Thomas Skov & Rasmus Bro

The Food Technology group is a section of the Department of Food Science, KVL, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

Exercises for SENSABLE

Introduction

The data accompanying these exercises was obtained in order to investigate if bad licorices could be differentiated from good licorices. The data is available at www.models.kvl.dk and is called *E-nose_data*.

The problem of bad licorices was revealed through several complaints from consumers that reported a burned taste of the licorices. The licorice company therefore initiated an investigation to find a method that could distinguish between bad and good licorices. A third group of fabricated bad licorices (licorices dried for a longer time) was included in the data set, to mimic the observed burned taste of the bad licorices.

An electronic nose combined with multivariate tools (chemometrics) showed applicable for this purpose. Due to the time-consuming data analysis with several pre-processing steps and model investigations an innovative GUI (Graphical User Interface), 'SENSABLE - Analysis of sensor based data', was made.

Data set

The **X** matrix is arranged as: Samples × Time × Sensor. See Figure 1.

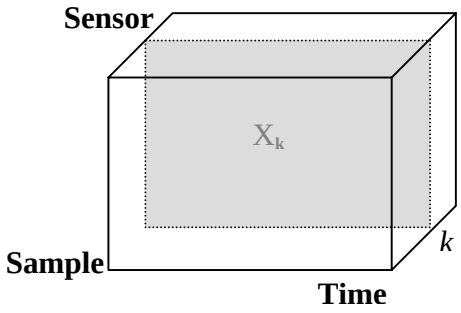


Figure 1. Structure of data from the electronic nose, with the *k* slab (*k* sensor) shown.

The sample mode consists of 6 good licorice samples, 6 bad licorice samples and 6 fabricated bad licorice samples. The time mode is a continuous time scale where the sensor signal has been measured over 120 sec every ½ sec. The first time is the baseline signal (i.e. signal of carrier gas). Twelve sensors, all based on Metal Oxide Semiconductor (MOS) technologies, were used to register the volatile compounds from the samples. This means that the size of **X** will be 18 × 241 × 12.

Y-matrix: Because we know that the samples belong to three specific groups a discriminant variable is included as a kind of external information. Ideally this is done to maximize the separation between the groups and to minimize the variation within each group. The discriminant Y-matrix is shown in **Text 1**.

A typical response of the twelve sensors from a general sample is shown in Figure 2.

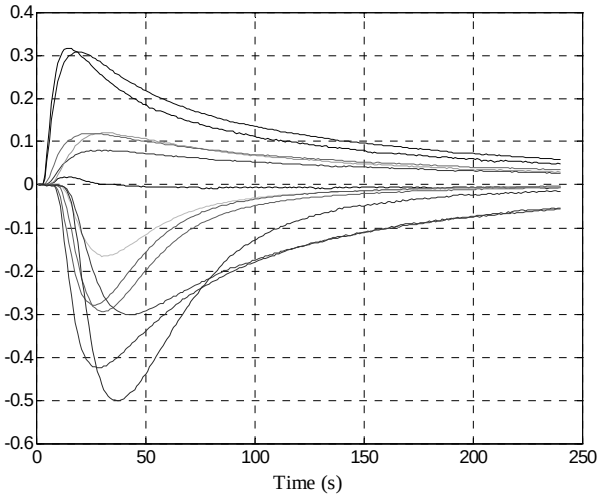


Figure 2. Baseline corrected sensor signal ($(S-S_0)/S_0$) for one licorice sample for the twelve sensors.

Exercises

Download data

Download the data set (*E-nose_data*) from www.models.kvl.dk. The data is also available when you download the SENSABLE file as described in the following.

Get and open SENSABLE

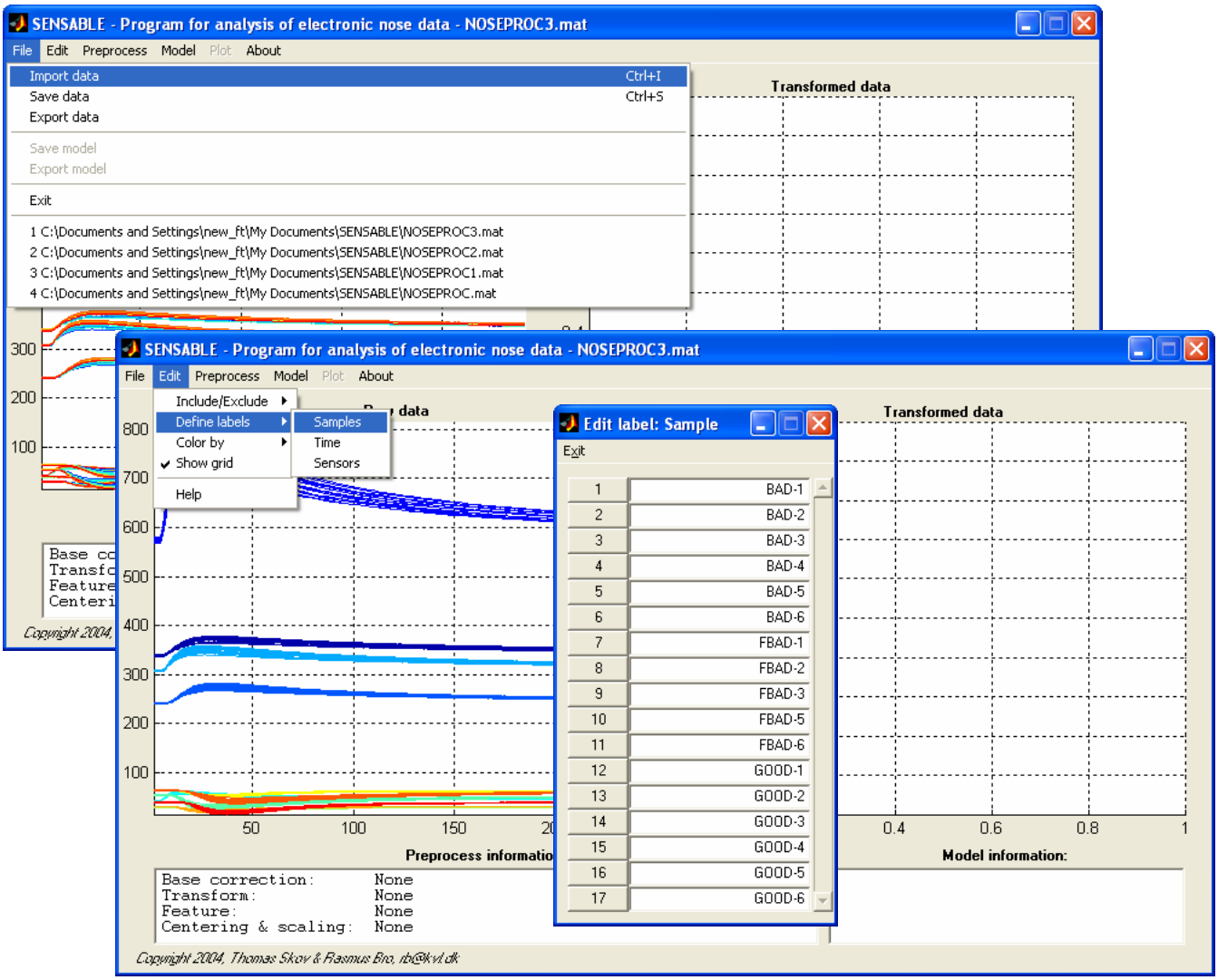
Download SENSABLE from http://www.models.kvl.dk. Extract the zipped files onto your computer. Open MATLAB and type *sensable* in the command window and the SENSABLE program will open.

Import and Edit data

Use the '*File*' menu in SENSABLE to import your own data or use the *E-nose_data* (we recommend you to try the *E-nose_data* first, to be able to use all the implemented exercises given here).

Notice that the name of the data file (MATLAB format) is not limited, but the file must contain a structure array named NoseData. Some strict rules apply for the structure of NoseData, and these are further outlined in the '*help*' file. ASCII files can also be imported but this is not included in this demonstration.

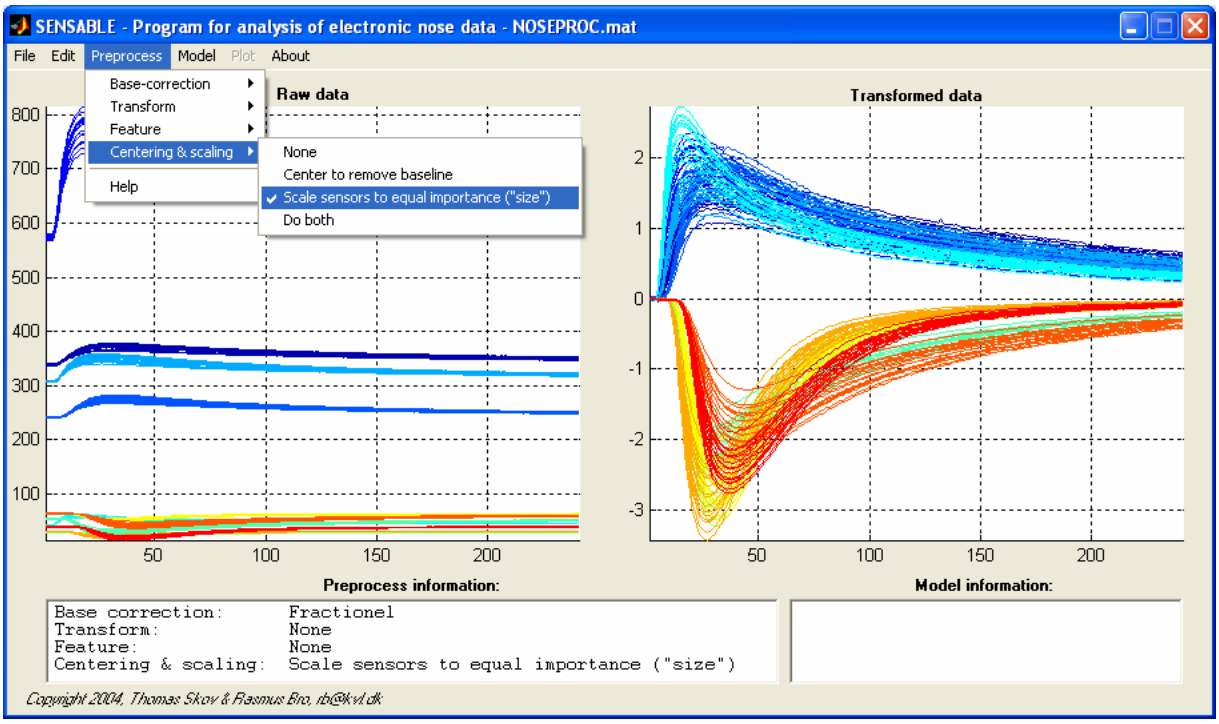
The data can be coloured in the '*Edit*' menu where you also can exclude/include samples, times and/or sensors and redefine sample and/or variable names.



Preprocess

The different processing techniques should be evaluated and the effect demonstrated and understood. In the screen-dump below the menu '*Preprocess*' is shown. The left figure shows the raw data coloured after sensors.

The right plot shows the pre-processed data and this figure is updated each time you select a new/different pre-processing step. In the '*Preprocess information*' box you can see what kind of pre-processing you have chosen.

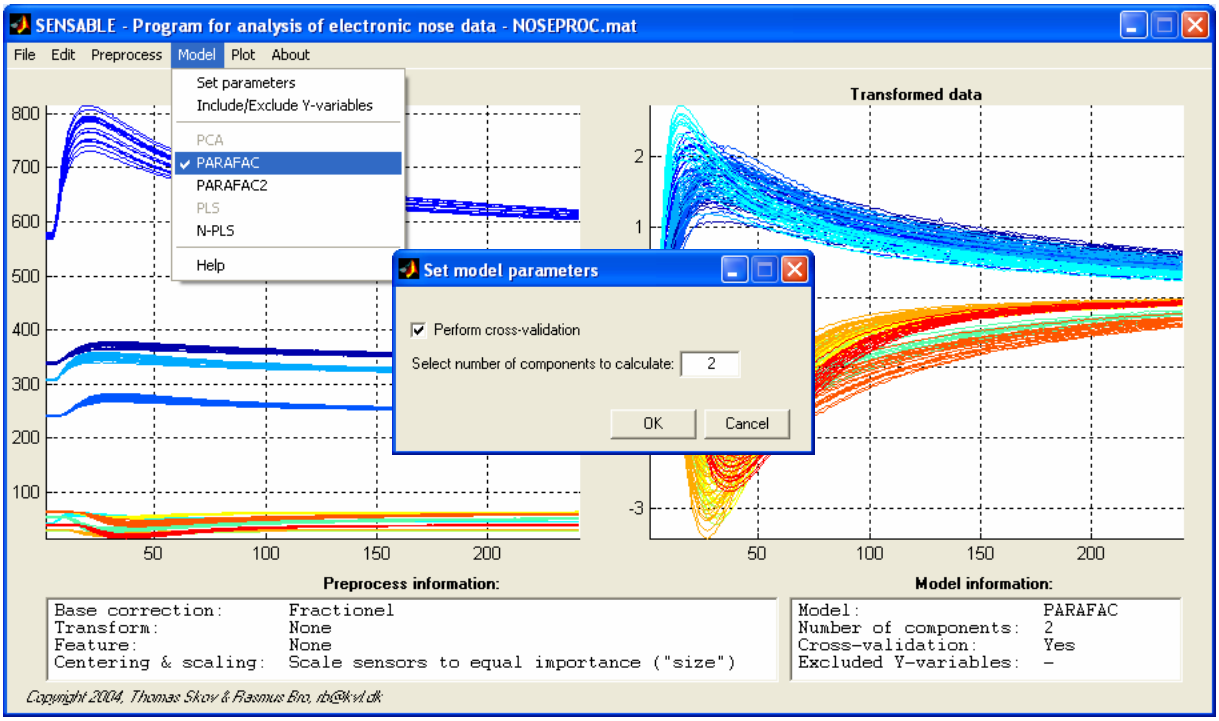


How to do it!

- Import data named E-nose_data into SENSABLE
- Examine the raw data (try zoom function) and try to colour both according to samples and sensors to get an idea of the structure of the data.
- Try different combinations of pre-processing techniques. Try to figure out what the different steps will do before executing the commands. E.g. what does it mean to subtract the baseline from the raw data and how will the first derivate look?
- Discuss if centering and/or scaling of the data are a good idea and why?
- Try to exclude variables from one or more modes to see the effect of reducing the size of the data!
- Try to identify outliers from the raw/pre-processed data. One of the sensors shows a pertubated and strange signal – which one? Will you include this sensor for the further analysis?

Model

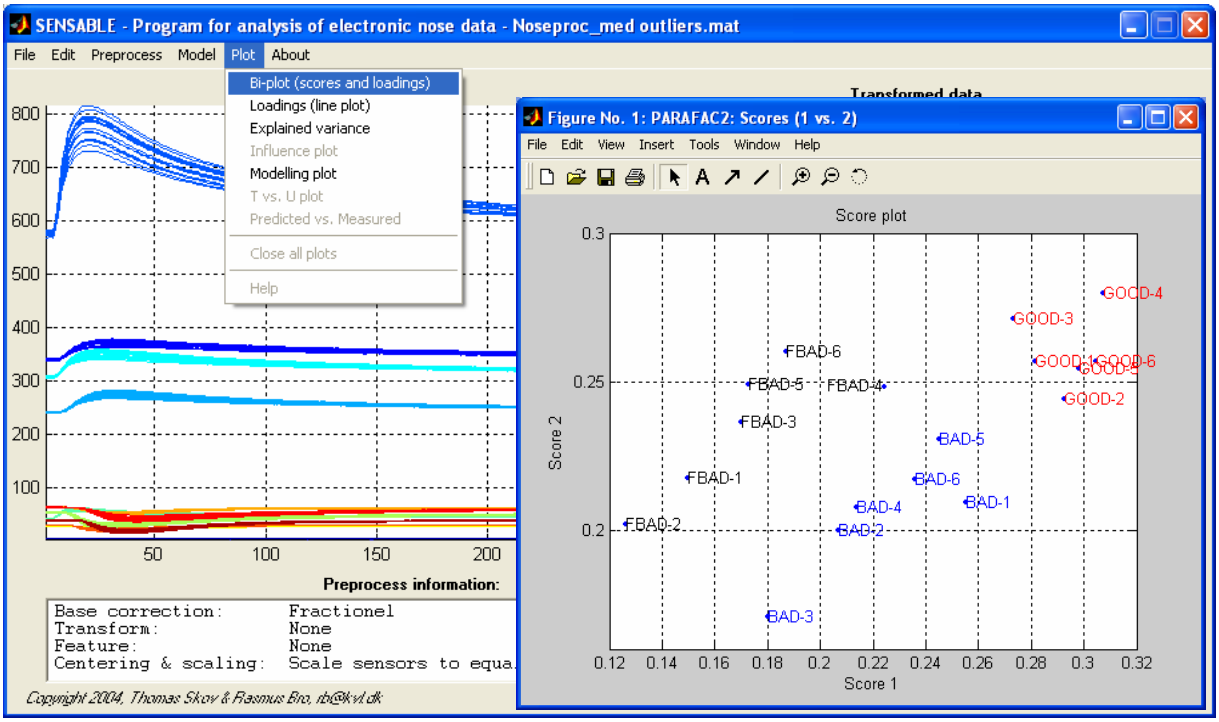
Before starting this step, make sure that you have an idea on how the different chemometric models work. Try to make some initial thoughts of which model(s) might be the proper choice(s) for the data. In the screen dump below you can see the menu '*Model*' is shown. Before choosing the model you should set the model parameters as indicated in the box in the middle. After you have calculated a model the selected features of the model will be given from the '*Model information*' field.



Plots

After the calculation of the model the functions in the '*Plot*' menu becomes available and it is now possible to select the highlighted plots. Notice that the choice of plots differs from model to model, but the philosophy behind SENSABLE is that the plots available should also be the most appropriate plots for each model.

In the following screen dump you can see the four highlighted model from the '*Plot*' menu. The first plot choice gives three plots but only the score plot is shown. The possibility to colour the samples are not implemented in this version of SENSABLE yet.



In order to perform explorative analysis, evaluate the different chemometric models to see how they work in practice. Study the different plots available and make sure you are familiar with the terms of scores and loadings and how they are connected.

Two-way chemometric models

Start with the basic chemometric models PCA and PLS where you try to separate the different subgroups of samples based on the intrinsic relationship in the raw data (PCA) or by using a supervised technique, where external knowledge is introduced in the model (PLS). A more detailed explanation of scores and loadings, model characteristics and the finding of outliers will be put forward during the calculations of the two-way models. However, the same guidelines and rules can be used for the more complex three-way models.

How to do it!

PCA & PLS

For the first model calculation choose the following pre-processing steps:
Base-correction: Fractional
Transformation: None
Feature: Absolute maximum
Centering & scaling: Do both
Outliers: You should exclude sensor from both the PCA and PLS model. Why?

Figure out the dimensions of the data set after the pre-processing steps before you carry on. Which modes have been reduced?

PCA
Try to fit a **PCA model** with five components using cross-validation.

Investigate the model through the available plots from the 'plot' menu. How many components should be included in the model? Investigate this with the explained variance plot. What is the difference between the PCA models that gives the fitted and Xvalidated bars, respectively (i.e. what is the consequence if the model is NOT cross-validated)?

Before deciding on the final number of components use the influence plot and the score plot to locate any outliers. If you think one sample might be an outlier – remove it (remember to recalculate the model and go through the plots again as described above) – or keep an eye on this sample when you proceed with the data analysis.

Can the samples be separated into the three known groups?

PLS
To make the separation of the three groups more visible you can use the external information that the 18 samples belong to three different groups. This can be done with a PLS model.

Calculate a **PLS model** with five components using cross-validation.

You should investigate the predicted vs. measured plot, and explained variance plots of X and Y before the score and loadings plot. Why? Try to investigate variable 1 (bad samples) for components 1-5 to see how well the data are predicted. Remember that the Y-data should be ones for the bad samples and zero for the two other groups. Hint **Text 4**.

Three-way chemometric models

For the model calculation choose the following pre-processing steps:
Base-correction: Fractional
Transformation: None
Feature: None
Centering & scaling: See below (*outliers*)

Outliers: You may want to exclude sensor 1 before calculating the models. Why? Why not? (for hint see **Text 2**). If you decide to include the noisy signal of sensor 1 you might want to reconsider centering the data. Why? (for hint see **Text 3**).

Figure out the dimensions of the data set after the pre-processing steps before you carry on. Take a closer look at the time profiles of the different sensors and notice the difference in magnitude and shape of the profiles.

Recap of theory: The strict (tri-)linearity of the PARAFAC1 is very advantageous when each sample is characterised by the same underlying structure, but the model will often be too restricted to deal with severely shifted time profiles e.g. as in the sensor signals in Figure 2. As seen the time profile does not only change in magnitude but also in shape. The effect of shifted time profiles and the consequence of using a more suitable chemometric model will be illustrated in the following.

How to do it!

PARAFAC

Calculate a PARAFAC model using two components and with cross-validation.

Two components are found to be adequate for both PARAFAC and PARAFAC2 so you should not investigate this further.

Evaluate the score plot for the samples. Are the samples separated more efficient than using two-way chemometrics (PCA)? Why – why not?

PARAFAC2

Use the same context as for the PARAFAC model above. Make sure that you know the difference between PARAFAC and PARAFAC2 before carrying on with PARAFAC2 calculations!

Evaluate the loadings of for the three modes and pay attention to the enhanced complexity of mode two. What is the possible advantage of making this mode more complex?

Evaluate the score plot for the samples. Are the samples separated more efficient than using PARAFAC? Why – why not?

N-PLS

Even though PLS did not give a usable model we will try to calculate an N-PLS model using five components and with cross-validation. Once again you should investigate the predicted vs. measured before interpreting the score plot. Has the model become better using the three-way structure of the data? Why - why not?

Question about SENSABLE

If you during these exercises run into problems understanding the different options given, have problems running the program, or find bugs that we haven't discovered ourselves, feel free to contact us for further information. You can reach us at thsk@kvl.dk.

Information Texts

Text 1
A discriminant variable consists of ones when the sample belongs to the specific group and zeros when the sample does not belong to the group. This illustrated below:

Group	Sample	Discriminant variable		
		I	II	III
1	1	1	0	0
	2	1	0	0
	3	1	0	0
2	4	0	1	0
	5	0	1	0
	6	0	1	0
3	7	0	0	1
	8	0	0	1
	9	0	0	1

Text 2
(NB. With sensor 1 a better separation of the three groups can be achieved, but the signal from sensor 1 looks very pertubated and the manufacture of electronic nose claims that this can be caused by a bad connection in the sensor. This could be the reason to exclude sensor 1, as the reproducibility of the signal might be impaired.

Text 3
The very low and pertubated signal of sensor 1 could be a problem if the data is centered, as the noise is enlarged dramatically. Try to calculate both models - with and without centering – and investigate how the samples are separated.

Text 4
The evaluation of a PLS2 models begins with the inspection of the predicted vs. measured plots, to see if the model is able to predict the given Y variable. If no correlation exists, then it makes no sense to make conclusions based on e.g. the score plot. As can be seen the correlation between X and Y matrix is around 0.75 using three components with an RMSEP value of 0.32 (remember that size of Y-variable is either 1 or 0). This indicates that the model is not very good in predicting the Y-value and thus is make no sense to make any further conclusions from this model (NB the same low correlation and high RMSEP values is observed for the other two Y-variables).