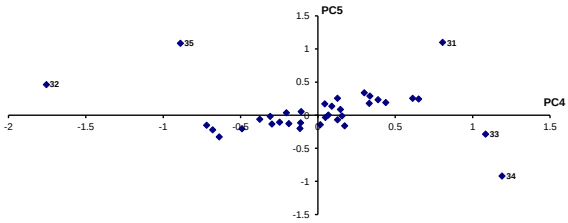

Fig. 2 Plot of the observations with respect to the last two principal components of the simulated data.

EXAMPLES

This section demonstrates three examples which show that different test statistics may indicate different potential outliers. If the same observations are identified by most distinct test statistics, they will be classified as clear outliers. However, for the observations that are detected to be outliers only by some test statistics, they may be other types of outliers, for instance they may be extreme in any variables. Therefore, these observations should be further considered for their behaviour. The first example, one data set which is simulated using SAS program, has 35 observations consisting of 5 variables and the last 5 observations are outliers that are not apparent with respect to the original variables. A scatter plot of pair of the last two principal components may be useful in identifying this type of outliers as seen in Fig. 1 and Fig. 2. The values of R^2_{2z} and R^2_{3z} in Table 1, defined in equations (1) and (2), respectively, can detect these outliers clearly, but they cannot be seen by the first two principal components.

Therefore the simulation result in this example indicates that R^2_{2z} and R^2_{3z} are efficient as they can be used to detect the outliers which are normal as a single variable, as same as observations 31–35.

The next example consists of a set of data for credit analysis that comes from a financial institution in Thailand. There are 10 financial variables and 65 observations, where x_1 = total assets turnover (time), x_2 = quick ratio, x_3 = current ratio, x_4 = equity

Table 1 The values of R^2_{2z} and R^2_{3z} of the potential outliers of the simulated data.

2 minor principal components		3 minor principal components	
Obs. No.	R^2_{2z}	Obs. No.	R^2_{3z}
31	5.18	31	54.03
32	4.25	32	20.27
33	10.62	33	25.78
34	18.52	34	21.36
35	9.85	35	8.39

Table 2 The values of R^2_{2z} and R^2_{3z} of the potential outliers of the financial data.

2 minor principal components		3 minor principal components	
Obs. No.	R^2_{2z}	Obs. No.	R^2_{3z}
2	5.61	2	5.55
8	6.27	3	9.51
38	12.30	38	18.81
53	8.12	53	4.34

ratio (time), x_5 = gross profit margin (%), x_6 = net profit margin (%), x_7 = return on asset ratio (%), x_8 = return on equity ratio (%), x_9 = inventory turnover, and x_{10} = fixed assets turnover (time). Note that some variables do not have units. Table 2 gives the values of R^2_{2z} and R^2_{3z} for the potential outliers on each statistic.

By examining values of R^2_{2z} and R^2_{3z} for each observation, it is found that some observations that may be possible outliers, or the potential outliers of R^2_{2z} and R^2_{3z} are the same as seen in Table 2 and also in Fig. 4. Therefore, observations 2, 8, 38, and 53 become extreme by the last few principal components. The observations contradict the correlation structure among all ten variables. However, different analyses may be capable of identifying different potential outliers.

Table 3 The observation number of the potential outliers of the data on milk.

2 minor principal components		3 minor principal components	
R^2_{2z}	d^2_{2i}	R^2_{3z}	d^2_{2i}
1	1	1	1
2	2	2	2
41	41	41	41
44	44	44	44
			74

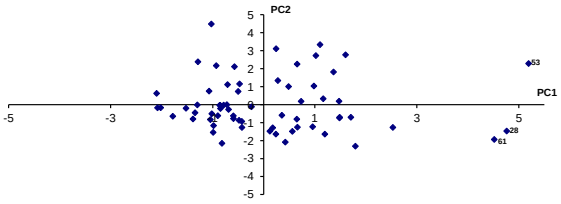

Fig. 3 Plot of the observations with respect to the first two principal components of the financial data.

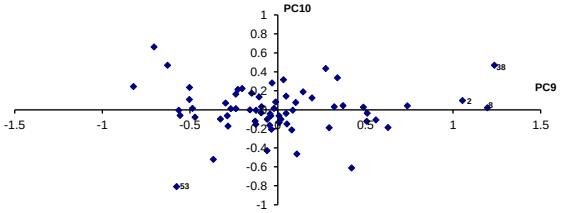

Fig. 4 Plot of the observations with respect to the last two principal components of the financial data.

Note that when the observations with respect to the first two principal components are plotted (Fig. 3) the extreme observations are not the same as in Fig. 4, except observation 53.

The last example comes from Daudin, Duby, and Trecourt[11] who give data on the composition of 85 containers of milk, on each of which 8 measurements were made. The variables are y_1 = density, y_2 = fat content (g/l), y_3 = protein content (g/l), y_4 = casein content (g/l), y_5 = cheese dry substance measured in the factory (g/l), y_6 = cheese dry substance measured in the laboratory (g/l), y_7 = milk dry substance (g/l), and y_8 = cheese produced (g/l)[12]. In order to compare results of the test statistics from this study with other test statistics to identify potential outliers of multivariate outliers using only minor principal components, the test statistic[13] d^2_{2i} is compared in the following example. The test statistic is declared by 2 and 3 minor principal components in this case.

$$d^2_{2i} = \sum_{k=p-q+1}^{p} \frac{z^2_{ik}}{l_k} \sim \chi^2_{(q)} \qquad (3)$$

where z_{ik} is the value of the kth principal component for the ith observation, l_k is the variance of the kth principal component, q is the number of minor principal components, and p is the total number of principal components. Table 3 shows observations which are detected to be possible outliers in each of the test statistics with 2 and 3 minor principal components.

Although the set of the four potential outliers are

the same for R^2_{2z}, R^2_{3z}, and d^2_{2i}, there is also observation 74 that can be found to be a possible outlier for d^2_{2i} in 3 minor principal components case. This example demonstrates that the potential outliers from each test statistics may indicate different observations. It may be difficult to decide which observations are in fact the outliers. However, this comparison of the results of all test statistics shows minor differences. Note that, according to Ref. 11, there is only one outlier from the forward plot[12], which is observation 69. This result is not apparent from R^2_{2z}, R^2_{3z}, and d^2_{2i}, because observation 69 clearly shows to be the outlier in a scatter plot between y_5 and y_6. The set of the potential outliers from R^2_{2z}, R^2_{3z}, and d^2_{2i} cannot be obviously seen by each variable because it is not apparent on a plot of one or two variables but it does not conform with the correlation structure of the data. Therefore, outliers detected by the forward search are different from those concerned in this study.

Certainly, it is possible that the outliers which are detectable from a plot of the first few principal components are those which inflate variances and covariances[14]. Similarly, an observation that inflates a covariance or correlation between two variables will often be extreme with respect to one or both of these variables looked at individually. It is not an outlier with respect to the correlation structure. This observation would appear as an outlier on one of the first few principal components. Besides, one obvious question which is raised in this example is how many outliers are there? The answer is, surely, unknown because nobody can know previously which observations are the outliers. As a result, our test statistics can be used to identify the potential outliers with respect to the correlation structure as well, as we cannot detect these outliers from other methods, such as calculating a statistical distance and graphical methods.

CONCLUSIONS

In this paper, the new test statistics that can be used to detect outliers using minor principal components are proposed. The test statistics are formed based on PCA and basic normal distribution properties. The outliers that do not conform with the correlation structure of the remainder of the data will be detected by using the last few principal component. Furthermore, the test statistics from this study are slightly different from the test statistic d^2_{2i}, which also comes from Chi-square distribution, but has different degrees of freedom. Thus R^2_{2z} and R^2_{3z} can be used easily for detecting outliers in multivariate data. They can be alternative test statistics to detect outlier in multivariate data for statisticians. However, R^2_{2z} and R^2_{3z} cannot be used to tell which outliers have a large effect or are "influential", because not every outlier needs to be influential. A recommendation for further study is to use other approaches to find test statistics which can deal with outliers that are different from the normal correlation structure of the data.

REFERENCES

1. Grubbs FE (1969) Procedures for detecting outlying observations in samples. *Technometrics* **11**, 1–21.
2. Barnett V, Lewis T (1994) *Outlier in Statistical Data*, 3rd edn, Wiley, Inc., New York.
3. Jolliffe IT (2002) *Principal Component Analysis*, 2nd edn, Springer-Verlag Inc., New York.
4. Hadi AS (1992) Identifying multiple outliers in multivariate data. *J Roy Stat Soc B* **54**, 761–71.
5. Munoz-Garcia J, Moreno-Rebollo JL, Pasual-Acosta A (1990) Outliers: a formal approach. *Int Stat Rev* **58**, 215–26.
6. Davies L, Gather U (1993) The identification of multiple outliers. *J Am Stat Assoc* **88**, 782–92.
7. Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. *Comput Stat Data Anal* **52**, 1694–711.
8. Jobson JD (1992) *Applied Multivariate Data Aalysis*. Springer-Verlag Inc., New York.
9. Rencher AC (2002) *Methods of Multivariate Analysis*, 2nd edn, John Wiley & Sons, Inc., New York.
10. Rencher AC (1998) *Multivariate Statistical Inference and Applications*, John Wiley & Sons, Inc., New York.
11. Daudin JJ, Duby C, Trecourt P (1988) Stability of principal component analysis studied by the bootstrap method. *Statistics* **19**, 241–58.
12. Atkinson AC, Riani M, Cerioli A (2004) *Exploring Multivariate Data with the Forward Search*, Springer-Verlag Inc., New York.
13. Hawkins DM (1980) *Identification of Outliers*, Chapman and Hall, London.
14. Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81–124.