Analytical Methods

Pre-processing in vibrational spectroscopy, a when, why and how

Journal:	*Analytical Methods*
Manuscript ID:	AY-TRV-12-2013-042270.R1
Article Type:	Tutorial Review
Date Submitted by the Author:	28-Apr-2014
Complete List of Authors:	Rinnan, Asmund; University of Copenhagen, Department of Food Science

ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

Pre-processing in vibrational spectroscopy, a when, why and how

Åsmund Rinnan,[*a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Pre-processing is nothing without scattering. If your spectra are from nice aqueous solutions with only fully dissolved particles, there is no light scattering, and as such, pre-processing is not necessary. However, and this is important, scatter could also be defined as unwanted variation in your data with a different source than light scatter. Sometimes it is possible to remove these unwanted variations from your data through pre-processing methods designed to remove scatter. In this paper I would like to take you into my world of pre-processing. Through three different examples I will discuss and tell what kind of information the pre-processing can tell the user about the data, as well as some common pitfalls.

1 Introduction

For the application of multivariate data analysis to work optimally, it is vital to pre-process the data in a correct manner. If this is not done, there will be a mix-up between the information which is sought and the noise which one are not interested in. Noise does not only constitute of random deviations in the measurements themselves. It can also contain systematic variations in the samples which is not of interest to the analyst. One such variation is the light scattering; created by particles which are illuminated. This effect is nearly non-existent for liquid samples (although suspensions will show scatter), while solid samples are prone to show scattering. Pre-processing of spectra has focused a lot on NIR spectra, but the same methods can readily be used for other spectroscopic techniques such as IR and Raman.

This article will focus on the practical issues with regards to pre-processing, and some issues one should be aware of while using these methods. For a detailed description of pre-processing methods I would like to refer to Rinnan et al.[1] and Boulet and Roger[2]. For a good discussion regarding the difference between SNV and MSC I would like to refer to Fearn et al.[3]. It should be noted that this paper is not an exhaustive review of pre-processing techniques, but should rather be seen as a tutorial.

2 Materials and methods

In order to show how to perform pre-processing, and to explain the reasoning behind each decision three different datasets will be used, one from each of the three techniques: NIR, IR and Raman.

[a] *Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark. Fax: +45 35 33 32 45; Tel: +45 35 33 35 42; E-mail: aar@food.ku.dk*

2.1 Software and datasets

All calculations and plots are made in Matlab (version 2012a, The Mathworks, Natwick, MA, US). The codes used are all in-house, but are all common chemometric tools, implemented in most software packages. The NIR and Raman datasets can be found at http://www.models.life.ku.dk/datasets (last accessed: april 28 2014).

2.2 NIR - Soil samples

The samples herein are made up of 108 soil samples measured by NIR in the range 400-2498nm with a 2nm resolution using a NIRS6500 instrument (Foss A/S, Hillerød, Denmark) with a round cup and equipped with a ring (Microsample inserts, Part No. IH-0337). Both the background spectra and the samples were measured as the average across 32 scans. Soil organic matter (SOM) was measured using a reference method, and the samples contain from 42.9%-95.9% SOM. More details on the dataset can be found in Rinnan and Rinnan[4].

2.3 IR - Milk

This dataset contain a total of 105 milk samples measured in triplicates on the MilkoScan (Foss A/S, Hillerød, Denmark)[b]. The reference which is of interest for this dataset is the fat content in the milk, given in %w/w. The samples have been measured in the range of 5000-800cm^{-1}, and water was used as the internal reference in the spectrophotometer.

2.4 Raman - Pork fat

A total of 105 pork fat samples were taken from a total of 16 pork carcasses, taken from different depth of the fat from the skin. The samples were measured on a RamanRcn1 instrument (Kaiser Optical Systems Inc., MI, USA) equipped with a

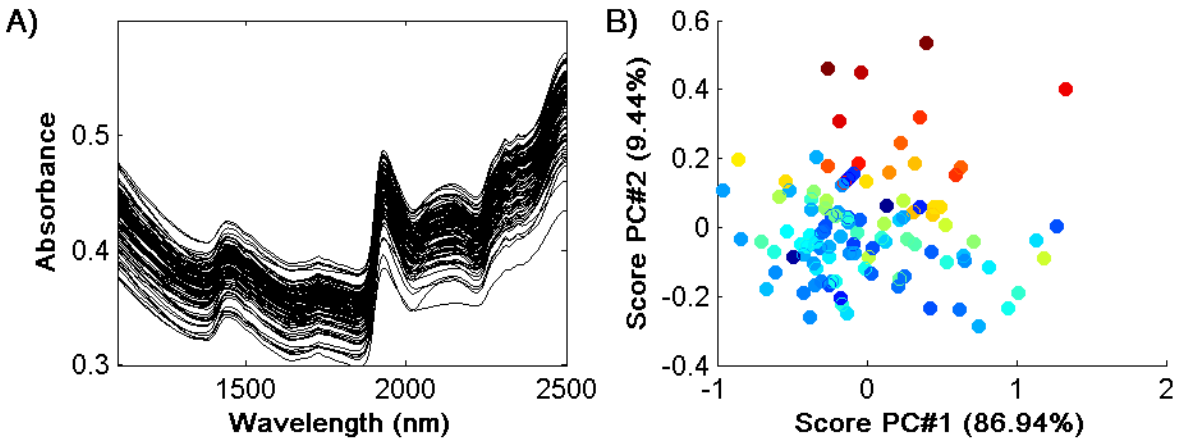

Fig. 1 Raw NIR soild data. Data inspected in two ways: A) by plotting the raw data, and B) through a PCA, where the sample is colored according to their SOM content (red = low, dark blue = high).

785nm near-infrared diode laser (Invictus, Kaiser Optical Systems Inc., MI, USA). The spectra were recorded in the range of 1800-200 cm^{-1}, as a sum of 16 measurements with 1s exposure. The iodine value was calculated on the basis of GC-MS measurements. More details with regards to the data can be found in Lyndgaard et al.[5].

3 Results and discussion

The following section will go into more detail with regards to practical suggestions and hints on how to correctly perform pre-processing to spectroscopic data. As said before, this is not aimed at giving an exhaustive introduction to all exotic variants of the most common methods, but I will rather focus on the most typical of all pre-processing techniques.

3.1 NIR - Soil samples

The first natural step in all data analysis is to investigate your raw data. This could either be done by plotting the spectra as they come from your spectroscopic instrument, or it could be based on a preliminary Principal Component Analysis (PCA)[6] of the data.

Figure 1A shows the raw un-treated NIR spectra of 108 soil samples, while the same data is shown in a PCA in Figure 1B. As can be seen from Figure 1A it is evident that there is a rather large variation along the y-axis. However, it is difficult to assess whether this is due to chemical information or if this is more physical in nature. By looking at the score-plot, shown in Figure 1B, it becomes evident that the main variation of these data are not the SOM content. A quick inspection of the SOM range (42.9-95.9%), it is clear that SOM should be the main variation in the data. This is clearly *not* the case. Since I know that these samples are soil samples I have a suspicion that the large variation seen in Figure 1A is due to scatter rather than chemical information. The natural next step would be to try a few different pre-processing techniques and see what effect these have on the data.

A natural pre-processing technique to apply to these data would be the derivative. Most often a first or a second order derivative is used. The first derivative will remove any offset difference between the data and the second derivative will furthermore remove any slope effect in the data. However, before I show the results of the derivation, I would like to get into a bit more detail with regards to how the different derivation techniques work.

As just stated, the use of derivation as a pre-processing technique for NIR data is quite common. There are two typical ways of estimating the derivative: Norris-Williams derivation (NW)[7] and Savitzky-Golay derivation (SG)[8]. The former is in many ways a simplification of the latter. In NW the smoothing performed to the data is according to a 0th order polynomial (the average only), while for SG this smoothing function can be set to any polynomial order (two is though the most common). The second parameter which should be set for NW is the gap size. This truly is a bit of a curiousity as there is nothing clear in the spectra which should indicate that you need a gap size for anything. However, upon looking closer at the equations used in NW, it becomes apparent that the gap-size has a smoothing effect on the calculated derivative. In equations 1-3 I have shown how the first derivative is calculated for variable number 4 in a spectrum, using a five point smoothing and a gap size of 2. The reason for showing these equations is

[b] *The data was kindly made available by Per Waaben Hansen, Foss A/S, Hillerød, Denmark.*

to emphasize that for many applications of NW, the derivative is a simple finite difference operation, and little or no smoothing actually takes place.

$$(x_4)' = x_{s5} - x_{s3} \quad (1)$$

$$(x_4)' = \frac{x_3 + x_4 + x_5 + x_6 + x_7}{5} - \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} \quad (2)$$

$$(x_4)' = \frac{x_6 + x_7 - x_1 - x_2}{5} \quad (3)$$

Where x_{sn} indicates the smoothed version of variable n of the original data, and x_n is the value of the raw data for variable n. The apostrophe (') indicates the derivative. As can be seen from equation 3 the derivative is the difference between the points 6 + 7, minus the similar values for 1 + 2. If the gap size was only one, the derivative would be:

$$(x_4)' = \frac{x_7 - x_2}{5} \quad (4)$$

The road to this equation follows the same steps as shown in equation 1 and 2. The difference in equation 3 and 4 is only in the number of points used in calculating the derivative. I.e. the larger the gap-size, the more numbers are used in the estimation of the derivative, and thus a greater smoothing effect has been achieved. The total smoothing effect is thus a combination of the window size *and* the gap size used. The smoothing effect can, though, *never* be higher than the window size.

SG on the other hand estimates a polynomial on the window size which is used for the smoothing of the data. The derivative is subsequently estimated from this fitted polynomial. The window size thus has a direct and straight forward effect on the estimated derivative.

In both NW and SG, there is a challenge with regards to the end-points. In NW, these end-points are simply removed from the data. In the example given in equation 3 above, the three first, and the three last points would thus be lost. If the number of variables is large (which is normally the case nowadays), this loss in information is neglible. The same accounts for SG-derivatives. However, Proctor and Sherwood[10] and later Gorry[9] both suggested to use an assymetric window for estimating the derivative at the end points. This can, however, have detrimental effect on the subsequent multivariate modeling. The reason can most easily be appreciated by looking at how the end-points will look like if a 2nd order polynomial was used in the smoothing, and the 2nd derivative is what is sought for. The general behaviour of the 2nd derivative can be seen in equations 5-7.

$$f(x) = b_0 + b_1 x + b_2 x^2 \quad (5)$$

$$f'(x) = b_1 + 2b_2 x \quad (6)$$

$$f''(x) = 2b_2 \quad (7)$$

Where b_n are constants, x are the original measured data points and $f(x)$ denotes the function of x which describes the absorbance. As can be seen here, any estimate of the 2nd derivative for any of the points used to estimate this polynomial will *all* be equal to $2b_2$. This means that the first number of variables will all have the same estimate for the derivative. The same goes for the last number of variables. The exact number of variables which will be identical depends only on the window size used. The larger the window, the more variables of identical estimated derivative. Now, let us consider where a spectrophotometer normally has the *worst* signal-to-noise ratio. This is typical at the edges, as the manufacturer has tried to push the system to give the maximum amount of output. I have nothing against that, but by including the same numerical value for the first couple of points, in effect, means that the first point has a higher influence in the subsequent multivariate data analysis, as the value has been duplicated a number of times (equal to half the windows size, rounded down). This does not make any spectroscopic (or mathematical) sense, and I would therefore suggest that the users should keep the end-points as missing values, and thus reduce the number of variables according to the window size used. I.e. each spectra will loose window size - 1 variables compared to the raw data.

If the user decides to use a higher order polynomial for the fitting (i.e. 3rd or higher), these effect becomes different, as the estimates of the derivative are not identical anymore. For a 3rd order polynomial the end points form a line, with a 4th order polynomial the end points form a polynomial etc. However, and this is general, no matter what polynomial order was used in the fitting; the polynomial parameters are estimated based on the end-points of the spectrum. These points normally have a lower signal-to-noise ratio than the rest of the spectrum (as discussed above). It means that you will let one fitting (based on low signal-to-noise ratio) to decide very many points in your smoothed spectrum. So, even though the effect of the noisy end-points is somewhat less if a higher order polynomial has been used in the fitting, I would still recommend to not perform this asymmetric fitting.

After this discussion on ways of estimating the derivative I would now like to inspect what effect the 2nd derivative has on the spectra, and the subsequent PCA.

By a close inspection of Figure 2A it can be seen that there are no values for the first and the last variables in the spectra. This is simply because I have deleted these variables after the Savitzky-Golay preprocessing as to not give the end-points too much influence in the PCA (see discussion above). Figure 2A also shows that the baseline variation has been minimized. It is easily appreciated, by looking at Figure 2B that the preprocessing has transformed the data so that the variation in the SOM is one of the major variations in the data, and not as in Figure 1B where there only was a slight tendency that

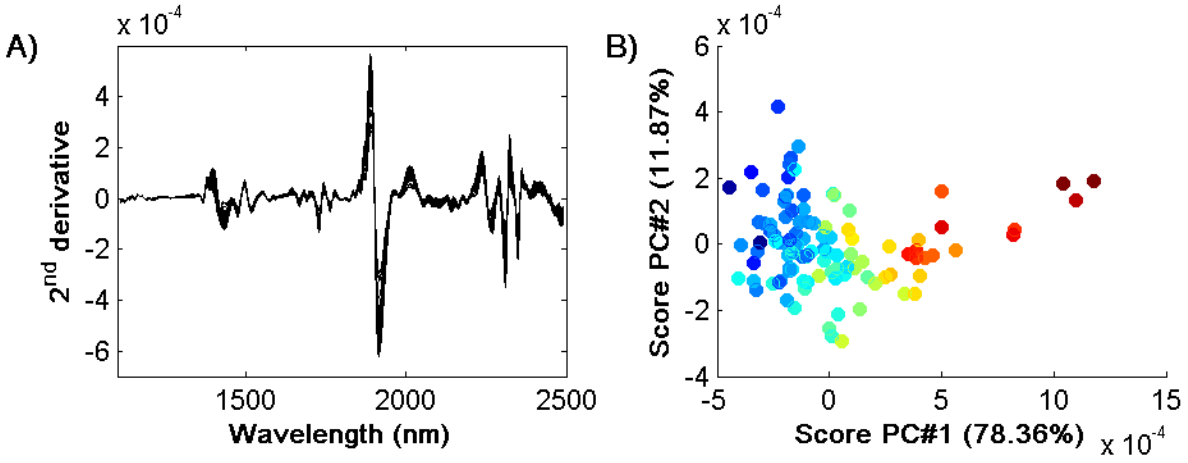

Fig. 2 NIR soil data. Data are treated by Savitzky-Golay using a window size of 9, 2^{nd} order polynomial fitting and calculating the 2^{nd} derivative. A) The SG preprocessed data, and B) PCA on the same data, where the sample is colored according to their SOM content (red = low, dark blue = high).

the variation in the SOM content was modelled along the 2nd PC. This is in accordance with the data as SOM clearly is the major component in soil. The above example clearly shows that by the use of the appropriate pre-processing, the chemical information inherent in the data can be brought forward.

3.2 IR - Milk

Normally it is not necessary to pre-process IR spectra, as they show a lot less scatter than NIR, which in general needs pre-processing and only seldom can be left as they are. However, there are cases with IR spectra where it is better to do a bit of pre-processing than to keep the spectra as is. The following data are an example of that. Furthermore, I would like to use this chapter to inform the reader on what information can be gathered from the subsequent regression analysis on the effect the pre-processing both has on the data, and where noise and information is present.

The raw data is shown in Figure 3, and there are some areas in the spectra which look suspicious. However, if the knowledge of the user is limited with regards to how the spectrophotometer works, it would be natural to include all variables in a subsequent data analysis. Based on this data I will create two partial least squares regression (PLS)[11] models, one which is based on mean-centered data, and one which is based on autoscaled data. One can readily argue that one should *never* autoscale spectroscopic data. Well, now I do it anyhow. The performance of the two models give a root-mean-squared-error of cross-validation (RMSECV) of 0.12 at four PLS components, and RMSECV = 0.07 at four PLS components, respectively for the mean-centered and the autoscaled data. Thus, using the "wrong" preprocessing technique the RMSECV value becomes a lot lower. Why is this? If we look back at Figure

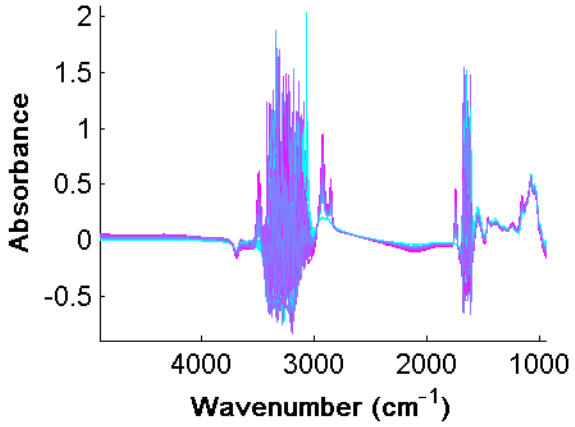

Fig. 3 IR milk raw data. Data inspected by plotting the recorded spectra including all wavenumbers. The spectra are colored according to the fat content (pink - high, cyan - low).

3 we can appreciate that the first model will focus on the areas around 3100 cm^{-1} and 1600 cm^{-1}, while the autoscaled version will *let* all variables have the same influence on the model. Thus, by weighting down the mentioned areas, as will be the case for the autoscaled data, the subsequent PLS model will be better. This indicates that these areas are detrimental for the regression model, and should be omitted. By investigating these areas it becomes evident that these are due to the water bands. As water is the reference, hardly any signal is transmitted through the sample and the reference in these areas. Thus two small, very uncertain, numbers are divided by each other to find the absorbance, and the signal-to-noise ratio will be very bad for these regions.

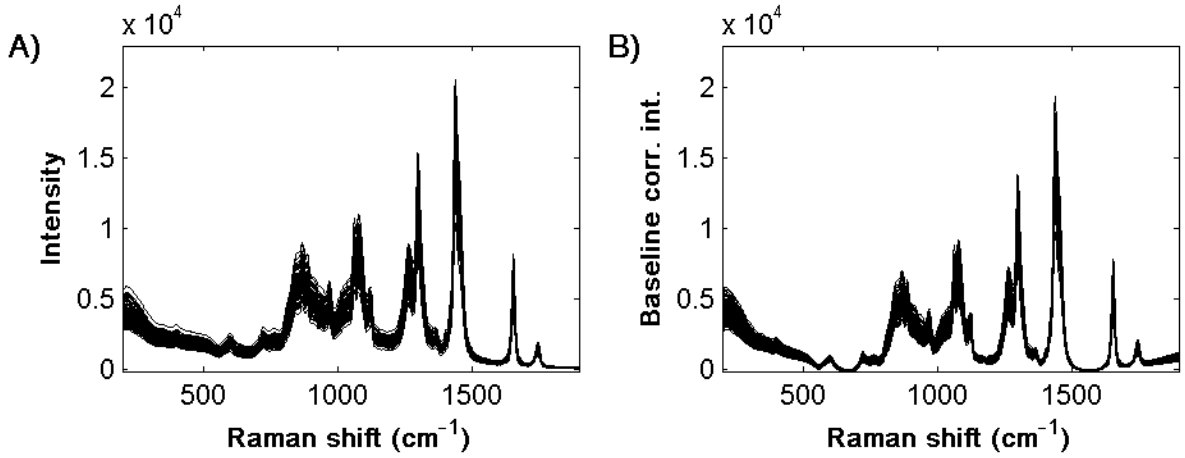

Fig. 4 The Raman pork data, plotted A) prior to pre-processing, and B) after 1st order polynomial baseline correction.

After removal of these areas and also the flat regions from 5000-3800cm^{-1} and 2800-1800cm^{-1}, and furthermore the areas in the fingerprint area below 1400cm^{-1}, the model should be able to give a good prediction of the fat content. The remaining of the spectra which has been kept has a high correlation to the fat-content. We can emphasize this by making two PLS models with different pre-processing, on this reduced data range: only mean-centering, and another with multiplicative scatter correction (MSC)[12], followed by mean-centering. The first model achieves a RMSECV = 0.048 at one PLS component, while the MSC version achieves a RMSECV = 0.63 at one PLS component. In other words, the error in the MSC model is more than an order of magnitude higher than the corresponding mean-centered model. How can this be? The answer can be found in how MSC performs the correction. As the correction factors in MSC are found by plotting the average calibration spectra versus the raw data, it is of importance what part of the spectra which is used in order to estimate these corrections factors. As mentioned by Geladi et al.[13] the correction parameter should optimally be calculated based on the baseline of the spectra. However, as this often is problematic to define, the whole spectra are very often used (also the case in most software). In the case above, the remaining peaks in the spectra mainly contain information with regards to the fat content of the sample. Thus the MSC correction *removes* this information from the spectra, and the subsequent PLS model would then have great difficulties with predicting the fat content since that information is hardly present in the spectra anymore. The correlation between the correction factors and the fat content is as high as $r^2 = 0.996$ for the slope correction, clearly indicating that the height of these peaks are important in the prediction of fat, and should not be minimized.

3.3 Raman - Pork fat

Raman spectroscopy is troubled by fluorescence in the spectra (just as fluorescence spectroscopy is troubled with Raman scattering). This is a typical love/ hate relationship which I do not want to get closer into here. What is important is that the fluorescence in Raman spectra has to be handled in order to extract the chemical important information from the Raman spectra. Lieber and An-Jansen[14] proposed a method of iterative baseline correction, where a baseline is fitted much like spectral detrending introduced by Barnes et al.[15]. However, Lieber and An-Jansen[14] suggested to perform this in an iterative manner, where the measurement points in the raw data which appears above the calculated baseline is not included in the next estimation of the baseline. This is performed until no more measurement points are removed in the process. The final baseline is then subtracted from the spectra. The baseline can in principle be calculated using any function, but a polynomial is the one which is most often used.

As can be seen from Figure 4 the first order polynomial baseline correction removes some of the variation in the baseline. By looking at the areas of the Raman spectra which indicate baseline (the lower areas in the spectra), the samples are in general well collapsed. It doesn't seem that the peak ranges have changed very much after the pre-processing. From a visual inspection of the effect of the pre-processing it seems that the pre-processing has performed satisfactory. However, as the goal of the analysis of these samples is the prediction of the Iodine Value (IV), it is important that the subsequent PLS-models perform as good as possible. In order to investigate this four different pre-processing methods were investigated and a 10 segmented partial random cross-validation (making sure that the range is nicely covered in each segment) was performed 20 times (same segmentation for all pre-processing

methods), and the performance of the methods was used as the basis for the selection of pre-processing method.

Figure 5B shows that the Standard Normal Variate (SNV)[15] correction is the superior of the four methods. Spectral detrending and baseline correction are in principle identical, while using no pre-processing leads to a model with one more PLS component. As the general goal of pre-processing is an improved and simpler model, I deem that the degree of freedom spent in the none vs. baseline/ detrend is the same, and thus I would prefer the no pre-processing model before any of these. However, SNV performs even better; not does it only lead to even fewer latent factors in the model, but the RMSECV is also constantly lower, i.e. a significant gain in the RMSECV going from the no pre-processed model data to the SNV corrected data.

This shows that even though baseline correction theoretically is the best method to be used for Raman spectra, it is not always the case (also higher order polynomials were tested). It is a good idea to additionally investigate a few other pre-processing techniques.

4 Summary

In this paper, I have discussed the use of some of the most common pre-processing techniques through three spectroscopic examples. I have shown some typical pitfalls and discussed the use of pre-processing of spectroscopic data.

References

1 Å Rinnan, F van den Berg, and S B Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, 28:1201–1222, 2009.

2 J-C Boulet, and J-M Roger. Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61–69, 2012.

3 T Fearn, C Riccioli, A Garrido-Varo, and J E Guerrero-Ginel. On the geometriy of SNV and MSC *Chemometrics and Intelligent Laboratory Systems*, 96:22–26, 2009.

4 R Rinnan, and Å Rinnan. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. *Soil Biology & Biochemistry*, 39:1664–1673, 2007.

5 L B Lyndgaard, K M Sørensen, F van den Berg, and S B Engelsen. Depth profiling of porcine adipose tissue by Raman spectroscopy. *Journal of Raman Spectroscopy*, 43:482–489, 2012.

6 S Wold, K Esbensen, and P Geladi. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52, 1987.

7 K H Norris, and P C Williams. Optimization of Mathematical Treatments of Raw Nrear-Infrared Signal in the Measurement of Protein in Hard Red Spring Wehat. I. Influence of Particle Size. *Cereal Chemistry*, 61(2):158–165, 1984.

8 A Savitsky, and M J E Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

9 P A Gorry. General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method. *Analytical Chemistry*, 62(6):570–573, 1990.

10 A Proctor, and P M A Sherwood. Smoothing of Digital X-ray Photoelectron Spectra by an Extended Sliding Least-Squares Approach. *Analytical Chemistry*, 52(14):2315–2321, 1980.

11 P Geladi, and B R Kowalski. Partial Least-Squares Regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

12 H Martens, S A Jensen, and P Geladi. Multivariate linearity transformations for near infrared reflectance spectroscpy. in: *Proceedings of the Nordic Symposium on Applied Statistics* (editor: O H J Christie), Stokkland Forlag, Norway, 205–234, 1983.

13 P Geladi, D MacDougal, and H Martens. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Applied Spectroscopy*, 39:491–500, 1985.

14 C A Lieber, and A M An-Jansen. Automated Method for subtraction of Fluorescence from Biological Raman Spectra. *Applied Spectroscopy*, 57: 1363–1367, 2003.

15 R J Barnes, M S Dhanoa, and S J Lister. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, 43:772–777, 1989.

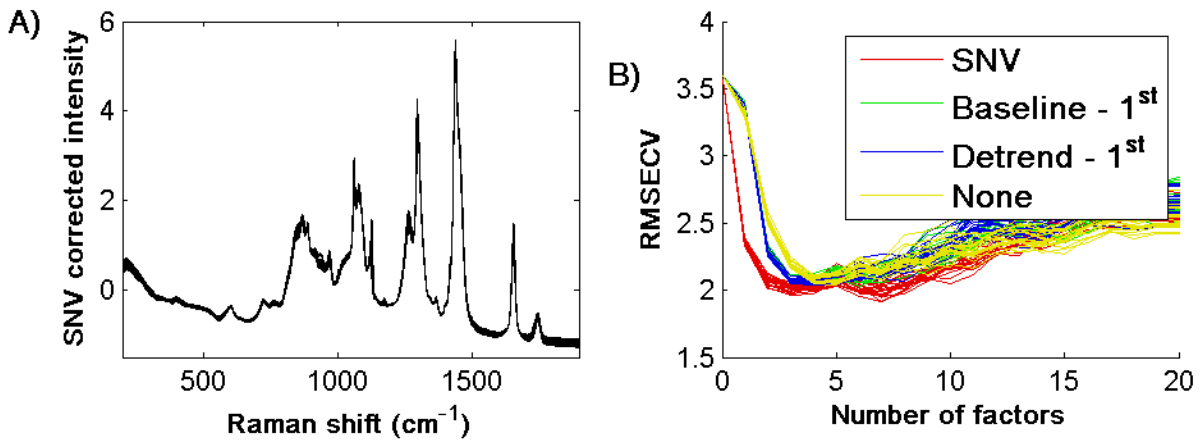

Fig. 5 The Raman pork data. A) The SNV corrected data. B) The number of latent factors plotted versus the RMSECV for 20 cross-validated models for data pre-processed by SNV (red), baseline correction (green), detrending (blue) and none (yellow).