Cluster PCA for Outliers Detection in High-Dimensional Data

George Stefatos and A. Ben Hamza
Concordia Institute for Information Systems Engineering
Concordia University, Montréal, QC, Canada
{g_stefat, hamza}@encs.concordia.ca

***Abstract*— We introduce a new method to detect multiple outliers in high-dimensional datasets using the concepts of hierarchical clustering and principal component analysis. The proposed algorithm is computationally fast and robust to outliers detection. A comparative study with existing techniques is performed on both low and high dimensional datasets. Our experimental results demonstrate an improved performance of our algorithm in comparison with existing multivariate outlier detection techniques.**

I. Introduction

With fast automated data collection tools and larger databases, there is a tremendous amount of information that is stored for future analysis [1]. Consequently, the need to develop tools and techniques to extract relevant information has made research areas such as data mining increasingly important [2]. Outlier detection is among these research areas that has attracted considerable attention in recent years [3]. Outliers are defined as abnormal data points which deviate from the normal variability found in a dataset. These outliers are often of primary interest in both chemical and engineering related processes [4]. For example, in geochemical exploration, outliers can often identify important mineral deposits [5].

In recent years, various techniques have been proposed for outlier detection in both univariate and multivariate settings [5–7]. These methods typically fall under two categories: supervised and unsupervised approaches. The supervised approaches compare new observations under an existing well defined model, whereas the unsupervised approaches classify each observation under normal and extreme variation based on a certain distance [8].

In this paper, we present a new distance-based approach which we refer to as cluster principal component analysis (cluster PCA). The goal of the proposed method is to identify outliers in both low and high dimensional datasets by combining the concepts of hierarchical clustering [9] and PCA [10] in order to improve the performance while keeping the complexity and computation time relatively low.

The rest of this paper is organized as follows. In the next section, we describe the problem formulation. In Section III, we briefly review some related work. Section IV introduces the proposed cluster PCA algorithm. In Section V, experimental results are presented to demonstrate the performance of the proposed approach in comparison with existing techniques. Section VI concludes the paper.

II. Problem Formulation

Mining information from a dataset containing multiple dimensions is becoming very common [11]. Many manufacturing and service businesses use independent univariate techniques to study each dimension. The univariate approaches may lead faulty results since it takes little or no account of the covariance that exist between the observations [12]. Using multivariate control charts, it is possible to maintain a specific error rate, while taking advantage of cross correlation between the variables, and the process can be analyzed for its stability without the complication of maintaining many control charts at once. Multivariate quality control provides a way for engineers and manufacturers to test their products in an environment that provides many advantages over univariate models. It is inherently more complex than univariate statistical process control, but it may be a more realistic representation of the data since in the real world processes do not usually have only one variable that is measured independent of all other variables in a system [13]. This is not trivial when a dataset contains multiple outliers. Even a small percentage of outliers can distort the results and render the outcome misleading or useless. To overcome this problem, statisticians have recently proposed robust methods to estimate key parameters such as the mean and covariance/correlation matrix without the negative effect of outliers [14]. Techniques such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) have proven their robustness but are limited to small moderate dimensions [4, 6].

For multivariate data, the process of finding meaningful outliers becomes inherently more complex [11]. For example, in the field of chemometrics, datasets containing thousands of dimensions are not uncommon. For these types of applications, projection pursuit (PP) and PCA may be used to process and analyze such large information [4].

Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]^T$ be an $m \times p$ data matrix of m vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, where each observation $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is a row vector with p variables (dimensions). Hotelling's T^2 statistic, also referred to as Mahalanobis distance, is defined as

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \quad \text{and} \quad S = \frac{1}{m-1} \sum_{i=1}^{m} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

1-4244-0991-8/07/$25.00 ©2007 IEEE

are the sample mean and covariance matrix respectively.

III. Related Work

In this section, we will briefly review some multivariate outlier detection techniques that will used for comparison with our proposed approach.

A. Minimum Volume Ellipsoid (MVE)

The MVE method is based on the concept of finding the smallest ellipsoid containing at least half of the observations. It is achieved by randomly selecting $p+1$ samples to calculate the mean and covariance matrix. The Mahalanobis distance is then calculated in order to find the median distance of the m observations. This distance will then be used as a magnification factor to the volume of the ellipsoid. The algorithm is repeated n times until the $p+1$ observations containing the smallest volume is found. These $p+1$ observations will be used to compute the robust mean and the robust covariance matrix. The MVE approach has been proven to be statistically sound and robust to multiple outliers [7, 15].

B. Minimum Covariance Determinant (MCD)

The MCD method is widely considered as the current best-performing technique for outlier detection in low dimensional data [6]. It randomly selects a subset of the data containing $p+1$ observations to calculate the mean and covariance matrix. The Mahalanobis distance is then calculated for all m observations in order to select ℓ observations with the smallest distance, where $\ell \geq m/2$. The determinant of the covariance matrix of the selected observation is then calculated. This process is repeated n times until the smallest determinant is found. The mean and covariance matrix of the ℓ observations containing the smallest determinant are considered to be robust [3, 6, 7, 15].

IV. Proposed Method

Most statistical-based techniques use the covariance matrix as the basis for detecting outliers in datasets [6]. For example, the MCD and MVE methods use the volume (determinant) of the covariance matrix to identify the subset of observations that are considered robust to calculate both the mean and the covariance matrix. This process might result very lengthy because the optimum subset might require $n = m!/(m-h)!$ permutations, where h is cardinality of the optimum subset of observations. Also, the determinant of the covariance matrix can only be computed if $p < h$, otherwise the determinant will be equal to zero. This is the main reason why some of the more robust algorithms are limited to only a small value of p (i.e. small number of dimensions) [4].

Our proposed algorithm, however, does not depend on a the number of permutations and can be applied to both high (large p) and low dimensional (small p) datasets. Our approach is based on hierarchical clustering in combination with PCA to obtain a robust subset of observations that are used to identify outliers. PCA is a method for transforming the observations in a dataset into new observations which are uncorrelated with each other and account for decreasing proportions of the total variance of the original variables. Each new observation is a linear combination of the original observations. Standardizing the data is often preferable when the variables are in different units or when the variance of the different columns of the data is substantial. The standardized data matrix is given by

$$Z = (X - \mathbf{1}\,\bar{\boldsymbol{x}})D^{-1/2} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m]^T,$$

where $\mathbf{1} = (1, \ldots, 1)^T$ is a $n \times 1$ vector of all 1's, and $D = (\mathrm{diag}(S))^{1/2}$ is the diagonal standard deviation matrix.

The Euclidean distance matrix between all the standardized observations is given by

$$d_{ij} = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|, \quad 1 \leq i, j \leq m.$$

Initially, each observation can be considered a cluster of its own until all the h observations are eventually integrated in one big cluster. If $m > p$, the optimal subset contains $h = \lfloor (m+p+1)/2 \rfloor$ observations, and if $m < p$ then the optimal subset contains $h = \lfloor \alpha m \rfloor$, where $\alpha \in (1/2, 1)$ is a parameter and $\lfloor x \rfloor$ denotes the floor function that returns the largest integer less than or equal to x. A smaller value of α tends to increases the robustness of the algorithm whereas a higher value of α tends to give better estimates of the uncontaminated data [4, 15].

Given a set of h observations to be clustered and an $h \times h$ distance matrix $D = (d_{ij})_{1 \leq i,j \leq h}$, the hierarchical clustering is performed as follows [9]:

1) Assign each observation to a cluster so that we have h clusters, each containing one cluster. Let the distances between the clusters be the same as the distances between the observations they contain.
2) Find the closest pair of clusters and merge them into a single cluster, so that we have one cluster less.
3) Compute distance between the new cluster and each of the old clusters. The distance between one cluster and another cluster is equal to the shortest distance from any member of one cluster to any member of the other cluster.
4) Repeat steps 3) and 4) until all observations are clustered into a single cluster of size h.

Once the sample $H = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_h]^T$ of the h observations are selected, we compute its robust sample mean $\overline{\boldsymbol{z}}_{\boldsymbol{H}}$ and its robust sample covariance matrix S_H. Then, we apply the eigendecomposition on S_H, that is $S_H = A \Lambda A^T$, where A is a matrix of eigenvectors (principal components) and Λ is a diagonal matrix of eigenvalues. We may select the most significant k principal components according to the following criteria

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 90\%$$

which can be used as reasonable cut-off value. The robust distance in the PCA subspace is then defined as

$$T_i^2 = (\boldsymbol{z_i} - \overline{\boldsymbol{z}}_{\boldsymbol{H}}) A_k {\Lambda_k}^{-1} A_k^T (\boldsymbol{z_i} - \overline{\boldsymbol{z}}_{\boldsymbol{H}})^T$$

where $A_k = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)$ and $\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$.

V. Experimental Results

In this section, we test the performance of the proposed cluster PCA method on two datasets with different dimensions, and we also compare the results with the previous methods discussed in the related work section.

The datasets are generated using the following Gaussian mixture model

$$(1-\varepsilon)N_p(0,\Sigma)+\varepsilon N_p(\widetilde{\mu},\Sigma), \tag{1}$$

where ε is the percentage of outliers, N_p denotes a p-variate Gaussian distribution, and $\widetilde{\mu}$ denotes the mean shift [4, 16]. Therefore each observation is normally distributed at a certain mean and varied randomly at certain standard deviation. Moreover, the simulated data is restricted between $\pm 2\sigma$ in order to have better control of the uncontaminated data and to also avoid any extreme case scenario.

We varied the values of $m, p, \varepsilon, \widetilde{\mu}$ for different settings and repeated the experiment 50 times in order to achieve the best estimates. A detailed description of the experiment can be summarized as follows:

- Two datasets $X_{m\times p}$ with $(m,p) = (100,4)$ and $(m,p) = (50,100)$ are generated using the Gaussian mixture model defined in Eq. (1).
- The percentage of outliers ε was set to 0%, 10%, 20%, 30%, and 40%.
- The mean shift $\widetilde{\mu}$ was set to 0.1, 0.15, 0.2, 0.25, and 0.30. The standard deviation was consistently set to $\sigma = 0.1$ in order to get different signal to noise ratios.
- We tested the performance of the algorithms under two criteria: (i) the percentage of correct outliers detection (# of correct outliers found / total # of outliers) and (ii) the false alarm ratio (# of observations found as outliers but are not / total # of good observations). We also tested the performance for both consecutive mean shift and scattered mean shift across the observations.
- The cardinality h of the optimum subset of observations was set to $\lfloor h = \lceil (m+p+1)/2 \rfloor$ when $m \geq p$, and to $h = \lfloor m/2 \rfloor$ when $m < p$.

A. First dataset

In this subsection, we perform a simulation study on the first Gaussian-mixture generated dataset $X_{m\times p}$ with $(m,p) = (100,4)$, and we compare the performance of the proposed approach with the MVE and MCD methods.

1) Performance of the MVE algorithm: The performance of the MVE algorithm is shown in Fig. 1 through Fig. 4, where it can be noted that this procedure works well under large mean shifts and small percentage of outliers. Also, it performs quite well in the scattered mean environment particularly when $\widetilde{\mu} \approx 0.1$. Moreover, the false alarm ratio is contained under 10% in both environments.

2) Performance of the MCD algorithm: The MCD approach is considered the most robust algorithm for low dimensional data as shown in Fig. 5 and Fig. 7, where we clearly see that is works very well across all percentages of errors and mean shifts. We also obtain similar results for both

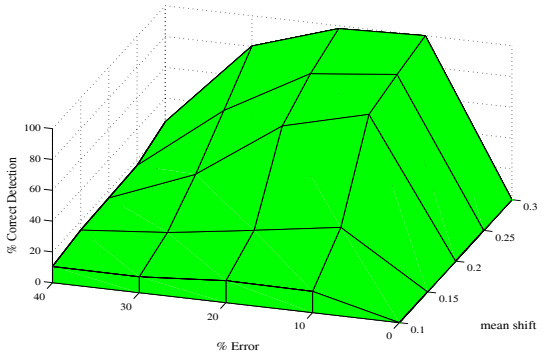

Fig. 1. MVE outlier detection performance in a consecutive mean environment.

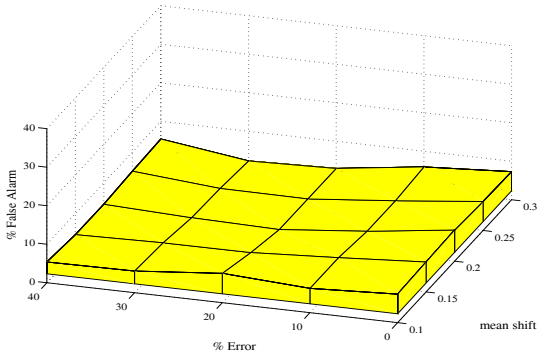

Fig. 2. MVE false alarm performance in a consecutive mean environment.

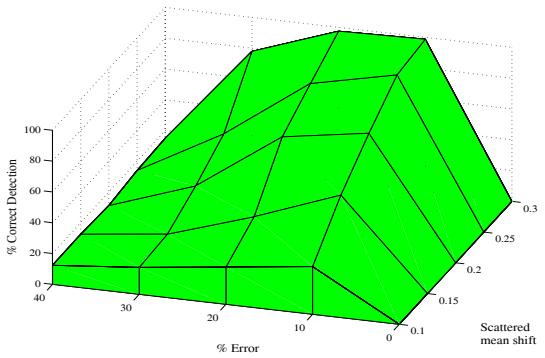

Fig. 3. MVE outlier detection performance in a scattered mean environment.

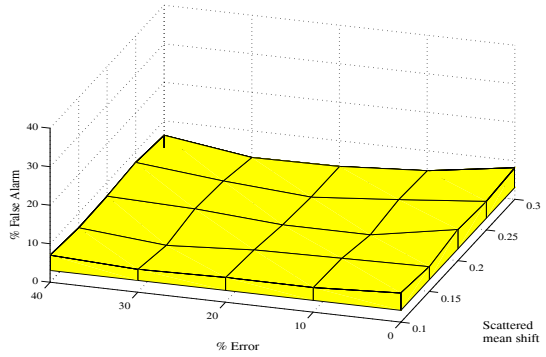

Fig. 4. MVE false alarm performance in a scattered mean environment.

scattered and consecutive mean environments. Moreover, the false alarm ratio is relatively high as shown in Fig. 6 and Fig. 8 where the performance is between 8% to 37%.

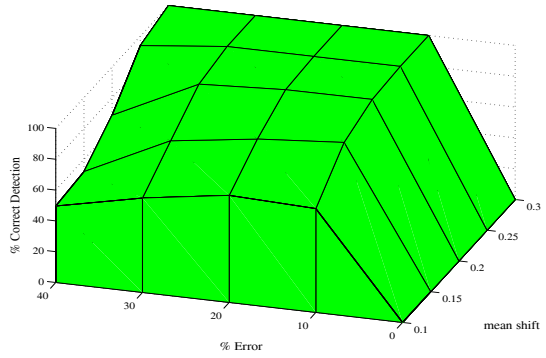

Fig. 5. MCD outlier detection performance in a consecutive mean environment.

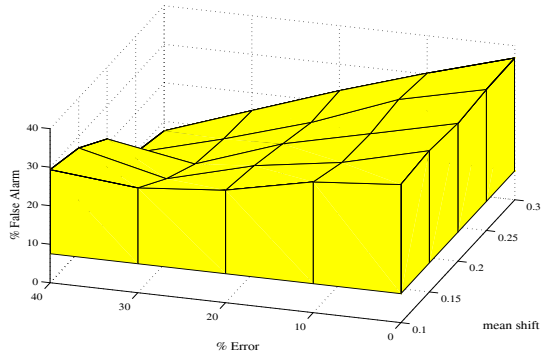

Fig. 6. MCD false alarm performance in a consecutive mean environment.

3) Performance of the cluster PCA algorithm: In cluster PCA, we varied the number of principal components k in such a way that 80% and 100% of the total variance would be selected. Fig. 9 through Fig. 11 show the results in consecutive as well as scattered mean shift environments. The false alarm ratio as shown in Fig. 10 and Fig. 12 lies at worst a the 20% range depending on the selected parameters. Also, it is important to note that the more k components are

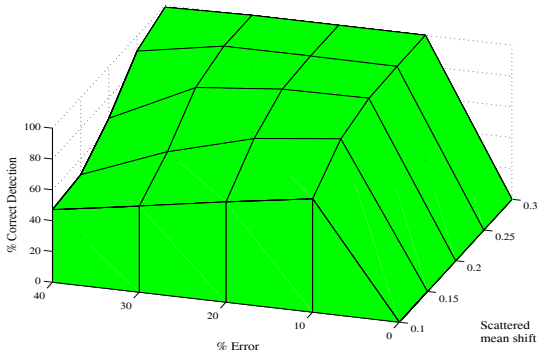

Fig. 7. MCD outlier detection performance in a scattered mean environment.

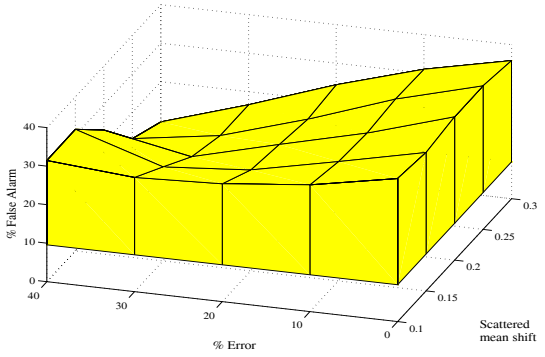

Fig. 8. MCD false alarm performance in a scattered mean environment.

selected, the better the outlier detection performance is but at the cost of higher false alarm ratio. This is mainly due to the fact that as more information (variance) is added, there is also more noise added. It is worth pointing out that the cluster PCA has a similar outlier detection performance to MCD, but it has a much better false alarm ratio.

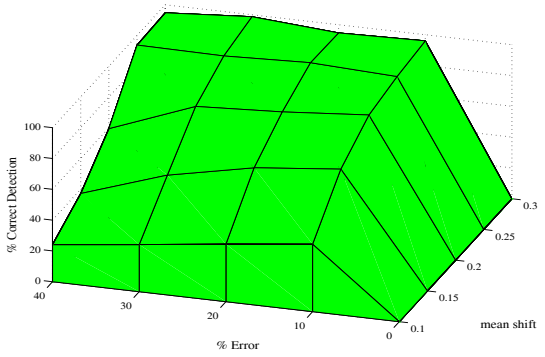

Fig. 9. Cluster PCA outlier detection performance in a consecutive mean environment (80% of variance used).

B. Second dataset

The main limitation of the MVE and MCD algorithms is their inapplicability to datasets having more variables than observations, that is when $p > m$. Our proposed cluster PCA

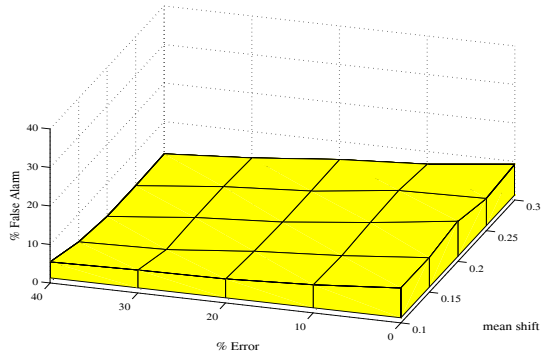

Fig. 10. Cluster PCA false alarm performance in a consecutive mean environment (80% of variance used).

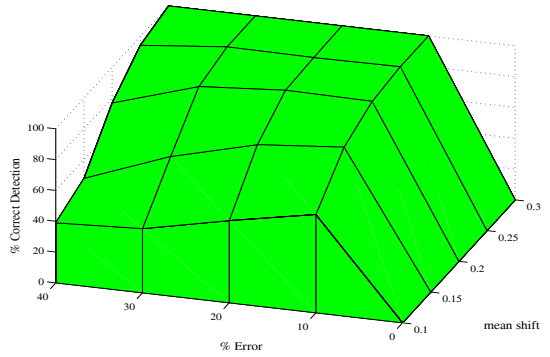

Fig. 11. Cluster PCA outlier detection performance in a consecutive mean environment (100% of variance used).

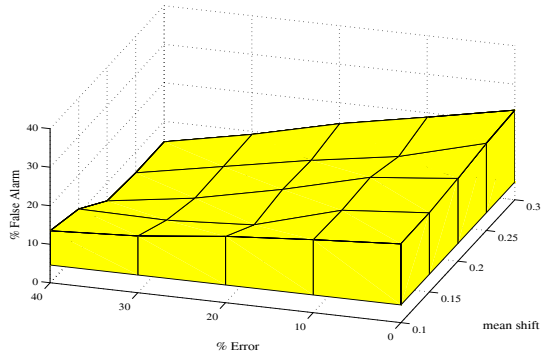

Fig. 12. Cluster PCA false alarm performance performance in a consecutive mean environment (100% of variance used).

is, however, applicable to such datasets as will be shown in the sequel. To this end, we generated a Gaussian-mixture dataset $X_{m\times p}$ with $(m, p) = (50, 100)$.

The performance of cluster PCA with different percentages of variance is depicted in Fig. 13 through Fig. 15, where it can be clearly observed that for high dimensional data the more components k are selected the more accurate the outlier detection performs. Also, the performance is similar in both consecutive and scattered mean shift as illustrated in Fig. 17. Moreover, the false alarm ratio is virtually zero across all parameters and environments as shown in Fig. 14, Fig. 16 and Fig. 18.

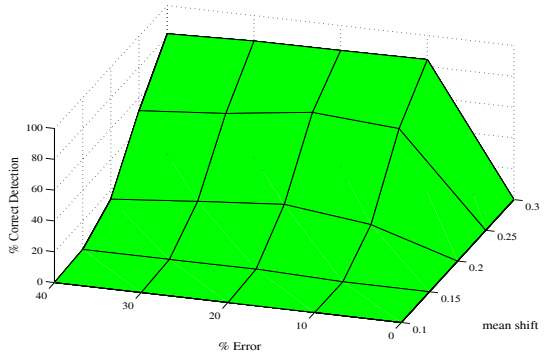

Fig. 13. Cluster PCA outlier detection performance in a consecutive mean environment (80% of variance used).

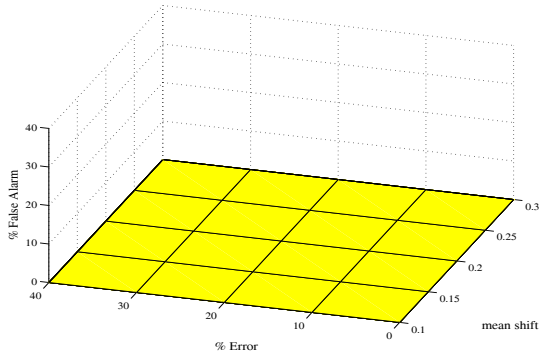

Fig. 14. Cluster PCA false alarm performance in a consecutive mean environment (80% of variance used).

VI. Conclusions

In this paper, we introduced a new multivariate outlier detection algorithm by combining hierarchical clustering and principal component analysis. The core idea behind our proposed technique is to use clustering analysis to determine an optimal subset of observations that is used to calculate the robust mean and the robust covariance matrix of the standardized data, followed by applying the PCA algorithm in order to robustly detect the outliers. The experimental results clearly show a much improved performance of the proposed approach in comparison with existing methods.

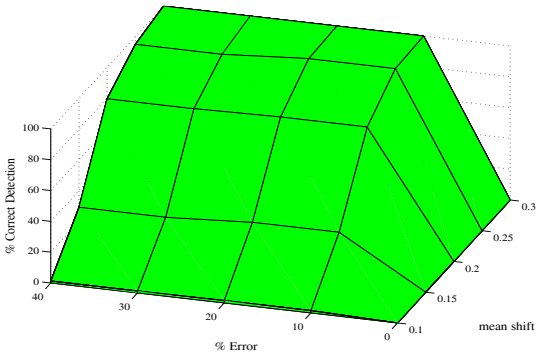

Fig. 15. Cluster PCA outlier detection performance in a consecutive mean environment (99% of variance used).

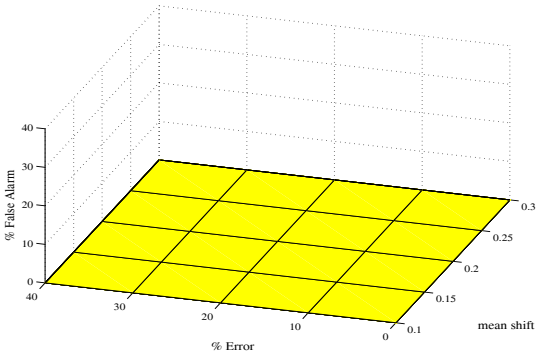

Fig. 16. Cluster PCA false alarm performance in a consecutive mean environment (99% of variance used).

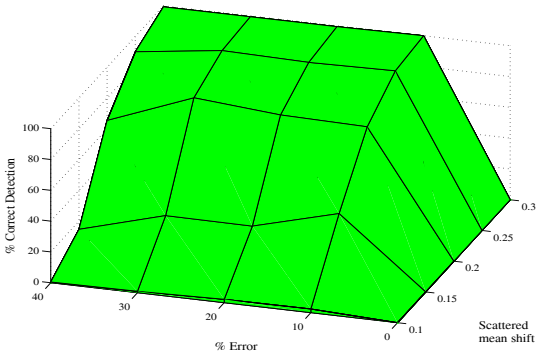

Fig. 17. Cluster PCA outlier detection performance in a scattered mean environment (99% of variance used).

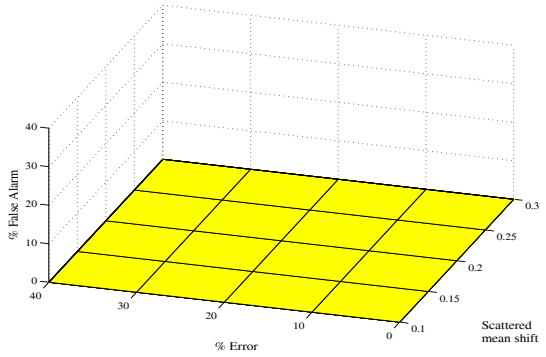

Fig. 18. Cluster PCA false alarm performance in a scattered mean environment (99% of variance used).

REFERENCES

[1] M.S. Chen, J. Han, and P.S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and Data Engineering* vol. 8, pp. 866-883, 1996.

[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[3] P.J. Rousseeuw, and K.V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212-223, 1999.

[4] M. Hubert, P.J. Rousseeuw, and K.V. Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, pp. 64-79, 2005.

[5] P. Filzmoser, "A multivariate outlier detection method" *Proc. International Conference on Computer Data Analysis and Modeling*, vol. 1, pp. 18-22, 2004.

[6] W.J. Egan and S.L. Morgan, "Outlier detection in multivariate analytical chemical data," *Analyical Chemistry*, vol. 70, pp. 2372-3279, 1998.

[7] J.A. Vargas, "Robust estimation in multivariate control charts for individual observations," *Journal of Quality Technology*, vol. 35, no. 4, October 2003.

[8] F. Angiulli, S. Basta, C. Pizzuti. "Distance-based detection and prediction of outliers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 145-160, 2006.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, 2nd Edition, Wiley Interscience, 2000.

[10] I.T. Jolliffe, *Principal Component Analysis*, New York: Springer, 1986.

[11] C.C. Aggarwal, and P.S. Yu, "Oulier detection for high dimensional data," *Proc. ACM SIGMOD*, 2001.

[12] D.C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, 2005.

[13] K. Yang and J. Trewn, *Multivariate Statistical Methods in Quality Management*, Mc Graw Hill Professional, 2004.

[14] F.A. Alqallaf, K.P. Konis, and R.D. Martin, "Scalable robust covariance and correlation estimates for Data Mining," *Proc. ACM SIGKDD*, 2002.

[15] P.J. Rousseeuw and A.M Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, NY, 1987.

[16] S. Engelen, M. Hubert, and K. Vanden Branden, "A comparison of three procedures for robust PCA in high dimensions," *Austrian Journal of Statistics*, vol. 34, pp. 117-126, 2005.