# A new approach for modelling sensor based data

Thomas Skov* & Rasmus Bro

Department of Food Science (IFV), The Royal Veterinary and Agricultural University (KVL)
Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
*E-mail: thsk@kvl.dk

## Abstract

Data from sensor based analytical instruments (e.g. electronic nose) can be arranged in three dimensions as Sample × Time × Sensor. Due partly to limitations in software, partly to general practice, the dimensions of the data are often reduced in the time mode and traditional two-way chemometric models (e.g. principal component analysis – PCA) are used for the exploration of the data. Thus the internal relationship between the different modes (samples, time and sensors) is destroyed and the result can be an incomplete data analysis.

A new approach to handle the time information is introduced combining new and more advanced multi-way chemometric models with traditional pre-processing techniques as an alternative to PCA-like methods. As the pre-processing is an essential but often time-consuming part of the data analysis, a semi-automated approach has been used to make comparison of multiple analyses simple.

The main feature of this new approach is the exploration of the total time profiles such that potentially relevant information is not discarded by feature extraction before the actual data analysis. Multi-way modelling is evaluated and for sensor based data, it is shown that the so called PARAFAC2 multi-way model that handles shifted profiles offers some advantages compared to PCA and alternative multi-way methods.

The principles are exemplified with an example from the licorice industry. An algorithm and graphical user interface has been made available at www.models.kvl.dk to ease testing this new approach.

**Keywords**: Electronic nose, data analysis, feature extraction, PARAFAC2

# 1. Introduction

A number of instrumental analytical techniques are used in food research as well as in laboratories of the food-associated industry (e.g. spectroscopy, gas chromatography, electronic noses/tongues). The raw data from these analyses are rarely used directly. Rather the data are manipulated using for example baseline correction, peak identification and feature extraction (slope, area etc.). These operations take time and offer no guarantee that the data is properly pre-processed and that the available and useful information is extracted. Many of the analytical techniques also require a non-trivial and often expert based post-processing of the data, which in many situations takes unnecessarily long time and uses an excess of manpower (e.g. identification and validation of peaks in gas chromatography). Often these techniques are combined with traditional univariate exploration of the data with extensive focus on a small (local) part of the total variables.

The electronic nose is a more or less automated analytical method that is based on headspace generation (volatile compounds) from a sample. The samples can often be analysed with little or no prior sample preparation, which makes the electronic nose very suitable for at-line analyses. The commercial electronic noses are often based on sensors with different selectivity and sensitivity – a hybrid sensor array system, which gives a very powerful analytical instrument [1]. The low selectivity of the sensors makes the electronic nose more appropriate for classification and separation tasks, than more specific analyses on the volatile compounds in the sample. This is especially so for complex food analyses where flavour plays an important part; e.g. coffee powders [2], olive oils [3], apple juices [4] and cheddar cheese [5].

The pre-processing step and pattern recognition system (e.g. principal component analysis – PCA) is often an integrated part of the electronic nose and thus a fast (initial) evaluation of data is possible. However, the choice of pre-processing steps and pattern recognition system is in general very restricted. The raw data from electronic noses is often arranged in three dimensions (e.g. *Samples* × *Times* × *Sensors*) (Figure 1) and thus can not be directly analysed with the two-way based integrated software. As a result of these limitations important information from the raw data may well be missed or changed leading to a suboptimal final interpretation. In this paper it is shown that it is not necessary to transform the informative measured time profile into univariate features before meaningful data analysis can be performed. By maintaining the time information, it is anticipated that at least

in some situations more quantitative and qualitative information can be extracted from the data.

The aim of this article is to show the steps in the data analysis of sensor based data and provide new models to such data. The steps are divided into a *pre-processing* function combined with a mathematical *processing* method e.g. multivariate data analysis (chemometrics) and a *post-processing* step.

An illustration of the efficiency of selecting the appropriate combination of pre-processing steps and a suitable chemometric model will be given using an example from a concrete situation.

# 2. Theory

## *2.1. Data analysis*

Data-analysis is usually done on pre-processed data. The rationale for pre-processing is to get the most efficient representation of the data, especially efficient with respect to the subsequent modelling of the data. The combination of pre-processing steps is evaluated through appropriate multivariate models, which again are often evaluated by means of graphical illustrations, quantitative classification or regression models. Thus these three steps strongly interact in finding the appropriate analysis approach for a given data set.

The interchanging of pre-processing steps is not straightforward as one option may influence the other – e.g. if the raw data needs to be scaled to give a reasonable solution this does not necessarily imply the same need for data transformed by taking first derivatives of the time profile.

## *2.2. Pre-processing*

This section describes some of the most common pre-processing techniques that are used for data from sensor based analytical instruments.

### Baseline correction

Several baseline correction methods can be used to pre-processes the typical sensor response curve [6-9]. A frequently used baseline correction is the difference between the maximum signal (or minimum) and the baseline signal, $\Delta S$ or $S\text{-}S_0$. The following baseline correction methods for MOS sensors (Table 1) have been suggested [9,10]. In Figure 2 an example of baseline correction using the fractional approach - $(S\text{-}S_0) / S_0$ is shown together with the corresponding raw data.

### Transformation

The structure of sensor based data often makes a “transformation” step feasible. As examples of transformations, for sensor based data, the use of logarithmic transformations has been proposed and also the use of first or second derivatives may result in an easier interpretation of the data by removing trivial but disturbing baseline offsets or slopes [7,8].

### Feature extraction

When analysing the sensor response curve there are several ways to extract the information that is located in different parts of the response curve (e.g. adsorption, maximum/minimum or desorption parameters) [7]. A simple technique to reduce the number of features is to select a subset of the available features (e.g. 10 of the 240 measurements of the sensor signal). The discarded features should then either contain no valuable (or little) information in the specific application or have a large correlation to some of the chosen features [8].

Several extraction features (individual/combinations) should be investigated to estimate their discrimination power. In general though, the quality of a feature extraction depends highly on both the application and the instrument.

**Centering and scaling**

Centering aims at removing constant terms in the data in order to make the data compatible with the model. Centering across the first mode (e.g. the sample mode) of a three-way array can be done by matricizing the array to an $I \times JK$ matrix and then center as in ordinary two-way analysis, where the column mean is subtracted from every column element [11]:

$$x_{ijk}^{Centered} = x_{ijk} - \frac{\sum_{i=1}^{I} x_{ijk}}{I}$$

Scaling provides that each variable is scaled to an equal variation and thus, each variable will have the same opportunity to affect the model [11]. Scaling within the third mode (e.g. the sensor mode) is done by scaling each of the sensor variables to a unit mean square [11]:

$$x_{ijk}^{Scaled} = \frac{x_{ijk}}{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{x_{ijk}^2}{IJ}}}$$

*2.3. Processing/modelling step*

2.3.1. Notation

The notation and terminology to describe matrices (2-way arrays) and higher order arrays is adapted from [12] and [13]. Scalars are indicated with lower-case italics (e.g. $x_{ijk}$) and vectors with bold lower-case characters (e.g. **y**). In chemometrics, the data is often described with the

symbol X. Ordinary two-way arrays (matrices) are denoted **X** (bold-face) whereas higher order arrays are denoted $\underline{\mathbf{X}}$. The *ijk*th element of a three-way array $\underline{\mathbf{X}}$ is denoted $x_{ijk}$ where the indices run as follows: $i = 1\ldots\ldots,I$; $j = 1,\ldots..,J$; $k = 1,\ldots..,K$.

The three-way array will be denoted $\underline{\mathbf{X}}$ ($I \times J \times K$) where $I$ is the number of samples, $J$ the number of time measurements and $K$ the number of sensors. Thus the *ijk*th element in $\underline{\mathbf{X}}$ corresponds to the measured sensor signal from sample $i$, measured at time $j$, with sensor $k$.

2.3.2. PCA and PARAFAC models

PARAFAC1 (often called PARAFAC) was developed in 1970 [14] in psychometrics, but has found widespread use in chemometrics. To explain the PARAFAC1 model it is feasible to start out explaining some of the basic properties of the PCA model. PCA is a *bilinear* model. Thus, in PCA each component is the outer product of two vectors (scores pertaining to samples and loadings pertaining to variables). The model is linear in the scores for fixed loadings and vice versa and therefore bilinear. If a given data set can be modelled by an $R$-component PCA model, then the $I \times J$ matrix **X** can be approximated as $\mathbf{TP}^{\mathrm{T}}$ where **T** is the score matrix ($I \times R$) and **P** the loading matrix ($J \times R$). One particular view of PCA will be helpful in the following. If the rows of **X** are different samples and the columns are holding the measurements at different ($J$) times then the PCA model specifically states that *any* measured time profile – row of **X** – can be modelled as a linear combination of a few ($R$) loadings. Thus PCA tries to describe the different time profiles as different linear combinations (defined by the scores) of a few underlying latent variables (the loadings). This is an essential and important property of PCA in this context. It is illustrated in Figure 3A where it can be seen that when the same phenomenon occurs in different proportions in different samples, this is easily handled by a one-component PCA model.

Sometimes though, the same phenomenon does not only change in magnitude but also in shape. E.g. the maximum of the peak shifts position (Figure 3B). This is a dramatically different situation because the variables (times in this case) essentially change meaning from sample to sample. PCA and most other multivariate methods can not handle such shifts in an efficient manner. Every type of shift must basically be modelled by a separate component which is not a meaningful and parsimonious way to model the data. The PARAFAC2 model to be discussed will be shown to be able to handle this type of variations.

If data is arranged in a three-way array (e.g. Samples × Times × Sensors) it is necessary to matricize (i.e. unfold) the data in order to fit a PCA model. This can be done as shown in Figure 4. A PCA model of such unfolded data is often referred to as unfold PCA. However, the unfolding may obscure the natural relationship between variables of different dimensions. The use of multi-way chemometric methods (e.g. PARAFAC1 and PARAFAC2) is often more appropriate and offers several advantages compared to PCA.

PARAFAC1 is similar to PCA. Instead of a bilinear model handling matrices, it is a trilinear model handling three-way arrays (or even higher-way arrays). This is illustrated in Figure 5. As PCA, a PARAFAC1 model provides scores for the sample mode and loadings for the time mode, but also an additional set of loadings for the sensor mode. Such a model has interesting properties in itself [12,14] but as for PCA, the PARAFAC1 model assumes that the underlying basic variations (latent variables) remain constant in shape across different samples. If this holds and the underlying phenomena are additive in nature, then the PARAFAC1 model directly and uniquely estimates the underlying phenomena [15]. This strict (tri-)linearity of PARAFAC1 is very advantageous when applicable [12] but the model will often be too restricted to deal with severely shifted time profiles e.g. as also seen in batch production data [16].

PARAFAC2 is a development of the original PARAFAC1 model which aims at handling shifted or more generally varying profiles in a more efficient manner than PARAFAC1 (and PCA) can. In Figure 6 the graphical description of the PARAFAC2 model is shown. In this figure, scores (A) and sensor loadings (C) are of a similar structure as in PARAFAC1. However, in the time mode where it is assumed that there are shifts, there is not only one loading matrix as there would be in a PARAFAC1 model. Rather, for each sensor, $k$, an individual time loading matrix is given. Hence, the PARAFAC2 model allows that every sensor (or alternatively sample) can have its own distinct set of time loadings. This is clearly different from the assumptions e.g. in the PCA model where all samples are assumed to be governed by the same set of time loadings.

The matrix formulation of the PARAFAC1 model [17] and PARAFAC2 model [18-20] are shown below. The multi-way models are somewhat difficult to grasp from linear algebraic expressions because linear algebra is intrinsically related to two-way matrices. However,

these expressions are important for specifying the exact nature of the PARAFAC2 model. The models are mostly fitted in a least squares sense.

PARAFAC1: $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T + \mathbf{E}_k$

PARAFAC2: $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T + \mathbf{E}_k$

where $\mathbf{X}_k$ is the *k*th frontal slab of the three-way array and $\mathbf{D}_k$ is a diagonal matrix holding the *k*th row of **C** in its diagonal. **A**, **B**/$\mathbf{B}_k$**,** and **C** are parameters to be estimated and $\mathbf{E}_k$ residuals. The major difference between the two models is that PARAFAC2 allows the loading matrix for mode B to be different for the *k* sensors ($\mathbf{B}_k$).

The PARAFAC1-solution cannot be rotated without a loss of fit and hence only one best-fit solution (unique) is possible. For the PARAFAC2 model to be valid and to retain uniqueness, all the cross-products matrices of the $\mathbf{B}_k$ matrices are constrained to be constant for all *k*. This can be formulated as [19]:

$$\mathbf{B}_k^T\mathbf{B}_k = \mathbf{B}^T\mathbf{B}, \qquad k = 1,\ldots\ldots K \tag{3}$$

This means that for every sensor *k*, a set of profiles $\mathbf{B}_k$ (e.g. elution profiles [18]) is estimated under the constraint that the cross-products of the profile matrices are identical.

### *2.4. Post-processing/graphical presentation*

The appropriate selection of pre-processing steps combined with the correct chemometric model provides a solution that is ready to be further analysed using post-processing techniques. The most obvious post-processing step is the graphical illustration of the model parameters, which will be described in the following section.

Data from sensor based analytical instruments can be arranged as *Sample/Sensor* × *Time* × *Sensor/Sample.* The results of a model fitted to such data directly can be presented in terms of parameters (scores or loadings) from a certain mode. The sample and sensor modes are illustrated with score and loading plots (scatter plots), respectively, whereas the time profiles are best shown as sensor signal vs. time. In Table 2 the number of elements in the plots from the different chemometric models are compared.

It can be seen that the use of the multi-way chemometric models gives a simpler loading plot of the sensor mode as compared to a corresponding PCA model of the data rearranged into a matrix. The plots of mode B differ for the different multi-way models and the special features of the models are illustrated in this mode. PARAFAC2 that allows a certain freedom in one mode (e.g. mode B) contains more loading elements, but the simple structure of the scores A and loadings C has been preserved.

Another way to evaluate the models parameters is to use different classification technique (e.g. K Nearest Neighbours (KNN) or Soft Independent Modelling of Class Analogy (SIMCA) [21]. In this paper both visual interpretation of scatter plots and results from a KNN classifier will be used to validate the chemometric models.

# 3. Experimental

## *3.1. Data origin*

In the food industry great emphasis has been put on the standardization and control of the food production. By minimizing the variations similar products can be produced even under varying conditions. Deviations will occasionally occur and it is then important to be able to recognize this and to take the proper actions. In the spring 2002, Pingvin Lakrids A/S initiated an investigation to find a fast, easy, and more or less automated instrumental analytical technique that could help them differentiate between normal licorices and deviating licorices. For this purpose an electronic nose showed promising results.

## *3.2. Instrumentation*

Analyses were conducted using an α-FOX-3000 electronic nose manufactured by Alpha M.O.S (Multi Organoleptic System) (Alpha M.O.S, Toulouse, France, http://www.alpha-mos.com/) with a sensor array system consisting of twelve Metal Oxide Semiconductive sensors (MOS) distributed in two chambers. Chamber one: SY/LG, SY/G, SY/AA, SY/Gh, SY/gCTI, and SY/gCT; chamber two: T30/1, P10/1, P10/2, P40/1, T70/2, and PA2. The signal of sensor one (SY/LG) was very noisy and perturbed and pre-analyses showed that this sensor disturbed the interpretation of the results. The data from sensor one (SY/LG) was therefore not included in the data set.

The presence of two additional sensors in each chamber permitted control of the internal temperature and the relative humidity. The temperatures in chamber one and two were 95°C and 65°C, respectively, and the humidity was adjusted to 0 % RH in both chambers.

## *3.3. Samples*

A commercial licorice product manufactured by Pingvin Lakrids A/S, Avedøre, Denmark was used for the analysis. From this three different licorice characteristics were evaluated: 1) Good licorices (GOOD), 2) Bad licorices (deviating licorices) (BAD) and 3) Fabricated bad licorices (FBAD). The latter samples were prepared by drying the good licorices for a longer time and were included because of an observed burned taste in the bad licorices and this was tried resembled by a further drying of the good licorices.

Equilibrium was established between the headspace and the licorice sample and from the headspace 1 ml gaseous sample was transferred to the sensors. An acquisition time of 120 sec

measuring the sensor signal (S) every ½ sec was applied for each measurement. Between the measurements the baseline of the sensors were re-generated with carrier gas and the baseline signal ($S_0$) measured. The experiment was completely randomised and each licorice characteristic was evaluated with six replicates.

Initial data analysis using unfold PCA and visual inspection of raw sensor signals revealed that sample FBAD-4 was an outlier and thus this sample is removed from the data set. Thus one element in $\underline{\mathbf{X}}$ corresponds to the measured sensor signal from sample $i$, measured at time $j$, with sensor $k$, where $i = 1\ldots\ldots,17$; $j = 1,\ldots..,241$; $k = 1,\ldots..,11$.

*3.4. Data handling*

The pre-processing and modelling steps were performed using a Graphical User Interface (GUI) program - SENSABLE. SENSABLE is a newly developed processing tool for sensor based data that can be downloaded at http://www.models.kvl.dk/source/. In Appendix A, a screen dump of the SENSABLE program is shown. The current version of SENSABLE requires MATLAB at least version 6.x.

The different chemometric models can also be calculated using N-way toolbox version 2.10 and the PARAFAC2 toolbox for MATLAB® version 6.5 (The MathWorks Inc., Natick, MA, USA). The toolboxes can be downloaded at http://www.models.kvl.dk/source/.

The K Nearest Neighbours (KNN) classification is performed calculating the distance to the three nearest neighbours using Euclidian distance in scores. Given the low number of samples in each group, it was not tested to increase the number of neighbours. The scores of three test samples – one from each licorice group randomly selected – are predicted (not included in the calibration) and the distances to the calibration samples calculated yielding a classification of each sample. This exclusion and classification of three samples was repeated several times.

## 4. Results

The effect of different pre-processing steps is evaluated and validated using an appropriate chemometric model. The combinations of pre-processing steps and the model will be assessed according to explained variance of the model and according to how well the samples are separated in a score plot (visually evaluated). Note though, that percentage fitted variation in itself does not provide a useful measure of appropriateness. For mathematical reasons, a model with more degrees of freedom (e.g. unfold PCA compared to PARAFAC1) will always fit better. The problem is that the better fit will mostly be fit to noise if the mathematically simpler model (PARAFAC1) is already fitting the systematic variation. The degree of overfit can be evaluated by comparing the fit with the cross-validated fit. The simplest possible yet adequate model is desired. Beforehand, PARAFAC1 can be anticipated to be adequate and hence sufficient, if the underlying time profiles do not change shape only magnitude across different sensors or samples. Unfold PCA on the other hand will be appropriate if the time profiles are unrelated across different sensors or samples. PARAFAC2 is an intermediate model that allows the time profiles to vary in shape but not to be absolutely unrelated.

The focus will be on separating the normal licorices (good) from the deviating licorices (bad), and not as much to separate the good or deviating licorices from the fabricated bad licorices as the latter two are expected to be more similar and of less interest to separate. Care must be taken, as e.g. removing outliers from the data will affect the pre-processing of the data. Upon removal of outliers the data must thus be re-preprocessed and a new model calculated before the actual model evaluation.

Initially the data are assumed to be well analysed with a multi-way chemometric model and the PARAFAC2 model is a likely candidate due to observed shift between the different sensors signals. In the PARAFAC2 model orthogonality constraints are applied for the sensor and the sample mode which helps avoiding numerical problems. All models presented in this article are partly validated using cross-validation by splitting the sample mode in seven randomly chosen intervals.

The number of components selected for each model is determined from a thorough model inspection using the following parameters: explained variance, size and structure of residuals compared to modelled data and raw data, split-half analysis, and visual interpretation of the

scores and loadings. While simple quantitative diagnostics such as cross-validation results seem easier to use, it is important to realize that these only rarely provide correct answers when used for exploratory models as those in this paper. It is always necessary to use additional means such as the above for overall validation. In the following, the basic workflow for achieving a good model is presented.

### *4.1. Pre-processing steps*

#### 4.1.1. Baseline correction and centering/scaling

In Table 3 PARAFAC2 models for different baseline corrections are evaluated. As shown in Figure 2 the baseline subtracted (fractional pre-scaled) sensor signals are of widely different scales. This indicates that scaling the sensor mode is useful to give all the sensors the same opportunity to influence the model.

As seen from Table 3 several combinations of centering/scaling and baseline pre-scaling methods provide an efficient separation of the samples. Thus one could argue that selecting just one of these combinations will give a sufficient pre-processing. However, the pre-processing steps interact and thus must be interchanged before the most efficient combination can be selected. Depending on the data size and structure this process can be time-consuming and thus a standard pre-processing is often selected. With good insight and experience such prior determined pre-processing will often be useful, but problems may arise when new types of samples or sensors are included in the analysis.

The use of integrated software facilitates comparison of many different models. This is helpful when exploring the usefulness of new models as those presented here. In the following a good combination of pre-processing, transformation and model is selected to show the effect of transformation and feature extraction. All combinations of pre-processing steps have been evaluated to make sure that the optimal selection from Table 3 is still optimal after the evaluation of the last two pre-processing steps.

An example of the separation using the fractional baseline correction and centering and scaling is shown in Figure 7A. Figure 7A shows that the good samples are well separated from the deviating samples, but also that it is possible to separate the deviating samples from the fabricated bad licorices. As expected the latter two licorice characteristics appear more similar to each other than to the good licorices.

Transforming the data with logarithmic and similar transformations are not likely relevant for these data (initial results not shown confirm this).

4.1.2. Feature extraction

The concept of feature extraction is to extract the relevant information from the data and to reduce the number of parameters in the data set, ultimately leading to more robust and simple models. In Table 4 an evaluation of some feature extraction methods is shown.

From Table 4 it can be seen that using a subset of the time variables gives a model that is explaining the data well and separates the samples as efficient as the model using all time variables (score plot is not illustrated, but is very similar to Figure 7A). The use of only the maximum of each sensor response curve gives a separation of the samples that is quite similar to the three-way models (Figure 7B). Other types of feature extraction makes the data array smaller and simpler, but offer a less sufficient separation of the samples.

To further evaluate the separation of the three groups a KNN classification is performed and the results are shown in Table 5.

Table 5 indicates that despite an almost perfect classification of the test samples some information about the samples is lost using only the maximum. But the use of the maximum makes it possible to use ordinary PCA and gives a good indication that the samples can be separated; however, the results using PARAFAC2 on the full data set are better though. This concept of using only one variable from the time mode (i.e. reducing the dimension of the data array) is often seen in analyses using commercial electronic noses because the integrated software limits the data to two dimensions.

*4.2. Processing step (modelling of pre-processed data)*

As was seen from Figure 7 both two-way and three-way chemometric models are capable of separating the samples into the three groups.

The major difference between the PARAFAC1 and PARAFAC2 is that PARAFAC2 is modelling each sensor separately ($B_k$ loadings), whereas PARAFAC1 estimate a common time profile loading (B loading) that is used for all sensors, as illustrated in Figure 8.

The scale showing the size of the $B_k$ loadings is arbitrary and can not be directly compared to the real time profiles. Both component one and two of the PARAFAC2 model contributes to the $B_k$ loadings, but it is seen that the contribution from component one is rather noisy and much smaller in magnitude than the loadings of the time profiles given in component two.

PARAFAC2 includes eleven different time profile loadings and this gives a superior separation of the samples as seen from Figure 7A. However, for each component some of the time profiles seem to differ only on the magnitude level and thus a model like PARAFAC1 might be adequate to describe the fewer common time profile shapes.

From pre-analyses (e.g. split-half and core consistency analyses) it was found that the more restricted PARAFAC1 model also needs two components to give the optimal (unique) solution. However, as shown in Figure 9 the samples are not separated in the same efficient way as for PARAFAC2 (and PCA).

The time profile loadings for PARAFAC1 in Figure 8 also shows that the two components have shapes that are very similar to the real sensor response curves (Figure 2). This suggests that the information of the shifted time profiles is not included in the model, but that only an average sensor time profile is allowed for each component. This could indicate that more components are needed to describe the shifted behaviour of the sensor signals. However, the model gets worse including more than two components - thus PARAFAC1 seems to be a too restricted model for this data.

The insufficient separation of the samples using PARAFAC1 indicates that this kind of data is not strictly tri-linear, but rather that the time profile of each sensor should be modelled independently, as is the case using PARAFAC2. The result from the PCA model also suggests that most information of the different samples is located in the peak, and in particular in the maximum of the signal. However, ignoring the three-way structure of the original data tends to give a less sufficient separation of the samples and suggests that important information also can be found in the rest of the time profile curve.

## 5. Conclusion

In this article the processing of sensor based data has been outlined and PARAFAC2 has been suggested for enabling analysis of the raw data even when there are time shifts. Refraining from extracting features is expected to make it possible to extract more information in situations where more subtle details of the data are needed for proper quantification or classification. The extent, to which this is feasible, in general, remains to be explored in various contexts.

Choosing appropriate pre-processing and model requires knowledge of the data structure and the model. For sensor based data the PARAFAC2 model that handles shifted profiles offers some advantages compared to related methods such as PCA. However, the structure of the PARAFAC2 model requires a more careful interpretation of the model parameters and an extensive post-processing procedure could be essential to the concluding interpretation.

The electronic nose combined with the correct data pre-processing, processing and post-processing was found to have the potential to be a fast, easy to use and suitable analytical methods in the application, but some important hurdles must be passed before this method can be introduced as an at-line method. Most notably, more experience is needed in using the multi-way approach for analyzing sensor data.

## Acknowledgements

The authors are grateful to the Danish Technical University (DTU), Lyngby, Denmark for placing the electronic nose at our disposal and for providing expertise on this field. We also wish to thank Pingvin Lakrids, Toms Group A/S, Avedøre, Denmark for letting us use their facilities to produce the samples.

# References

[1] P.N. Bartlett, J.W. Gardner, Sensors and Sensory Systems for an Electronic Nose. Kluwer Academic Publishers, London, UK, 1992.

[2] C. Gretsch, J. Delarue, A. Toury, P. Visani, R. Liardon, Detection of aroma above a coffee powder: limits and perspectives of electronic sensors. ASIC Conference - 17th International Scientific Colloquium on Coffee, Nairobi, Kenya, 1997.

[3] Y.G. Martín, J.L.P. Pavon, B.M. Cordero, C.G. Pinto, Classification of vegetable oils by linear discriminant analysis of Electronic Nose data, Analytica Chimica Acta 384 (1999) 83-94.

[4] R.N. Bleibaum, H. Stone, T. Tan, S. Labreche, E. Saint-Martin, S. Isz, Comparison of sensory and consumer results with electronic nose and tongue sensors for apple juices, Food Quality and Preference 13 (2002) 409-422.

[5] H.-Y. Chung, J.A. Partridge, B.R. Harte, Discrimination of light oxidized off-flavors in milk and cheddar cheese using solid-phase microextraction - gas chromatography and the olfactory sensing technique. Proceedings of the 13th IAPRI Conference on Packaging (2002) 1144-1155.

[6] J.W. Gardner, M. Craven, C. Dow, E.L. Hines, The prediction of bacteria type and culture growth phase by an electronic nose with a multi-layer perception network, Measurement Science & Technology 9 (1998) 120-127.

[7] S. Roussel, G. Forsberg, V. Steinmetz, P. Grenier, V. Bellon-Maurel, Optimisation of electronic nose measurements. Part I: Methodology of Output Feature Selection, Journal of Food Engineering 37 (1998) 207-222.

[8] T. Eklöv, P. Mårtensson, I. Lundström, Selection of Variables for Interpreting Multivariate Gas Sensor Data. Analytica Chimica Acta 381 (1999) 221-232.

[9] C. Distante, M. Leo, P. Siciliano, K.C. Persaud, On the study of feature extraction methods for an electronic nose, Sensors and Actuators B 87(2) (2002) 274-288.

[10] R. Gutierrez-Osuna, H.T. Nagle, B. Kermani, S.S. Schiffman, Handbook of Machine Olfaction, WILEY-VCH Verlag GmbH & Co., Weinheim, Germany, 2003.

[11] R. Bro, A.K. Smilde, Centering and scaling in component analysis, Journal of Chemometrics 17 (2003) 16-33.

[12] R. Bro, PARAFAC – Tutorial and Applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149-171.

[13] H.A.L. Kiers, Towards a standardized notation and terminology in multiway analysis, Journal of Chemometrics 14 (2000) 105-122.

[14] R.A. Harshman, Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis, UCLA Working Papers in Phonetics 16 (1970) 1-84.

[15] N.D. Sidiropoulos, R. Bro, On the uniqueness of multilinear decomposition of N-way arrays, Journal of Chemometrics 14 (2000) 299-240.

[16] B.M. Wise, N.B. Gallagher, E.B. Martin, Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch, Journal of Chemometrics 15 (2001) 285-298.

[17] R.A. Harshman, M.E. Lundy, PARAFAC: Parallel Factor Analysis, Computational Statistics & Data Analysis 18 (1994) 39-72.

[18] R. Bro, H.A.L. Kiers, C.A. Andersson, PARAFAC2 - Part II. Modelling chromatographic data with retention time shifts, Journal of Chemometrics 13 (1999) 295-309.

[19] H.A.L. Kiers, J.M.F ten Berge, R. Bro, PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model, Journal of Chemometrics 13 (1999) 275-294.

[20] R.A. Harshman, PARAFAC2: Mathematical and technical notes. UCLA Working Papers in Phonetics 22 (1972) 30-44.

[21] S. Wold, C. Albano, W. J. Dunn, III, U. Edlund, K. H. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjöström, Multivariate data analysis in chemistry, in: B. R. Kowalski (Ed.), Chemometrics - Mathematics and Statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Netherlands, 1984.

## Figure legends

Figure 1. Example of three-way data set, X from an electronic nose.

Figure 2. Left: Raw/Real sensor signal and Right: Baseline corrected sensor signal - (S-S0)/S0: of one sample and eleven MOS (Metal Oxide Semiconductive) sensors.

Figure 3. Illustration of the principles of data that can be modelled with PCA (A) and of data that show shifted profiles and thus can not be modelled using PCA (B). X is the two-way data; T is the score vector and P the loading vector.

Figure 4. Three-way nose data of dimension *I* (samples) × *J* (times) × *K* (sensors). The data can be matricized to two-way data by concatenating every frontal slab next to each other. Note that sensors in the second (*J*) mode and time in the third (*K*) mode is then confounded in the resulting matrix. The data can also alternatively be unfolded in the other modes.

Figure 5. Graphical explanation of the principles of the PCA model (A) and PARAFAC1 model (B) using a two-component model as an example.

Figure 6. Graphical description of the PARAFAC2 model. Data: Sensors × Time × Samples. Mode A: sensors, mode B: time profiles, and mode C: samples.

Figure 7. Score plots for PARAFAC2/PCA models. (A): PARAFAC2 model of the fractional baseline corrected data centering the sample mode and scaling the sensor mode (Data size: 17×241×11). (B) PCA model of data where the maximum of each sensor signal is extracted (Data size: 17×11).

Figure 8. $B_k$ loadings (time profile loadings) for each sensor of the PARAFAC2 model component 1 (A) and component 2 (B). B loadings for a PARAFAC1model with two components (C). From pre-analyses the number of components is found to be the most optimal describing the pre-processed data for the given model.

Figure 9. Score plot for samples for a PARAFAC1 model with 2 components.

## Table legends

Table 1. Baseline correction formulas for MOS sensors (modified from [9])

Table 2. Characteristics of plots from the different chemometric models (in brackets are shown the number of elements of an $I \times J \times K$ array in the different plots). Unfold PCA is PCA performed on the three-way array rearranged (unfolded) to a matrix.

Table 3. Baseline correction methods and use of centering and/or scaling evaluated from explained variance and score plots from PARAFAC2 models. Two components are included in all models. Note that explained variance is in percentage of pre-processed data and hence not directly comparable for different models.

Table 4. Feature extractions based on fractional pre-scaled data evaluated from PARAFAC2 model (N>2) and PCA models (N=2). Two components are used in each model and the sample mode is centered and sensor mode scaled.

Table 5. Results from KNN classification calculating the distance from each test sample to the three nearest neighbours. The test set consists of one sample randomly selected from each licorice group. Sixty samples were evaluated for each model.

# Figures

Figure 1

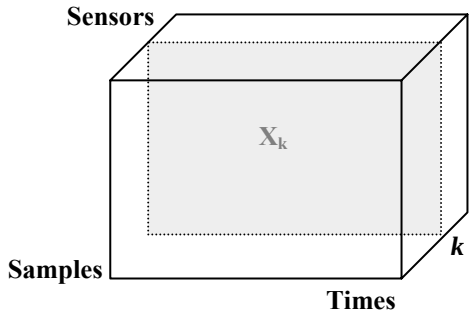
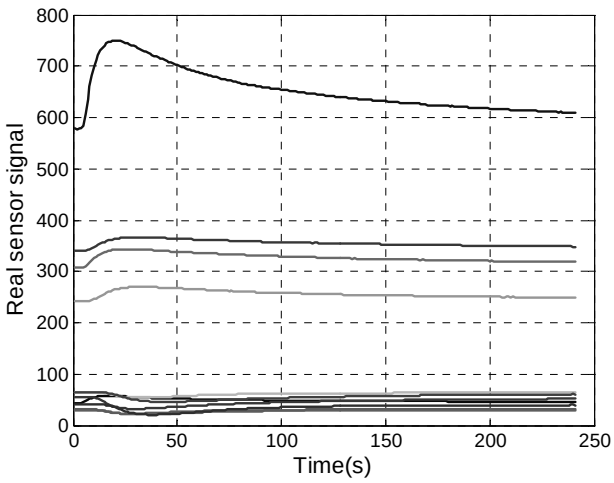



Figure 2



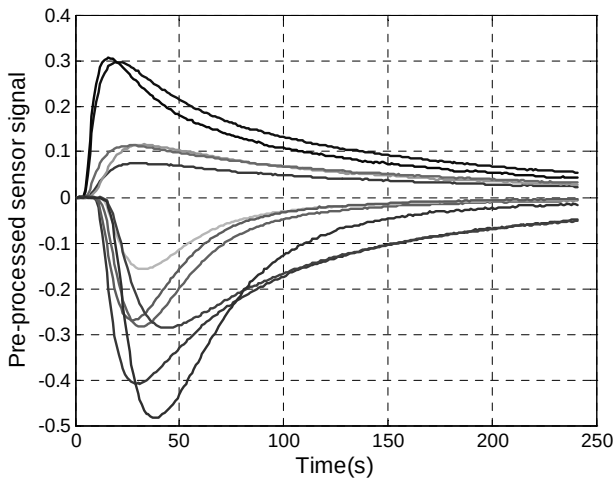



Figure 3

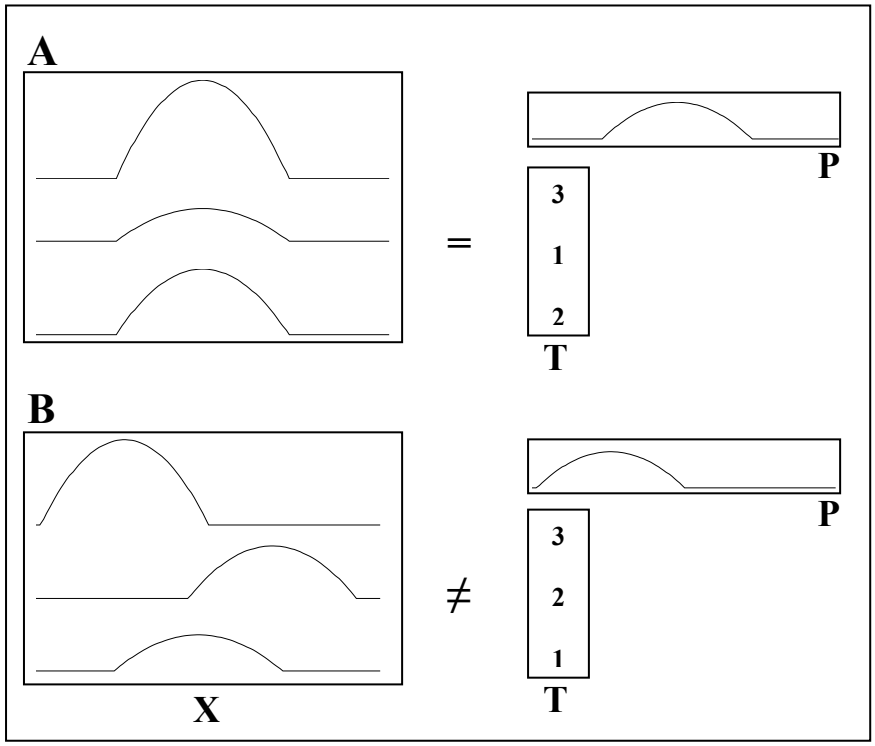



Figure 4

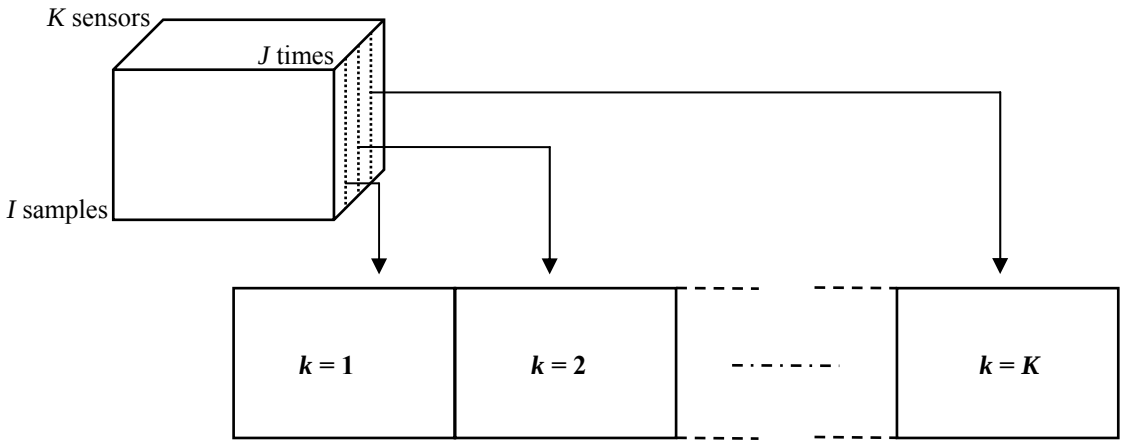


Figure 5

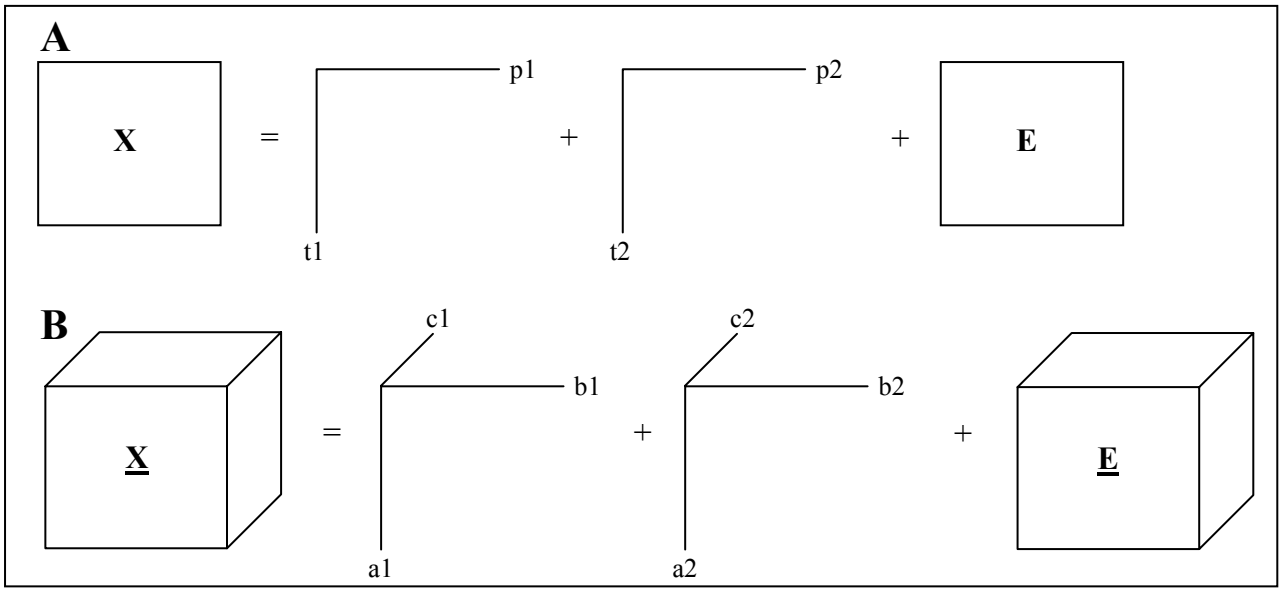



Figure 6

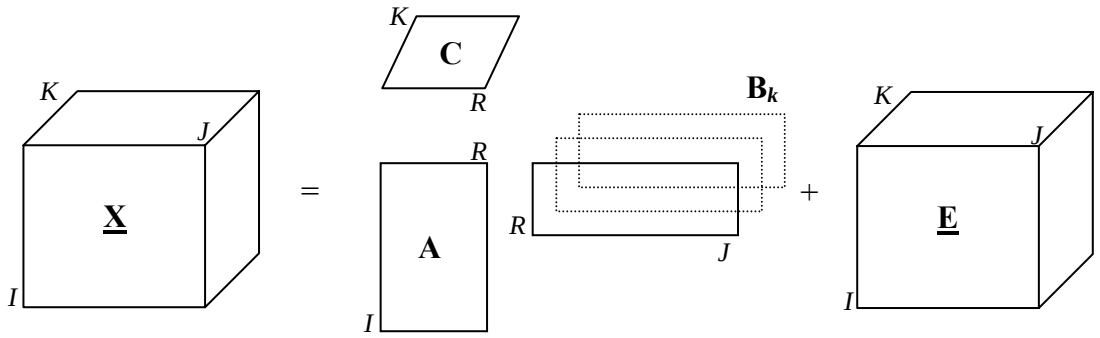



Figure 7

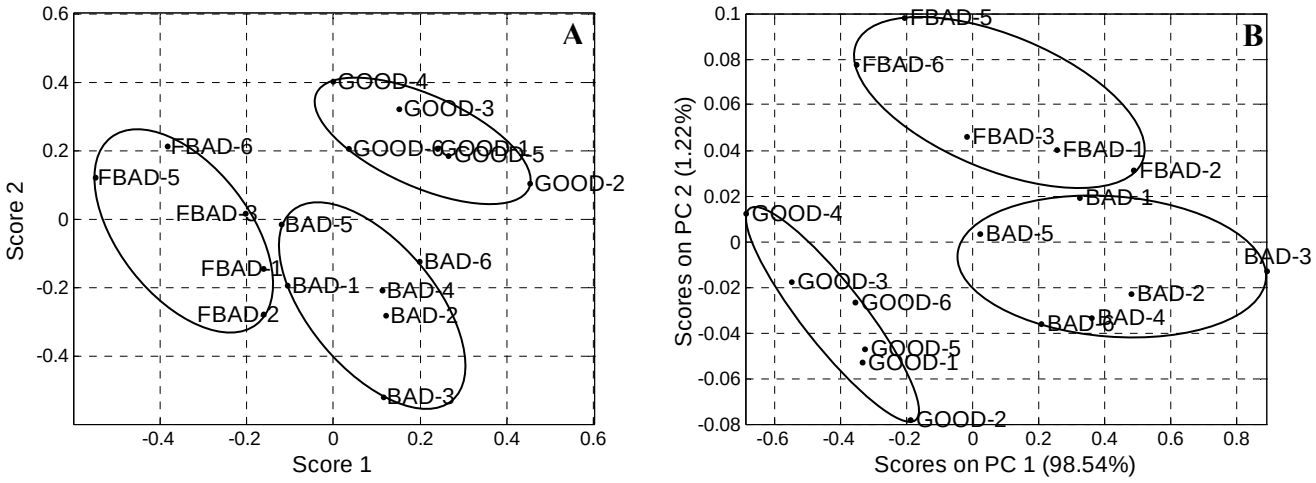


Figure 8

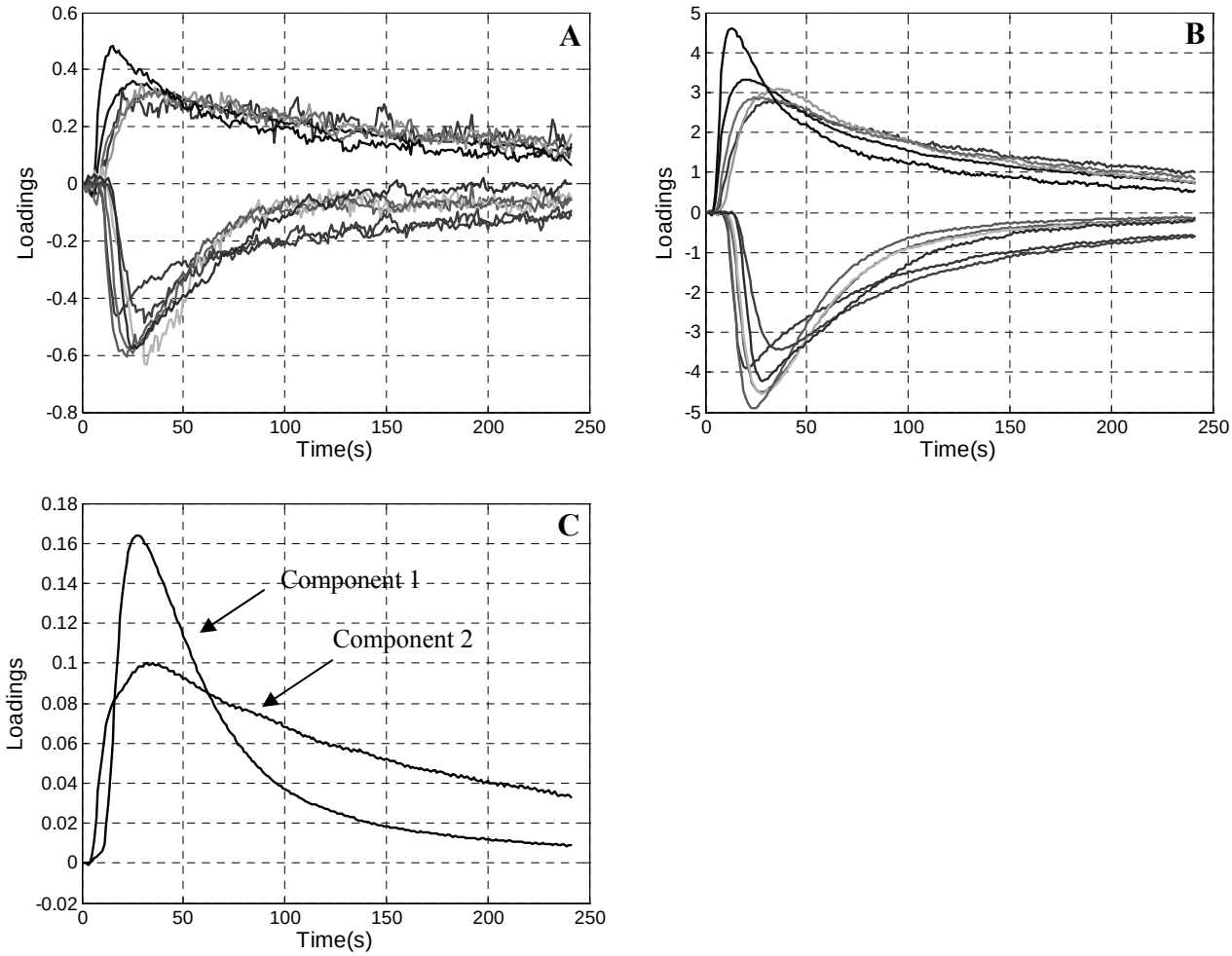
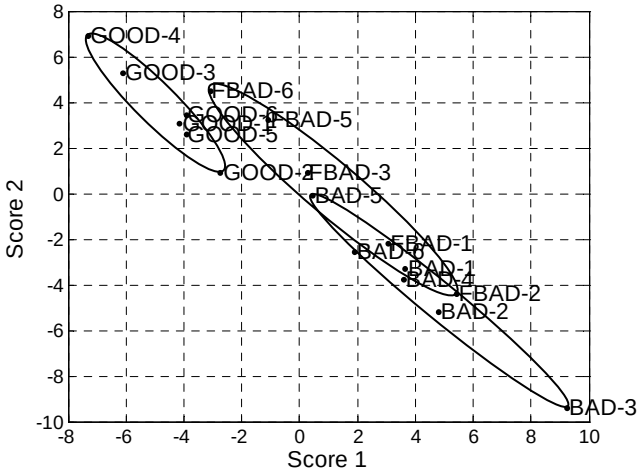



Figure 9



# Tables

Table 1

| Method | Formula [1] |
|---|---|
| Difference | $S(t) - S_0$ |
| Relative | $S(t) / S_0$ |
| Fractional | $(S(t) - S_0) / S_0$ |

[1] $S_0$ is the baseline signal (signal of the carrier gas) and S(t) the signal after a certain time.

Table 2

| Model | Mode A | Mode B (time) [1] | Mode C |
|---|---|---|---|
| Unfold PCA | Score plot (*I*) | Loading plot (*J* × *K*) [2] | |
| PCA [3] | Score plot (*I*) | Loading plot (*K*) | |
| PARAFAC1 | Score plot (*I*) | Sensor signal vs. Time (*J* × R) | Loading plot (*K*) |
| PARAFAC2 | Score plot (*I*) | Sensor signal vs. Time (*J* × *K* × R) | Loading plot (*K*) |

[1] R is the number of components in the model
[2] Mode B is nested within mode C giving a complex loading plot with many loading elements
[3] Extracting one feature of mode B reduces the data size to *I* × *K*

Table 3

| | Centering and/or scaling | | | | Model characteristics | | |
|---|---|---|---|---|---|---|---|
| **Method** | **None** | **Centering** | **Scaling** | **Centering + scaling** | **Explained variance** | **Separation of samples [2]** | |
| Difference [1] $S-S_0$ | | | X | | 99.96 | + + | B-FB |
| | | | | X | 99.06 | + + | B-FB |
| Fractional $(S-S_0)/S_0$ | X | | | | 99.97 | - | |
| | | X | | | 99.30 | + + | B-FB |
| | | | X | | 99.96 | - | |
| | | | | **X** | **99.20** | **+ +** | |
| Relative $S/S_0$ | X | | | | 99.99 | + (+) | B-FB |
| | | X | | | 99.30 | + (+) | B-FB |
| | | | X | | 99.99 | + (+) | B-FB |
| | | | | X | 99.30 | + (+) | B-FB |

[1] Scaling is essential for this baseline correction method
[2] Separation of normal and deviating licorices evaluated from score plots with the first component plotted against the second. -: no separation, +: some separation, and + +: samples are well separated. Symbols in brackets indicate a less significant separation than the same symbol with no brackets. Letters - G: good, B: bad, FB: Fabricated bad – are used to indicate which groups that are not 'perfectly' separated.

Table 4

| Feature extraction | | Size of array | Model | Explained variance | Separation of samples [4] |
|---|---|---|---|---|---|
| All time variables [1] | | 17 × 241 × 11 | PARAFAC2 | 99.20 | + + |
| Subset of time variables [2] | d = 3 | 17 × 81 × 11 | PARAFAC2 | 99.21 | + + |
| | d = 7 | 17 × 35 × 11 | PARAFAC2 | 99.21 | + + |
| | d = 20 | 17 × 13 × 11 | PARAFAC2 | 99.21 | + + |
| | d = 40 | 17 × 7 × 11 | PARAFAC2 | 99.15 | + + |
| Absolute maximum | | **17 × 11** | **PCA** | **99.76** | + + |
| Max adsorption slope | | 17 × 11 | PCA [3] | 92.31 | - |
| Max desorption slope | | 17 × 11 | PCA [3] | 89.11 | - |

[1] PARAFAC2 model from Table 3 that is fractional pre-scaled, sample mode centered and sensor mode scaled
[2] Subset using every d variable in the time mode (data array **X**: 17 × 1:d:241 × 11). If the size of the time mode is not divisible with d, the number of features selected will be rounded to the lowest integer.
[3] Five components in the PCA model
[4] See Table 3.

Table 5

| | Classification | |
|---|---|---|
| **Model** | **Correct** | **Incorrect** |
| PCA | 57 | 3 |
| PARAFAC2 | 59 | 1 |

# Appendix

Screen dump from the SENSABLE program