A ROBUST PCA MODELING METHOD FOR PROCESS MONITORING

D. Wang and J.A. Romagnoli

Dept. of Chemical Engineering,
The University of Sydney, NSW 2006 Australia

Abstract: A robust method for dealing with the gross errors in the data collected for PCA model is proposed. This method, using M-estimator based on the generalized t distribution, adaptively transforms the data in the score space in order to eliminate the effects of the outliers in the original data. The robust estimation of the covariance or correlation matrix is obtained by the proposed approach so that the accurate PCA model can be obtained for the process monitoring purpose. Comparisons with the conventional PCA modeling and other robust outlier's replacement approaches are illustrated through a chemical engineering example.

Keyword: Principal component analysis, multivariate outliers, robust estimation, winsorization, process monitoring.

1 INTRODUCTION

Data-driven process monitoring based on multivariate statistic techniques is widely used in chemical industries with a large amount of measurements provided by the modern hardware. The measurement variables are usually highly correlated and the real dimensionality of the process variables is considerably less than that represented by the number of process variables collected. Monitoring this "data rich" process inevitably need dimensionality reduction techniques to grasp the driven force embedded in these measurements. By converting the large amount of data collected from the process into a few meaningful measures, one can assist the industrial operators in determining the status of the operations and in detecting and diagnosing the faults. Principal components analysis (PCA) is such a dimensionality reduction technique and it is heavily used in modeling the multivariate process for monitoring purpose (Kresta, et. al., 1991).

The performance of PCA model is based on the accurate estimation of the covariance or correlation structure of the data. The optimality of the conventional PCA is based on the assumption that the data are normal distributed around their locations with the scales. However, normal distribution usually dose not exist in real chemical engineering practice, it is hard to assure the normality even for high quality measurements. Specially, the frequent presence of gross errors and outliers violates the assumptions in the conventional approach (even through the data is auto scaled) and makes the results invalid (Hoo, K. A. et. al., 2002).

Several approaches can be employed to alleviate the outlier problem in PCA modeling. One of them is based on filter approaches to detect the gross outliers and delete them or replace them with some values before the conventional PCA is used. This pre-treatment approach is intuitive but it may suffer information and performance loss due to its subjective or ad hoc fashion. In addition, the multiple outliers are hard to be detected by using univariate techniques, which will result in the loss of efficiency. Another approach is based on the robust estimation of covariance or correlation matrices of the data. Some of the methods used are multivariate trimming (MVT), minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE) (Devlin et. al., 1981). In MVT, a certain percentage of the observations with highest Mahalanobis distance (MD) are removed and the covariance matrix is formed using the remaining observations. In MCD, a subset of data is formed by randomly selecting some percentage of the samples. The determinant of this subset of data is then computed. The mean and S.D. of the data subset with the minimum non-zero determinant are then used to calculate the covariance or correlation matrices. In MVE, the smallest set ellipse, which contains half of the data, is obtained. The mean and S.D. of the samples inside this ellipse are calculated and rescaled so that they estimate a multivariate normal distribution. Such techniques may be suffered with the disadvantage that ignoring the data which are believed to be "good" by process operators will inevitably result the information loss.

In order to maximally use the information provided in the data while lessening the effects of the outliers, other robust approaches have been investigated. In Hybrid projection pursuit (HPP), an M-estimator like formulation is used for weighting each observation in the data set according to its MD so that a weighted PCA is proposed with eliminating the 'discontinuity problems' in projection pursuit (Chen et. al., 1996). However, since HPP relies on the MD, the presence of multiple outliers may yield erroneous results (Hoo, K. A. et. al., 2002). Recently, a method of robust multivariate outlier replacement was developed for PCA modeling (Hoo, K. A. et. al., 2002). In this approach, a winsorization, which is a procedure that replaces the observations by its pseudo values, is carried out iteratively in score space obtained in PCA. The data, especially the outliers, are transformed into a

tight cluster of majority of data set so that the effect of outliers can be reduced. A Huber or Hampel like M-estimator is used in the winsorization process. Even through it is effective in eliminating the outliers, this approach could suffer the performance loss. Because by using Huber or Hampel like estimators, one has to specify the breakdown points, which are the degrees of the freedom in the estimators, and these parameters are difficult to be determined as *a priori*. Injudiciously specified parameters will result in performance loss of the method or erroneous results.

In this article, an adaptive robust PCA method is proposed with the aim of maximal use of the information in the data as well as robustness to the deviation from the ideality caused by the outliers. A winsorization procedure is employed in the score space as that in the approach by Hoo, K.A. et. al., but a partially adaptive M-estimator based on the generalized t (GT) distribution is used instead of Huber or Hampel like estimators. This GT based estimator is obtained directly from the data in score space and its influence curve fit the data adaptively. This will improve the performance and it is optimal in MLE sense. By using GT distribution, the data can adjust itself to the shape of its distribution, in such a way the advantages of both robustness and maximum likelihood estimation (MLE) retained.

The paper is organized as follows. In the next section, a brief overview of robust estimate and several robust estimators are introduced. Specially, robust estimator based on GT distribution and its robust properties are discussed. In section 3, after giving a brief introduction of PCA, the proposed robust PCA approach using adaptive GT based M-estimator is developed. The winsorization procedures of the approach are also highlighted in the section. In section 4, the proposed method is implemented and its performance is compared with that of conventional PCA and the robust PCA using Huber's M-estimator through the data collected from a chemical engineering simulation. Finally, the conclusions are given in the section 5.

2 ROBUST ESTIMATES

2.1 M-estimates

The essence of robust estimates can be explained by the simple one-dimension parameter estimation problem

$$y = f(x, q) + e \tag{1}$$

where y, x and ε are the dependent, independent and error variable, respectively. θ is the parameter to be estimated. After collecting a set of data, the parameter θ can be estimated by least squares method,

$$\hat{\theta} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i, \theta)\right)^2 \tag{2}$$

Under the assumption that the error e is normal distributed, the estimation of θ is optimal in the sense of maximum likelihood estimation.

However, if the error is not normal distributed, especially there are outliers in the data, the above estimate will be biased. This problem can be solved by designing a robust estimator, which is insensitive to the deviation of the assumption for the majority of data. The design of this estimator is usually converted into choosing the objective function $\rho(u)$ (not necessary the quadratic one as in the conventional approach) in the optimization problem

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \rho\left(y_i - f(x_i, \theta)\right) \tag{3}$$

where $\rho(u) = -\ln p(u)$, $p(u)$ is the PDF of the residual $u_i = y_i - f(x_i, \theta)$. Solving the following equations can also solve the above problem,

$$\sum_{i=1}^{n} \psi\left(y_i - f(x_i, \theta)\right) = 0 \tag{4}$$

where $\psi(u) = \rho'(u)$

To be robust, the objective function must give less weight to large value of u than its quadratic form u^2 so that the estimator will down-weight or ignore the contribution of the large errors in the data. In problem (3), a number of candidates $\rho(u)$ can be chosen as the objective function so that different robust estimators can be obtained. These kinds of estimators correspond to the M-estimators in robust statistics. In robust estimation, the ψ function or influence function defined by $\psi(u) = \partial\rho(u)/\partial u$ is the usual tool for comparing alternative M-estimators for their robustness. The ψ function measures the "influence" that a residual will have on the estimation process. Some suggested criteria for the ψ function are that: it is (a) bounded, (b) continuous, and (c) descending and identically zero outside an appropriate region. The motivation for these properties is that (a) a single "anomalous" observation would have limited influence on the estimator, (b) grouping or rounding of data would have minimal impact on the estimator, and (c) ridiculously large observations would have no impact on the estimator.

The typically used robust estimator is Huber's (Huber, 1981)

$$\rho(u) = \begin{cases} \dfrac{u^2}{2} & |u| \le k \\ k|u| - \dfrac{k^2}{2} & k < |u| \end{cases} \tag{5}$$

The influence function is in this case

$$\psi(u) = \begin{cases} -k & u < -k \\ u & -k \le u \le k \\ k & k < u \end{cases} \tag{6}$$

where, k is the parameter characterizing the degree of contamination, being used as tuning parameter for the estimator's performance. Other robust estimators can be found in the literature (Wang, et. al 2000).

Even though the above approaches are less sensitive to gross errors and outliers, the optimality of the estimation, in MLE sense, is still dependent on the suitability of the chosen function with respect to the actual distribution of the data. It is generally hard to characterize the distribution of the errors correctly without posteriori estimation. If the real errors do not follow the specified distributions, the performance of estimator may deteriorate and the estimation could be biased. Considering these disadvantages, a more flexible probability distribution function will be discussed next to describe the error distribution. It is designed to allow for a variety of thickness of tails, to capture the shape of distribution and to accommodate other distribution as much as possible as special cases. The corresponding estimator will then be robust by its ψ function and be efficient by estimating its distributional parameters from the data in the MLE sense.

2.2 The generalized T density and its robust properties

The proposed robust estimator for PCA modeling in this work is based on the assumption that the data in score space is following the generalized T distribution (GT) (Butler, et. al., 1990), which has flexibility to accommodate various distributional shapes

$$f_{GT}(u;\sigma,p,q)=\frac{p}{2\sigma q^{1/2}B(1/p,q)\left(1+\frac{|u|^p}{q\sigma^p}\right)^{q+1/p}} \quad -\infty<u<\infty \tag{7}$$

Where σ, p, q are distributional parameters, σ corresponds to the standard deviation, p and q are parameters corresponding to the shape of distribution. This density is symmetric about zero, uni-modal, and suitable to describe the error characteristics in most cases. By choosing different values of p and q, the GT distribution will accommodate the real shape of the error distribution. The larger the value of p or q, the "thinner" will be the tail of the density. Similarly, smaller values of p and q will be associated with "thicker" tails. The tails behavior and other characteristics of the distribution, depend upon these two distributional parameters, which will be estimated from the data (Wang et. al., 2003). In addition, the GT distribution defines a very general family of density functions and combines two general forms, which include most of stochastic specifications one meets in practice as special or limiting cases.

The robustness of the estimator based on a GT distribution can be discussed by investigating its ψ function. This ψ function, corresponding to the objective function $\rho(u,\sigma,p,q)=-\log f_{GT}(u;,\sigma,p,q))$ is given by

$$\psi_{GT}(u;\sigma,p,q)=\frac{(pq+1)sign(u)|u|^{p-1}}{q\sigma^p+|u|^p} \tag{8}$$

For finite q, this influence function is bounded and reaches a maximum for positive u at $u^*=\left((p-1)q\sigma^p\right)^{1/p}$ and has a maximum value of

$$\psi(u^*;\sigma,p,q)=\frac{(p+1/p)(p-1)^{[(p-1)/p]}}{\sigma q^{1/p}} \tag{9}$$

Furthermore, $\lim_{u\to\infty}\psi(u;\sigma,p,q)=0$, so this influence function exhibits a descending pattern. Consequently, "large" deviation will not have an impact on this estimator when q is finite. Also, for a given finite q, p control the behavior of $\psi(u;\sigma,p,q)$ near the origin. For example, if $p>2$, then this influence function will be less steeply sloped near the origin than the influence function for the t distribution with $2q$ degrees of freedom.

3 PCA AND ITS ROBUSTNESS BASED ON M-ESTIMATE WINSORIZATION

3.1 Principal Component Analysis

The cornerstone of data-driven process monitoring approach is the projection method of PCA. The philosophy of this technique is to reduce the dimensionality of the problem by forming a new set of variables. The method generates the new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. The first few principal components can capture the most of the variance in the data so that they are used as the model. The new data will be fitted by the model in order to see if the measures developed are in the normal range.

Let X be a $n\times m$ data matrix containing n process measurements of m variables $(m\le n)$. PCA decomposes the observation X as

$$X=TP^T=\sum_{i=1}^{m}t_i p_i^T \tag{10}$$

Mathematically, p_i and t_i can be calculated by finding the eigenvalues and their companion eigenvectors of covariance or correlation matrix S of data X,

$$\begin{aligned} S&=P\Lambda P^T \\ T&=XP \end{aligned} \tag{11}$$

where Λ is the diagonal matrix containing the ordered eigenvalues of S and P is the corresponding eigenvector matrix. In PCA, P is defined as loading matrix and T is defined to be the matrix of principal component scores. The loadings provide information as to which variables contribute the most to individual PCs and they are the coefficients on the principal component model; whilst the score matrix provide the information on the clustering of the samples and the identification of transitions between different operating regimes.

In general, if the process variables are collinear, the first k principal components can be used to explain sufficiently the variability in the whole data set with less information loss, and the determination of the number k can be obtained via several techniques such as scree test and cross-validation. It then follows that

$$X = T_k P_k^T + E = \sum_{i=1}^{k} t_i p_i^T + E \tag{12}$$

$$\hat{X} = T_k P_k^T = \sum_{i=1}^{k} t_i p_i^T \tag{13}$$

Once the PCA model is established, analysis and usage of these lower dimension orthogonal variables are preceded and the measures such as T^2 and SPE along with some visualization plots in score space can be employed for process monitoring.

3.2 Robust PCA Based on M-estimate Winsorization

PCA transform the data set by projection onto loading vectors to form score vectors which are uncorrelated. Hence, univariate concepts can be employed in the score space. The outliers present in the original data manifest themselves in the score space. By recurrently winsorizing the scores and replacing them with suitable values, it is possible to detect multivariate outliers and replace them by values, which conform to the correlation structure in the data. The concept of winsorization is briefly explained first and its application to robust PCA is then investigated.

Consider the linear regression problem

$$y = f(X, \theta) + \varepsilon \tag{14}$$

where $y = (y_1, y_2, ..., y_n)'$ is a $n \times 1$ vector of dependent variables, $X = (x_1', x_2', ..., x_n')'$ is a $n \times m$ matrix of independent variables, and θ is a $p \times 1$ vector of parameters, e is a $n \times 1$ vector of model error or residual. An estimation of parameter θ, $\hat{\theta}$, can be obtained by minimizing the function

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \rho\left(\frac{y_i - f_i(x_i, \theta)}{s} \right) \tag{15}$$

where s is an estimation of the scale of the distribution of residuals and ρ is objective function to be minimized.

With the parameter $\hat{\theta}$ estimated, the prediction or estimation of the dependent variable $y_i \, (i = 1, ..., n)$ is given by

$$\hat{y}_i = f_i(x_i, \hat{\theta}) \tag{16}$$

and the residual is given by

$$r_i = y_i - \hat{y}_i \tag{15}$$

In winsorization process, the variable y_i is transformed using pseudo observation according to specified M-estimates such as Huber's;

$$y_i^w = \begin{cases} \hat{y}_i - k s_i & r_i < -k s_i \\ y_i & |r_i| \le k s_i \\ \hat{y}_i + k s_i & r_i > k s_i \end{cases} \tag{17}$$

here the parameter k is the degree of freedom, which regulate the amount of robustness and s_i is the estimation of scale associated with r_i. Other robust estimates can also be employed, especially the one based on GT distribution:

$$y_i^w = \psi_{GT}(y_i; \sigma, p, q) \tag{18}$$

σ, p, q are the parameters which accommodate the shape of the residual distribution. These parameters can be estimated with the data y_i.

The technique of winsorization can be used in PCA to eliminate the effects of outliers in the following. The data value y in score space can be transformed into a new value y^w by winsorization as follows,

$$y_i^w = \psi(y_i), \qquad i = 1, 2, ...n \tag{19}$$

where y is any robust influence function discussed before. Using the winsorization process, the large values exhibited as outliers in the original data set will be brought closer to the other observations after they are transformed from the score space back to the original data space. A new PCA model is obtained using the new data set. This process is carried out iteratively until there is not much change in the loading vectors.

The advantage of using GT based robust estimate over Huber-like robust estimate is obvious. The GT based approach can accommodate the shape of the residual distribution so that it should be more effective when the winsorization is processed, because the estimate is optimal in MLE sense. Huber-like approach needs to pre-specify the robust parameter in an ad hoc manner, this may result in the inefficiency of the estimation.

The steps of the robust PCA based on GT winsorization are described as follow:

1) Scale the data matrix X^j (j is the iteration number, $j = 1,2,...$) using some estimates of scale and location (μ^j, σ^j). Calculate the correlation matrix S.
2) Apply PCA to the correlation matrix S and generate the PC loadings and scores
3) Fit the score data to the GT distribution and calculate its influence function y . Winsorize the score space variables using the transformation:
$$t_i^{j,w} = \frac{t_i^j + \psi(t_i^j)}{2} \quad i = 1,2,...,n$$
4) Reconstruct the actual data using the loading vector and winsorized score vector,
$$X^{j,w} = T^{j,w} P^{jT} = \sum_{i=1}^{m} t_i^{j,w} p_i^{jT}$$
5) Check for the convergence of the loading vectors, $\max\left(\left\|P^j - P^{j-1}\right\|\right) \le \varepsilon_s$ where ε_s is a user-defined threshold.
6) If convergence is not achieved at iteration j, then go back to step 1, otherwise stop.

4 SIMULATION STUDY

The heat-exchanger network example (Romagnoli and Sanchez, 2000) will be used to demonstrate the performance of the proposed robust PCA modeling method based on GT winsorization (Figure 1).

Process stream A is heated by process streams B, C and D at various junctions. The system has 16 measured variables which are either flow rates or temperatures. The open loop data are generated by adding Gaussian noise with zero mean and variances of 2% of their values on all the values when the process is operating at the normal conditions. 200 samples are generated and the sampling time is 0.1 hour.

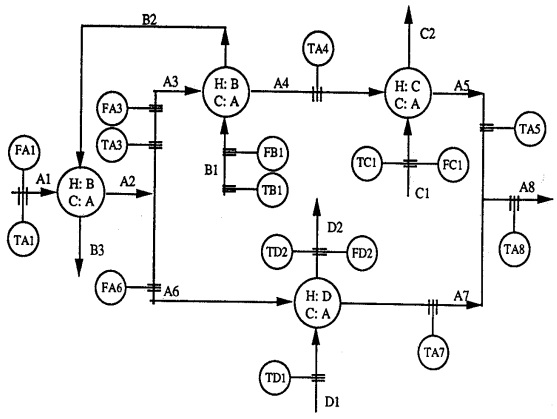

Figure 1 Heat Exchange Network

The data generated above are treated as real good data set and labeled as X^*. In order to compare the performance of the proposed method with the others, outliers are introduced by adding randomly to anywhere in X^* with the larger values (variances up to 10% of their median values) from different error distribution such as Gaussian, *t* and non-central *t* distribution. The case of non-central *t* distribution will be reported here. Figure 2 shows the measurement data corrupted by non-central *t* distribution. This corrupted data set is labeled as X.

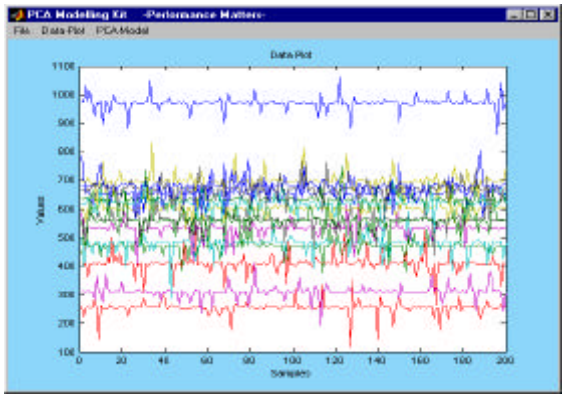

Figure 2 Data Set X

The performance criterion which will be used to compare the efficiencies of the various PCA methods is the mean-squared error (MSE):

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\lambda_i^* - \lambda_i\right)^2 \qquad (20)$$

where, is the eigenvalue vector of data matrix , is the eigenvalue vector of reconstructed data matrix by different PCA approaches with corrupted data , is the dimension of the variables or the number of principal components chosen. The MSE is constructed such that better performance is obtained if its value is driven toward zero.

Figure 3 shows the normal plot of the data . If the data fell on the straight lines, then their distribution is assumed to be normal. Clearly, it shows that the data are not normal distributed.

Three PCA methods are applied to the data , the results are shown in the tables. Table 1 lists the explained variation in the data by each eigenvalue along with the cumulative percentage of the explained variation. If the selection of the number of principal components is based on a requirement that 85% of the variance be explained, then eight principal components are required for original data. However, for the same criterion seven principal components are required based on the conventional PCA with the corrupted data . The robust PCA approaches can recover the real variation explained by the principal components so that they diminish the effects of outliers in the data. The proposed GT based winsorization has better performance than the winsorization

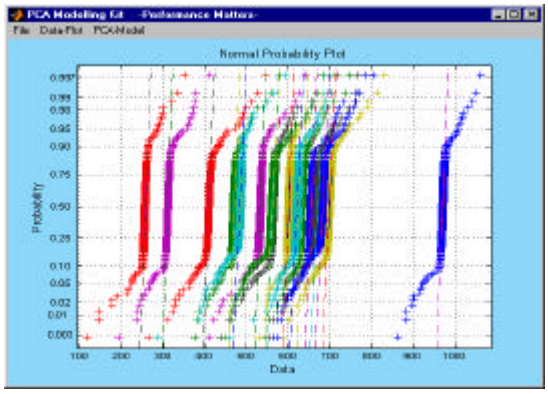

Figure 3 Normal Plot of Data

using Huber estimator. This is due to that GT winsorization can accommodates the distribution of the data, while winsorization based on Huber estimator relies on *a priori* parameters so that it is not adaptive. Judiciously specifying the threshold in Huber-like winsorization may improve the performance. The normal plots of the filtered data are given in Figure 4. It is shown that after the winsorization based on the GT, the distributions of the data are normal so the correct results can be obtained by using PCA with the data X. The MSE criterion is listed in table 2, which shows that the proposed approach has the best performance. It is also observed in table 1 that for the outliers-free case, the proposed robust PCA approaches still have acceptable performance.

5 CONCLUSIONS

A Robust PCA modeling method based on winsorization in score space using adaptive robust estimator was developed and presented. The effects of outliers in the data can be eliminated by the method while the effectiveness as well as the robustness is kept by using GT-like estimator. The performance of the proposed method is compared with the others using a chemical example. The usage of the approach in process monitoring such as faults detection, identification and diagnosis is promising.

6 REFERENCES

Butler, R. J., McDonald, J. B., Nelson, R D., White, S. B.(1990), "Robust and Partially Adaptive Estimation of Regression Models", The Review of Economics and Statistics, 1990, 321

Chen, J., A. Bandoni and J. A. Romagnoli (1996), "Robst PCA and Normal region in multivariate statistical process monitoring", AIChE. J. 42(12), 3563-3566

Devlin S. J., Guanadesikan, R. and Kettenring, J. R. (1981), "Robust Estimation of Dispersion matrices and principal components", J. of Amer. Stats. Asso. 76, 354-362

Hoo. K.A., K.J. Tvarlapati, M.J. .Piovoso and R. Hajare, (2002) "Amethod of robust maultivariate outliers replacement", Comp. Chem. Eng. 26. 17-39

Huber, P. J., (1981), "Robust Statistics", *John Wiley & Sons, New York, NY.*

Kresta, J. V. J. F. MacGregor and T. E. Marlin (1991), "Multivariate statistical monitoring of process operations", Can. J. of Chem. Eng., 69, 35.

Romagnoli, J. A. and M, C. Sanchez, (2000), "Data Processing and Reconciliation for Chemica Process Operations", *Academic Press.*

Wang, D., Safavi A. and Romagnoli J. A., (2000), "Wavelet-based adaptive robust M-estimator for nonlinear systems identification", AIChE Journal, 46, 8. 1607-1615.

Wang, D. and Romagnoli, J. A. (2003), "A Framework of Robust Data Reconciliation Based on a Generalized Objective Function", to appear on J. of Industrial and Engineering Chemistry Research.

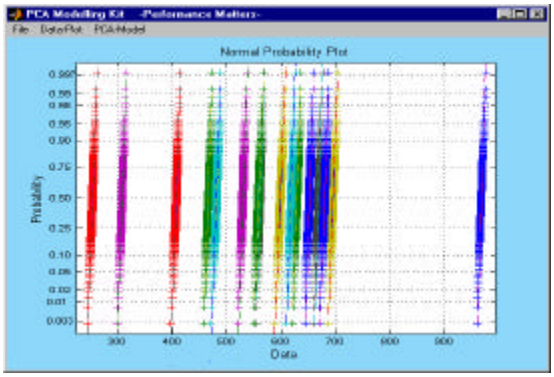

Figure 4 Normal Plot Data after Winsorization

Table 1. Results of PCA

Num. of PCs	Explained variance(%) Data X^*	Cumulative explained variances(%)	Explained variance(%) X	Cumulative explained variances(%)
	conventional			
1	17.6918	17.6918	21.3087	21.3087
2	16.1712	33.8630	16.2429	37.5515
3	15.6300	49.4930	14.5289	52.0805
4	10.0487	59.5417	11.9913	64.0718
5	9.2196	68.7613	9.3763	73.4482
6	8.3142	77.0755	6.9700	80.4182
7	6.7640	83.8395	5.4210	85.8392
8	5.4261	89.2656	4.3469	90.1862
	Huber's			
1	18.7511	18.7511	19.3563	19.3563
2	17.1491	35.9002	15.3889	34.7452
3	11.8246	47.7249	11.6123	46.3576
4	10.2274	57.9523	10.8261	57.1837
5	9.2651	67.2174	9.2336	66.4173
6	7.5333	74.7507	7.6582	74.0755
7	6.0656	80.8163	6.7601	80.8356
8	4.7430	85.5593	5.9184	86.7540
	GT			
1	17.2848	17.2848	17.3761	17.3761
2	16.4029	33.6877	14.3915	31.7676
3	13.0628	46.7505	14.2124	45.9801
4	10.2440	56.9946	11.4886	57.4687
5	9.2602	66.2548	9.3367	66.8054
6	7.3126	73.5674	8.1144	74.9198
7	6.9677	80.5351	6.6013	81.5211
8	6.1711	86.7062	5.1630	86.6841

Table 2. MSE values of different methods

	Conventional PCA	Huber's Winsorization	GT Winsorization
MSE	4.7826	2.6002	0.9374