ELSEVIER

Chemometrics and Intelligent Laboratory Systems 52 (2000) 87–104

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Multiple outlier detection for multivariate calibration using robust statistical techniques

Randy J. Pell *

Analytical Sciences Laboratory, The Dow Chemical Company, 1897S Building, Midland, MI 48667 USA

Received 14 February 2000; received in revised form 24 May 2000; accepted 25 May 2000

Abstract

Outliers that are incorporated into a multivariate calibration model can significantly reduce the performance of the model. In the case of multiple outliers, the standard methods for outlier detection can fail to detect true outliers and even mistakenly identify good samples as outliers. Robust statistical methods are less sensitive to outliers and can provide a powerful tool for the reliable detection of multiple outliers. This paper examines the use of robust principal component regression (PCR) and iteratively reweighted partial least squares (PLS) for multiple outlier detection in an infrared spectroscopic application.

Keywords: Robust regression; Iteratively reweighted PLS; Least trimmed squares; PLS; PCA; Resampling by half mean; Outliers

1. Introduction

Building high-quality multivariate calibration models depends on the execution of several steps [1]. One of the most important of these steps is outlier detection. So important is outlier detection in multivariate calibration that Martens and Næs [2] devote an entire chapter to this subject in their book. They discuss the standard diagnostic tools for both calibration and prediction outlier detection. While these tools are quite powerful for detection of prediction outliers and single calibration outliers, they are inefficient at best and can be misleading at worst for the detection of multiple outliers in the calibration phase. Detection of multiple outliers in calibration using the standard diagnostics can be severely limited due to masking and swamping. Masking is the case where one or more outliers are incorrectly identified as normal samples because other outliers are masking their presence. Swamping is the case where normal samples are made to appear to be outliers.

It may seem improbable that multiple outliers would be present in data collected for a calibration experiment, given the high degree of control normally accompanying such experiments. This may be the case for laboratory calibration experiments; but for data collected from experiments in which the analyst has less control, it is reasonable to expect that outliers will be more prevalent. The application of multivariate analysis techniques to data collected from on-line analytical instrumentation or data from collections of univariate process sensors (e.g. tem-

* Tel.: +1-517-638-4967; fax: +1-517-636-4882.
E-mail address: pellrj@dow.com (R.J. Pell).

0169-7439/00/$ - see front matter
PII: S0169-7439(00)00082-4

perature, pressure, flow) represent two such cases of reduced control over the data generation conditions. Others have recognized outlier detection as an essential step in the application of multivariate analysis to process data [3]. With the increased use of multivariate techniques for process analysis data [4], there is a need for reliable multiple outlier detection methods for multivariate calibration.

There are two approaches to outlier detection. The first approach is to fit the data with least squares, construct regression diagnostics and then remove the outliers. The second approach is to construct estimators that fit the majority of the data and examine the residuals from this fit to detect outliers. As Rousseeuw and Leroy [5] have noted, the two approaches have the same goal but they are performing the steps in the opposite order. The second alternative often makes use of robust statistics and it is a better choice given that many regression diagnostics are affected by the outliers they are supposed to detect.

Robust statistical methods are constructed so that they provide reliable results even with outliers present in the data. Methods have been developed to address exploratory and regression analysis problems. Jolliffe [6] reviews some of the early methods for robust principal component analysis (PCA). These methods include robust estimates of the elements of the covariance or correlation matrix [7], adjustments to the internal workings of the singular value decomposition (SVD) algorithm [8], and application of projection pursuit [9]. Within the robust estimators, there are certain methods that are very robust and are known as high breakdown point (HBP) estimators. Many of the HBP estimators, including least median of squares (LMS) and least trimmed squares (LTS) for robust regression and the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) for multivariate analysis, can be attributed to Rousseeuw [10]. A more thorough discussion of these methods can be found in Rousseeuw and Leroy [5]. An update for the program PROGRESS has recently been provided and a publication describes the implementation of least quantangle of squares (LQS) and LTS [11]. The LMS and MVE have been supplanted by their creator in favor of LTS and MCD both of which have had faster computational algorithms introduced [12,13].

Rousseeuw and van Zomeren [14] used MVE to identify leverage points and LMS to distinguish between good and bad leverage points. There are interesting comments and a rejoinder accompanying this paper. After the introduction of LMS, there has been considerable discussion on the computational aspects [15–17], observed instabilities [18] and the exact fit property inherent to all HBP estimators [19]. Procedures for improving or supplementing LMS have included confirmatory analysis using multiple deletion diagnostics [20], stepwise confirmation [21] and forward searching with stalactite plots [22]. Alterations to LMS and MVE have included sequential construction of outlier-free subsets [23], new estimators derived from elemental sets [24], methods that separate data into clean subsets and subsets with potential outliers [25] and one step general M-estimators based on HBP initial estimators [26].

The statistical community seems to generally consider data structures that have more objects than variables and are full rank. Data sets that are most prevalent in the chemometrics literature have many more variables than objects and are rank deficient. In the chemometrics literature, several researchers have addressed the robust analysis of rank deficient data. Walczak and Massart [27] adapted the robust techniques of ellipsoidal multivariate trimming (MVT) and LMS to make principal component regression (PCR) robust. Walczak [28] proposed a new procedure based on a genetic algorithm for multiple outlier detection. Wang et al. [29] proposed the use of maximum sum of binary coded residuals (MASBR) as an alternative to the sum of squared residuals for a novel robust regression approach. Three different groups have proposed methods for making partial least squares (PLS) robust [30–32]. Liang and Kvalheim [33] provide a tutorial on robust methods. Hove et al. [34] proposed a new method for robust latent-structure decomposition. More recently, Egan and Morgan [35] have suggested a very understandable and easy to program method for finding outliers in multivariate chemical data. For those interested in an overview of robust statistics, see Rousseeuw [36]; for application, see Singh [37]. Ryan [38] also provides a very readable chapter on robust regression.

In this paper, the detection of multiple outliers in multivariate calibration is explored. Standard diagnostics for outlier detection are compared with two

robust analysis approaches. Robust PCR and iteratively reweighted PLS are tested for detection of multiple outliers in a spectroscopic calibration application.

2. Theory

2.1. Standard diagnostics for sample outlier detection in multivariate calibration using PLS

The standard diagnostics for outlier detection in the calibration phase of a multivariate calibration experiment have been thoroughly discussed by Martens and Næs [2]. They include diagnostics for ''inside'' the model space such as sample leverage or Mahalanobis distance (proportional to sample leverage) and those for ''outside'' the model space such as an F-test on the spectral residuals. In the calibration phase, the difference between known and predicted concentrations can be computed and used as an additional diagnostic for outlier detection. The concentration residuals can be scaled by the standard deviation of the residuals and a function of the sample leverage to generate either studentized or leverage-corrected concentration residuals. While these diagnostics can work well for single outliers, especially if the outlier is removed when the diagnostic is computed (as in cross-validation), they can show poor performance if there are multiple outliers. For this paper, sample leverage and studentized concentration residuals are used for detection of outliers for the standard approach.

2.2. Robust PCR

PCR can be thought of as a two-step process [1]. Step 1 consists of a PCA of the response matrix, **R**, to generate surrogate variables called scores. Step 2 is a multiple linear regression of a concentration vector onto an appropriately truncated score matrix. Robust PCR, as presented by Walczak and Massart [27], makes each of the steps robust.

2.2.1. Robust PCA

There are two main approaches for making PCA robust. One uses robust estimates of location and dispersion to replace the standard mean and covariance measures. The other approach detects and removes outliers before the mean and covariance measures are formulated. There are many approaches for making the covariance matrix robust. Joliffe [6] reviews some of those approaches and Rousseeuw and Leroy [5] discuss robust estimation of multivariate location and covariance matrices.

Many of the approaches put forth in the literature rely on complex and often very computer intensive calculations to carry out the analysis. Recently, Egan and Morgan [35] proposed two methods that can be used to identify outliers in multivariate data without excessive computations. After the outliers are removed, the PCA is performed. Their methods, resampling by half-means (RHM) and smallest half-volume (SHV), are easy to understand and simple to program. A simulation study [35] showed that SHV and RHM have similar breakdown properties as the methods of ellipsoidal MVT and MCD. The methods of Egan and Morgan [35] can readily handle rank deficient data while other methods cannot. For this paper, RHM was implemented to detect outliers in the response matrix, **R**, in order to remove those samples from consideration before the PCA step. A brief review of this method is given below.

The RHM method is based on the estimation of the distribution of vector lengths via resampling. A certain percent of the data is sampled without replacement. The number of resampling experiments should be on the order of two to three times the number of samples in the calibration set. The mean and standard deviation of each variable for a given sample are computed. All of the data files are then scaled using the mean and standard deviation computed from this sample. The vector lengths for all of the scaled samples are computed giving a distribution of vector lengths for each resampling experiment. A fixed percentage of the longest vectors are examined for each resampling experiment and the number of times a sample appears in this set of longest vectors over the course of many resampling experiments is recorded. For this paper, those samples that appear in the upper 5% of vector lengths one or more times are considered extreme. The PCA decomposition is performed without those extreme samples but the extreme samples are projected onto the PCA space and the scores from the extreme samples are used with the rest of the calibration samples in the regression step discussed below.

2.2.2. Robust multiple linear regression

Robust regression estimators are often distinguished using the concept of breakdown point and efficiency. The breakdown point is the smallest fraction of anomalous data that can render the estimator useless. For example, the mean has a zero breakdown point because it only takes one point to be moved to an arbitrary location for the estimate to be made useless and, thus, the smallest fraction approaches zero as the number of samples grow large. The median has a breakdown point of 50%. Rousseeuw and Leroy [5] offer a more formal definition of breakdown. Breakdown points vary considerably for different classes of estimators and can be adjusted as desired for some estimators. Ryan [38] suggests that rather than using only one very HBP estimator, it may be better to use several breakdown points and compare the results.

Efficiency is the ratio of the mean square error from a robust estimator to the mean square error from an ordinary least squares estimator when applied to a data set that is sufficiently normal and has no influential points [38]. An estimator with an efficiency approaching 1 is most desirable. Many of the robust estimators have poor efficiency and, therefore, the use of robust estimators for detection of outliers followed by ordinary least squares analysis for determining the final model may be the most reasonable approach.

Ryan [38] classifies robust estimators into three classes. The M-estimators first introduced by Huber [39] minimize a symmetric function of the residuals. Unfortunately, this estimator is greatly influenced by X-outliers and has only a $1/n$ breakdown. Bounded influence estimators, frequently referred to as GM- or generalized M-estimators, were developed to over-

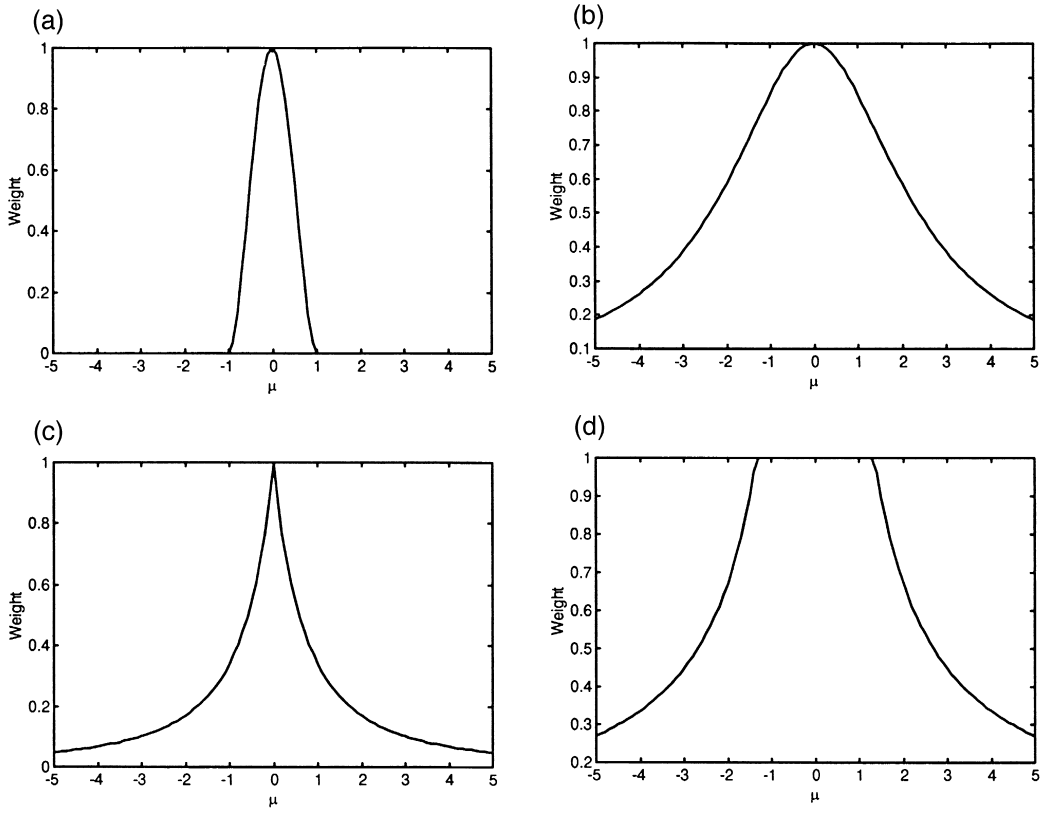

Fig. 1. (a) Bisquare weight function vs. μ, (b) Cauchy weight function vs. μ, (c) Fair weight function vs. μ, (d) Huber weight function vs. μ. The symbol μ represents the residual divided by the MADM of the residuals.

Table 1
Functions used for the iteratively reweighted PLS study

Function name	Function	Constant c
Bisquare	$[(1-\mu^2)/c^2]^2$, $\lvert\mu\rvert < c$ 0, otherwise	4.685
Cauchy	$1/[1+(\mu/c)^2]$	2.385
Fair	$1/(1+\lvert\mu/c\rvert)^2$	1.345
Huber	1, $\lvert\mu\rvert \le c$ $c/\lvert\mu\rvert$, otherwise	2.795

come the X-outlier problem of M-estimators. These also can have low breakdown points. HBP estimators are the third class of estimators. As the name implies, these estimators have an HBP, up to 50%. Rousseeuw and Leroy [5] have developed most of these estimators and they include the LMS, LTS and S-estimators. For this paper, the LTS estimator is used with multiple breakdown points as suggested by Ryan [38]. A brief review of the LTS estimator is given in the Appendix.

2.3. Iteratively reweighted PLS

Iteratively reweighted least squares is a type of M-estimator. As the name implies the method uses a weighted regression step. The sample weights are developed based on the concentration residuals and a user selected function. The sample weights are initially set to 1 and are updated after each iteration. Cummins and Andrews [30] adapted this approach for PLS with one difference. Instead of using a fitted concentration residual, they used a cross-validated concentration residual to develop the sample weights. They optimized both the weights and the rank of the model simultaneously. For the work reported in this paper, the weights were optimized at a fixed rank.

There are many weight functions from which to choose. For this paper, four weight functions were tested including bisquare, Cauchy, Fair and Huber. Fig. 1a–d displays the weight functions vs. μ. The symbol μ represents the residual divided by the me-

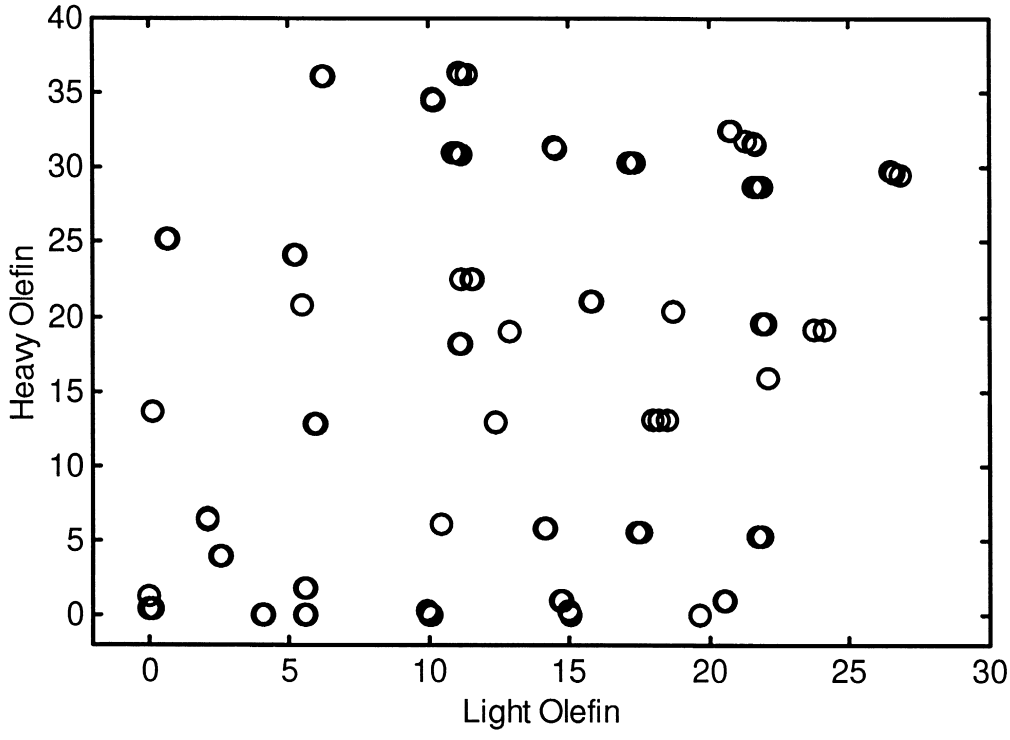

Fig. 2. Heavy olefin vs. light olefin wt.%.

dian absolute deviation from the median (MADM) of the residuals. Table 1 summarizes the functional form for the weight functions. The c in the equations is a tuning constant and the values used in this study are reported in column 3 of Table 1. The bisquare functional form was found in Ref. [40] and the Cauchy, Fair and Huber were taken from Ref. [30].

3. Experimental and methods

3.1. Experimental design, data collection and analysis

A five-level two-factor experimental design was specified for the calibration experiments. Reference concentration values were determined using gas chromatography. Quantitative analysis was performed using normalized external standard calibration. The final wt.% concentrations of the light and heavy olefin are shown in Fig. 2.

Infrared spectral measurements were collected using an Analect PCM 4000 FTIR spectrometer. Spectra were collected from 4000 to 400 cm^{-1} at 4 cm^{-1} resolution using a 0.028-in. pathlength high-pressure transmission cell. Three spectra were collected for each design point giving a total of 116 spectra. Fig. 3 displays a representative collection of the spectra used for this calibration experiment. The region from 1760 to 1992 cm^{-1} was used for the model construction. A one-point baseline correction was performed at 2120 cm^{-1}. All calculations were implemented in MALAB™ for WINDOWS™ version 5.3 (The MathWorks, Natick, MA).

3.2. Standard PLS

The standard NIPALS decomposition method was used for the PLS analysis [2]. The data were mean centered. Outliers were flagged at ± 2.5 studentized residuals and 2.0 times the average leverage. These potential outliers were further studied in relationship to their position within the calibration design in order to make a final judgement on whether or not to remove them from the analysis. Leave-one-out cross-validation was used to compute RMSEP values used to assess model performance. The standard PLS outlier detection techniques are compared to the two robust techniques.

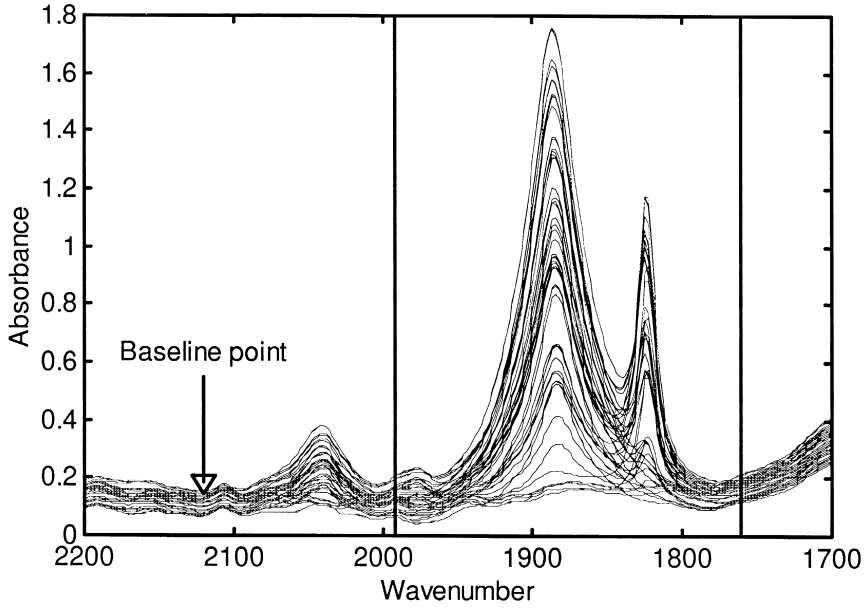

Fig. 3. Absorbance vs. wavenumber for spectra used in the calibration experiment. The solid vertical lines indicate the region used for calibration and the arrow indicates the point where the one-point baseline was taken.

3.3. Robust PCR

A PCA model was developed using the samples identified by the RHM method. Score values for all of the data were then computed using this PCA model. The concentrations were regressed onto the scores that were augmented with a column of ones so as to estimate an intercept. Intercept adjustments were performed at each elemental regression step as suggested by Rousseeuw and Leroy [5]. No other scaling was performed. Five thousand elemental regressions were performed at each factor level. Each regression experiment was repeated at 50–10% breakdown in 10% increments. Robust estimates of the model concentration residuals were computed (see Eq. 6 of Appendix) and used to calculate standardized concentration residuals. A critical value for the robust-scaled standardized residual was set to ± 2.5. Once outliers were identified using the robust PCR outlier diagnostics, these samples were removed from the analysis and the standard PLS regression model was built using the cleaned data set. Leave-one-out cross-validation was used for the validation of the final PLS model.

In general, PLS and PCR models can be different, especially when there is significant variance in X not related to y. For the data used in this study, the results from the PLS and PCR models are quite similar. For brevity, only the results from the PLS models will be presented with the understanding that the corresponding PCR models are not significantly different.

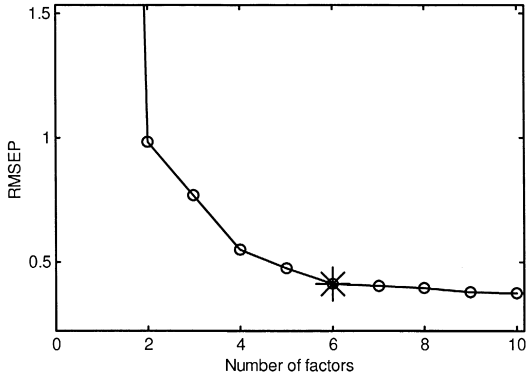

Fig. 4. RMSEP vs. factor number for the standard PLS analysis.

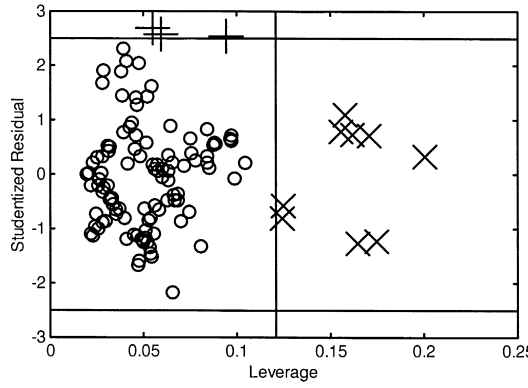

Fig. 5. Studentized residual vs. leverage for standard PLS analysis. The samples marked with a '+' are extreme in studentized residuals and those marked with an 'X' are extreme in leverage.

It is an interesting question as to whether or not, in general, outliers identified from a robust PCR analysis are relevant to a PLS analysis. It is this author's opinion that the PLS and PCR methods are not so different as to give rise to method specific outliers and that outliers found using a robust PCR approach are relevant to a PLS analysis. This is certainly the case for the data in this study, but it is beyond the scope of this paper to address this issue, in general. Perhaps others will find this question of sufficient interest to study.

3.4. IRPLS

Implementation of the iteratively reweighted PLS algorithm followed that of Cummins and Andrews [30] except for one modification. Instead of allowing the rank to change as the weights change, the rank is fixed until the weights converge. The convergence of the weight vectors is assessed by computing the mean of the absolute value of the relative difference of successive weight vectors, excluding zero weights. The convergence criterion is set to $1e-5$ or 20 iterations. In order to speed up the analysis, a leave-out-one-group cross-validation scheme is used. If anything, this should provide a somewhat less optimistic RMSEP than the leave-one-sample-out cross-validation [41]. Based on the concentration data, it was determined that 33 groups existed. The number of samples in a group varied from nine to one with the majority having three samples. Final model per-

formance was assessed using the RMSEP values generated from the leave-one-group-out cross-validation. The IRPLS analysis was conducted on both the mean-centered and nonmean-centered data. Mean centering of the data for the IRPLS analysis tended to provide for poorer performance than without mean centering and, thus, only the nonmean-centered results are presented. The degradation in performance with mean centering is consistent with the proposal that mean centering data does not make sense if outliers are potentially present due to the fact that the mean is a nonrobust estimate of location. The four weight functions used for this analysis are shown in Table 1 and are displayed in Fig. 1a–d. The μ in the equations is computed from the residuals divided by the MADM of the residuals. The c in the equations is the tuning constant and the values used in this study are reported in column 3 of Table 1.

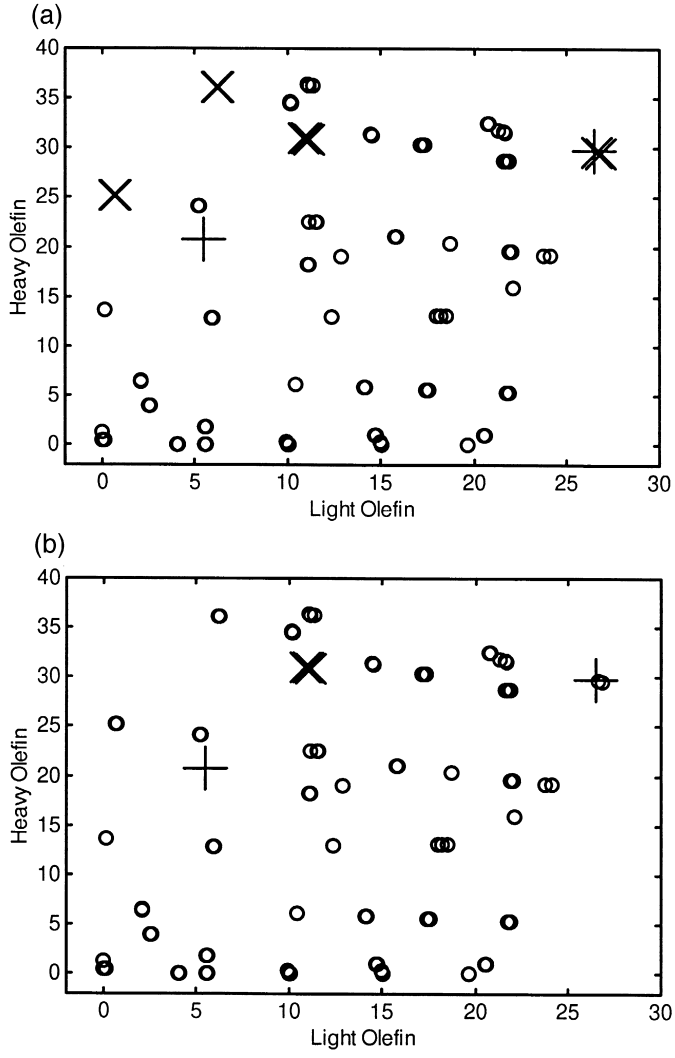

Fig. 6. (a) Heavy olefin vs. light olefin with 12 potential outliers highlighted for the heavy olefin analysis using a six-factor model. High studentized residuals are marked with a '+', those with excessive leverage are marked with an 'X', (b) Heavy olefin vs. light olefin, outliers that were removed are marked.

4. Results

4.1. Presentation

A detailed analysis of the heavy olefin model construction and a more abbreviated analysis of the light olefin model construction are given below. The analysis is separated into the standard PLS analysis, the robust PCR analysis and the IRPLS analysis.

4.2. Heavy olefin analysis

4.2.1. Standard PLS analysis

Fig. 4 displays the RMSEP vs. factor number for the standard PLS analysis using leave-one-out cross-validation. As may be seen, there is no anomalous behavior in this plot that might suggest the presence of outliers. A rank of 6 is reasonable for this model with an RMSEP of 0.41. Fig. 5 displays the studentized residuals vs. leverage for the six-factor model. There are 12 potential outliers, three in studentized residual, marked with '+' and nine with excess leverage, marked with an 'X'. Removing the 12 points and repeating the analysis is the most straightforward way to proceed. However, a more detailed examination of the position of the potential outliers in the concentration space is necessary to make a final assessment.

Fig. 6a displays the calibration design with the samples showing high studentized residuals marked with a '+' and those with high leverage marked with an 'X'. Samples with high leverage but at the edges of the concentration design space were allowed to remain in the model. Fig. 6b displays the final calibration design. The three samples marked with a '+' and the three samples marked with an 'X' were removed from the analysis. The PLS analysis was repeated on the reduced data set and a very similar RMSEP vs. factor number plot was obtained. Again, a six-factor model was selected and only slight improvement of the RMSEP (0.37 vs. 0.41) was found. With this second model, the same six high leverage points were found to have high leverage again, and one sample was found to be slightly high with respect to the studentized residuals. This was deemed to be the final model for the application of standard PLS to the heavy olefin analysis.

4.2.2. Robust PCR analysis

The first step for a robust PCR is a robust PCA. For robust PCA, outlier samples are identified first and then the PCA analysis is performed. To identify the outlier samples, the method of RHM was used [35]. Fig. 7 displays the number of times a sample had a vector length appear in the upper 5% of vector lengths vs. the heavy olefin wt.% vs. the light olefin wt.%. There are 17 samples that have one or more counts in the upper 5% of vector lengths and are, thus, identified as potential outliers. Those samples at the lowest concentrations of both components are the most extreme. Other samples identified by this procedure are found at the edges of the calibration design. These 17 samples were excluded from the PCA decomposition but were projected onto the PCA space with the rest of the calibration samples so they could be used in the multiple regression step.

The next step for the robust PCR is the multiple linear regression of the concentration values onto all of the scores from the PCA plus a column of ones to estimate an intercept. The step is made to be robust by use of the LTS regression method. This method was applied at each rank setting from 1 to 10 and at breakdown settings of 50% to 10% in 10% increments. At each rank and each breakdown point setting, a PCR error estimate is given. The robust error estimate is more closely related to a prediction error than a fit error, because only as many samples as there are, parameters are used to estimate the robust model, while all of the samples are used to estimate the error.

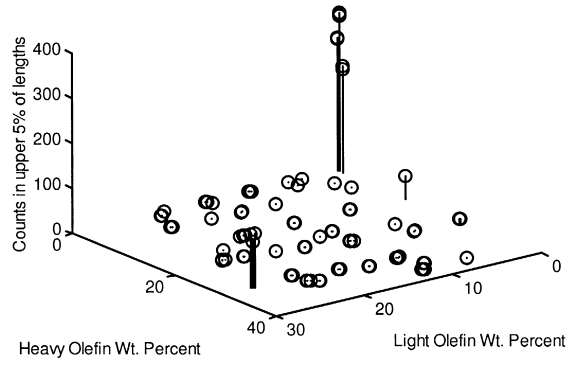

Fig. 7. Number of counts in the upper 5% of vector lengths vs. heavy olefin wt.% vs. light olefin wt.%.

Table 2
PCR error estimate for heavy olefin at ranks 1 to 10 and breakdown points of 50% to 10%

Rank	Breakdown				
	50%	40%	30%	20%	10%
1	12.15	11.55	11.39	11.33	11.24
2	1.02	0.89	0.98	0.79	0.85
3	0.70	0.66	0.66	0.77	0.80
4	0.56	0.51	0.45	0.42	0.48
5	0.18	0.16	0.15	0.15	0.42
6	0.16	0.16	0.14	0.15	0.41
7	0.16	0.16	0.16	0.16	0.41
8	0.16	0.16	0.15	0.30	0.40
9	0.21	0.18	0.15	0.27	0.41
10	0.22	0.14	0.13	0.32	0.37

Table 2 displays the PCR error estimate for models from rank 1 to 10 and breakdown points from 50% to 10% for the heavy olefin analysis. From these data, an optimal model is selected. It is accepted that a model with lower rank for an equivalent prediction error is better than a higher rank model. It is also reasonable to believe that a model found at a lower breakdown point (a model using more of the data) giving equivalent prediction error is better than another model found at a higher breakdown point (a model using less of the data). Given these criteria, the optimal model is found by simultaneously searching from low to high rank and low to HBP for a minimum in prediction error, or more practically for insignificant differences in prediction error. Here, a rank 5 model with a breakdown point of 20% and error estimate of 0.15 is chosen. This error estimate is consistent with a reference value error of approximately 0.2.

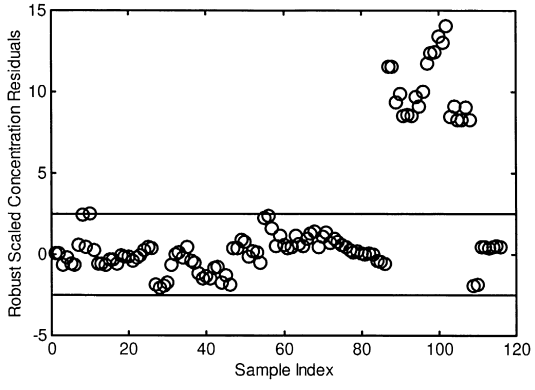

Fig. 8. Robust scaled concentration residuals vs. run order number, heavy olefin analysis.

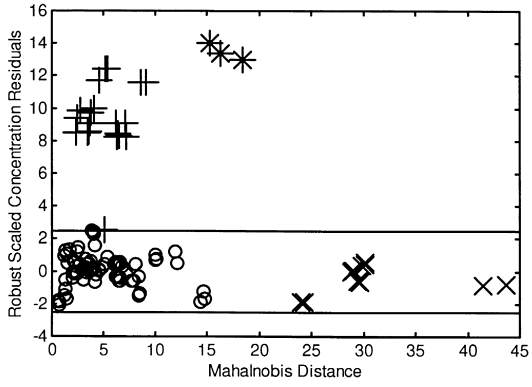

Fig. 9. Robust scaled concentration residuals from LTS analysis vs. robust Mahalanobis distance from PCA analysis. The samples marked with an 'X' are the 17 samples that were identified as outliers using the RHM analysis in the PCA step. Samples marked with a '+' are samples identified as outliers using the LTS analysis and samples marked with both an 'X' and a '+' were found to be outliers by both methods.

Fig. 8 displays the robust-scaled concentration residuals vs. sample run order for the five-factor model with 20% breakdown. There are 23 samples deemed to be potential outliers using a scaled concentration residual critical value of ± 2.5. All of those samples except one occur between and including run numbers 87 to 108. The nature of these samples is discussed below.

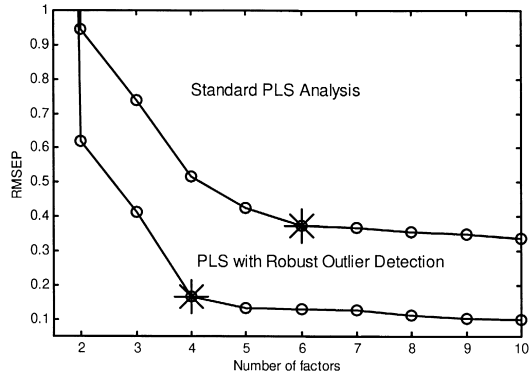

Fig. 10. RMSEP vs. factor number for standard PLS analysis and PLS analysis of data after removing the 23 samples found to be outliers by the robust PCR technique. The '*' symbol indicates the rank choice.

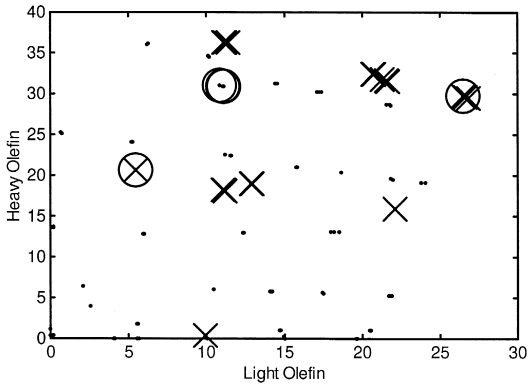

Fig. 11. Heavy olefin vs. light olefin wt.%. Samples marked with an 'O' were those found to be outliers using the standard PLS diagnostics and those marked with an 'X' were found to be outliers using the robust PCR approach.

A plot of the robust scale concentration residuals from the LTS analysis vs. a robust Mahalanobis distance from the robust PCA provides an overview of the two step robust PCR analysis. Fig. 9 displays this plot for the heavy olefin analysis. The two horizontal lines represent ± 2.5 robust-scaled concentration residuals. The samples marked with an 'X' are the 17 samples that were identified as outliers using the RHM analysis. The samples marked with a '+' are the 23 samples found to be beyond the 2.5 robust-scaled concentration residuals by the LTS regression method. There are three samples that are marked with both an 'X' and a '+' because they were found to be outliers by both methods. Fourteen of the samples found to be extreme when using the RHM technique were found to fit the five-factor PCR model well. The

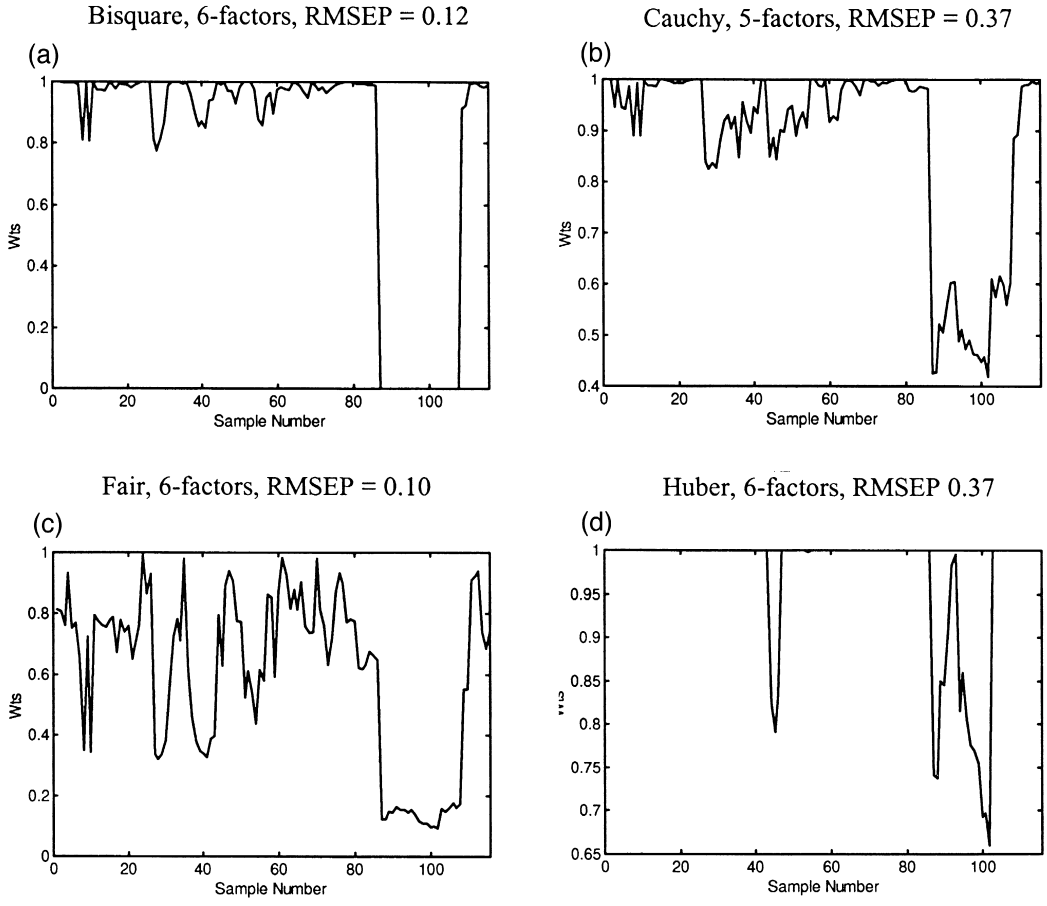

Fig. 12. Sample weight vs. sample order for heavy olefin analysis: (a) Bisquare, (b) Cauchy, (c) Fair, (d) Huber.

large Mahalanobis distances for these samples are consistent with their position in the concentration space. These samples are good leverage points and will be retained for the final PLS model. Three of the 17 samples were not fit well by the five-factor PCR model. Given these considerations, a final PLS model was constructed with 93 samples (the 23 samples from the LTS analysis were excluded). The RMSEP vs. factor number plot for the model with 93 samples is shown in Fig. 10 along with the RMSEP vs. factor number plot for the standard PLS analysis. A four-factor PLS model was selected as optimal giving an RMSEP of 0.17.

A comparison of the samples that were removed from the analysis using the standard PLS diagnostics and those removed using the robust PCR is shown in Fig. 11. The samples marked with an 'O' were those found to be outliers using the standard PLS diagnostics and those marked with an 'X' were found to be outliers using the robust PCR approach. There are few samples in common between the two exclusion sets. If the robust results are to be believed, then there are clearly instances of masking and swamping in this data set.

4.2.3. Iteratively reweighted PLS analysis

For the iteratively reweighted, PLS analysis four weight functions were tested including bisquare, Fair, Cauchy and Huber (see Table 1 and Fig. 1a–d). The performance of each weight function was assessed using a leave-one-group-out cross-validation. Fig. 12a–d displays the weight vector using the bisquare, Cauchy, Fair and Huber weight functions at the optimal model rank. Table 3 displays the optimal rank and RMSEP values for the heavy olefin analysis for each of the four weight functions.

Table 3
Analysis summary for heavy olefin

Method	Factors	RMSEP	Number of samples
PLS full data set	6	0.41	116
PLS std. outlier detection	5	0.35	110
PLS with robust PCR outlier detection	4	0.17	93
IRPLS (bisquare weight)	6	0.12	NA
IRPLS (Cauchy weight)	5	0.37	NA
IRPLS (Fair weight)	6	0.10	NA
IRPLS (Huber weight)	6	0.37	NA

NA = Not applicable.

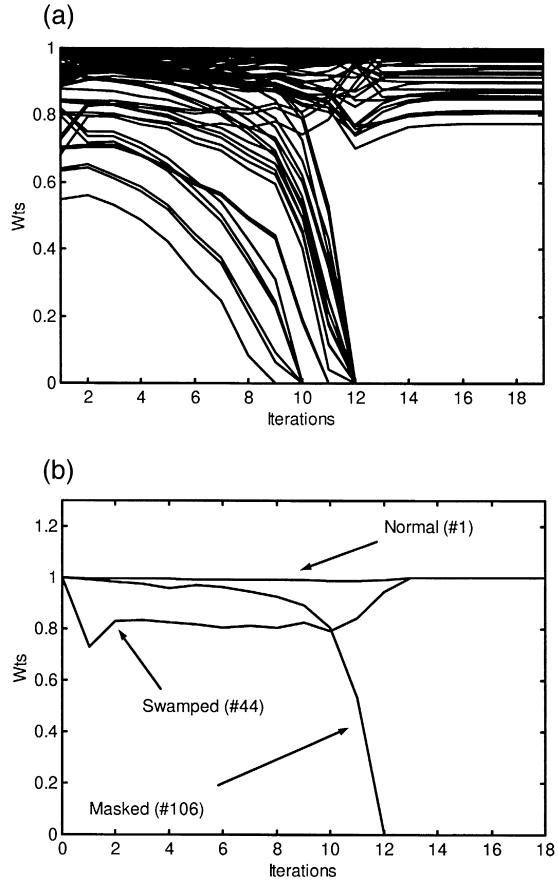

Fig. 13. (a) Sample weight vs. iteration number for bisquare function, heavy olefin analysis, six-factor model, (b) Sample weight vs. iteration number for three specific samples using the bisquare weight function, heavy olefin analysis, six-factor model.

The bisquare and Fair weight functions provide RMSEP values lower than the standard PLS applied to the data with the outliers removed using the robust PCR approach. The Huber and Cauchy weight functions provided larger RMSEP values. The RMSEP values from the bisquare and Fair weight functions may be too good to be true given that the reference error estimate is 0.2. The bisquare weight function clearly distinguishes the outlier samples by assigning a weight of zero (see Fig. 12a). Those samples assigned a weight of zero are consistent with the samples identified as outliers using the robust PCR ap-

proach. The other weight functions showed similar behavior but are not as clear as the bisquare function in distinguishing the outliers. The weight functions may have shown different behavior with different tuning constants, and this makes their use somewhat problematic in that another parameter has to be optimized.

Fig. 13a displays the weight for each sample as a function of iteration number for the bisquare function and a six-factor model. The weights stabilize by iterations 9–13. The weight functions for three specific samples are shown in Fig. 13b. Sample #1 is considered to be a normal sample with a weight that starts out near one and does not vary much from this value over the iterations. Sample #44 is an example of a swamped sample. Early in the iterations, the sample is down-weighted relative to the normal samples; but by iteration number 10, the weight begins to increase and by iteration #14 has reached a value consistent with normal samples. Sample #106 is an example of a sample that is masked. Early in the iteration number, this sample has a weight consistent with the normal samples. At iteration number 10, the weight for this sample shows a sharp decrease and finally is set to zero at iteration number 12.

The final results for the heavy olefin analysis are summarized in Table 3. The lowest RMSEP value was found with the IRPLS analysis using the Fair weight function. This result may be somewhat suspect given that the reference value error estimate for the heavy olefin analysis is 0.2. The results from a PLS analysis using the samples found to be outliers

Table 4
PCR error estimate for light olefin at ranks 1 to 10 and breakdown points of 50% to 10%

Rank	Breakdown				
	50%	40%	30%	20%	10%
1	1.28	1.38	1.38	1.23	1.08
2	0.80	0.66	0.64	0.59	0.63
3	0.48	0.42	0.44	0.47	0.51
4	0.54	0.42	0.34	0.39	0.54
5	0.58	0.59	0.45	0.53	0.49
6	0.52	0.38	0.35	0.43	0.47
7	0.54	0.37	0.53	0.42	0.34
8	0.56	0.43	0.52	0.66	0.36
9	0.56	0.46	0.42	0.42	0.39
10	0.56	0.57	0.57	0.43	0.39

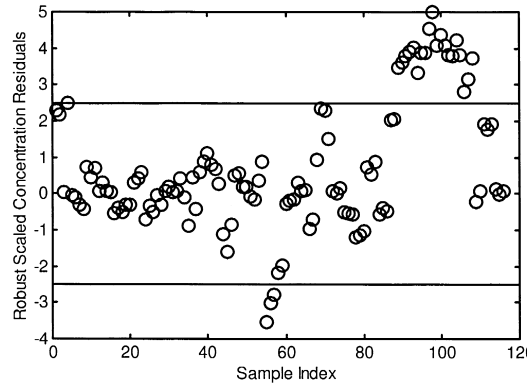

Fig. 14. Robust scaled concentration residuals vs. run order number, light olefin analysis.

by the robust PCR method had an RMSEP of 0.17. It required the elimination of 23 samples as outliers. The standard PLS analysis after application of the standard outlier diagnostics showed a similar RMSEP value to the results before the outliers were eliminated. That RMSEP value, 0.35, was very close to the largest RMSEP value, 0.37, which was the same for the IRPLS results using the Cauchy and Huber weight functions.

4.3. Light olefin analysis

4.3.1. Standard PLS analysis

The standard analysis of the light olefin data was somewhat more complicated than for the heavy olefin analysis. Outliers were identified using studentized

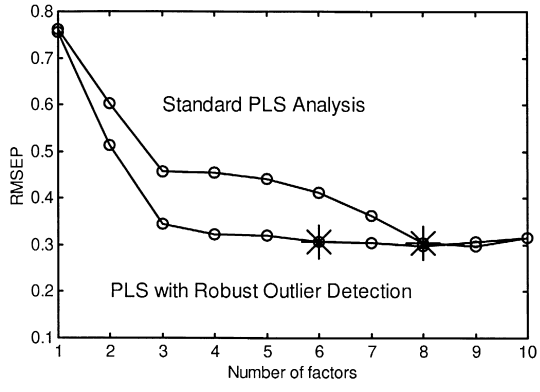

Fig. 15. RMSEP vs. factor number for the light olefin analysis using the standard PLS diagnostics and the robust regression diagnostics.

Table 5
RMSEP and rank for optimal models for the light olefin analysis

Method	Factors	RMSEP	Number of samples
PLS full data set	9	0.34	116
PLS standard outlier detection	8	0.30	103
PLS with robust PCR outlier detection	6	0.31	93
IRPLS (bisquare weight)	7	0.32	NA
IRPLS (Cauchy weight)	7	0.34	NA
IRPLS (Fair weight)	7	0.16	NA
IRPLS (Huber weight)	7	0.37	NA

residual vs. leverage plots with consideration of where the samples were relative to the calibration design as was done for the heavy olefin analysis. Four outlier identifications and removal steps were required for the analysis of the light olefin data. This type of behavior is not uncommon when using standard outlier diagnostics, thus making for a tedious and time-consuming process. An eight-factor model is reasonable giving an RMSEP of 0.3. The error in the reference method is estimated to be 0.4 so the 0.3 value found here at eight factors might be somewhat optimistic.

4.3.2. Robust PCR

The robust PCA step is the same as that for the heavy olefin analysis and is summarized in Section 4.2.2. The results for the robust regression step are shown in Table 4. In this case, the model with four factors and a 30% breakdown appears to be optimal. Fig. 14 displays the scaled robust concentration

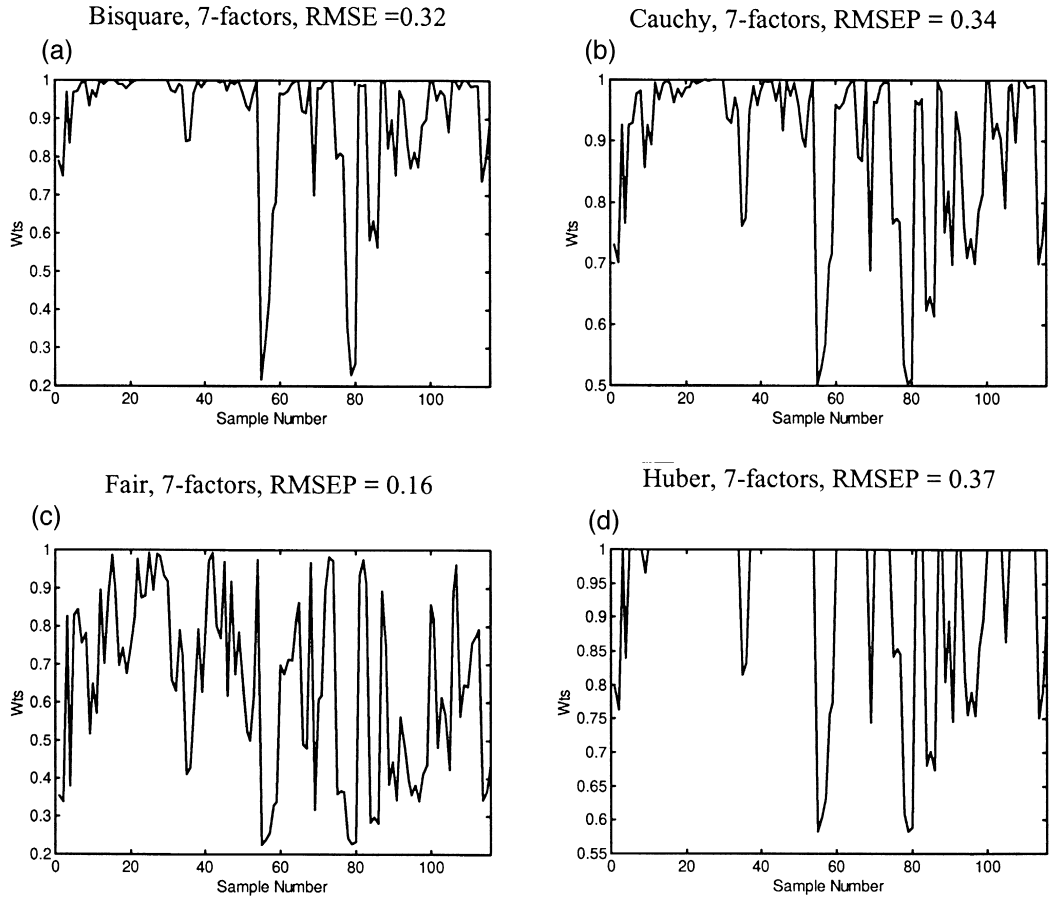

Fig. 16. Sample weight vs. sample order for light olefin analysis: (a) Bisquare, (b) Cauchy, (c) Fair, (d) Huber.

residuals vs. the sample run order for the four-factor model with a 30% breakdown. In this case, there are 23 samples identified as potential outliers. The trend in the outliers is similar to that found for the heavy olefin analysis with 55–57 and 89–108 found to be beyond the ± 2.5 robust-scaled concentration residuals. Fig. 15 displays the RMSEP vs. factor number for the PLS analysis using the standard diagnostics and the PLS analysis after application of the robust diagnostics. For this analysis, there is very little improvement in the RMSEP using the robust PCR analysis over the standard diagnostics.

4.3.3. Iteratively reweighted PLS analysis

The results for the iteratively reweighted PLS analysis are summarized in Table 5. Fig. 16a–d displays the final weight vectors for the light olefin analysis. In this case, a rank 7 model is appropriate for each of the weight function models. The Fair weight function provides the lowest RMSEP value, substantially below the reference method error of 0.4. A summary of the light olefin analysis results is displayed in Table 5.

5. Discussion

5.1. Heavy olefin analysis

The heavy olefin analysis showed considerable improvement in the leave-one-out cross-validation RMSEP values when using either the robust PCR to identify outliers or IRPLS to down weight outliers. The outliers that were identified using the robust PCR analysis and the IRPLS analysis agreed very closely, especially for the bisquare weight function results.

It is always good practice to attempt to find a physical reason for outliers that are found using statistical techniques. For this analysis, the gas chromatography reference method calibration runs were reviewed. It was found that between runs #86 and #87, a 40% change was introduced to the light olefin response factor. Because the chromatography calibration method made use of a normalized calculation, a change in any one of the response factors will introduce a change in the calculated amounts for all of the components. This change in the light olefin response factor corresponded precisely with the beginning of the most extreme outliers identified using the robust approaches.

The response factors were not changed for the reminder of the calibration experiment. If one assumes that the reason for the outliers was due to an anomalous response factor, then why do the outlier samples end at sample 108? The reason for the change in the nature of the samples from outlier to nonoutlier at sample 108 to 109 for the heavy olefin can be explained as follows. From samples 109 to 116, the concentration of the heavy olefin was 1% or less and, thus, the difference in the concentration due to the change in response factor was approximately 0.1 wt.%. Therefore, even though these samples may still be outliers because of the change in the response factor, they were not detected because the difference in concentration level introduced by the response factor change was not detectable. Samples 100–102 show a similarly small difference between concentration levels due to the difference in response factor, so why did they appear as outliers? Recall from Fig. 9 that three of the samples that were deemed to be outliers due to concentration residuals were also found to be extreme by the resampling method and have large Mahalanobis distances. It was exactly these three samples, 100–102, that had the large Mahalanobis distances. Thus, it is believed that these three samples were outliers due to spectral differences and not necessarily due to reference value problems.

5.2. Light olefin analysis

The outliers for the light olefin analysis using the robust PCR agreed with the outliers identified in the heavy olefin analysis. The outliers identified using IRPLS analysis for the light olefin were not consistent with the robust PCR analysis nor were they consistent among the different weight functions. There was apparently considerable improvement in the prediction error using the Fair function with the IRPLS method over the standard PLS and the robust PCR results. The errors from the IRPLS were well below the estimated reference method error so overfitting must be considered as a possibility.

Why the nature of the samples for the light olefin analysis went from outlier to normal samples from sample number 108 to 109 is not clear. Samples 109 and 110 had zero light olefin concentration and, thus,

could explain the change; but samples 111 to 116 had significant concentrations of light olefin, and if a bad response factor was to blame, then they should have appeared as outliers.

5.3. Robust PCR vs. iteratively reweighted PLS

The robust PCR approach identified outliers in a consistent manner for the light and heavy olefin analysis. These identifications are supported by original laboratory references concerning significant changes made to a gas chromatographic response factor. The removal of these outliers provided for significant improvement in calibration model performance over a model constructed using the standard diagnostics. The robust PCR technique distinguish between outliers in the spectral space and in the concentration space. It is relatively easy to program although calculation times can be significant.

The IRPLS approach was consistent with the robust PCR approach for outlier identification for the heavy olefin calibration but showed significant ambiguity in the light olefin analysis. The prediction performance results were better than the robust PCR results, but were better than the estimated reference method error and may be indicative of an overfitting problem. Different weight functions as well as different tuning constant for the same weight function can give different results as shown here and discussed in Ref. [30]. These data-dependent features make the iteratively reweighted approach less attractive.

6. Conclusions

There is increasing interest in the application of multivariate calibration techniques to process measurements where there can be enormous potential value to industry. In these applications, the analyst often has less control over the execution of the experiment and there is correspondingly a higher chance for outliers to be present in the data. Standard multivariate calibration outlier diagnostics are ill-equipped to identify multiple sample outliers in the calibration phase of model building. Robust statistical methods offer an alternative to standard regression analysis that can efficiently detect multiple outliers. Robust PCR for the data studied in this report appears to offer a more reliable approach than IRPLS as well as having fewer adjustable parameters.

Unfortunately, there are few commercial products available that implement robust regression solutions. Without commercially available products and success stories for these techniques, there will be little widespread use of robust analysis. It is hoped that this report demonstrates the power of robust analysis for detection of multiple outliers and the gross inadequacies of the standard diagnostics.

Acknowledgements

The author acknowledges Barbara Kirsch and Kevin Winnet for providing the data. Dr. Mia Hubert is thanked for helping me to understand the details of the least trimmed squares calculations. Mary Beth Seasholtz is thanked for her constructive criticism of the manuscript.

Appendix A

It is well understood that the least squares approach for regression relies on the minimization of the sum of the squared residuals. For the LTS method, the sum of the squared residuals is replaced by a trimmed sum of squared residuals. The squared residuals are ordered from low to high and the sum is computed from the lowest residual to a depth of h given by Eq. 1:

$$h = [(1-\alpha)I] + [\alpha(A+1)], \quad (1)$$

where the square brackets, [ ], refer to the integer portion only; I is the number of samples; A is the number of parameters in the model (for PCR implemented here, A is the number of factors plus one because an intercept is estimated) and α is the desired breakdown point. The breakdown point of the estimator can be varied from 0% to 50% by including more or less squared residuals in the sum.

A key aspect of the high breakdown estimators is how the residuals are generated. The residuals from a standard least squares fit to all of the data cannot be used, because if there are outliers present, then the residuals will be influenced. Instead, a subsample of the data (equal to the number of parameters being es-

timated) is used to estimate the parameters of the model using least squares and then residuals for all of the data from this model are computed and used to calculate the trimmed sum of squared residuals. This subsampling of the data set is repeated many times in order to guarantee with some level of certainty that an uncontaminated subsample of the data is found. The subsample giving the smallest trimmed sum of squared residuals is a solution, although it may not be the exact LTS solution. Recently, Rousseeuw and Van Driessen [12] have introduced a fast algorithm for LTS computations. From this solution, a preliminary estimate of the concentration residual error scale, S^0_{LTS} can be estimated using Eq. 2:

$$S^0_{\mathrm{LTS}} = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h}\left((r)^2\right)_{i:n}}. \tag{2}$$

The $\sum_{i=1}^{h}\left((r)^2\right)$ is meant to indicate a summation of the ordered squared residuals from 1 to a depth h. The constant $d_{h,n}$ is computed using Eq. 3:

$$d_{h,n} = \frac{1}{\sqrt{1-\frac{2n}{hc_{h,n}}\phi\left(\frac{1}{c_{h,n}}\right)}}. \tag{3}$$

The function ϕ is the standard normal density function shown in Eq. 4:

$$\phi(x) = \sqrt{\frac{1}{2\pi}}\,\mathrm{e}^{\left(\frac{-x2}{2}\right)}. \tag{4}$$

The constant $c_{h,n}$ is computed using Eq.5:

$$c_{h,n} = \frac{1}{\Phi^{-1}\left(\frac{h+n}{2n}\right)}. \tag{5}$$

And $\Phi(x)$ is the standard normal distribution function $\Phi(x) = P(X \le x)$. The constants $c_{h,n}$ and $d_{h,n}$ are chosen to make the scale estimator consistent at the Gaussian model. A final scale estimate is computed using Eq. 6:

$$S_{\mathrm{LTS}} = \sqrt{\frac{\sum_i w_i r_i^2}{\sum_i w_i - p}} \tag{6}$$

with p equal to the number of parameters and r_i equal to the ith residual and:

$$w_i = \begin{cases} 0 \text{ if } \left|\dfrac{r_i}{s^0_{\mathrm{LTS}}}\right| > 2.5 \\ 1 \text{ otherwise} \end{cases}$$

As may be seen from Eq. 6, the final scale estimate is robust to outliers as well. Samples with residuals beyond ± 2.5 robust-scaled residuals are suspect.

Using Eq. 1, the depth of the sum can be computed for a desired breakdown point. Multiple breakdown points are used in this paper in order to avoid the consequences of the exact fit property possessed by all HBP estimators. The consequences of this property are that the estimator can follow the bad data points rather than the good data points. Having a large number of data points can mitigate the consequences, but having a higher number of dimensions will exacerbate it. In order to guard against this problem, Ryan [38] has suggested that multiple breakdown points be used and the results compared. This is the approach used in this paper. Others have suggested that the exact fit property of high breakdown point estimators is a desirable property that can be used to identify meaningful structure in the data [42]. The details appearing in this appendix were taken from Rousseeuw and Hubert [11] and from personal communications with Dr. Hubert.

References

[1] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics A Practical Guide, Wiley, New York, 1998.

[2] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1989.

[3] M.J. Piovoso, K.A. Kosanovich, J.P. Yuk, Process data chemometrics, IEEE Trans. Instrum. Meas. 41 (2) (1992) 262–268.

[4] J. Workman Jr., D.J. Veltkamp, S. Doherty, B. Anderson, K. Creasy, M. Koch, J.F. Tatera, A.L. Robinson, L. Bond, L.W. Burgess, G.N. Bokerman, A.H. Ullman, G.P. Darsey, F. Mozayeni, J.A. Bamberger, M.S. Greenword, Process analytical chemistry, Anal. Chem. 71 (1999) 121R–180R.

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

[6] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.

[7] S.J. Devlin, R. Gnanadesikan, J.R. Kettenring, Robust estimation of dispersion matrices and principal components, J. Am. Stat. Assoc. 76 (1981) 354–362.
[8] K.R. Gabriel, C.L. Odoroff, Resistant lower rank approximation of matrices, Technical report 83/02, Department of Statistics University of Rochester, 1983.
[9] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, J. Am. Stat. Assoc. 80 (1985) 759–766.
[10] P.J. Rousseeuw, Least median of squares regression, J. Am. Stat. Assoc. 79 (1984) 871–880.
[11] P. Rousseeuw, M. Hubert, Recent developments in PROGRESS, L1-Statistical procedures and related topics, in: Y. Dodge (Ed.), The IMS Lecture Notes-Monograph Series vol. 31 (1997), pp. 201–215.
[12] P.J. Rousseeuw, K. Van Driessen, Computing LTS regression for large data sets, Technical Report, University of Antwerp, 1999.
[13] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[14] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, J. Am. Stat. Assoc. 85 (1990) 871–880.
[15] G.E. Dallal, P.J. Rousseeuw, LMSMVE: a program for least median of squares regression and robust distances, Comput. Biomed. Res. 25 (1992) 384–391.
[16] D.M Hawkins, J.S Simonoff, High breakdown regression and multivariate estimation, Appl. Stat. 42 (1993) 423–441.
[17] G.A. Watson, On computing the least quantile of squares estimate, SIAM J. Sci. Comput. 19 (1998) 1125–1138.
[18] T.P. Hettmansperger, S.J. Sheather, A cautionary note on the method of least median squares, Am. Stat. 46 (1992) 79–83.
[19] L.A. Stefanski, A note on high-breakdown estimators, Stat. Probab. Lett. 11 (1991) 353–358.
[20] A.C. Atkinson, Masking unmasked, Biometrika 73 (1986) 533–541.
[21] W. Fung, Unmasking outliers and leverage points: a confirmation, J. Am. Stat. Assoc. 88 (1993) 515–519.
[22] A.C. Atkinson, Fast very robust methods for the detection of multiple outliers, J. Am. Stat. Assoc. 89 (1994) 1329–1339.
[23] A.C. Atkinson, H.M. Mulira, The stalactite plot for the detection of multivariate outliers, Stat. Comput. 3 (1993) 27–35.
[24] M.S. Mayo, J.B. Gray, Elemental subsets: the building blocks of regression, Am. Stat. 51 (1997) 122–129.
[25] A.S. Hadi, J.S Simonoff, Procedures for the identification of multiple outliers in linear models, J. Am. Stat. Assoc. 88 (1993) 1264–1272.
[26] C.W. Coakely, T.P. Hettmansperger, A bounded influence, high breakdown, efficient regression estimator, J. Am. Stat. Assoc. 88 (1993) 872–880.
[27] B. Walczak, D.L. Massart, Robust principal components regression as a detection tool for outliers, Chemom. Intell. Lab. Syst. 27 (1995) 41–54.
[28] B. Walczak, Outlier detection in multivariate calibration, Chemom. Intell. Lab. Syst. 28 (1995) 259–272.
[29] J. Wang, Y. Xie, R. Yu, Maximum sum of binary-coded residuals (MASBR) regression as a robust procedure for treatment of spectral data, J. Chemom. 9 (1995) 373–387.
[30] D.J. Cummins, C.W. Andrews, Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation, J. Chemom. 9 (1995) 489–507.
[31] I.N. Wakeling, H.J.H. Macfie, A robust PLS procedure, J. Chemom. 6 (1992) 189–198.
[32] J.A. Gil, R. Romera, On robust partial least squares (PLS) methods, J. Chemom. 12 (1998) 365–378.
[33] Y. Liang, O.M. Kvalheim, Robust methods for multivariate analysis — a tutorial review, Chemom. Intell. Lab. Syst. 32 (1996) 1–10.
[34] H. Hove, Y. Liang, M. Kvalheim, Trimmed object projections: a nonparametric robust latent-structure decomposition method, Chemom. Intell. Lab. Syst. 27 (1993) 33–40.
[35] W.J. Egan, S.L. Morgan, Outlier detection in multivariate analytical chemical data, Anal. Chem. 79 (1998) 2372–2379.
[36] P.J. Rousseeuw, Tutorial to robust statistics, Chemom. Intell. Lab. Syst. 5 (1991) 1–20.
[37] A. Singh, Outliers and robust procedures in some chemometric applications, Chemom. Intell. Lab. Syst. 33 (1996) 75–100.
[38] T.P. Ryan, Modern Regression Methods, Wiley, New York, 1997.
[39] P.J. Huber, Robust regression: asymptotics, conjectures and Monte Carlo, Ann. Stat. 1 (1973) 799–821.
[40] P.W. Holland, R.E. Welsch, Robust regression using iteratively reweighted least squares, Commun. Stat., Part A. Theory Meth. 6 (1977) 813–827.
[41] H.A. Martens, P. Dardenne, Validation and verification of regression in small data sets, Chemom. Intell. Lab. Syst. 44 (1998) 99–121.
[42] V.J. Yohai, R.H. Zamar, High breakdown-point estimates of regression by means of the minimization of an efficient scale, J. Am. Stat. Assoc. 83 (1988) 406–413.