ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/10575010

Introduction and Application of Secured Principal Component Regression for Analysis of Uncalibrated Spectral Features in Optical Spectroscopy and Chemical Sensing

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 2003

Impact Factor: 5.64 · DOI: 10.1021/ac020758w · Source: PubMed

CITATIONS
23

READS
35

2 AUTHORS, INCLUDING:

Boris Mizaikoff
Universität Ulm
310 PUBLICATIONS **4,799** CITATIONS

SEE PROFILE

Available from: Boris Mizaikoff
Retrieved on: 29 January 2016

Anal. Chem. **2003,** *75,* 3050−3058

Introduction and Application of Secured Principal Component Regression for Analysis of Uncalibrated Spectral Features in Optical Spectroscopy and Chemical Sensing

Frank Vogt* and Boris Mizaikoff

Georgia Institute of Technology, School of Chemistry and Biochemistry, Atlanta, Georgia 30332-0400

In this study, a novel chemometric algorithm for improved evaluation of analytical data is presented and applied to three spectroscopic data sets obtained by different analytical methods. This so-called secured principal component regression (sPCR) was developed for detecting and correcting uncalibrated spectral features newly emerging in spectra after finalizing the PCR calibration, which may result in major concentration errors. Hence, detection and correction of uncalibrated features is essential. Furthermore, detected uncalibrated features provide qualitative information for sensing and process monitoring applications indicating problems in the process flow. After conventional PCR calibration, sPCR analyzes measurement data in two steps: The first step investigates whether the obtained data set is consistent with the calibration model or not. If spectroscopic features are found that cannot be modeled by the principal components, they are extracted from the measurement spectrum. This corrected spectrum is then evaluated by conventional PCR. In the Experimental Section, sPCR was successfully applied to three data sets obtained by different spectroscopic measurements in order to corroborate general applicability of the proposed concept. For each data set, one of several substances was excluded from the calibration acting in the sPCR assessment as uncalibrated absorber. The test sets consisted of disturbed and undisturbed samples. A total of 109 out of 110 test samples were correctly classified as disturbed or undisturbed by an uncalibrated absorber. It was confirmed that the extracted disturbance spectra are in accordance with the spectra of the uncalibrated analytes. The concentration results obtained with sPCR were found to be equivalent to conventional PCR results in the case of undisturbed samples and more precise for disturbed samples.

Principal component analysis (PCA) and principal component regression (PCR)[1−3] are now standard techniques for calibration of spectrometers and for evaluation of unknown measurement spectra. Conventional PCA/PCR determines an appropriate calibration model and uses this model to evaluate spectra by projecting them onto the principal components (PCs). This projection results in score vectors from which concentration values can be calculated. However, if new spectral features emerge during measurements after the calibration is finalized, this model is not valid anymore and conventional PCA/PCR evaluation fails without notification since these new features are not contained in the PCs. In the remainder of this study, such new spectral elements are denoted as "uncalibrated" features. Occurrence of uncalibrated spectral features is a commonly encountered event during application of optical measurement techniques and especially problematic when user-interactive data control is not feasible or not possible. This is the case for a broad variety of spectroscopic and sensing applications such as process monitoring tasks or continuous control of legal emission and immission limits of pollutants. There are two main reasons why uncalibrated spectral features must be detected and corrected: First, measured data have to be corrected for uncalibrated features in order to avoid deriving wrong concentrations. Second, in on-line process monitoring applications, recognition of such disturbances is particularly relevant since they may indicate quality problems or even process failures. During emission monitoring, detection of new absorbers in the samples, e.g., exhaust gas, may indicate new sources of pollution. Hence, such information is particularly valuable for tracking the origin of pollutants. For immission measurement applications, it is very unlikely that all possible analytes can be calibrated due to the multitude of compounds in varying composition. Hence, data evaluation has to be inherently prepared for processing of uncalibrated analytes.

In ref 4 an expanded classical least-squares algorithm is presented for synthetic correction of unmodeled spectral features caused by temperature drifts. However, this approach is limited to disturbances resulting in linear unmodeled influence, such as temperature in this example, and it is assumed that the underlying shape of the disturbance does not change. In contrast, Vogt et al. have developed universal algorithms[5,6] without such restrictions. These algorithms do not require a priori information for detecting and correcting uncalibrated spectral features. Preliminary results for detection and correction of localized uncalibrated spectral

* Corresponding author. Phone: +1 (404) 894-4030. Fax: +1 (404) 894-7452. E-mail: FRNVogt@aol.com.

(1) Marbach, R.; Heise, M. *Chem. Intell. Lab. Syst.* **1990**, *9*, 45−63.
(2) Martens, H.; Næs, T. *Multivariate Calibration*, 2nd ed.; John Wiley & Sons: New York, 1991.
(3) Egan, W.; Brewer, W.; Morgan, W. *Appl. Spectrosc.* **1999**, *53*, 218−225.
(4) Haaland, D. *Appl. Spectrosc.* **2000**, *54*, 246−254.
(5) Vogt, F.; Tacke, M. *Proc. SPIE-Int. Soc. Opt. Eng.* **2000**, *4201*, 12−23.
(6) Vogt, F.; Mizaikoff, B. *J. Chemom.* **2003**, *17*, 225−236.

10.1021/ac020758w CCC: $25.00 © 2003 American Chemical Society
Published on Web 05/07/2003

characteristics have been discussed in ref 5. Extensive theory of the so-called *secured PCR* (sPCR)[6] algorithm has recently been proposed and been applied successfully to simulated spectroscopic data sets.

This study focuses on the practical application of sPCR and the correction of disturbed experimental data sets resulting from various spectroscopic techniques. Three spectroscopic data sets have been analyzed by means of the sPCR algorithm and compared to the results obtained by conventional PCR. Spectra obtained from gaseous and from liquid-phase samples in the ultraviolet and the mid-infrared wavelength region have been investigated in order to demonstrate the wide applicability of sPCR.

In general, there are two main categories of uncalibrated spectral features: (1) disturbances introduced by the measurement device itself and (2) disturbances due to unexpected changes of the analyzed samples. Device problems can be resolved by a new calibration. In particular, drifts can be corrected by application of so-called pseudo principal components.[7] However, unexpected changes of the sample matrix demand thorough analysis of the entire measurement procedure. On account of this circumstance, recognition of calibration model failures is of broad interest for analytical chemistry.

While an appropriate approach is being developed, three main requirements have to be met. First, the qualitative analysis must be independent from sample composition; i.e., no assumption about the shape, position, and strength of uncalibrated spectral features can be made. Second, this analysis must be performable by a computer without human interaction in order to enable automated on-line monitoring. Finally, in the undisturbed case, sPCR must result in concentrations equivalent to conventional PCR.

In the remainder, matrices will be noted in capital boldface letters and vectors in small boldface letters. Transposed objects are indicated by superscript T; subscript cal or meas specify calibration or measurement items. To introduce the used notation PCR is summarized briefly: During the calibration process K spectra of calibration samples are measured at N different wavenumber positions. Usually, calibration spectra are written in the rows of a calibration matrix $\mathbf{X}_{cal}$. If $N > K$, however, it is advantageous to write them in columns of $\mathbf{X}_{cal}$, since the computation effort can be decreased considerably.[9,10] Mean centering[2,8] of calibration spectra and calibration concentrations is performed as a preprocessing step in order to subtract a common background spectrum. Calculation of PCs is accomplished by a singular value decomposition (SVD)[11,12] of $\mathbf{X}_{cal}$:

$$\mathbf{X}_{cal(N\times K)} = \mathbf{P}_{(N\times K)}\cdot\mathbf{S}_{(K\times K)}\cdot\mathbf{Z}^{\mathbf{T}}_{(K\times K)} = \mathbf{P}_{(N\times K)}\cdot\mathbf{T}^{\mathbf{T}}_{(K\times K)} \quad (1)$$

The orthonormal PCs are contained in the columns of

$$\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K] \quad (2)$$

the corresponding scores of the calibration spectra are contained in the columns of $\mathbf{T}^{\mathbf{T}} = \mathbf{S}\cdot\mathbf{Z}^{\mathbf{T}}$. Due to noise contained in the calibration spectra, the rank of $\mathbf{X}_{cal}$ usually is equal to min(K,N), although there are only $R \le \min(K,N)$ spectroscopic meaningful PCs. The most difficult part of the PCA is the decision about the true "spectroscopic" dimension R of the calibration model. If all K PCs would be included, the PCA calibration model would usually be overfitted and the results would be downgraded.[13] Hence, R is usually determined by cross-validation.[2,14] After defining R, $\mathbf{P}_{(N\times R)}$ and $\mathbf{T}^{\mathbf{T}}_{(R\times K)}$ are abridged to the number of relevant PCs without changing the notation in the following. After this calibration, new mean-centered measurement spectra $\mathbf{x}_{meas(N\times 1)}$ can be analyzed by a multivariate least-squares fit in order to obtain scores $\mathbf{t}_{meas}$:

$$\mathbf{x}_{meas} = [\mathbf{p}_1 \cdots \mathbf{p}_R]\cdot\mathbf{t}_{meas\ (R\times 1)} + \epsilon$$

$$\mathbf{t}_{meas} = \underbrace{\left(\mathbf{P}^{\mathbf{T}}\mathbf{P}\right)^{-1}}_{=\mathbf{1}}\cdot\mathbf{P}^{\mathbf{T}}\cdot\mathbf{x}_{meas} = \mathbf{P}^{\mathbf{T}}\cdot\mathbf{x}_{meas} \quad (3)$$

In the second line of eq 3, advantage was taken of $\mathbf{P}$ being orthogonal.[11,12] From the scores vector $\mathbf{t}_{meas}$, a concentration vector $\mathbf{y}_{meas(P\times 1)}$ of P calibrated analytes is derived then. The residual spectrum

$$\epsilon = \mathbf{x}_{meas} - \mathbf{P}\cdot\mathbf{t}_{meas} = [\mathbf{1}_{(N\times N)} - \mathbf{P}\cdot\mathbf{P}^{\mathbf{T}}]\cdot\mathbf{x}_{meas} \quad (4)$$

cannot be modeled by the PCs. If there are no uncalibrated absorbers present in a sample, the residual spectrum obtained from this sample will consist of noise only. However, if uncalibrated absorbers emerge, ϵ will feature nonrandom components. Hence, the proposed sPCR algorithm takes advantage of ϵ containing additional information for detecting uncalibrated spectral features which is disregarded by conventional PCR.

SPCR ALGORITHM

1. Basic Idea of sPCR. The sPCR algorithm as developed and discussed in detail in ref 6 will be qualitatively explained here. Since uncalibrated spectral features appear after the calibration is finalized, the calibration step of sPCR is the same as a conventional PCR calibration. These PCs (2) are determined once and used for the sPCR algorithm until a new calibration is performed.

To avoid major concentration errors due to uncalibrated spectral features, the novel sPCR algorithm augments the conventional one-step PCR evaluation of measurement spectra by a qualitative data pretreatment. This first qualitative sPCR step analyzes whether the measured data are consistent with the PCs at all. If uncalibrated spectral features are found in a measurement spectrum, the algorithm subtracts them and evaluates the corrected spectrum by the second quantitative step, a conventional PCR. Once a disturbance is found in a measurement spectrum

(7) Vogt, F.; Rebstock, K.; Tacke, M. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 175−178.
(8) Draper, N.; Smith, H. *Applied Regression Analysis*, 3rd ed.; John Wiley & Sons: New York, 1998.
(9) Vogt, F.; Tacke, M. *Chemom. Intell. Lab. Syst.* **2001**, *59*, 1−18.
(10) Vogt, F.; Tacke, M. *J. Chemom.* **2002**, *16*, 562−575.
(11) Golub, G.; Van Loan, C. *Matrix Computations*, 2nd ed.; Johns Hopkins University Press: Baltimore, MD, 1989.
(12) Press: W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992.
(13) Mandel, J. *Am. Stat.* **1982**, *36*, 15−24.
(14) Davis, A. *Spectrosc. Eur.* **1998**, *10/2*, 24−25

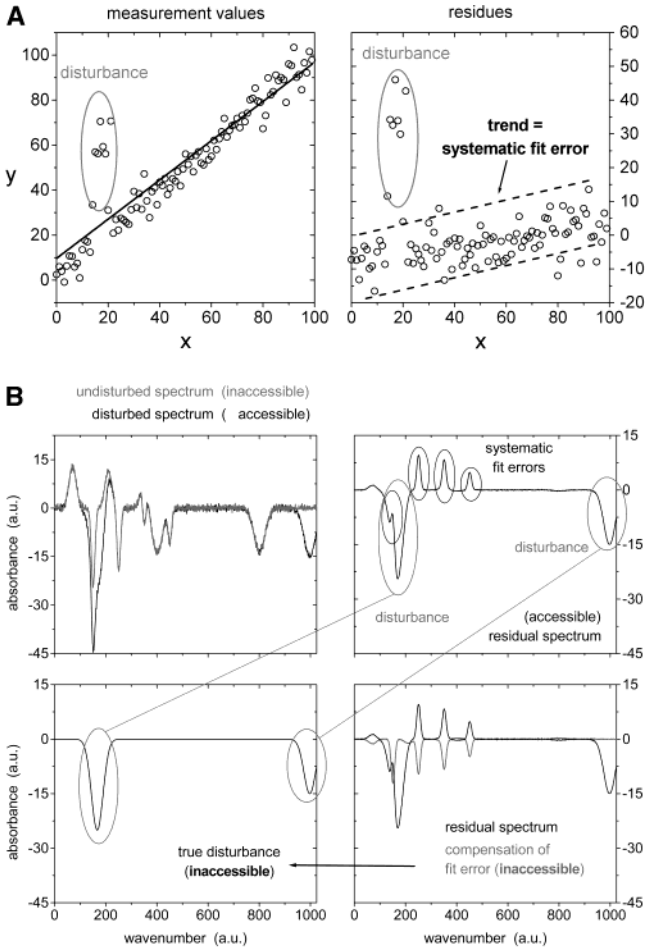

Figure 1. (A) (Left graph) simulated disturbed univariate data and linear fit function ($y = mx + b$); (right graph) results of a localized disturbance on the residues: large residual values (encircled) and a trend in the residues. (B) (Upper left graph) comparing a undisturbed with a disturbed simulated spectrum; (upper right graph) residual spectrum ϵ (4) consisting of the true disturbance and systematic fit error **e**; (lower right graph) residual spectrum and inaccessible compensation of the systematic fit errors—this compensation is estimated by the sPCR algorithm in order to determine the disturbance; (lower left graph) inaccessible true disturbance obtained from adding the compensation to the residual spectrum.

and extracted, this information can be used for tracking the origin of the disturbance. Since this qualitative extraction step does not involve concentration information on the investigated analytes, it can also be applied to a PCA only.

The central idea sPCR is based on is to make use of the residual spectrum ϵ (4), i.e., that part of the measurement spectrum which cannot be modeled by PCs (2). Conventional PCR does not exploit this additional information since it assumes that the residual spectrum consists of noise only. This assumption, however, is violated if uncalibrated absorbers are contained in the samples. The basic idea behind analyzing residues is introduced by utilizing a simulated univariate example (Figure 1A): Measured values that are inconsistent with the model (compare the encircled points in the left graph) are found in the residues (right graph) along with a trend in the residues due to systematic fit errors caused by the disturbed values—and noise of course. The same types of residuals are present in the multivariate case, i.e., unavoidable measurement noise, measurement values inconsistent with the model, and systematic fit error **e** introduced by the disturbed measurement values. The upper left graph of Figure 1B shows a mean-centered[8] simulated[6] spectrum with and without disturbance. The true disturbance, which is shown in lower left graph of Figure 1B, cannot be extracted and subtracted from the disturbed measurement spectrum since there is no viable way to split up a disturbed spectrum into an undisturbed part and the disturbance as this resembles a one-equation−two-unknowns problem. However, there is additional information available enabling an approximation of the disturbance, namely, the residual spectrum (4) (upper right graph of Figure 1B). Based on analyzing the residual spectrum, an estimation of the disturbance spectrum can be derived by the qualitative step of sPCR. In the undisturbed case, all relevant measured spectral features can be modeled and the residual spectrum consists of noise only. If uncalibrated features are present in a measurement spectrum that cannot be modeled by the PCs, the residual spectrum contains nonrandom disturbances (encircled in gray in the upper right graph of Figure 1B). Furthermore, they cause systematic fit errors, which are also part of the residual spectrum (encircled in black). The basic idea of extracting uncalibrated features is to analyze the spectral features of the residual spectrum whether they are nonrandom disturbances or systematic fit errors. If the residual spectrum is compensated for systematic fit errors, the true disturbance remains. This compensation of the residual spectrum for systematic fit errors is shown in the lower right graph of Figure 1B along with the residual spectrum. The nonrandom disturbances, i.e., the uncalibrated features, are determined by adding the compensation and the residual spectrum. However, this compensation for systematic fit errors cannot be determined directly since splitting up of the residual spectrum into true disturbance and systematic fit errors is not possible as this also resembles a one-equation−two-unknowns problem. However, an approximation of the compensation is feasible and was already developed.[6]

2. Realization of sPCR. After this demonstrative discussion on the ideas behind sPCR, details on the algorithm will be given in this paragraph. The approximation of the uncalibrated spectral features is based on analyzing the residual spectrum piecewise inside narrow spectral windows. For this purpose, the wavenumber axis is split up into a number of nonoverlapping narrow spectral windows whose widths have to be chosen empirically. As derived in ref 6, the window must contain at least R measurement points. Since a good wavenumber resolution of the detection algorithm is wanted, the window width should not exceed R much. For the experimental examples investigated in the study, one (compare section 1 in Results and Discussion), three (compare section 2 in Results and Discussion), and five (compare section 3 in Results and Discussion) relevant PCs, respectively, were determined. For all three examples, the window width was empirically selected to contain five measurement points.

Inside all of these spectral windows, the residual spectra (4) are analyzed individually whether they contain real spectral features or noise only. If real spectral features are found, it has to be decided whether they are due to uncalibrated spectral features or due to systematic fit errors. This decision is based on comparing this part of the residual spectrum with the parts of the PCs located inside the same wavenumber window (Figure 2): If the part of the residual spectrum inside this window is found

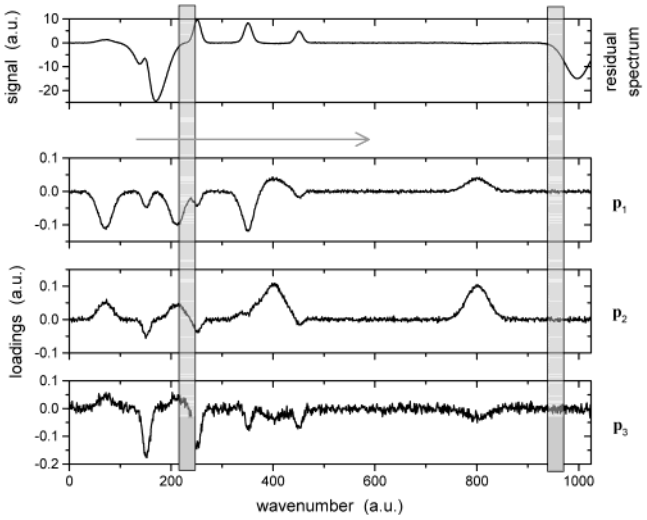

Figure 2. Scanning the residual spectrum ϵ (4) for systematic fit errors with a narrow window shown in gray. By means of the window at 250 wavenumber units, a systematic fit error **e** is found since the spectral feature of the residual spectrum is linear dependent from the PCs $\mathbf{p}_1 \cdots \mathbf{p}_3$ (2) inside this window—the window at 950 wavenumber units contains a uncalibrated disturbance since this part of the residual spectrum is linear independent from the PCs inside the considered wavenumber range.

to be linear dependent from the parts of the PCs inside the window, it represents a systematic fit error **e**, which has to be compensated (e.g., Figure 2, window at 250 arbitrary wavenumber units). If this part of the residual spectrum is linear independent from the corresponding PCs parts, it is considered an uncalibrated feature (e.g., Figure 2, window at 950 arbitrary wavenumber units). After determining the systematic fit errors in the residual spectrum, an appropriate compensation has to determined (Figure 1B, lower right graph, gray curve). Calculating such a compensation includes extracting the actual shape of the feature to be compensated. The shape of the feature to be compensated is already defined by the residual spectrum. The following discussion of calculating a compensation is supported by an example given in Figure 3. From the residual spectrum (gray) plotted in the top left graph of Figure 3, the parts shown in black were found to be linearly dependent from the PCs demanding for compensation. In general, there are three different kinds of compensations, which are indicated in the top right graph of Figure 3 by means of (*), (**), and (***), respectively: (*) indicates a feature that can be compensated in the easiest way. Since this systematic fit error is not overlapping with true disturbances, it must just be reflected at the wavenumber axis. (**) was erroneously extracted for compensation. Hence, its influence in the compensation should be kept as small as possible. The reader should bear in mind, however, that there is no way to find out that features are determined wrongly. (***) is an example that is difficult to compensate since this systematic fit error is overlaid by the true disturbance. Apparently, reflecting it at the wavenumber axis like the (*) example is not the correct method since this part is not supposed to be compensated to zero for it is partly a true disturbance. This spectroscopic "dent must be padded". As proposed in ref 6, an approximative way for compensating all three types of systematic fit errors in the same way is to use so-called reflection lines, which are shown as dash−dotted lines in the top right graph of Figure 3. The basic idea behind these reflection lines is to draw a straight line between the two undisturbed measurement points located immediately to the left and the right of an extracted systematic fit error, i.e., the so-called boundary points (indicated by gray arrows in the top right graph of Figure 3). This procedure is done separately for every single spectral feature needing compensation. The difference of reflection line minus considered part of the systematic fit error is an estimate of the compensation for a specific spectral feature (e.g., (*), (**), or (***)). After determining the compensations for all extracted systematic fit errors, a complete compensation is derived (bottom left graph of Figure 3). Now, the estimate of the disturbance **d** is calculated be adding the compensation of the systematic fit error **e** to the residual spectrum ϵ:

$$\mathbf{d} = \epsilon + \mathbf{e} \tag{5}$$

After the uncalibrated disturbance **d** (bottom right graph of Figure 3) is estimate, it is subtracted from the measured spectrum $\mathbf{x}_{\text{meas}}$ resulting in a corrected measurement spectrum

$$\mathbf{x}_{\text{corr}} = \mathbf{x}_{\text{meas}} - \mathbf{d} \tag{6}$$

which is then evaluated by a conventional PCR in order to determine the concentrations of the calibrated analytes.

3. **Assessing the Extracted Disturbance.** After a disturbance spectrum is extracted, it must be analyzed whether it contains only noise or relevant spectroscopic features. If characteristic spectroscopic features are found, it is necessary that sPCR provides an indication that corrections were made, which should be part of the analysis output together with corrected concentration results. The user should be notified that (1) relevant disturbances were found and (2) that the measured data have been modified. Hence, a figure of merit has to be defined by means of which it can be decided automatically whether pure noise was found or whether real uncalibrated spectral features are contained in a measurement spectrum. According to this, the decision must be based on spectral features only and it must take into account the noise level of a certain application. A decision threshold needs to be defined exceeding which flags a measurement spectrum as disturbed.

The residual spectra of the calibration spectra are one appropriate source of data for deriving such a decision threshold since they reflect the measurement noise and are free of uncalibrated absorbers by definition. It is assumed that the noise level remains stable over time and wavenumber. All calibration spectra $\mathbf{X}_{\text{cal}}$ (1) are affected by noise, which is transferred to the orthonormal PCs $\mathbf{p}_1 \cdots \mathbf{p}_K$ (2); however, the major part of the noise is contained in the $K - R$ "noise PCs" $[\mathbf{p}_{R+1} \cdots \mathbf{p}_K] = \mathbf{P}_{\text{noise}}$. The noise part $\mathbf{x}_{\text{noise}}$ of a calibration spectrum $\mathbf{x}_{\text{cal}}$ is determined by projecting a calibration spectrum onto the noise PCs:

$$\mathbf{P}_{\text{noise}}\mathbf{P}^{\text{T}}_{\text{noise}} \cdot \mathbf{x}_{\text{cal}} = \mathbf{x}_{\text{noise}} \tag{7}$$

Since the noise of the calibration and the measured spectra are usually uncorrelated, the noise would be greatly underestimated if measured spectra instead of calibration spectra would be used in eq 7. From $\mathbf{x}_{\text{noise}}$, a measure for the decision threshold is

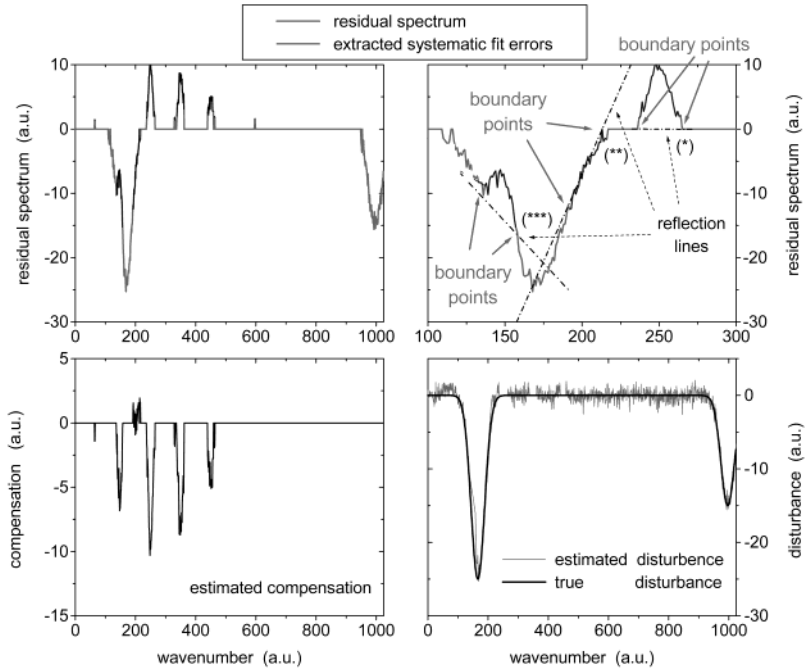

Figure 3. (Top left) residual spectrum and systematic fit errors **e** which need compensation; (top right) enlarged top graph with reflection lines necessary for calculating the compensation; (bottom left) estimated compensation which is added then to the residual spectrum resulting in an estimated disturbance spectrum (5); (bottom right) comparing the true, simulated disturbance and the estimated disturbance.

deduced be the standard deviation of the noise spectrum elements, i.e.: $\mathrm{sd} = (\mathrm{var}(\mathbf{x}_{\mathrm{noise}}))^{1/2}$. Based on this, the decision threshold is defined empirically as

$$\mathrm{th} = 3\mathrm{sd} = 3\sqrt{\mathrm{var}(\mathbf{x}_{\mathrm{noise}})} \tag{8}$$

Then for every single disturbance spectrum **d** (5) the value $\mathrm{sd}_{\mathbf{d}} = (\mathrm{var}(\mathbf{d}))^{1/2}$ is determined and compared to this threshold (8). If $\mathrm{sd}_{\mathbf{d}} \geq \mathrm{th}$, a relevant disturbance spectrum was found. If $\mathrm{sd}_{\mathbf{d}} < \mathrm{th}$, it is assumed that the extracted disturbance **d** (5) consists of noise only. The factor 3 introduced in (8) ensures that fluctuations in the noise do not immediately result in a disturbance warning.

To use all available information, (8) is not only derived from one calibration spectrum but from all calibration spectra. That is, eq 7 is successively applied to all calibration spectra $\mathbf{x}_1^{\mathrm{noise}} \cdots \mathbf{x}_K^{\mathrm{noise}}$. From these K noise spectra of N elements each, an extended noise vector

$$\mathbf{x}_{\mathrm{extended}}^{\mathrm{noise}} = (x_{1,1}^{\mathrm{noise}} \cdots x_{1,N}^{\mathrm{noise}}, x_{2,1}^{\mathrm{noise}} \cdots x_{2,N}^{\mathrm{noise}}, \cdots, x_{K,1}^{\mathrm{noise}} \cdots x_{K,N}^{\mathrm{noise}})^{\mathrm{T}}_{(K\cdot N\times 1)}$$

is derived. The threshold for uncalibrated spectral features used in this study is very similar to the one given in (8). It is defined as

$$\mathrm{th} = 3\sqrt{\mathrm{var}(\mathbf{x}_{\mathrm{extended}}^{\mathrm{noise}})} \tag{9}$$

Alternatively to the subtraction (6) of any extracted disturbance spectrum regardless of whether true disturbances were found, the subtraction can also be controlled by this assessment result. In this case, the extracted disturbance spectrum would only be subtracted from the measurement spectrum, if a relevant uncalibrated spectral feature unequal to random noise were found.

EXPERIMENTAL SECTION

In the present study, experimental spectra are analyzed that were obtained with different spectrometers by different research groups. Analyzing this variation of data sets impressively demonstrates the broad applicability of the proposed method for handling uncalibrated disturbances in experimental spectroscopic data sets. Details on the measurement techniques, sample preparation, measurement parameters, and results of measurement series can be found in refs 16−18. For this study, all three calibrations were repeated but left out one of the analytes at a time. These excluded analytes were then handled as uncalibrated absorbers. Since not all test samples contained the excluded analytes, disturbed and undisturbed samples were available. It is of importance to investigate both cases since sPCR is supposed to be equivalent to PCR in the undisturbed case and superior in handling disturbed samples. Although in the Results section only one analyte per data set will be assumed as uncalibrated disturbance, the algorithm is in theory able to handle a certain number of uncalibrated substances. This is due to its design analyzing spectroscopic features independently from their origin. It makes no difference for the algorithm whether one analyte introduces several unknown absorption bands or whether there are several uncalibrated analytes with one feature each.

The first data set was determined using mid-infrared (MIR) attenuated total reflection (ATR) spectroscopy[15] in the range of 2000−800 cm^{-1} for analyzing methanol and acetone concentrations of <10% volume in water.[16] The spectra of this data set contain 934 measurement points each. Data set 2 was gained by optical

(15) Harrick, N. J. *Internal Reflection Spectroscopy*; John Wiley & Sons: New York, 1979.

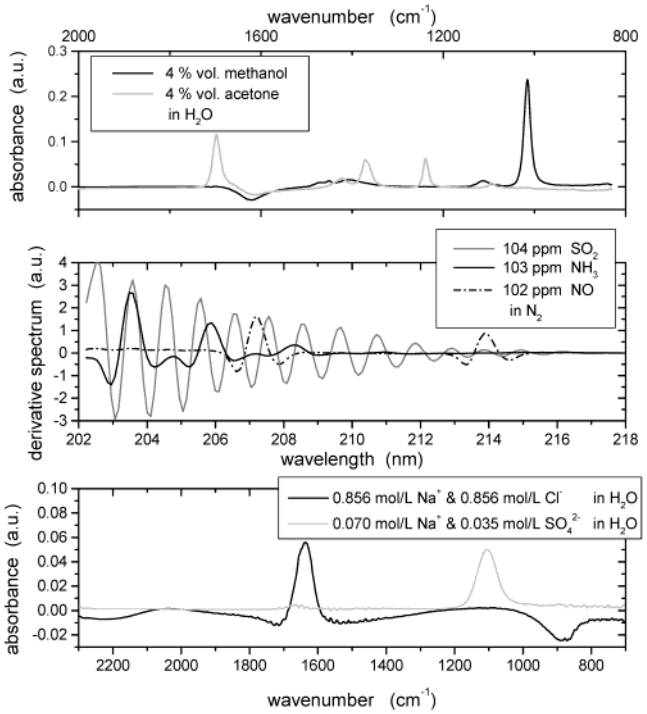

Figure 4. (Top) ATR absorbance spectra of methanol (calibrated) and acetone (uncalibrated)—data set 1;[16] (middle) second-derivative spectra of the considered analytes SO_2 and NH_3 (calibrated) as well as NO (uncalibrated)—data set 2;[17] (bottom) changes of the ATR spectrum of water introduced by dissolved NaCl (one of the calibrated analytes) and Na_2SO_4 (the sulfate ion acts as uncalibrated absorber)—data set 3.[18]

UV second-derivative spectroscopy analyzing gaseous mixtures of SO_2, NH_3, and NO diluted in N_2 at concentration levels of <150 ppm.[17] These spectra consist of 150 data points each covering the wavelength region 202−218 nm. The third spectra set with 827 measurement points at a time was gained from salinity investigations determining the concentrations (<1 mol·L^{-1}) of Na^+, Cl^-, K^+, Br^-, Ca^{2+}, Mg^{2+}, and SO_4^{2-} ions in water by means of MIR ATR spectroscopy.[18] For this purpose, the wavenumber region 2300−700 cm^{-1} was analyzed. In the last example, it was shown that although single atom ions have no bonds giving rise to IR absorption bands, the MIR spectrum of water is changed in dependence of the dissolved ionic species nature and their concentrations. These changes were used for concentration evaluation of the different ion species.

All results presented in the remainder of this study were determined by means of a self-developed chemometrics program using the Microsoft Visual C++ 6 compiler (Microsoft, Redmond, WA).

RESULTS AND DISCUSSION

1. Application of sPCR to MIR ATR Spectroscopy of Methanol Samples. ATR spectra of 4% (v) methanol and 4% (v) acetone dissolved in water are shown in the top graph of Figure 4. Methanol was considered as the target analyte; acetone was selected to serve as uncalibrated absorber. For the calibration, five ATR spectra were used containing 1, 2, 4, 8, and 10% (v) methanol dissolved in water. For experimental details, please refer to ref 16. From these five methanol spectra, one relevant principal component was extracted. For assessment of the sPCR algorithm, 53 samples disturbed by 1−10% (v/v) acetone were used as well as six undisturbed samples containing methanol only. Three examples are elaborated in Figure 5: In the upper left graph, an undisturbed spectrum of 6% (v) methanol is plotted together with the found disturbance. As expected for this undisturbed case, the extracted disturbance is negligible. If methanol and the interfering acetone are present at considerable concentrations, i.e., 6% (v) each (upper right graph of Figure 5), the major features of acetone are found and subtracted from the disturbed measurement spectrum. Comparing the undisturbed and the disturbed 6% (v) methanol spectra after this correction reveals that the methanol features are equal. However, there is a small difference found around 1635 cm^{-1}. This difference is due to different water concentrations in the two samples resulting in different magnitude of the water absorptions (O−H bending vibration). The disturbed sample contains a total of 12% (v) analytes compared to 6% (v) present in the undisturbed case. Water was included implicitly into the calibration using aqueous methanol samples and is hence calibrated into the PC model. There is another water absorbance below 900 cm^{-1} (librational or L_2 band[19]) that is distorted due to the proximity to the cutoff frequency of the used mercury−cadmium−telluride detector at 670 cm^{-1} causing increased noise levels. In a third example (lower right graph of Figure 5), no methanol at all is present but only the interfering (uncalibrated) acetone. Even in this case, the uncalibrated acetone features are extracted from the measurement spectrum resulting in an almost flat methanol spectrum, which is in accordance with the true methanol concentration of 0% (v). All 59 test samples were correctly identified by sPCR based on the classification threshold (9) to be disturbed and undisturbed, respectively, and plotted in the graphs of Figure 6 (note the different *y*-axis scaling). The disturbance spectra found for the undisturbed samples (upper graph of Figure 6) only show some noisy features in the wavenumber regions of strong water absorbance (1635 and >900 cm^{-1}), around the strongest acetone absorbance (1700 cm^{-1}), and at the methanol band (1050 cm^{-1}). However, if the *y*-axis scaling of this graph is compared to one of the disturbed cases (lower graph of Figure 6) or to the single-analyte spectra (Figure 4, top graph), it can be concluded that these remaining features are negligible. The disturbances extracted from truly disturbed samples always show the acetone features. The strength of the disturbance spectrum is dependent on the acetone concentration. Hence, it can be stated that for no sample in this data set a nonexistent disturbance spectrum was wrongly generated by sPCR and that no disturbance was overseen. Since the methanol and acetone spectra are not substantially overlapping in the considered wavenumber range of 2000−800 cm^{-1}, the mean absolute concentration errors of this series are found to be 0.4% (v) for both, the conventional PCR and the sPCR. Nonetheless, the conventional PCR is not able to detect and extract the spectral features of the uncalibrated acetone.

(16) Vogt, F.; Karlowatz, M.; Jakusch, M.; Mizaikoff, B. *Analyst* **2003**, *128*, 397−403.
(17) Vogt, F.; Klocke, U.; Rebstock, K.; Schmidtke, G.; Wander, V.; Tacke, M. *Appl. Spectrosc.* **1999**, *53*, 1352−1360.
(18) Vogt, F.; Kraft, M.; Mizaikoff, B. *Appl. Spectrosc.* **2002**, *56*, 1376−1380.
(19) Zelsmann H. R. *J. Mol. Struct.* **1995**, *350*, 95−114.

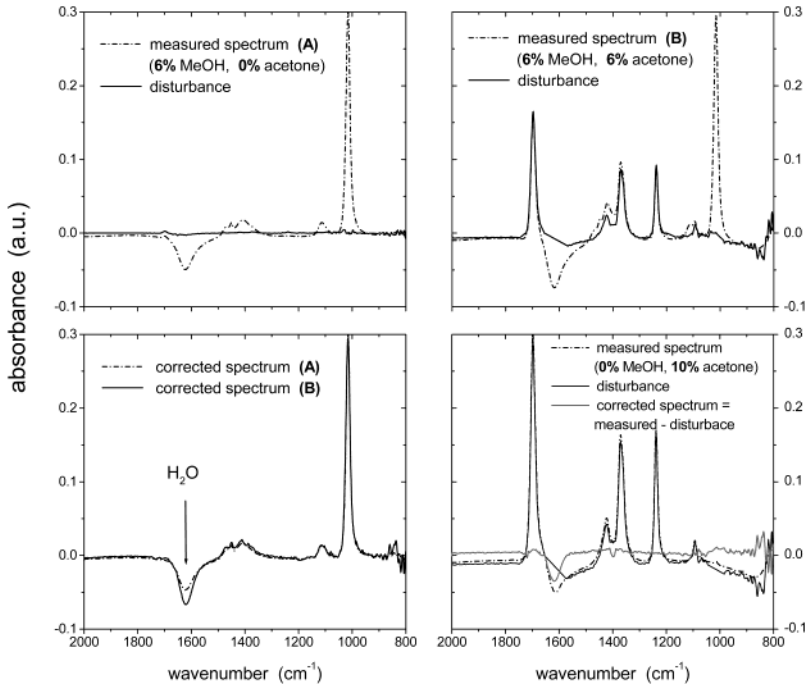

Figure 5. (Upper left) undisturbed methanol spectrum and determined, negligible disturbance (5); (upper right) methanol spectrum disturbed by acetone and found disturbance; (lower left) comparing the corrected undisturbed methanol spectrum; (upper left graph) with the corrected (6) disturbed one (upper right graph); they are equivalent aside from a different water absorbance at 1635 cm^{-1}; (lower right) measurement spectrum of acetone only, found disturbance, and corrected methanol spectrum (6).

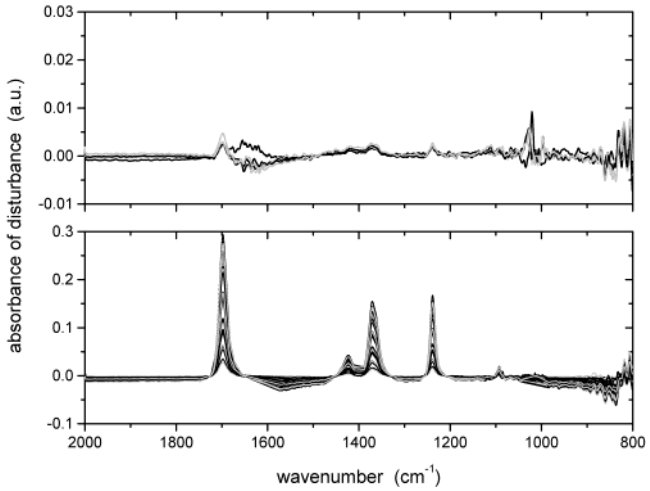

Figure 6. (Upper graph) disturbances (5) found in undisturbed spectra of data set 1 (lower graph) and in spectra disturbed by varying acetone concentrations (compare Figure 4, top graph). The *y*-axis of the upper graph is stretched 10 times compared to the lower graph.

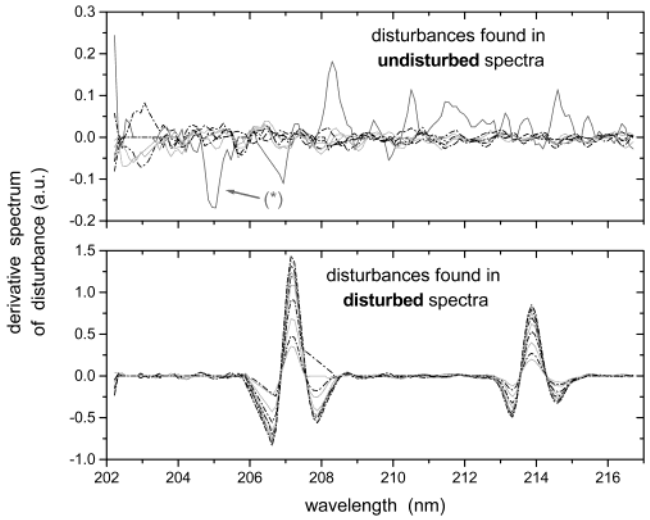

Figure 7. (Upper graph) disturbances (5) found in undisturbed second-derivative spectra of data set 2 (lower graph) and in second-derivative spectra disturbed by NO (compare the graph in the middle of Figure 4). The *y*-axis of the upper graph is stretched 5 times compared to the lower graph.

2. Application of sPCR to UV Second-Derivative Spectra. Second-derivative UV spectra of the three considered gases SO_2, NH_3, and NO are shown in the graph in the middle of Figure 4. Experimental details can be found in ref 9. These analytes are strong absorbers in the analyzed wavelength range of 202−218 nm; especially SO_2 and NO as well as SO_2 and NH_3 show strongly overlapping spectral features. Seven samples containing SO_2 and NH_3 diluted in N_2 at different concentrations were used as calibration set; NO was excluded from the calibration and serves as uncalibrated disturbance. On account of the high data quality, three PCs were found to be sufficient to model mixture spectra of SO_2 and NH_3. For assessment of the sPCR algorithm, 10 mixtures of SO_2 and NH_3 disturbed by 10−102 ppm NO were used as well as 8 undisturbed samples containing the calibrated analytes only. Seven of the eight undisturbed samples were determined (9) to be free of uncalibrated absorbers. One undisturbed sample (cp. gray (*) in the upper graph of Figure 7), however, was considered to be error affected based on the threshold (9). After visual inspection of this derivative spectrum, a small wavelength drift was found compared to the calibration spectra. On account of the comblike structure of the SO_2 derivative spectrum, non-random disturbances were introduced by this shift. Hence, the algorithm is also capable of dealing with such spectrometer-related disturbances. None of the disturbance spectra found in undis-

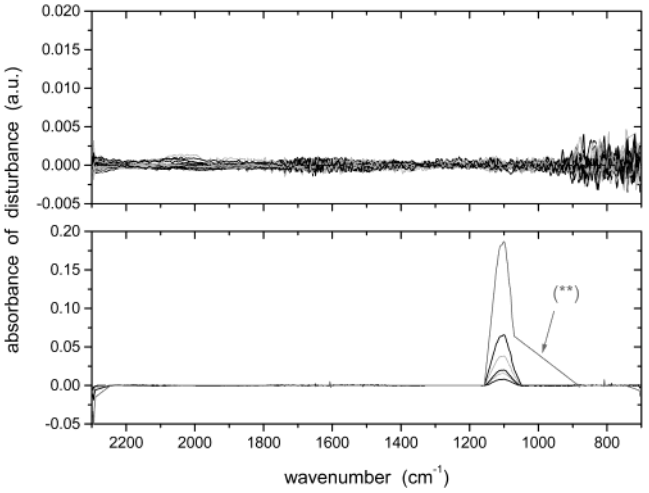

Figure 8. (Upper graph) disturbances (5) found in undisturbed samples of data set 3 (lower graph) Disturbances found for the samples disturbed by sulfate ions, The *y*-axis in the upper graph is stretched by a factor of 10 compared to the lower graph.

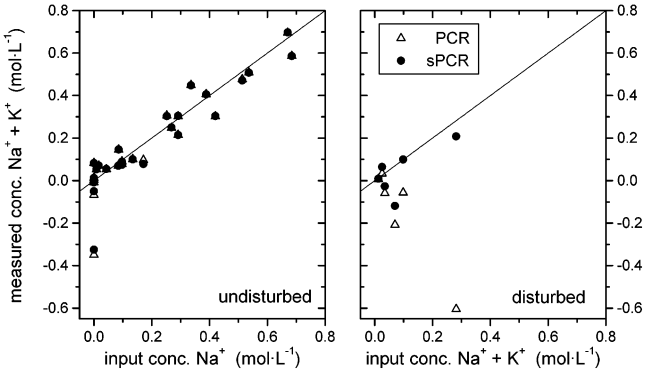

Figure 9. Results of $Na^+ + K^+$ ion concentrations[18] obtained by conventional PCR and novel sPCR. (Left graph) most results of undisturbed samples are equivalent for both algorithms. (Right graph) sPCR could achieve improved results for the six samples disturbed by dissolved Na_2SO_4.

turbed samples show spectroscopic features; just noise was extracted. All 10 disturbed samples were found as disturbed by sPCR. No disturbance was overseen since there is no featureless flat disturbance spectrum in the lower graph of Figure 7. From comparing these disturbance spectra with the pure NO derivative spectrum plotted in the graph in the middle of Figure 4, it can be concluded that sPCR determines the uncalibrated NO features very reliably, even if there are strongly overlapping calibrated structures. The strength of these interfering features is also found to be related to the NO concentration. From the 18 test samples, the mean absolute errors of the SO_2 and NH_3 concentrations could be determined since the input concentrations of all analytes were known. Applying PCR resulted in a mean SO_2 error of 1.8 ppm; an error of 0.9 ppm was obtained using sPCR—the mean error could be reduced by 50%. Since the spectral features of NO are not overlapping too much with NH_3, PCR and sPCR resulted in equivalent mean errors of 3 ppm.

3. Application of sPCR to Salinity Analysis. In this application, the change of the water absorbance due to the presence of ion species and their concentration was evaluated. Details on the experiments are provided in ref 18. Since the influence of Na^+ and K^+ on the water spectrum could not be discriminated at the considered concentration level of <1 mol·L^{-1}, both ionic species were evaluated as a sum parameter $Na^+ + K^+$.[18] The calibration for this task was based on 10 aqueous solutions of Na^+, Cl^-, K^+, Br^-, Ca^{2+}, and Mg^{2+} ions. Five PCs were selected for the calibration model. The test set consisted of 33 spectra of which 27 contained the calibrated ion species only at varying concentrations. The remaining six spectra were disturbed by dissolved Na_2SO_4, i.e., Na^+ and SO_4^{2-} ions, as uncalibrated features. By means of the classification threshold (9) derived in section 3 in sPCR Algorithm, all 33 spectra were correctly classified as disturbed and undisturbed, respectively. The extracted disturbance spectra are shown in Figure 8 split up in two graphs considering undisturbed and disturbed samples, respectively. As an example for the equivalence of PCR and sPCR in the undisturbed case and the superiority of sPCR for disturbed samples, the concentration results for $Na^+ + K^+$ are given in Figure 9. In the left graph, the measured concentrations are plotted versus the input concentrations in the undisturbed case determined by conventional PCR and the novel sPCR. In three cases, minute differences of the PCR and sPCR results were obtained; however, for the majority of the samples, equal concentrations were determined. In the right graph of Figure 9, the results of the disturbed samples are depicted. The superiority of the improved sPCR algorithm over the conventional PCR in the case of disturbances is obvious as considerable improvements of the concentration results could be achieved.

4. Assessment of the Benefit and Outlook. In this study, the theoretical concept of secured principal component regression[6] was applied to experimental spectroscopic data sets for the first time with the aim of demonstrating the feasibility and reliability of this method to find and correct uncalibrated spectral features. To demonstrate broad applicability, different experimental data sets were used, which were obtained by different working groups with different equipment. The broad practicability of sPCR in spectroscopy was revealed by successful application of the algorithm to gaseous- and liquid-phase samples, which were analyzed in the UV and the mid-infrared wavelength region, respectively. It was shown that for different experimental data sets disturbances can be extracted and corrected with high reliability. Despite the positive results presented in this experimental study, one has to keep in mind that the disturbance spectrum is estimated and not analytically calculated. As was shown in previous works of the authors,[6] in particular, disturbances that substantially overlap with calibrated features of similar broadness are difficult to estimate. This may lead to imprecisely estimated disturbances and subsequently to an inexactly corrected measurement spectrum (see, for example, the wavelength regions marked with (**) in Figures 5, 7, and 8). This effect was also observed for synthetic spectra applied in ref 6 during investigation of the algorithms' performance. However, as was demonstrated previously, an overall major improvement of the corrected concentrations compared to the conventional approach could be achieved nonetheless. On account of the algorithm design, only uncalibrated spectral features can be found that are different from calibrated features. If a disturbance has the same shape and is located at the same wavelength region as calibrated features, it would be considered as calibrated.

Due to its broad applicability, we currently evaluate whether sPCR algorithms can also be applied to automated peak deconvolution for chromatographic measurement techniques for qualitative peak characterization. In this case, the wavelength axis would be replaced by a time axis. While sPCR would not affect concentration evaluation, the detection of unexpected chromatographic peaks may facilitate and accelerate interpretation of complex chromatograms. PCA calibration would include chromatograms of all known analytes resulting in a set of principal components. A new chromatogram would then by analyzed by sPCR for uncalibrated peaks.

Certainly the most versatile field of application is sensor technology. Specifically, spectroscopic sensors probing an entire wavelength region thereby providing the capacity of multicomponent analysis will benefit from advanced data evaluation strategies,[20] chemical sensors,[21] and in particular spectroscopic chemical sensors based on interaction of analyte molecules with a molecular recognition interface covering the transducer surface require considerate compensation for drifts due to, for example, swelling of the sensing membrane and occurring uncalibrated spectral features. Development of appropriate automated compensation and data evaluation techniques based on advanced multivariate calibration models will be among the key technologies facilitating widespread implementation of chemical sensor technology in continuous environmental or process monitoring applications.[16] The examples in this study based on mid-infrared evanescent field spectroscopy demonstrate that multicomponent analysis with a single sensor system is substantially assisted by the developed sPCR algorithm. Recently, the application of chemical IR sensors has been extended to sustain harsh environmental conditions and even a deep sea environment.[23−25] Remote measurement conditions particularly demand for automated data evaluation and robust calibration procedures, as frequent system recalibration is not feasible. It is expected that application of this sensor concept in combination with multivariate data evaluation based on the developed sPCR algorithm will greatly facilitate the deployment of remotely operated spectroscopic sensing systems in the field of environmental and process analysis.

CONCLUSION

In general, chemometric algorithms require a set of calibration samples that contains all spectroscopic information being analyzed during the measurement process. Hence, uncalibrated absorbers contained in the analyzed samples seriously affect derived concentrations. This circumstance is especially detrimental for process monitoring and remote sensing applications leading to errors in process control based on continuous concentration readings. Furthermore, unexpected analytes are an indication for problems in the process flow. The present study introduces secured principal component regression as a novel chemometric data evaluation method preventing ambiguous concentration prediction. The sPCR algorithm was developed in order to detect and correct uncalibrated spectral characteristics serving two main purposes: (A) correction for concentration errors and (B) extraction of the spectrum of the interfering analyte for further analysis of the origin of the disturbance.

By means of three different spectroscopic data sets measured at gaseous and liquid samples in the ultraviolet and the mid-infrared wavelength region, the general applicability of this algorithm was demonstrated. Furthermore, it is shown that sPCR is equivalent to conventional PCR in the undisturbed case and superior to PCR in case of disturbances. The spectra of uncalibrated analytes could be reliably extracted even in cases of strong spectral overlap with calibrated substances. It is expected that this algorithm will find widespread application for automated multivariate data evaluation in the field of process analysis, environmental monitoring, and remote sensing.

ACKNOWLEDGMENT

This work was funded in part by a research grant (Vo 895/1-1) received from the Deutsche Forschungsgemeinschaft (German Research Foundation) and by the U.S. Department of Energy (DE-FC26-00NT40920). The authors also acknowledge helpful discussions with Maurus Tacke, FGAN-Research Institute of Optronics and Pattern Recognition, Ettlingen/Germany.

Received for review December 11, 2003. Accepted February 25, 2003.

AC020758W

(20) Mizaikoff, B.; Lendl, B. In *Handbook of Vibrational Spectroscopy*; Chalmers, J. M., Griffiths, P. R., Eds.; John Wiley & Sons: New York, 2002; Vol. 2, pp 1560−1573.

(21) Janata, J.; Josowicz, M.; Vanýsek, P.; DeVaney, D. M. *Anal. Chem.* **1998**, *70*, 179R−208R.

(22) Holst, G.; Mizaikoff, B. In *Handbook of Fiber Optic Sensing Technology: Principles and Application*; Lopez-Higuerra, J. M., Ed.; John Wiley & Sons: New York, 2002; pp 729−749.

(23) Mizaikoff, B. *Meas. Sci. Technol.* **1999**, *10*, 1185−1194.

(24) Kraft, M.; Karlowatz, M.; Mizaikoff, B.; Stück, R.; Steden, M.; Ulex, M.; Amann, H. *Meas. Sci. Technol.* **2002**, *13*, 1294−1303.

(25) Kraft, M.; Jakusch, M.; Karlowatz, M.; Katzir, A.; Mizaikoff, B. *Appl. Spectrosc.*, in press.