PERGAMON

Control Engineering Practice 8 (2000) 531–543

www.elsevier.com/locate/conengprac

The application of principal component analysis and kernel density estimation to enhance process monitoring

Q. Chen[a], R. J. Wynne[b,*], P. Goulding[c], D. Sandoz[a]

[a]*Control Technology Centre Ltd., School of Engineering, University of Manchester, Manchester, M13 9PL, UK*
[b]*School of Engineering, Sheffield Hallam University, Sheffield, S1 UK*
[c]*Mechanical Engineering Division, School of Engineering, University of Manchester, Manchester, UK*

Received 9 September 1998; accepted 6 October 1999

Abstract

This paper discusses the application of kernel density estimation (KDE) and principal component analysis (PCA) to provide enhanced monitoring of multivariate processes. Different KDE algorithms are studied and assessed in depth in the context of practical applications so that one bandwidth selection algorithm is recommended for process monitoring. The results of the case studies clearly demonstrate the power and advantages of the KDE approach over parametric density estimation which is still widely used. Statistical summary charts are suggested to raise early warning of faults and locate the physical variables which are the prime indicators of the faults. © *2000 Elsevier Science Ltd. All rights reserved.*

Keywords: Multivariate processes; Principal component analysis; Kernel density estimation; Process monitoring

1. Introduction

The requirement for improved productivity and energy utilisation, reduced levels of manning, and for safer processes has led to increased research activity in fault detection and isolation. Many companies already have well-established condition monitoring regimes for critical plant items that together with predictive maintenance have resulted in significant improvements in plant utilisation. However, classical condition monitoring methods can themselves lead to data overload and with the increasingly easy access to plant data there is a temptation to monitor ever more variables. In a recent discussion within the steel industry it came to light that 94 variables were being monitored in one plant in an attempt to isolate recurring faults. The success of this level of monitoring is dependent on the skill of the plant engineer to manually reduce the number of variables so that sensible deductions can be made about potential fault conditions.

One approach that has gained rapid acceptance within the process industries is principal component analysis (PCA) (MacGregor & Kourti, 1995; Martin, Morris & Zhang, 1996). This method provides a means of reducing the dimensionality of the data so that only that part of the data that reflects the most significant variability is retained for further analysis. Whilst PCA does provide an effective tool for determining the critical features of the data, further interpretation of the data through the use of the residuals and Hotelling's T^2 analysis, for example, is based on the assumption that the probability density function (PDF) of original data can be described by a known density estimate such as a Gaussian or normal density estimate. In the vast majority of cases this will clearly not be true and the form of the density estimate cannot be realised with much precision, if at all, using a parametric approach due to the nature of the process data. To overcome this limitation to data interpretation it is necessary to use data-driven techniques such as non-parametric empirical density estimates using kernel extraction (Chen, Bandoni & Romagnoli, 1996). Martin and Morris (1996) also use kernel extraction but their approach is based on the 'M^2 statistic'. Whilst this leads to improved density estimates the problem of bandwidth selection, although recognised, was not considered in any detail in that paper.

In kernel density estimation the selection of the bandwidth is of extreme importance and provides a significant

* Corresponding author. Fax: 0114-225-3433.
E-mail address: r.j.wynne@shu.ac.uk (R. J. Wynne).

PII: S0967-0661(99)00191-4

challenge when applying this approach in the case of complex process data. In the context of this work the bandwidth is also known as the smoothing parameter as it has the effect of smoothing the estimated density function. There are several methods available for bandwidth selection and these are discussed at length in the statistical literature, referenced below, but they need to be assessed from the applications point of view. Three of the most important methods for bandwidth selection are considered in detail and their performance reviewed in relation to process monitoring applications. The three methods are:

- Mean integrated square error (MISE) cross validation (Bowman, 1984)
- Adaptive mean integrated square error (ADMISE) cross validation (Silverman, 1986)
- Biased asymptotic mean integrated square error (BAMISE) cross validation (Wand, 1995a)

The review of performance is based on the results of their application to a benchmark data set (Jackson, 1991) and data from an industrial process. The outcome and a contribution of this study is that the BAMISE technique is shown to yield superior results to the other two.

Once a density estimate and associated confidence limits have been established the next challenge is to provide the multivariate information in a form that is readily assimilated. Therefore, to ensure that no information contributed by the selected principal components is ignored, a chart of the density function, in effect a quality control chart (QC chart), is introduced for monitoring the state of overall statistical control of the process. If a warning signal is issued by the QC chart, the squared prediction error chart or error contribution chart (EC chart) can then be used to trace the variable(s) causing the abnormal operating condition.

The results from a waste melter show that the procedures proposed in this paper work well. A fault (crack on the melter) is detected by the QC chart about two hours earlier than more conventional methods and before the fault develops to its final stage. The EC chart is then used to indicate that the fault can be traced through monitoring a particular temperature sensor. In the paper principal components are initially reviewed and this is followed by an introduction to the use of kernel density estimates. The results of applying the technique to the melter are then discussed and finally a procedure is described for validating the density estimates.

2. Principal components

The PCA technique determines combinations of variables that describe major trends, or variation, in a data set. For a data matrix, $\mathbf{X}_0$, with m rows (of measurements or samples) and n columns (of variables), the covariance matrix of $\mathbf{X}_0$ is defined as

$$\mathrm{cov}(X_0) = \frac{\mathbf{X}_0^t \mathbf{X}_0}{m-1} = \Sigma,$$

where the columns of $\mathbf{X}_0$ have been centred to have a mean of zero and scaled to a standard deviation of one, then the result will be the correlation matrix Σ. PCA decomposes this data matrix, $\mathbf{X}_0$, as the sum of the outer product of vectors, $\mathbf{t}_i$ and $\mathbf{p}_i$ plus a residual error matrix $\mathbf{E}$ (Wise & Gallagher, 1996):

$$\mathbf{X}_0 = \mathbf{t}_1\,\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\,\mathbf{p}_2^{\mathrm{T}} + \dots + \mathbf{t}_k\,\mathbf{p}_k^{\mathrm{T}} + \mathbf{E}, \tag{1}$$

where $\mathbf{t}_i$ are the score vectors and $\mathbf{p}_i$ are the loading vectors. The scores describe the relationship between the samples and the loading matrix the relationship between the variables. This can be summarised as

$$\mathbf{X}_0 = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}. \tag{2}$$

It is now possible to generate a new data set based on the number of principal components retained (k). In general, k will normally be much smaller than the number of variables in the original data.

Once the principal components have been calculated, and those of interest retained, it is possible to calculate values to determine whether the process is in control or not. Although the retained principal components together with the loading vectors do not constitute a 'model' of the system, it is possible to treat them as such because together they describe the most important process variation and can be recombined to produce estimates of the original data matrix, $\mathbf{X}_0$. Furthermore, it is possible to project a new data set onto a 'model' created from normal process operation to determine whether a process disturbance has occurred during the collection of the new data set.

3. Kernel density estimation and bandwidth selection

KDE refers to a class of data-driven techniques for the nonparametric estimation of density functions. It is a powerful tool that can extract an empirical distribution density function from a given sample from a population under consideration. See the work of Silverman (1986) and Wand and Jones (1995a) for further insight into this approach to density estimation.

Consider a kernel function $K(\mathbf{x})$ and a sample $D_0 = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ from a population distribution density $f(\mathbf{x})$, then the density estimator of the sample can be expressed as

$$\hat{f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)), \tag{3}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_d)^{\mathrm{T}}$ is the d-dimension vector, which comprises the variables of the system, and $\mathbf{H}$ is the smoothing matrix or bandwidth matrix. In practice, the form of the kernel function is not very important, and the normal distribution kernel function will be used in all cases. The critical issue in forming density estimates, however is the *bandwidth* (smoothing parameter or window width) selection which is now considered.

3.1. Leave-one-out cross validation

Once a model structure has been defined, a cross validation method (Stone, 1974) can be used to determine the model parameters. Ideally, this is based on the prediction error between a test data set and the model fitted to a training data set that does not contain any part of the test data. The model parameters are determined by minimising the prediction error. However, one of the most widely used cross-validation methods is the leave-one-out algorithm in which the test data set contains all but one observation set. Leave-one-out cross validation is at the heart of most bandwidth selection algorithms for kernel density estimates.

3.2. Mean integrated squared error (MISE) cross validation

Based on the leave-one-out scheme, the MISE cross validation (Bowman, 1984) for a univariate density function can be extended to multivariate problems with little modification. The first requirement is the choice of bandwidth matrix, **H**, to be used. There are three possible approaches to this.

(1) For a problem with d-variables, a full symmetrical bandwidth matrix

$$\mathbf{H} = \begin{bmatrix} h_{11}^2 & \dots & h_{1d}^2 \\ \vdots & \cdots & \vdots \\ h_{d1}^2 & \dots & h_{dd}^2 \end{bmatrix}$$

which is a positive-definite matrix with $d(d+1)/2$ parameters, in which $h_{ik}^2 = h_{ki}^2$;

(2) A diagonal matrix with only d parameters, $\mathbf{H} = diag(h_1^2, h_2^2, \dots, h_d^2) = \mathbf{h} \otimes \mathbf{h}$, where $\mathbf{h} = (h_1, h_2, \dots, h_d)^{\mathrm{T}}$. Operator $\otimes$ performs element to element product;

(3) A diagonal matrix with one parameter, $\mathbf{H} = h^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix.

Approach one is the most accurate but is unrealistic in terms of computational load and time. Approach three is the simplest and the method used in the case of univariate data and can be adopted with slight modification. However, it can cause problems in some situations due to the loss of accuracy by forcing the bandwidths in all dimensions to be the same. This can introduce significant error in the density function shape. However, if the data are 'sphered' before being used to fit the density function i.e. re-scale the data to be of the same scale in all dimensions (Fukunaga, 1990), the third approach could be a reasonable and practical choice for its simplicity and much reduced computational load. Hence approach two appears to be the appropriate choice, as a compromise, but the computational load is still unacceptable if the dimensionality of the problem is high and the size of the sample is large.

Generally, the determination of **H** is based on minimisation of the global error criterion. In MISE cross validation the criterion is given by

$$MISE\{\hat{f}(:, \mathbf{H})\} = E\int [\hat{f}(\mathbf{x}, \mathbf{H}) - f(\mathbf{x})]^2 \, d\mathbf{x}, \tag{4}$$

where $\hat{f}(\mathbf{x}, \mathbf{H})$ is the fitted density function and $f(\mathbf{x})$ the real density function. Applying leave-one-out cross validation and using the multivariate normal kernel function Eq. (3) gives the equivalent error function (Bowman, 1984).

$$\begin{aligned} MISE\{\hat{f}(:, h)\} = & \frac{1}{n-1} N(0, 2h^2) \\ & + \frac{n-2}{n(n-1)^2} \sum_{i \neq j} N(\mathbf{x}_i - \mathbf{x}_j, 2h^2) \\ & - \frac{2}{n(n-1)} \sum_{i \neq j} N(\mathbf{x}_i - \mathbf{x}_j, h^2), \end{aligned} \tag{5}$$

where $\mathbf{H} = h^2\mathbf{I}$, $N(\mathbf{x}, h)$ is the multivariate Guassian density function, and n is the number of observations. The optimal bandwidth h_{opt} can be obtained by minimising $MISE\{\hat{f}(:, h)\}$ over h. The method is relatively straightforward and works well on data with reasonable variation, however, there are two problems with MISE cross validation. Firstly, it produces a degenerated solution when the smoothed data contains discrete clusters. The optimal bandwidth will always be very close to zero and in multivariate process systems, clusters are almost inevitable. Secondly, the smoothing parameter determined by the MISE cross validation is not stable, it changes dramatically with the number of observations used to fit the density function so that the convergence of the smoothing parameter cannot be guaranteed. These factors can result in the loss of confidence in the derived estimates, particularly in the case of small data sets (Silverman, 1986).

3.3. Adaptive MISE cross validation

In the case where the distribution has long 'tails', the MISE cross validation can have difficulties in coping with the observations contributing to the low-density areas. The adaptive kernel estimator is therefore considered to deal with the long-tail distributions by means of a two-stage procedure. The first stage is to get an approximate density estimate as a pilot. The pilot estimate is then used to determine the pattern of the adaptive bandwidths corresponding to each observation. The bandwidth pattern provides the guide line for the adaptive bandwidth of the density estimate. The second stage is then based on the bandwidth pattern established in the first stage, the bandwidth being obtained by minimising the $MISE\{\hat{f}(:,h)\}$. The strategy is given as follows:

(I) Specify a pilot density estimate $\tilde{f}(\mathbf{x})$ that satisfies $\tilde{f}(\mathbf{X}_i) > 0, \quad i = 1, 2, \dots, n$. The optimal bandwidth of multivariate normal distribution is both convenient and adequate to obtain the pilot density estimate because the optimal bandwidth of the adaptive density estimate is not sensitive to the variation of the *pilot density*.

(II) Determine the *local bandwidth factors* λ_i for each observation by

$$\lambda_i = \{\tilde{f}(\mathbf{X}_i)/g\}^{\alpha}.$$

(III) Define the adaptive kernel estimate $\hat{f}(\mathbf{x})$ by

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d\lambda_i^d}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h\lambda_i}\right),$$

where $K(u)$ is the kernel function.

In step (II) α is a sensitivity parameter the value of which is given in advance. The choice of its value will be discussed later. Compared with the simple MISE cross validation, adaptive MISE cross validation provides more flexibility and gives an improved density estimate when the population has a long tail. In the case of a uniformly distributed sample, the ADMISE and MISE produce very similar bandwidths and density estimates. However, ADMISE can still fail to produce an acceptable bandwidth when dealing with the data containing discrete clusters or small variation.

3.4. Biased asymptotic MISE cross validation

To improve the stability of the MISE cross validation methods, biased MISE cross validation is developed. This is based on the equation for the asymptotic MISE of univariate analysis. For multivariate problems, the criterion function, with the introduction of what is termed 'plug-in' by statisticians, is given by (Wand & Jones, 1995a)

$$BAMISE\{\hat{f}(:,\mathbf{H})\} = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}\mu_2(K)^2(vech\,\mathbf{H})^{\mathrm{T}}\hat{\Psi}_F(vech\,\mathbf{H}), \quad (6)$$

where the factors are defined as

$$R(K) = \int K(\mathbf{x})^2\,\mathrm{d}\mathbf{x}$$

$$\mu_2(K) = \int x_i^2 K(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

and $vech\,\mathbf{H}$ is the notation of a $d(d+1)/2\times 1$ vector containing the entries of matrix $\mathbf{H}$ that are on and below the main diagonal listed column by column. $\hat{\boldsymbol{\Psi}}_F$, the 'plug-in' factor, is a complex $d(d+1)/2\times d(d+1)/2$ matrix dependant on the Hessian matrix of $\hat{f}(\mathbf{x})$.

If a multivariate normal kernel and a single bandwidth scheme are adopted, then the criterion function is reduced to the form

$$\{\hat{f}(:,h)\} = \frac{1}{n(2\pi)^{d/2}h^d} + \frac{1}{4n(n-1)h^d} \times \sum_{i\neq j}\left[\sum_{k=1}^{d}\frac{(x_{ik}-x_{jk})^2}{h^2} - (2d+4)\sum_{k=1}^{d}\frac{(x_{ik}-x_{jk})^2}{h^2} + (d^2+2d)\right]N(\mathbf{X}_i-\mathbf{X}_j,h), \quad (7)$$

where x_{jk} is the value of kth variable at the jth sampling point.

It can be proved numerically that the optimal bandwidth of biased MISE cross validation, h_{BAMISE} is more stable than h_{MISE} in terms that its asymptotic variation is considerably lower. However this reduction in variation comes at the cost of an increase in bias (Wand & Jones, 1995b) but it provides much more consistent results in the case of data with small variation for which the MISE and ADMISE failed. This approach to bandwidth selection for KDE has been used to investigate the plant data.

4. Charting the statistical estimate

There are various ways to chart the statistical estimates to monitor the process performance. Which statistical estimates should be charted depends on the purpose and availability of replications. Frequently charted statistical estimates are SPE (squared prediction error), T^2 distance, etc. From a process monitoring point of view, there are only two important issues to be addressed: early alarm of the process departure from the normal operation and the location of the physical variables affected by the abnormal event. In this work, the fault is identified using a Schewhart chart. In addition, an error contribution bar chart is used to identify the measurement which indicates the presence of the fault.

4.1. Charting the estimated density function for early alarm

After obtaining the density function the estimated confidence intervals can be calculated and presented together with the operating points of the process being monitored. As a plot of two or three variables at most is possible, the confidence intervals are normally graphed for combinations of two PCs only. To monitor the complete performance of the plant all possible combinations of PCs must be charted. This would not be convenient for a multivariate density function with high dimensionality. Therefore, a single chart with comprehensive information from all selected PCs is more practical for process performance monitoring.

Using the concept of a Schewhart chart, it is possible to inspect the statistical behaviour of the operation over all selected PCs against observation time. With 95% and 99% confidence intervals obtained from the density function plotted on the chart, the values of density function are then also plotted against the observation time or observation indicator. The Schewhart chart then gives a clearer picture of the 'total operating performance' of the plant. This is referred to as a quality control (QC) chart (Fig. 10). From the QC chart, the following information can be observed:

(a) the historic pattern of the statistical behaviour of the process operation,
(b) the position of the current operating point in relation to the confidence intervals.

By combining two pieces of information, operators can see the plant moving out of control knowing that the information presented is based on the multivariate data set and not just one variable at a time as in the case of the application of classical SPC methods.

4.2. Tracing the fault indicators

The QC chart gives the operator a view of the state of statistical control of the process. When the operation is deteriorating, the operator can observe an unusual event occurring in the process as the data migrates outside the confidence intervals on the QC chart. However, from the QC chart, it is impossible to determine where within the plant the problem is located and which variable(s) are affected. To help operators trace the variables responsible for the fault, a bar chart of prediction error is presented.

As noted above, PCA uses singular value decomposition to determine a set of new independent variables, score vectors or PCs. Now let $\mathbf{X}$ be the measured data matrix consisting of N observations (number of rows) and M physical variables (number of columns). This can be decomposed as

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}, \tag{8}$$

where $\mathbf{T}$ is score matrix consisting of m selected PCs, $\mathbf{P}$ the loading matrix which is the PCA model parameter, and $\mathbf{E}$ the residual matrix. Then the squared prediction error (SPE), can be defined as

$$SPE_i(k) = [X_i(k) - \hat{X}_i(k)]^2/\sigma_i^2 \qquad i = 1, 2, \ldots, M, \quad k = 1, 2, 3, \ldots, \tag{9}$$

in which $X_i(k)$ and $\hat{\mathbf{X}}_i(k)$ are the values of the ith measured and predicted variables of sample k, and σ_i is the standard deviation of the SPE_i calculated upon the training data set for scaling purpose. $\hat{X}_i(k)$ is predicted using only the selected PCs. In the literature, the SPE is often defined as the sum of the SPE_i over the time indicated by k. The potential danger of this definition is that the accumulation of the squared error over the time history can lead to a false alarm in some circumstances. Therefore, it is suggested that the SPE_i values should be depicted in a bar chart with the physical variables at current operating point. If any unusual events occur in the operation, the prediction error or the SPE_i of the variables which are affected by the events will increase, which provides clear guidance to the operators in tracing the physical location of the fault in the process.

5. Comparison based on estimators using plant data

To assess the methods described in Section 3 from an applications point of view, the methods will be applied to two samples from different populations, one of which contains a known fault condition. The other data set has been used in the literature to provide a benchmark against which to assess the applicability of such statistical methods. In both cases the true density distribution of the populations are unknown. The only way to verify the density estimator is to compare the results from different methods and to investigate the probability of misclassification of the confidence intervals to provide an indication of the quality of the density estimator.

5.1. Benchmark data

The benchmark data is taken from Jackson (1991). The data consists of 100 sets of measurements to assess hearing loss in males. This is measured using an audiometer in which an individual is exposed to a signal of a given frequency with increasing intensity until the signal is just audible. Four frequencies are chosen for both right and left ears. The data used here provides a means of comparing the density estimates using the three bandwidth selectors. The test observations are not from a separate data set but the last 10 observations from a training data set.

All 100 observations are used to carry out the initial PCA calculation. Three principal components explaining 85% of the variance of the data are chosen. The density function and the 95 and 99% confidence intervals of the three principal components are obtained using each of the estimators MISE, ADMISE and BAMISE.

From the results shown in Figs. 1–4, it is clearly shown that there are four modes in the data. The density estimates from MISE cross validation and BAMISE cross validation is very close to each other (see Figs. 1 and 2). The ADMISE cross validation treats the observations differently via local bandwidths using a sensitivity parameter α, especially those in low-density areas where the

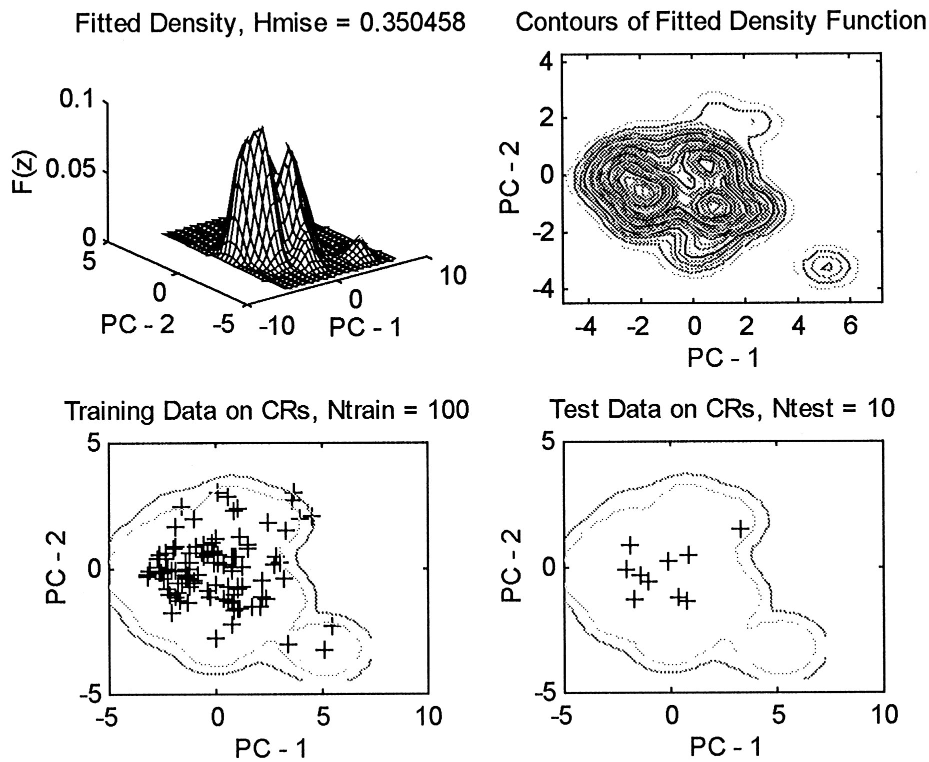

Fig. 1. Density function estimated with MISE cross validation.

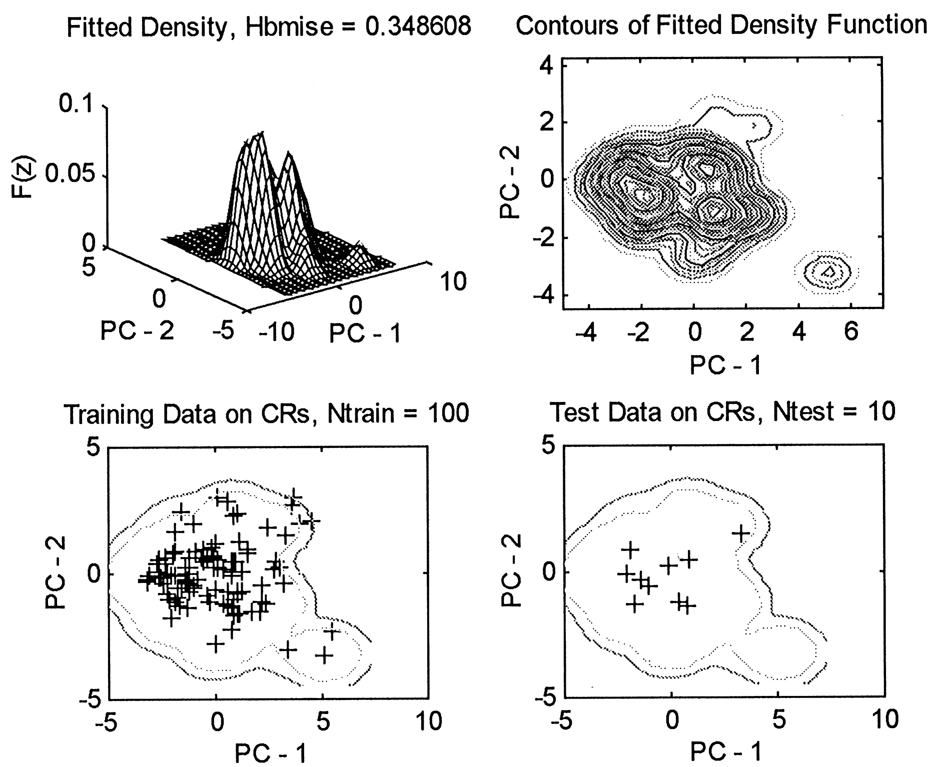

Fig. 2. Density function estimated with BAMISE cross validation.

centres are assigned a wider local bandwidth, which alters the overall shape of the density estimate. However, the adaptive cross validation technique with $\alpha = 0.5$ gives a slightly different bandwidth so that the confidence intervals that result are different to those previously obtained (Fig. 3). To investigate the sensitivity of the

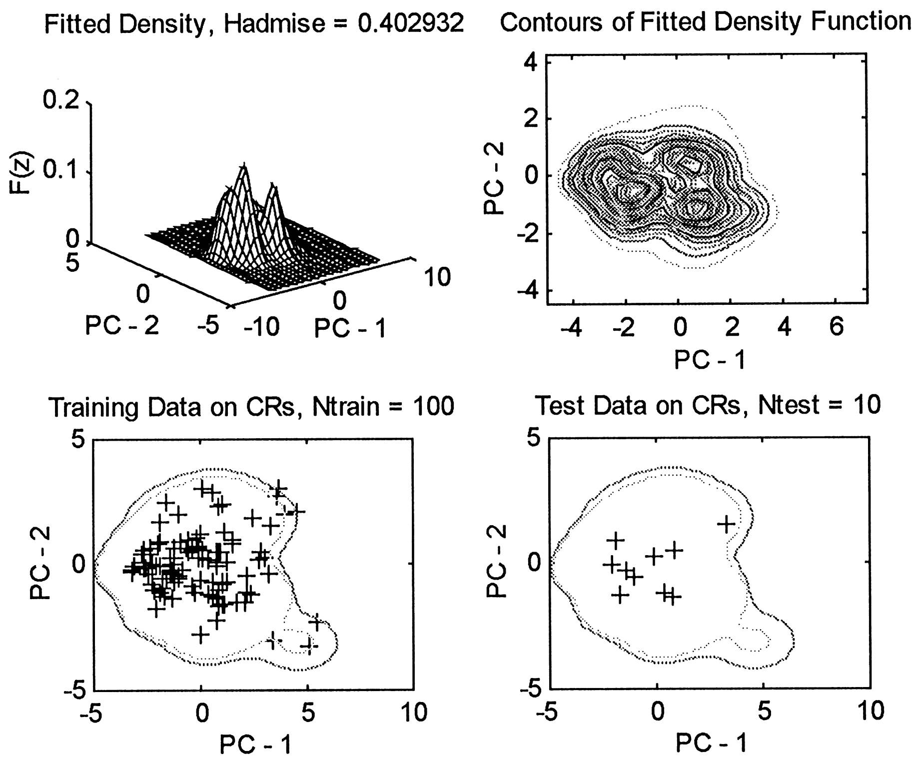

Fig. 3. Density function estimated with ADMISE with $\alpha = 0.5$.

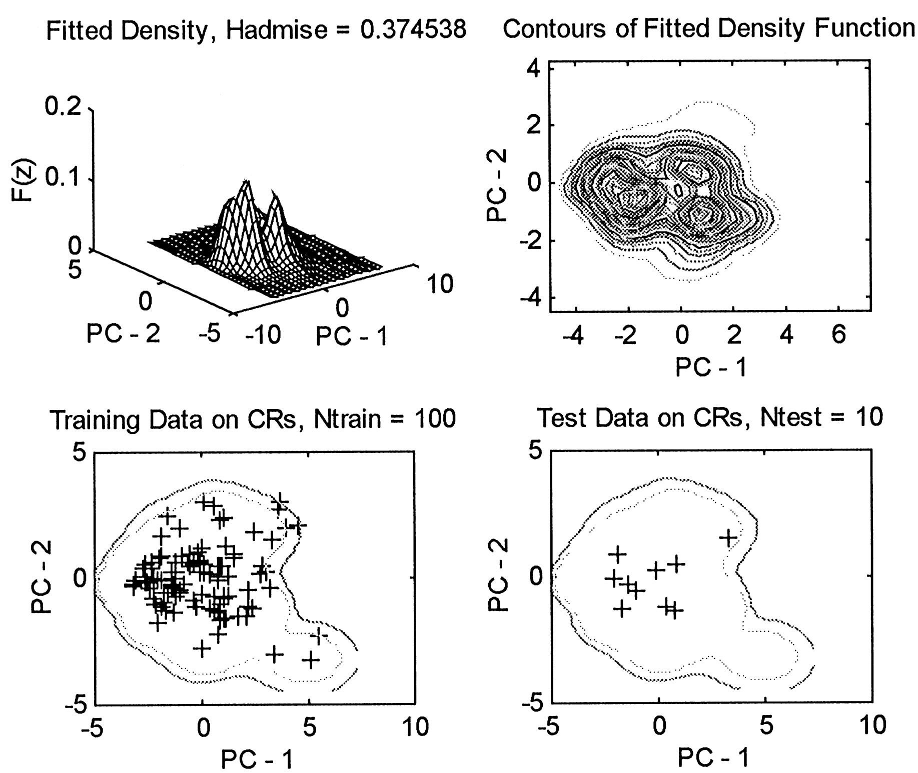

Fig. 4. Density function estimated with ADMISE with $\alpha = 1/d$.

parameter α, $\alpha = 1/d$ is used and the ADMISE cross validation procedure is repeated with all other settings unchanged. The results are shown in Fig. 4 which indicates that the sensitivity parameter has a particular affect on the estimated density function, and unlike when a constant e.g. 0.5 is used, the choice of $\alpha = 1/d$ gives a density estimate which looks very close to those obtained by MISE and BAMISE cross validation. From Fig. 4, it is clearly seen that the observations in the remote area from the modes are given much more attention than that in Fig. 3. Therefore, it could be argued that $\alpha = 1/d$ is a better choice for the sensitivity parameter, at least in the case of the audio data.

In the case of the pilot density, the results in Fig. 4 are obtained with the optimal bandwidth of a normal multivariate distribution. To test the conclusion that the bandwidth from the ADMISE is not sensitive to the pilot density, the density estimate from the MISE cross validation, which is as good as that of ADMISE with a pilot of multivariate normal distribution, is used as a pilot density for the ADMISE cross validation. The results of the test are shown in Fig. 5, which suggests that the conclusion is basically correct, at least in this case. One of the attractive properties of the ADMISE cross-validation method is that there are two 'tuning' variables, the pilot density and the sensitivity parameter, which can be adjusted to obtain the best result.

5.2. Industrial data

A waste product is reprocessed in a glass melter. Because of periodic heating-up and cooling-down operations it is possible for a crack to develop in the melter and early warning of such a condition is required. The original data from the melter includes the electrical power and temperature measurements. Three power and 10 temperature measurements are used to form the raw data matrix for PCA. The data set contains 1050 observations in which the last 50 were collected when a crack on the melter wall was developing.

Initially the first 300 observations are used to fit a PCA model and an estimate of the density function based on the first three PC's is obtained. The confidence limits obtained based on the training data are then applied to the test data, Fig. 6d, where the fault is clearly identified as the operating points migrate outside the confidence limits. A bandwidth is 0.613 was selected using BAMISE and gives an acceptably smooth density function. To verify the quality of the density estimate, both 500 and 1000 observations are used for the training data and the procedure is repeated resulting in bandwidths of 0.153 and 0.116. The results are shown in Figs. 7 and 8. Comparing the confidence intervals in Figs. 6–8 indicates that the confidence intervals based on a larger sample give a more detailed distribution though their overall structures remain very similar. The confidence interval resulting from each of the three cases give the same information about the fault except that the more detailed density estimate indicates the fault much earlier as the operating points migrate out of the 95% confidence interval at the bottom of a pocket at a much earlier stage (Figs. 7 and 8). From this point of view, adequate

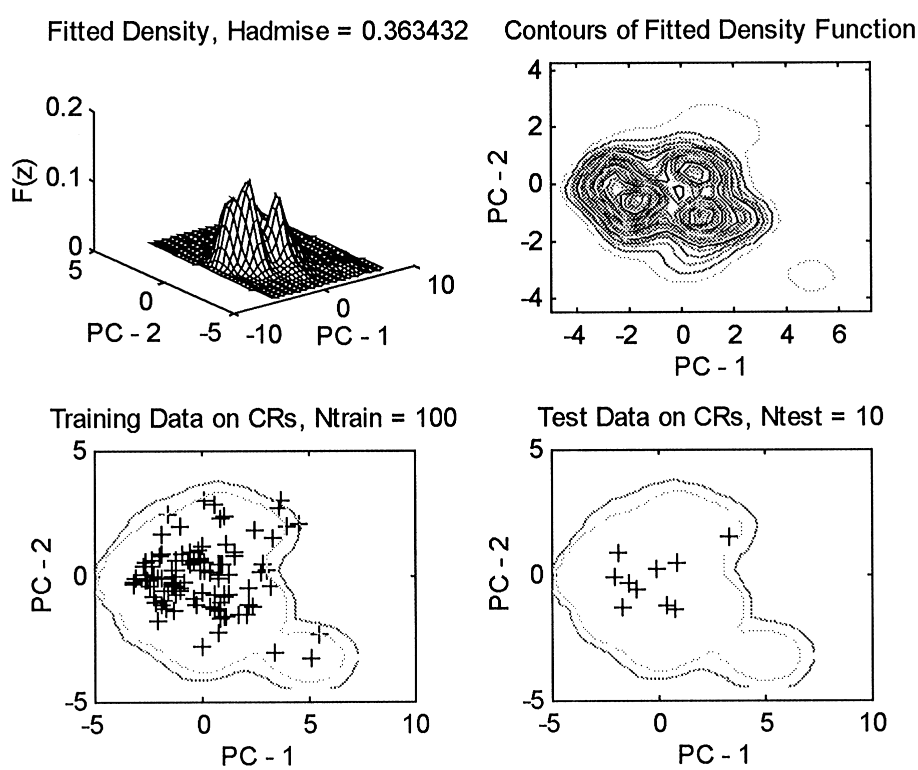

Fig. 5. Density function estimated with ADMISE with $\alpha = 1/d$ and a pilot density from the MISE cross validation.

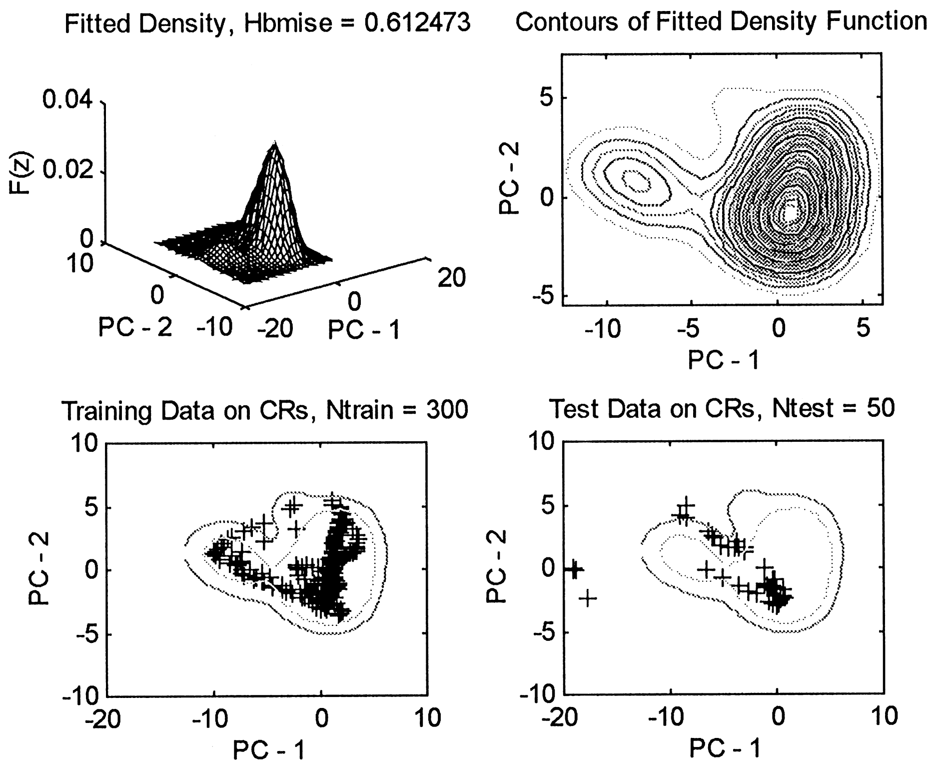

Fig. 6. Density function estimated with BAMISE with 300 training data points.

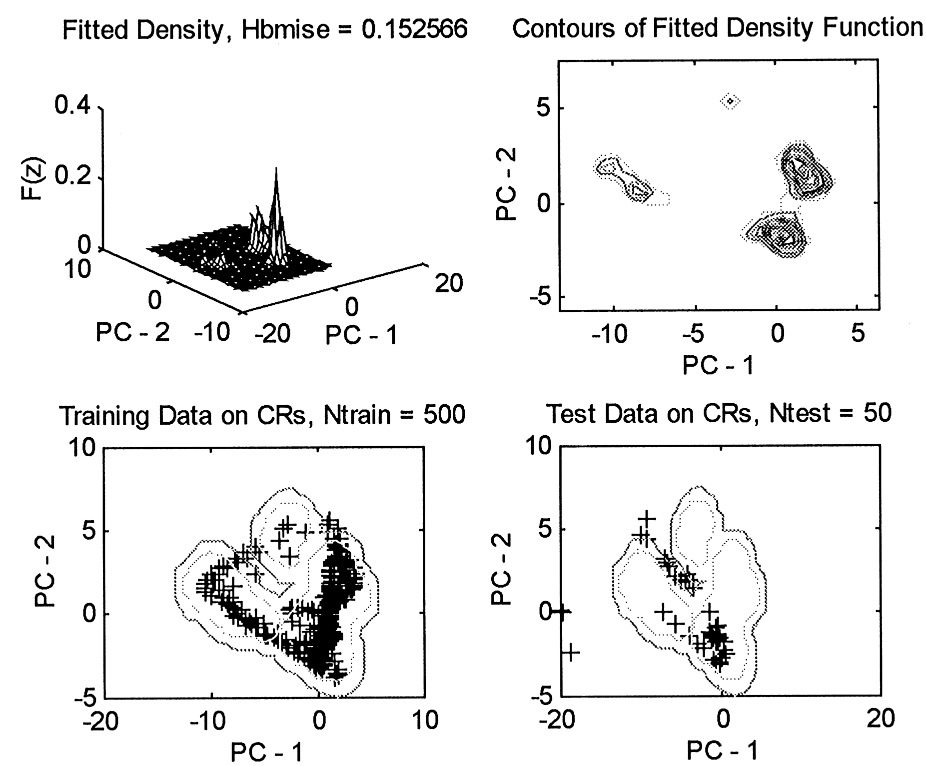

Fig. 7. Density function estimated with BAMISE with 500 training data points.

sample size is essential to extract crucial details of the distribution.

The confidence intervals estimated on the assumption of a normal distribution are shown in Fig. 9. Comparing this with the confidence intervals in Fig. 8, the KDE detects the fault about 50 min earlier than in the case of the elliptical confidence limits of the normal density function. The operating points migrate out of the 95%

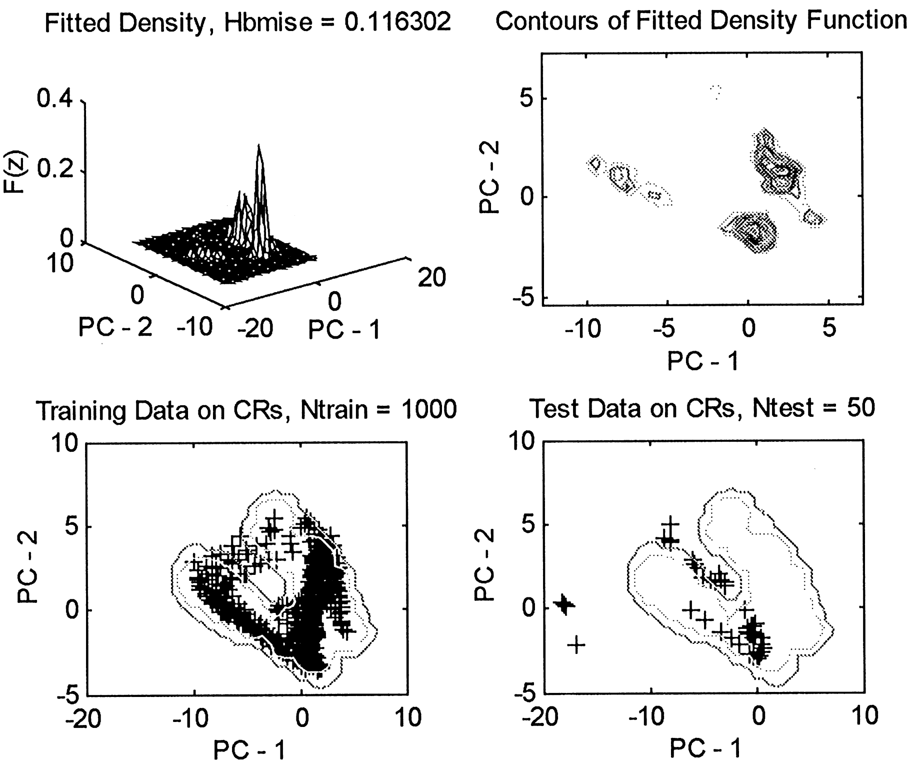

Fig. 8. Density function estimated with BAMISE with 1000 training data points.

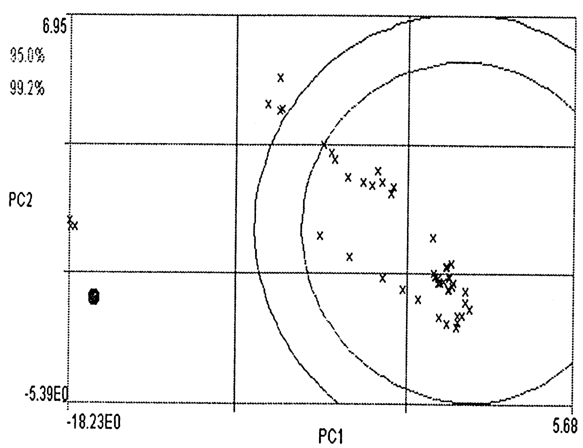

Fig. 9. Gaussian density function estimated with parametric method on 500 training data points.

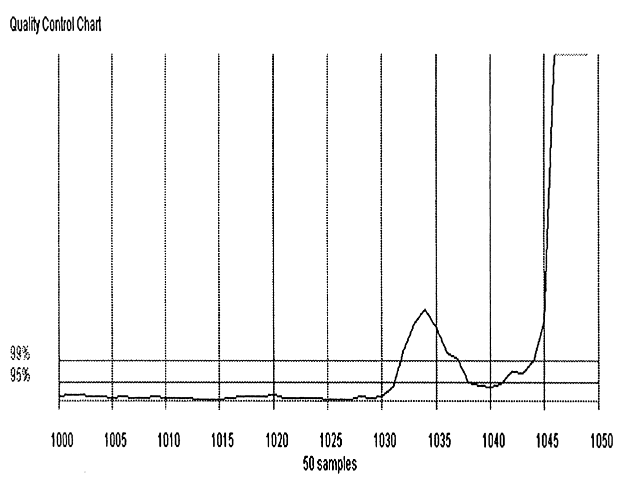

Fig. 10. QC chart of the kernel density function — melter data.

confidence interval at the bottom of a pocket indicating an abnormal event developing. This case clearly shows the advantage of the kernel density estimate over the conventional technique.

The MISE and ADMISE cross validation were also used but failed to obtain any sensible bandwidth values because these data contain clusters with very small variation in the data. This leads to the conclusion that the BAMISE cross validation is more robust and stable. It can cope with both clustered data and data containing relatively small variations. Since the BAMISE cross validation produces a positively biased bandwidth, the density estimate will in general be oversmoothed.

The QC chart of the melter is shown in Fig. 10. It clearly indicates that the process starts going out of control near observation 1030 and completely out of control at the end of the data set. This is consistent with the confidence intervals detecting the departure in Figs. 7 and 8. The alarm is raised by the QC chart more reliably than through the usual confidence intervals based on

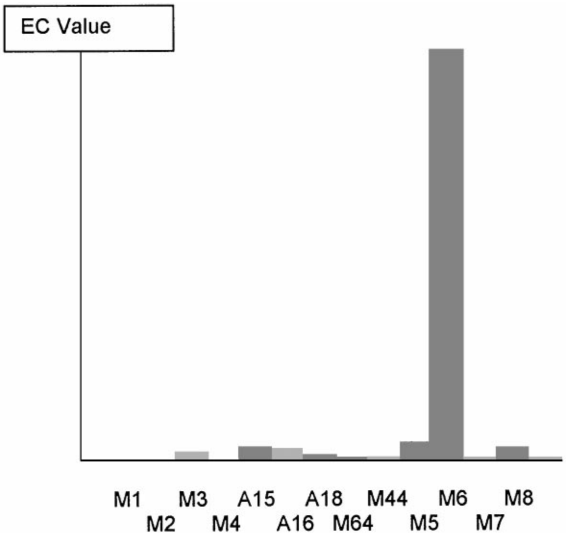

Fig. 11. EC chart of the PCA model — melter data.

a parametric normal distribution as the QC chart takes all the PCs into account.

Finally, investigation of the error contribution chart for the data is needed to trace the faulty variables. Starting from sample interval 1030, the SPE_i of variable represented by $M6$ is growing up rapidly. The EC chart of the SPE_i (Fig. 11) at sample interval 1049 clearly indicates the greatest variability of the variable $M6$ from the model prediction. Therefore the variable $M6$ is most closely associated with the fault.

6. Validation of the density estimators

To assess the validity of the estimated PDF, new data samples from the population are needed to test the estimator. However, it is not always possible to obtain reliable data. To obtain more normal operation data as validation data of the estimated PDF, one approach that is often used to derive an alternative set of data based on the original data set is known as bootstrapping, (Efron, 1981). However, simple bootstrap resampling is not adequate for this purpose because the simple bootstrap samples will only produce data that occurred in the original sample and any spurious fine structures introduced by a random error may well be reproduced in the simulated samples. To test the validity of the density estimate it is essential that the simulated data structure contains only the underlying 'true' structure of the observed data without sharing spurious details that have arisen from the random effects. Smoothed bootstrap resampling (Efron, 1982) is a method that avoids the problems of the simple bootstrap method without losing the power of bootstrap resampling.

6.1. Smoothed bootstrap

Assume that the estimated density function, $\hat{f}(\mathbf{x}, h)$, of the population has been obtained with one of the density estimators discussed above. The following procedure will then generate the bootstrap samples:

1. Randomly resample the original sample $\mathbf{X}_0$ with replacement to generate B bootstrap samples $\mathbf{X}_b^*$, $b = 1, 2, \ldots, B$.
2. Generate samples E_b, $b = 1, 2, \ldots, B$ from the kernel $K(\mathbf{x})$ used to estimate $\hat{f}(\mathbf{x}, h)$ and which is scaled to have variance matrix the same as the original data $\mathbf{X}_0$.
3. Form new test samples, $\mathbf{Y}_b^* = \bar{\mathbf{X}} + (\mathbf{X}_b^* - \bar{\mathbf{X}} + hE_b)/(1 + h^2)^{1/2}$, where $\bar{\mathbf{X}}$ is the mean of the original sample $\mathbf{X}_0$.

The bootstrap samples $\mathbf{Y}_b^*$ are generated based on the estimated density function and training sample, so they should have the same distribution as the original sample. If the bootstrap samples are depicted on the confidence intervals from $\hat{f}(x, h)$, there should be 95 and 99% simulated observations inside the 95 and 99% confidence intervals, respectively. The number of bootstrap observations may not be exactly 95 and 99% of the sample, but they should be close to those percentiles because any statistical replication of bootstrap samples should be normally distributed around that of the original sample. Obviously, the density estimate is one of the statistical replications. Now, the smoothed bootstrap samples $\mathbf{Y}_b^*$ can be used to validate the quality of the density estimate and the confidence limits by considering the misclassification of the confidence intervals from the simulated samples.

With the smoothed bootstrap sampling method, simulated samples from the estimated density function, $\hat{f}(\mathbf{x}, h)$, are generated and are then imposed on the confidence intervals obtained from the fitted PDF using one of the three bandwidth selectors. The number of simulated points inside the 95 and 99% confidence intervals, $N_{95\%}$ and $N_{99\%}$, are counted, respectively. The percentiles of the numbers over the sample size are then determined as

$$\alpha_{95\%} = \frac{N_{95\%}}{N}, \qquad \alpha_{99\%} = \frac{N_{99\%}}{N}. \tag{9a}$$

The misclassification probability is then given by:

$$\rho_{95\%} = 0.95 - \alpha_{95\%}, \qquad \rho_{99\%} = 0.99 - \alpha_{99\%}. \tag{10}$$

6.2. Probability of misclassification — audiometric data

To assess the three bandwidth selectors, the smoothed bootstrap method is used to generate a further 50 test samples, each consisting of 500 simulated operating points. Then the confidence intervals constructed from the density functions estimated using each of the three

Table 1
Misclassification probabilities of density estimators on audio data

Selectors	$\bar{\alpha}_{95\%}$	$\bar{\rho}_{95\%}$	$\bar{\alpha}_{99\%}$	$\bar{\rho}_{99\%}$
MISE	0.9524	− 0.0024	0.9911	− 0.0011
ADMISE	0.9560	− 0.0060	0.9927	− 0.0027
BAMISE	0.9545	− 0.0045	0.9930	− 0.0030

bandwidth selectors, respectively, are tested. Applying Eqs. (9) and (10) to the 'new' data generates the results shown in Table 1.

From the results it can be seen that each of the bandwidth selection techniques produced comparable levels of misclassification and all produced results around 0.5% probability of misclassification. This indicates that each of the estimators yields acceptable results for this data set and leads to the conclusion that this is a data set that contains neither clusters nor small variations and therefore the MISE bandwidth selector is adequate for such a data pattern.

6.3. *Probability of misclassification — industrial data*

The three bandwidth selectors, MISE, ADMISE and BAMISE are used in an attempt to extract the density function of the melter data. However, MISE and ADMISE selectors cannot produce any sensible results for this very dense data and only BAMISE cross validation produced an acceptable estimator.

The density function is fitted on the basis of the original sample of 500 observations. The values of $\alpha_{95\%}$ and $\alpha_{99\%}$ of the 30 smoothed bootstrap test samples of 1000 points were obtained and the average percentiles and misclassification probabilities over the 30 smoothed bootstrap samples are listed in Table 2.

The results show that the BAMISE selector gives a good density estimate which correctly captures the distribution information from the measured sample very accurately even if the sample has very small variation. From Fig. 12 it is clear that the simulated operating points have the same distribution pattern as the original sample, which is a sign of the success of smoothed bootstrap simulation.

Again the misclassification probabilities are better than 0.7 and 0.5%. Therefore fewer than 1% of normal operating points are misclassified as faulty points. As a result of this error the 95 and 99% confidence intervals are reduced to 94.3 and 98.5% confidence intervals, which is within the tolerance of most process measurements. This clearly demonstrates the power of the BAMISE selector in the case of clustered data which also has small variability.

Table 2
Misclassification probabilities of density estimate on melter data

Selectors	$\bar{\alpha}_{95\%}$	$\bar{\rho}_{95\%}$	$\bar{\alpha}_{99\%}$	$\bar{\rho}_{99\%}$
MISE	N/A	N/A	N/A	N/A
ADMISE	N/A	N/A	N/A	N/A
BAMISE	0.9430	0.0070	0.9850	0.0050

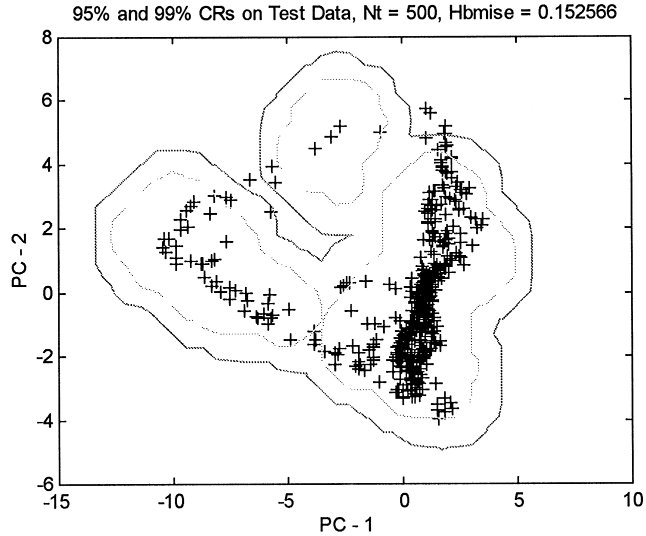

Fig. 12. Smoothed bootstrap sample on fitted confidence regions.

7. Conclusions

This study has provided further evidence of the benefits of using kernel density estimates for enhancing the interpretation of the results from PCA for process monitoring. The selection of the correct bandwidth for the estimators is a critical factor in the estimation process particularly in the case of highly clustered data with small variability. In this work it has been shown that the BAMISE cross validation procedure is the most effective of the techniques considered in this study when applied to such data.

The results from the analysis of the plant data using the kernel density estimators has shown that significant improvements can be made to existing multivariate statistical methods to provide plant operators with reliable early warning of a fault, based on a single quality control chart. The improvements in fault detection arise from the ability to use reliable density estimates to generate more realistic confidence limits and to generate the total plant performance data using a Schewhart like quality control chart.

Further work has demonstrated that additional improvements can be derived by using additive noise to increase the variability of highly clustered data. The results of this work will be reported shortly.

Acknowledgements

This work was supported by the EPSRC ROPA scheme.

References

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*, 353–360.

Chen, J., Bandoni, J. A., & Romagnoli, J. A. (1996). Robust PCA and normal region in multivariate statistical process monitoring. *A.I.Ch.E. Journal*, *42*(12), 3563–3566.

Efron, B. (1981). Nonparametric estimates of standard error: The jack-nife, the bootstrap and other methods. *Biometrika*, *68*, 589–599.

Efron, B. (1982). The bootstrap, the jacknife and other resampling plans. *CBMS-NSF regional conference series in applied mathematics*, vol. 38.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). London: Academic Press Inc.

Jackson, J. E. (1991). *A user's guide to principal components*. New York: Wiley.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414.

Martin, E. B., Morris, A. J., & Zhang, J. (1996). Process performance monitoring using multivariate statistical process control. *IEE proceedings: Control theory and applications*, *143*(2), 132–144.

Martin, E. B., & Morris, A. J. (1996). Non-parametric confidence bounds for process performance monitoring charts. *Journal of Process Control*, *6*(6), 349–358.

Silverman, B.W. (1986). *Density estimation*. London: Chapman & Hall.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistics Society Bulletin*, *36*, 111–147.

Wand, M. P., & Jones, M. C. (1995a). Multivariate plug-in bandwidth selection. *Computational statistics*, *9*, 97–116.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329–348.