

The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams

J. Luypaert^a, S. Heuerding^b, Y. Vander Heyden^a, D.L. Massart^{a,*}

^a ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussel, Belgium

^b Novartis Pharma AG, Pharmaceutical and Analytical Development, CH-4002 Basel, Switzerland

Received 17 May 2004; received in revised form 21 June 2004; accepted 21 June 2004

Available online 26 August 2004

Abstract

This work is part of a study in which the possibility of NIR combined with some chemometrical methods is investigated as a suitable technique to classify clinical study samples of a cream. In this study, the influence of different preprocessing methods on the removal of spectral variations due to some variance sources has been investigated. The applied preprocessing methods are standard normal variate (SNV), detrend correction, offset correction, and first and second derivation. The investigated variance sources are different batches of ingredients, different samples of the same batch, different days and different positions of the sample cup in the sample drawer of the instrument. A nested ANOVA design has been applied in order to quantify the variances introduced by these variance sources. Since ANOVA is a univariate technique, the necessary variable (wavelength) selection has been performed by the Fisher criterion. The best results, i.e. largest reduction of interfering variability and clearest distinction between different clinical study samples, are obtained with the second derivative spectra.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Near-infrared spectroscopy; ANOVA; Preprocessing methods; Creams

1. Introduction

Near infrared spectroscopy (NIR) is a fast and easy technique that allows both qualitative and quantitative analyses. In the pharmaceutical industry, NIR can be considered as a routine technique for the identification of raw materials [1–4]. More and more NIR applications for quantitative analyses and online process control are reported [5–8]. NIR can also be used as a fast technique in clinical trial studies [9,10] to determine whether or not a sample is a placebo and to which concentration class of the active compound it belongs. This makes NIR a suitable analysis technique in double blind studies.

This work is part of a study in which we investigate the possibility of using NIR as a classification technique for creams according to their active-compound concentration.

This classification is complicated by the fact that, apart from the active-compound concentration, other parameters also have an influence on the NIR spectra. The potentially influential parameters investigated in this study are batches of ingredients, samples of the same batch, days (time) and positions of the sample cup in the sample drawer of the instrument. Besides these studied parameters, also physical variables, like particle size, temperature, humidity, etc. will be reflected in spectral differences. The above mentioned influences justify the common use of spectral preprocessing methods to reduce the effect of these interfering variance sources, thereby increasing the part of the variance due to the concentration differences. The effect on the information content of some of these preprocessing methods was evaluated.

A similar study on tablets and capsules has been performed by Candolfi et al. [11]. Another paper reporting the study of different variability sources on NIR spectra of pharmaceutical drug products has been written by Borer et al. [12]. However these authors stress more on the effects of data

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

collection and treatment parameters of the spectral processing methods.

2. Theory

2.1. Preprocessing methods

Different spectral preprocessing methods are described in the literature. Those applied in this study are summarised below.

2.1.1. Offset correction [13]

The aim of an offset correction is to correct for a parallel baseline shift. This correction is performed by subtracting the mean of the first few (in this study the first five) variables from each spectrum individually:

$$x_{ij,0} = x_{ij} - \bar{x}_{i,1-5}$$

2.1.2. Detrend correction [14]

Detrend correction is applied to spectra in order to remove curvilinearity and baseline shifts. The $\log(1/R)$ values in NIR spectra, with R being the reflectance, often show an increasing trend between 1100 and 2500 nm. To correct for this effect, the baseline is fitted by a second degree polynomial and subsequently subtracted from the spectra:

$$x_{i,j,d} = x_{ij} - b_{ij}$$

with b_{ij} the baseline value of spectrum i according to the second degree polynomial at wavelength j .

2.1.3. First and second derivative [15]

Deriving spectra is used to separate overlapping peaks and to correct for baseline shifts. A drawback of deriving spectra is the enhancement of noise. In order to avoid this drawback, spectra are smoothed by using the Savitzky–Golay algorithm, which is a moving window averaging method: a window is selected where the data are fitted by a polynomial (second degree polynomial in this study). The central point in the window is replaced by the value of the polynomial. For the first derived spectra, a window of seven points and for the second derivative, one of 15 points is used.

2.1.4. SNV correction [16]

Standard normal variate (SNV) correction is applied to remove scatter interferences or scatter differences between the samples. To perform this correction, the mean of each spectrum (\bar{x}_i) is subtracted from the whole spectrum (x_i) and these centred values are divided by the standard deviation (s_i) of each spectrum:

$$x_{i,\text{SNV}} = \left(\frac{x_i - \bar{x}_i}{s_i} \right)$$

2.2. Nested ANOVA [17]

Since each measurement is subject to measurement errors, variance is introduced into the data. The physical parameters to which NIR is susceptible introduce additional variability. ANOVA (analysis of variance) is a statistical method that is used to estimate the degree of variance introduced into the measurements by a certain variance source. In this study, a ‘nested’ ANOVA design is used to estimate the contribution of each of these variance sources. In a nested design each of the variance sources is considered as hierarchically ordered: each of the higher level groups contains subgroups. The variance sources examined in this study are ‘concentration of active compound’, ‘different batches’, ‘different samples’, ‘measuring day’ and ‘position of the sample cup’.

When the contribution of each of the sources is estimated, one knows which is largest and hence requires the strictest standardization during measurement or calibration.

2.3. Wavelength selection [14]

Since ANOVA is a univariate technique and the NIR spectra are recorded at 701 measuring points, individual wavelengths have to be selected. These wavelengths are selected according to the Fisher criterion (FC). This criterion describes the ratio of between-class variance to within-class variance:

$$\text{FC} = \frac{\sum_{j=1}^k n_j (\bar{x}_{ij} - \bar{x}_i)^2}{\sum_{j=1}^k (n_j - 1) s_{ij}^2}$$

In this equation, $j = 1, 2, \dots, k$ is the number of classes, n_j the number of objects in class j , \bar{x}_{ij} the mean absorbance of the objects belonging to class j at the i th wavelength, \bar{x}_i the mean absorbance of the objects belonging to all classes at the i th wavelength, and s_{ij} the standard deviation of the absorbance of the objects belonging to class j at the i th wavelength.

It shows which variables have the highest discriminating power between the classes: wavelengths at which the variances between the classes are large and those within the classes are small result in high FC values. Those wavelengths are important for classification purposes.

Apart from the Fisher criterion, the loadings on the first two principal components (explaining the largest part of the variance in the data) are also considered in the wavelength selection.

3. Experimental

3.1. Material and methods

3.1.1. NIR spectrometer

The measurements are performed using a Bran&Luebbe InfraAlyzer 500[®] (Norderstedt, Germany). The spectra are acquired using the SESAME[®] software coupled to the

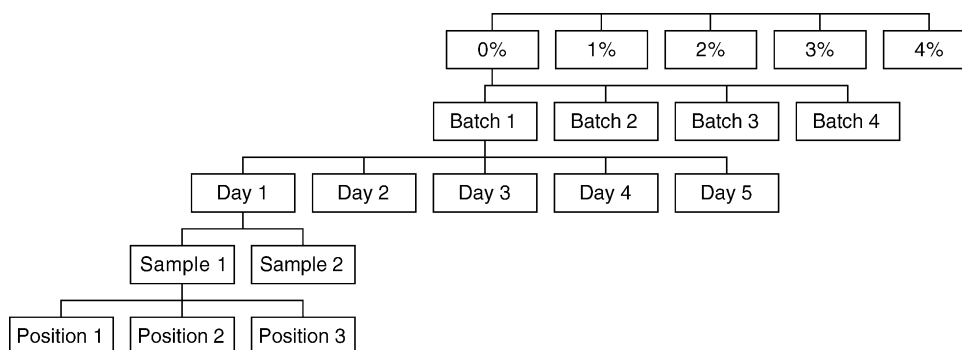


Fig. 1. Nested ANOVA design representing the hierarchical structure of the different variance sources: concentration of active compound, batches, measuring days, samples, cup position in the instrument drawer.

instrument. The samples are measured in a transfectance cup, which is used because of its reproducible path length. The spectra are recorded between 1100 and 2500 nm with a measuring point each 2 nm, which results in a total of 701 measuring points.

3.1.2. Computer calculations

All spectral calculations are performed with the Matlab[®] software (version 5.3, The MathWorks Inc., Natick, MA) using an in-house toolbox.

The nested ANOVA calculations are performed with Microsoft Excel[®] 97 software.

3.1.3. Samples

All cream samples were prepared in-house. As cream base, the *Cremor non ionicus aquosus* from the Formularium Nationale, fifth edition (FN V) [18], has been selected. The creams are composed of cera emulsificans cetomacrogolis FN V [18] (15.0 g), cera liquida (10.0 g), propyleneglycolium (10.0 g), acidum sorbicum (0.2 g), and aqua ad 100.0 g.

This cream has been prepared four times with different batches of ingredients in order to simulate the batch differences that can also be encountered in an industrial environment. For each of the batches, a placebo cream is prepared and four different concentrations of the model substance (herein called the active compound) are added: 1, 2, 3 and 4% (m/m). For the creams with a drug concentration below 4%, water was added so that the sum of the weights of active and water equals 4% of the final weight of the creams.

3.1.4. Variance sources

3.1.4.1. Concentration of active compound. The final aim of the study is to classify the creams according to their concentration of active compound. For this reason, we would prefer that the drug concentration of the creams is the major source of variance at the examined wavelength. The active compound is used in four different concentrations: 1, 2, 3 and 4%. Since in a clinical trial, the active compound has to be compared with a placebo product, a placebo cream is also made. In this way, five different concentration classes will be considered.

3.1.4.2. Different batches. It is known that the use of different batches of excipients can introduce variability into the end product and consequently into the NIR spectrum of the creams. In order to investigate this influence, four different batches of creams are prepared.

3.1.4.3. Different samples. From each of the creams (each batch, each concentration) two samples are measured on the same day. If the creams are homogeneous, this should not be a major variance source, although the filling of the measurement cup can introduce variability in the spectra.

3.1.4.4. Measuring day. Depending on the atmospheric conditions of the laboratory (temperature fluctuations, humidity, etc.) and also on the instrument, small between-day spectral differences can occur. The impact of the time effect on the measurements was examined by repeating the measurements on five different days.

3.1.4.5. Position of the sample. As reported by Candolfi et al. [11] the positioning of the sample can have a large influence on the spectral variance. Although in their case this variability was mainly due to scatter effects on the shiny shell of capsules, we considered it necessary to check the influence of the cup position in the instrument. Each sample is measured three times after rotation of the cup for 120°.

3.1.5. Nested ANOVA design

Taking into account the above-mentioned variance sources, the set-up of the nested ANOVA design is represented in Fig. 1.

4. Results and discussion

4.1. Spectra of active compound and the creams

The spectrum of the pure compound is presented in Fig. 2. The major absorbance bands are situated at 1688, 2270, 2314 and 2396 nm. These peaks are mainly due to C–H and C–C bonds.

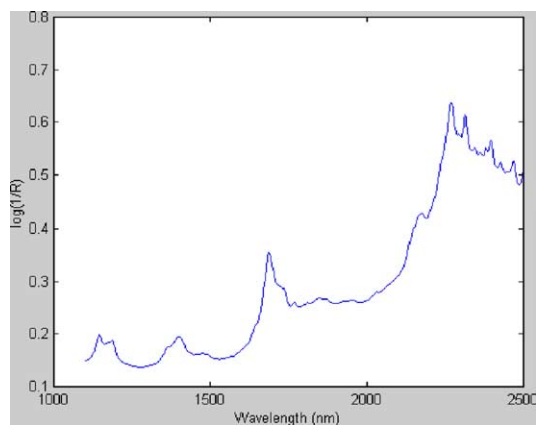


Fig. 2. NIR spectrum of the pure active compound.

The $\log(1/R)$ spectra of the creams are shown in Fig. 3a. The two major absorbance peaks (around 1450 and 1950 nm) are due to the presence of high amounts of water in the creams. Mainly at these peaks (and also at wavelengths between 2400 and 2500 nm), a concentration trend in the $\log(1/R)$ values can be noticed: the spectra of the placebo creams have higher values than the spectra of the drug containing creams. This is due to the way the creams are prepared: the placebo creams have a higher water content than the other creams because for all creams containing less than 4% of active, water has been added in order to adjust for the weight of the active compound (see Section 3.1.3).

As described before, the wavelengths are selected based on the FC and according to the loadings on the first principal components. For the raw spectra, 1686 (based on FC; see Fig. 4a), 2174, 2266 and 2308 nm (based on loadings) are selected. These wavelengths are situated very close to the absorption peaks of the pure compound. In order to make this univariate approach more robust to, among others, wavelength shifts, the sum of the absorbance values at the selected wavelengths and of the values at the two neighbouring wavelengths at both sides of the selected one were calculated.

Table 1 shows the ANOVA results for the selected variables. Generally, the position of the sample in the instrument and the samples itself are the smallest sources of variance for the selected wavelengths. The highest variance source between the different spectra is the concentration of active. Especially at 1686 nm (a peak corresponding to an absorbance peak in the pure compound spectrum), the drug concentration contributes for more than 60% to the total variance. For all selected wavelengths, batch and day variances are comparable. However, based on the $\log(1/R)$ values of the original spectra at the selected wavelengths, a clear classification according to the concentration of the active compound could not be made.

4.2. Effect of offset correction

The offset corrected spectra are shown in Fig. 3b. In this figure, the same trend according to the active concentration

as for the original spectra can be noticed. The loading plots are very similar to those of the original spectra and thus the selected wavelengths are also close to those chosen for the original spectra: 1686, 2174, 2270 and 2314 nm (all based on a combination of loadings on PC 1, PC 2 and on the Fisher criterion (Fig. 4b)). The results of the nested ANOVA calculations are represented in Table 1. Compared with the calculations for the original data, it can be observed that the fraction of variance due to the concentration differences of the active is slightly higher than for the original data. The explained variance due to batch differences of the creams, different samples and sample position is comparable to that for the non-treated spectra. Day differences contribute less to the total variance, especially at 1686 nm. This can be explained by the fact that the first points of the spectra taken on a given day are different from those taken on another day, and by applying an offset correction, these differences can be minimised.

4.3. Effect of detrend correction

The increasing baseline that is typical for NIR spectra is removed by applying detrend on the spectra (see Fig. 3c). In Fig. 3c, the trend according to the active concentration can again be observed, especially around the water peaks and also at wavelengths where the drug compound shows strong absorbances (around 1680 and 2270 nm). The FC plot (Fig. 4c) shows some pronounced peaks which means that at these wavelengths the spectral differences are mainly due to concentration differences of the creams. Compared to the original spectra and the offset corrected spectra, the FC values are 10-fold higher which implies that at these selected wavelengths the discrimination between the different concentration classes will be better. The wavelengths selected to perform the ANOVA calculations are 1180, 1686, 2270 and 2324 nm. The ANOVA results are shown in Table 1. It can be seen that at 1180 and at 2270 nm, the concentration differences represent respectively 97.4 and 98.6% of the variance. Consequently the contribution of the other considered variance sources can be neglected at these wavelengths. These findings can be matched with the high FC values at these wavelengths. At 2324 nm, the concentration contributes less to the total variance: 84.5% of the variance can be attributed to the concentration differences, 6.9% to the batch differences, 4.5% to the day differences, 3.1% to the sample differences and 1% to the position of the sample into the sample drawer.

It can be concluded that the detrend correction reduces the influence of the other factors in such a way that the spectral differences due to the active concentration are amplified.

4.4. Effect of standard normal variate (SNV)

Standard normal variate correction is commonly used to correct for scatter effects due to particle size differences between samples. Although the samples we are using are not powders, we wanted to evaluate its performance. The SNV corrected spectra are represented in Fig. 3d. Compared to

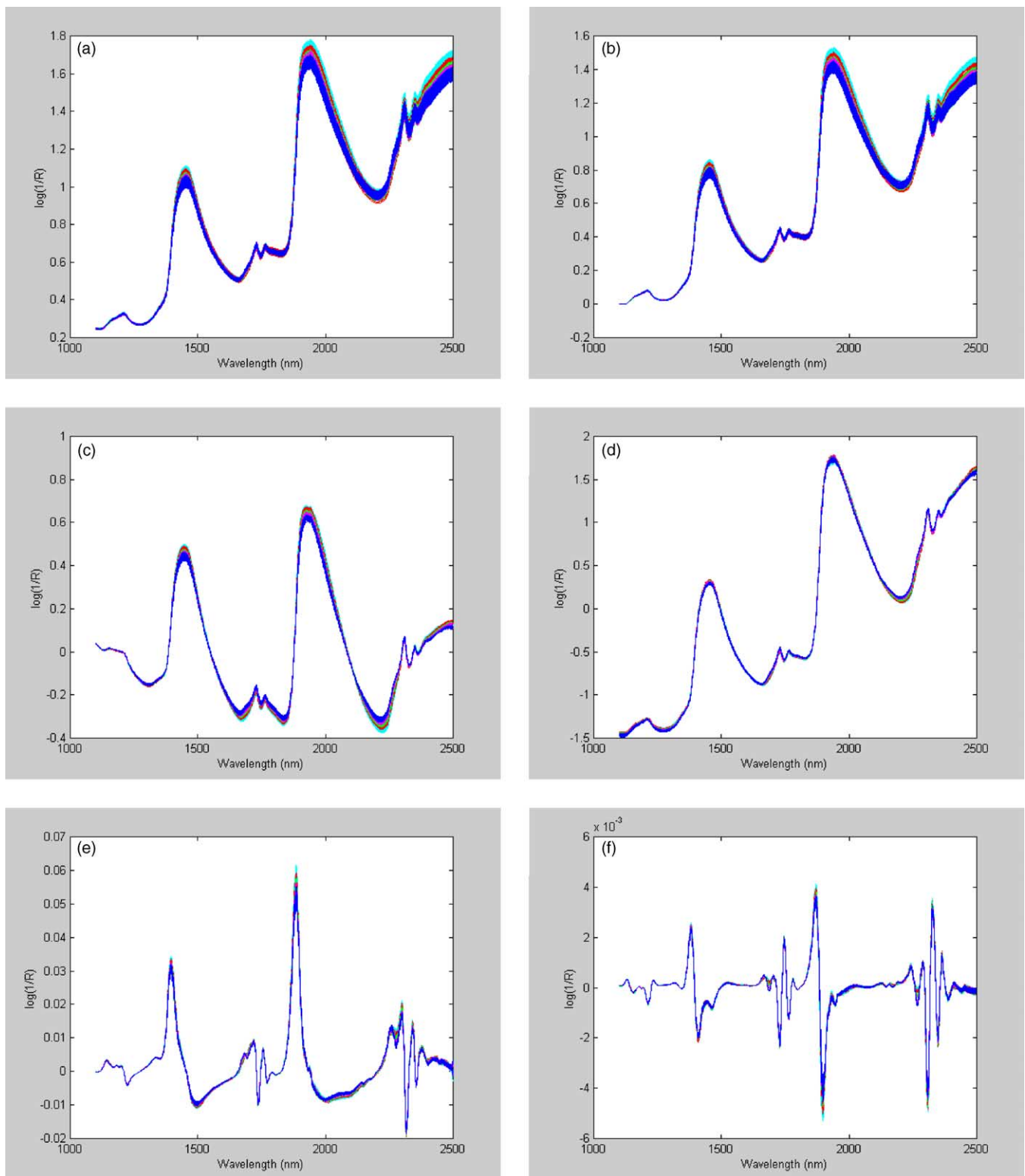


Fig. 3. Cream spectra after different preprocessing methods coloured according to the active concentration (0%: cyano; 1%: red; 2%: green; 3%: magenta; 4%: blue; for interpretation of the references to color in this figure legend, the reader is referred to the web version of the article). (a) Original spectra; (b) offset corrected spectra; (c) detrend corrected spectra; (d) SNV corrected spectra; (e) first derived spectra; (f) second derived spectra.

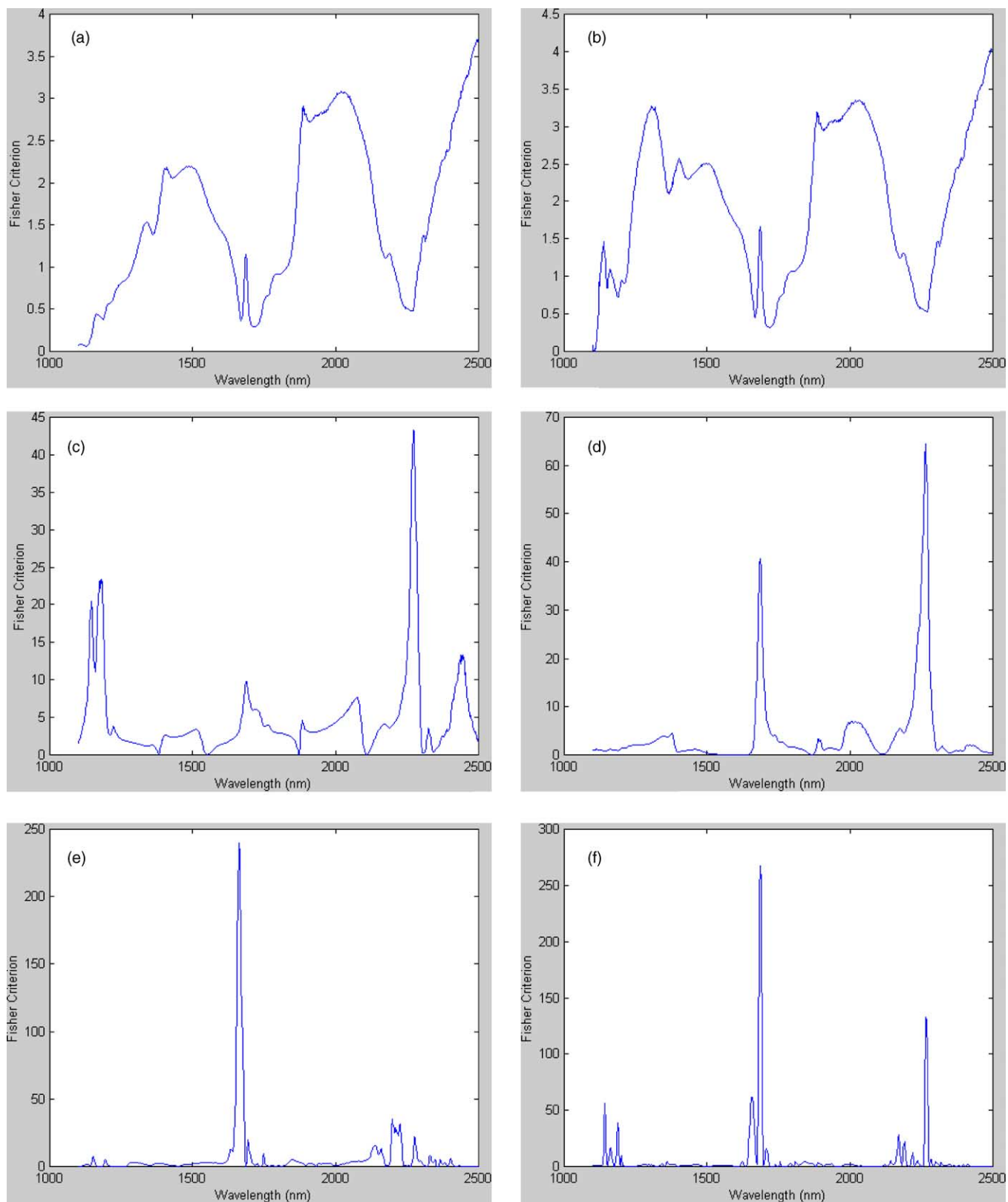


Fig. 4. Fisher criterion obtained from the preprocessed spectra: (a) original spectra; (b) offset corrected spectra; (c) detrend corrected spectra; (d) SNV corrected spectra; (e) first derived spectra; (f) second derived spectra.

Table 1

Variance contribution for each of the different influence factors at the selected wavelengths for the original, offset corrected, detrend corrected, SNV corrected, first and second derived spectra

Wavelength (nm)	Concentration (%)	Batch (%)	Day (%)	Sample (%)	Position (%)
Original spectra					
1686	61.2	15.6	13.3	7.7	2.2
2174	47.1	19.7	19.1	12.2	1.9
2266	28.6	24.7	27.9	16.4	2.4
2308	56.5	14.4	15.7	11.8	1.6
Offset corrected spectra					
1686	70.6	12.1	7.4	7.4	2.5
2174	50.1	19.3	16.5	12.1	2.0
2270	30.0	24.9	26.1	16.4	2.6
2314	57.6	14.5	14.8	11.4	1.7
Detrend corrected spectra					
1180	97.4	1.2	0.2	0.9	0.3
1686	92.7	1.8	2.4	2.5	0.6
2270	98.6	0.5	0.1	0.7	0.1
2324	84.5	6.9	4.5	3.1	1.0
SNV pretreated spectra					
1686	98.3	0.2	0.3	0.7	0.5
2174	89.9	2.9	0.3	5.6	1.3
2266	98.9	0.5	0.2	0.3	0.1
2314	70.9	13.9	5.5	8.7	1.0
First derivative					
1398	73.7	5.3	9.6	8.0	3.4
1662	99.6	0.1	0.1	0.1	0.1
2198	98.8	0.1	0.1	0.3	0.7
2270	91.3	2.8	2.4	2.9	0.6
Second derivative					
1142	99.1	0.1	0.0	0.1	0.7
1686	99.6	0.1	0.1	0.1	0.1
2174	98.3	0.1	0.1	0.2	1.2
2270	99.5	0.1	0.1	0.1	0.2

the original spectra, SNV is able to reduce a large part of the variance between the spectra, which can be observed visually. The FC plot (Fig. 4d) reveals two large peaks with an FC equal to 65 around 2266 nm and an FC equal to 40 at 1686 nm. ANOVA was applied at these wavelengths. Two other wavelengths, 2174 and 2314 nm, were selected based on their high loadings on PC 1. The ANOVA results are represented in Table 1. From these results it can be concluded that SNV has a positive influence on the results. At 1686 and at 2266 nm, the wavelengths selected based on their high FC values, the concentration differences between the creams explain more than 98% of the variance. At the two other wavelengths, selected based on high PC 1 loadings, the other variance sources are responsible for a higher percentage of variance: at 2174 nm, 89.9% of the variance can be explained by the concentration differences between the creams and the sample differences represent 5.6% of variance between the spectra. At 2314 nm, only 70.9% of the variance is explained by the concentration differences. The batch differences represent 13.9% of the total variance at this wavelength and the differences between the samples 8.7%. A possible explanation for the good results of the SNV preprocessing method may be that the size of the emulsion particles is not equally

distributed, resulting in different scatter effects, which are removed by SNV.

4.5. Effect of first derivative

Deriving spectra stresses spectral differences and splits overlapping peaks. This makes that the shape of the first derived spectra (see Fig. 3e) is different from that of the original spectra. It can be seen that the concentration classes now almost can be separated based on their $\log(1/R)$ values at the wavelengths between 1650 and 1700 nm and around 2270 nm where the active substance has high $\log(1/R)$ values (Fig. 5a and b). The FC plot (Fig. 4e) shows one main peak with an FC value of about 250 at 1662 nm. Nested ANOVA was applied at this wavelength. The selection of the other wavelengths was performed based on the PC 1 loadings. Nested ANOVA was thus applied on the results measured at 1398, 1662, 2198 and 2270 nm. Table 1 confirms our findings that 1662 nm is a selective wavelength to perform classification of the samples according to their active concentration: the variance between the spectra due to concentration class differences amounts to 99.6%. At 2198 nm, the contribution of the concentration differences to the total variance is 98.8%.

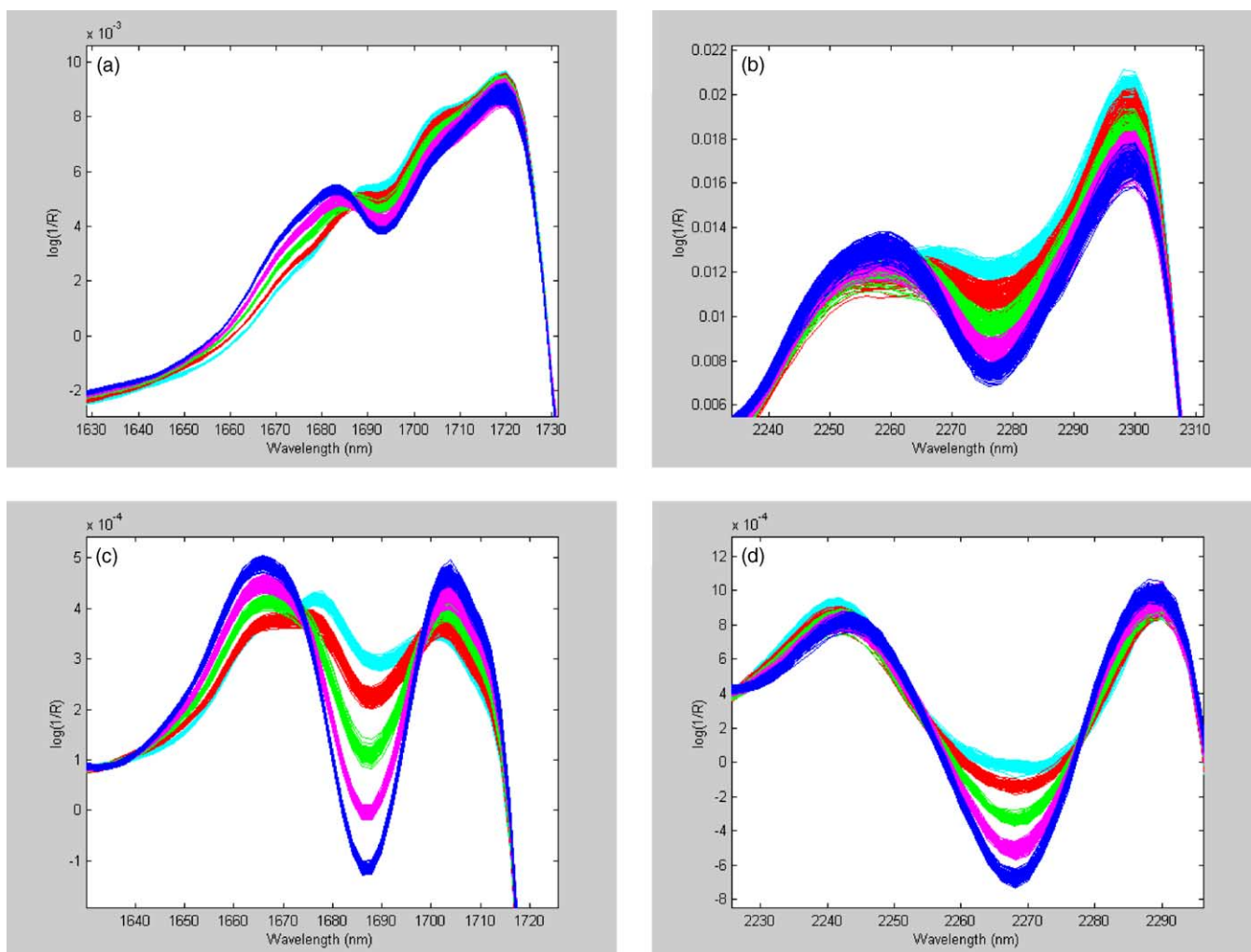


Fig. 5. (a) First derived spectra between 1640 and 1730 nm, (b) first derived spectra between 2240 and 2310 nm, (c) second derived spectra between 1640 and 1720 nm, (d) second derived spectra between 2240 and 2290 nm (0%: cyan; 1%: red; 2%: green; 3%: magenta; 4%: blue; for interpretation of the references to color in this figure legend, the reader is referred to the web version of the article).

These results show that deriving spectra filters out spectral differences introduced by batch, day, sample or sample positioning.

Although on the plot of the spectra (Fig. 5b), 2270 nm also seems to be an interesting wavelength to perform classification according to the drug content of the creams, the nested ANOVA results reveal that also batch, day and sample differences contribute to a small extent to the variance between the different spectra at this wavelength. 91.3% of the variance at this wavelength can be attributed to the differences in the drug content between the creams while the other above mentioned variance sources each contribute for approximately 2.5% to the variance between the spectra.

4.6. Effect of second derivative

The second derived spectra are shown in Fig. 3f. At 1686 nm, the different classes can be clearly separated (see

Fig. 5c). The region around 2270 nm (Fig. 5d) also shows a good separation according to the different concentrations of the active compound of the creams. In this region, the best separation can be obtained at 2268 nm. These two wavelengths are close to the absorbance peaks of the active spectrum: 1688 and 2270 nm (Fig. 2).

As could be expected, the FC plot (Fig. 4f) shows two major peaks, i.e. at these wavelengths. The same wavelengths also have high loadings on the first PC and thus will be considered for the nested ANOVA calculations. Two additional wavelengths (1142 and 2174 nm) are also selected based on their high FC values. At 1142, 1686 and 2270 nm, the variance almost exclusively (more than 99%) can be attributed to the active content of the creams (see Table 1). At 2174 nm, 98.3% of the variance was explained by the concentration differences between the spectra and 1.2% of the variance is due to the different positions of the sample cup.

5. Discussion

Comparing all results, it can be stated that all applied preprocessing methods reduce the variance contributions of the batch, day, sample and positional differences. While the offset correction results in a small reduction of these variance contributions, all other studied methods enhance the concentration differences to more than 90% of the total variance around a typical absorbance wavelength of the active compound (e.g. 1686 nm). For the first and second derivative spectra, the concentration determines almost the total variance at this wavelength (>99%).

Generally, the wavelengths selected according to the Fisher criterion give better results than those selected according to high PC 1 or PC 2 loadings. This is not unlogic because the FC criterion specifically selects those wavelengths where the discrimination between the (concentration) classes is highest, while the PC loadings only indicate wavelengths where the variance between the spectra is high not taking into account the origin of the variance source.

6. Conclusion

The results show that most of the applied preprocessing methods reduce the variance introduced in the original spectra by the considered variance sources. To highlight the spectral differences due to the concentration differences of the creams at the selected wavelengths and to reduce the influence of the other investigated variance sources, the best suited of the applied techniques is the second derivative of the spectra. Using this preprocessing technique, the contribution of the concentration to the total variance is higher than 99% at three out of the four selected wavelengths.

Acknowledgements

The authors would like to thank Professor Plaizier-Vercammen of the Department of Pharmaceutical Technology of the Vrije Universiteit Brussel (Belgium), for providing us with the materials needed for the preparation of the creams.

References

- [1] A. Candolfi, R. De Maesschalck, D.L. Massart, P.A. Hailey, A.C.E. Harrington, *J. Pharm. Biomed. Anal.* 19 (1999) 923–935.
- [2] M. Blanco, M.A. Romero, *Analyst* 126 (2001) 2212–2217.
- [3] E. Dreassi, G. Ceramelli, P. Corti, P.L. Perruccio, S. Lonardi, *Analyst* 121 (1996) 219–222.
- [4] K. Kramer, S. Ebel, *Anal. Chim. Acta* 420 (2000) 155–161.
- [5] A.C. Moffat, A.D. Trafford, R.D. Jee, P. Graham, *Analyst* 125 (2000) 1341–1351.
- [6] S. Vaidyanathan, S.A. Arnold, L. Matheson, P. Mohan, B. McNeil, L.M. Harvey, *Biotechnol. Bioeng.* 74 (2001) 376–388.
- [7] M. Andersson, M. Josefson, F.W. Langkilde, K.G. Wahlund, *J. Pharm. Biomed. Anal.* 20 (1999) 27–37.
- [8] J. Rantanen, E. Räsänen, J. Tenhunen, M. Käsäkoski, J.P. Mannermaa, J. Yliruusi, *Eur. J. Pharm. Biopharm.* 50 (2) (2000) 271–276.
- [9] A. Candolfi, W. Wu, D.L. Massart, S. Heuerding, *J. Pharm. Biomed. Anal.* 16 (1998) 1329–1347.
- [10] C. Tso, G.E. Ritchie, L. Gehrlein, E.W. Ciurczak, *J. Near Infrared Spectrosc.* 9 (2001) 165–184.
- [11] A. Candolfi, D.L. Massart, S. Heuerding, *Anal. Chim. Acta* 345 (1997) 185–196.
- [12] M.W. Borer, X. Zhou, D.M. Hays, J.D. Hofer, K.C. White, *J. Pharm. Biomed. Anal.* 17 (1998) 641–650.
- [13] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, *J. Pharm. Biomed. Anal.* 21 (1999) 115–132.
- [14] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, *Anal. Acta Chim.* 315 (1995) 243–255.
- [15] P.A. Gorry, *Anal. Chem.* 62 (1990) 570–573.
- [16] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, *J. Near Infrared Spectrosc.* 2 (1994) 43–47.
- [17] R.R. Sokal, F.J. Rohlf, *Biometry*, 2nd ed., W.H. Freeman and Company, New York, 1981.
- [18] *Formularium Nationale, Editio Quinta, Algemene Pharmaceutische Bond (APB), Brussel, 1977.*