Chemometrics and Intelligent Laboratory Systems 122 (2013) 93–102

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Optimized bucketing for NMR spectra: Three case studies

S.A.A. Sousa [a], Alviclér Magalhães [b], Márcia Miguel Castro Ferreira [a,*]

[a] Laboratory for Theoretical and Applied Chemometrics, Institute of Chemistry, University of Campinas — UNICAMP, Campinas, SP 13084-862, P.O.B. 6154, Brazil
[b] Department of Inorganic Chemistry, Instituto Nacional de Ciência e Tecnologia em Bioanalítica, University of Campinas — UNICAMP, Campinas, SP 13083-970, Brazil

ARTICLE INFO

Article history:
Received 8 June 2012
Received in revised form 17 December 2012
Accepted 16 January 2013
Available online 24 January 2013

Keywords:
OBA
Spectral alignment
Wine
Biodiesel
Brain tumor

ABSTRACT

The use of nuclear magnetic resonance (NMR) data coupled to chemometric methods has become increasingly popular in the last decade. However, a serious drawback of these approaches is the common misalignments of ^{1}H NMR spectra. To overcome this problem, bucketing or binning techniques have been used. In this work, an algorithm is proposed to perform an optimized bucketing that yields better results than the conventional bucketing implemented in some commercial software. The improvement proposed here for optimized bucketing deals with the bucket boundaries, which are defined by local minima from average NMR spectrum over all samples. Applicability of the new algorithm, named OBA (optimized bucketing algorithm), is demonstrated for real data sets in comparison to other alignment approaches and conventional bucketing.

1. Introduction

NMR spectroscopy is a powerful, versatile, non-destructive, non-invasive, and reproducible technique. Besides being adequate for structure elucidation, it can be used to analyze complex samples without physical separation. These advantages may be explored by means of suitable chemometric tools to provide several types of useful information in pattern recognition, adulteration detection, metabolic profiles, and other fingerprint applications [1–5].

Exploration of the chemical information encoded in all NMR data suffers from the misalignments that are frequent, especially in ^{1}H NMR spectra. They arise due to various factors, such as instrumental instabilities, pH, ionic strength, and temperature, among others, and they can lead the statistical analysis to incorrect interpretation of results, because the bi- or multi-linear assumptions of the chemometric approaches on the nature of the data are not considered properly [6–9].

In the literature, different methods have been proposed to correct these misalignments. One popular approach of low computational cost is called *binning* or *bucketing* [6]. Alternatively, more elaborated methods such as correlation optimized warping (COW) [10–14], dynamic time warping (DTW) [10,12], correlation-shifting (coshift) [15], and interval-correlation-shifting (icoshift) [7,15] have also been employed. Bucketing is theoretically simpler than the other cited methods.

Actually, bucketing performs a data reduction by grouping spectral responses, not being strictly a method to align data. In the conventional method, the spectra are divided into evenly spaced windows, named bins or buckets, whose width commonly ranges between 0.01 and 0.05 ppm. The intensities inside each bin are summed, so that the area under each spectral region is used instead of individual intensities. Therefore, a new smaller set of variables (each one is the result of the sum of intensities) is created and, as the width of the buckets is set to cover the chemical shift variability around the peaks, the misalignment tends to be overcome [6,16,17].

Eq. (1) summarizes the bucketing procedure applied to a data matrix $\mathbf{X}(I,J)$ with samples in rows, variables in columns and elements x_{ij}, where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. For each sample i, x_{ij} is an intensity in the raw signal at point j. The N parameter is the number of data points in each bucket and can be calculated by the ratio between the bucket width and the sampling interval (e.g., if the bucket width is 0.05 ppm and the sampling interval is 0.0005 ppm, the N parameter will be equal to 0.05/0.0005 = 100 points). The sampling interval changes between different experiments because it depends on how the acquisition of the *Free Induction Decay* (FID) is performed, that is, on the parameters set in the NMR experiment, such as, acquisition time, total digitalized data points and spectral width. The K parameter is the final number of buckets and equal to the integer part of the ratio J/N. Therefore, according to Eq. (1), the new variable domain axis k is created where new intensities z_{ik} are organized in a new matrix $\mathbf{Z}(I,K)$.

$$z_{ik} = \sum_{j=N*(k-1)+1}^{N*k} x_{ij} \quad k = 1, 2, \ldots, K \tag{1}$$

A drawback for this method is that some areas from the same resonance signal can appear in two or more bins, splitting the chemical information in question. This occurs because conventional bucketing

* Corresponding author. Tel.: +55 19 3521 3102; fax: +55 19 3521 3023.
E-mail address: marcia@iqm.unicamp.br (M.M.C. Ferreira).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.01.006

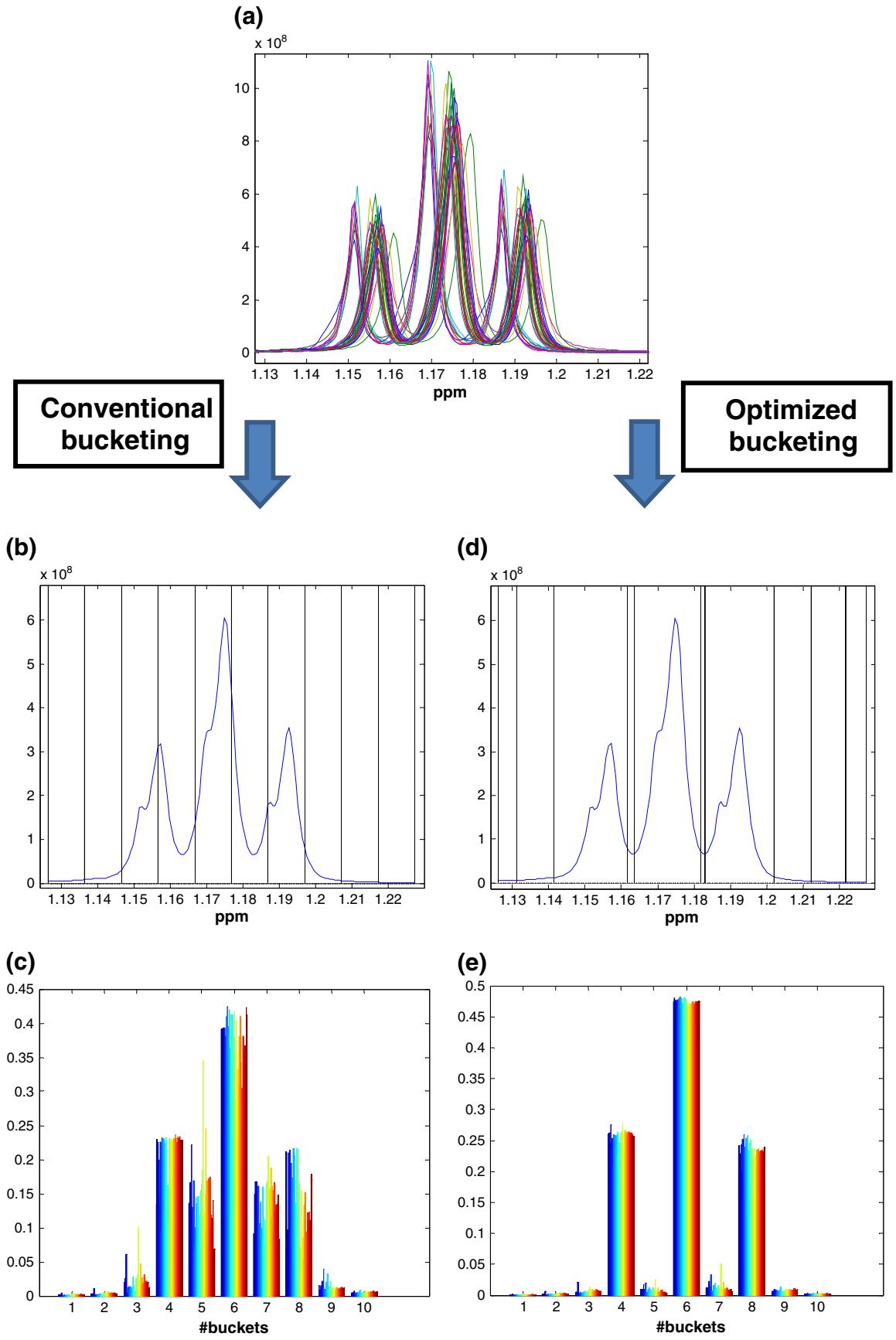

Fig. 1. Scheme of the conventional and optimized bucketing procedures. (a) Simulated NMR spectra with misalignments. (b) Average simulated spectrum and the bucket boundaries (vertical lines) delimited by conventional bucketing, with buckets of size 0.01 ppm. (c) Bucket values from each sample shown through a bar plot obtained by conventional bucketing. (d) Average simulated spectrum and the bucket boundaries (vertical lines) delimited by OBA, with initial buckets of size 0.01 ppm and slackness of 50%. (e) Bucket values from each sample shown through a bar plot obtained by OBA. In (c) and (e) the bars' height are related to the value of the integrals that were normalized to a total sum equal to one.

uses rigid boundaries. Despite this, several papers in the literature [1,2,18,19] effectively use this methodology.

Fig. 1a shows a set of simulated misaligned NMR spectra. Fig. 1b and c illustrates the conventional bucketing procedure. As can be seen in Fig. 1b, which presents the average simulated spectrum with the bucket boundaries denoted by vertical lines, the bucketing with 0.01 ppm width (Fig. 1b) is unable to properly isolate the signals. As result, in Fig. 1c, where the buckets' values for each sample are shown through the colored bars (the bar heights are related to the values of the integrals that were normalized to the total sum equal to one), five important variables are observed, containing the principal information on the data set, which actually has three signals. Hence, this could seriously hamper interpretations, for example, when principal component analysis (PCA) is used.

The drawback cited above can be overcome by having the bin boundaries adjustable to minima, in order to provide optimized buckets of different sizes. In fact, a similar type of solution has already been proposed in the literature, as for example, the methodology for binning implemented in the commercial software ACD/Labs™ (Toronto, Canada) named intelligent bucketing [16,17]. In this method, the software chooses integral divisions based on local minima, thus searching for a better way of slicing up the spectra and avoiding the problem of the conventional bucketing. However, the software is not open source and the method for finding minima has not been reported. In other work, Davis et al. [20] proposed a methodology named adaptive binning, where undecimated wavelet transform is used for denoising and to find all minima in a reference spectrum (maximum over each sample) performing the integration between these minima for each individual spectrum. However, in the decomposition a predefinition of both wavelet level and basis functions is necessary. Thus, there is a dependence on the number of levels in the decomposition, besides the threshold for denoising. Other alternatives for the traditional bucketing have been proposed in the recent literature, named Gaussian binning [21] and dynamic adaptive binning [22], but as the method proposed by Daves et al., these methodologies require a higher level of user expertise, being more complex than the algorithm presented here.

The aim of this work is to present a bucketing method that optimizes bucket sizes by setting their boundaries at the local minima determined through the average NMR spectrum and is designated as "optimized bucketing algorithm" (OBA). The proposed approach is simple, accessible to any user as an in-house routine built in Matlab code, and is available for free download at http://lqta.iqm.unicamp.br.

2. Methodology

2.1. The "optimized bucketing algorithm"

The optimized bucketing algorithm (OBA) that is being proposed is a modification of the conventional bucketing procedure. In order to define buckets with variable size, but common to all samples, the average spectrum $\bar{\mathbf{x}}^T$ is used, where T superscript represents the transpose operation and each element $\bar{x}_j$ is the mean of j-th column of $\mathbf{X}$. First, two parameters are defined: 1) the initial bucket width in ppm, which is converted into the number of points N by the algorithm, using the sampling interval calculated from the ppm axis, requested as input and 2) the *slackness*, which is given as percentage of N and defines how far the boundary can move while searching for the local minima in the average spectrum. The *slackness* is converted in the algorithm into the parameter s, expressed as $slackness * 0.01 * N$. Therefore, using as inputs initial bucket widths of 0.04 ppm, for example, in a data set with sampling intervals of 0.0004 ppm ($N = 0.04/0.0004 = 100$ points) and *slackness* of 50% ($s = 50 * 0.01 * 100 = 50$ points), the bucket sizes could range from 0.02 to 0.06 ppm (100 ± 50 points), depending where the local minimum is found. The outputs are the pretreated matrix with dimensions (I,K), the optimized boundaries and the resulting size of each bucket, both in ppm.

From the mathematical point of view, OBA can be reasoned as follows: Once the bucket width and the slackness are known, the vector $\mathbf{v}^T$ (Eq. (2)), whose elements define the bucket boundaries, is created. The first bucket starts in variable $j = 1$ and the last bucket ends at variable $j = J$ from the mean spectrum $\bar{\mathbf{x}}$ and these are the first and last elements of vector $\mathbf{v}^T$. The other elements of vector $\mathbf{v}^T$ are, in fact, the index q of the q-th element from $\bar{x}_q$ which corresponds to the local minimum in the region delimited by $\bar{x}_{N*t-s}$ and $\bar{x}_{N*t+s}$, where $t = 1, 2, \ldots, T$, with T being equal to the integer part of $(J/N) - 1$, as defined in Eq. (3).

$$\mathbf{v}^T = [1, \cdots, q, \cdots, J] \tag{2}$$

$$\bar{x}_q = \min\left(\bar{x}_{N*t-s} : \bar{x}_{N*t+s}\right) \tag{3}$$

The elements from $\mathbf{v}^T$ replace the integration limits in Eq. (1), thus providing the optimized bucketing, for each sample i, as shown in Eq. (4), where $\mathbf{v}(k)$ is the k-th element from vector $\mathbf{v}$. The new matrix $\mathbf{Z}(I,K)$ is obtained where the new intensities z_{ik} are organized and the new variable domain axis k is created.

$$z_{ik} = \sum_{j=\mathbf{v}(k)}^{\mathbf{v}(k+1)} x_{ij} \quad k = 1,\ 2,\ \ldots, K = \text{length}(\mathbf{v}) - 1 \tag{4}$$

Fig. 1d and e show a scheme for OBA when applied to the misaligned spectra. The number of buckets K is the same as before (Fig. 1c), but it is clear from Fig. 1d that the new algorithm was able to set the boundaries at the local minima (vertical lines). On both sides of the central peak, the buckets become narrow because in the search for local minima, the boundaries, initially set at non-optimized positions, tend to move close to the region between the peaks. As the result, in Fig. 1e, where the bucket values for each sample are shown through the colored bars, only three important buckets are observed, as expected, since the simulated data has only three peaks. The superior performance of the proposed methodology over the conventional bucketing procedure is easily visible.

An important issue to be considered in OBA is the choice of the best combination between the bucket width and the *slackness* for each data set. Visual inspection on the misalignment extents at the baseline could be of great help in defining these parameters. Also, some criterion, as for example, the variance explained in first principal components from a principal component analysis (PCA) or the simplicity value [11] of the bucketed matrices could give a reasonable estimate of the two input combinations. The simplicity value is related to how well a data set is aligned and this parameter could be used to evaluate the results from a given bucket width and slackness in an optimized bucketing procedure. Finally, it is advisable not to use very large bucket widths, because this approach has an inherent decrease of resolution at the chemical shift axis, thus there must be a compromise between the gain from correcting misalignments and the reduction in the number of variables. For the data sets used in this work, the input parameters have been determined by inspecting the baseline misalignments and the plots of the obtained buckets, in order to choose those with lower decrease in resolution.

Aiming to test the applicability of OBA and to compare it with other methods from literature, three NMR data sets were selected and they are described below. Two of these data sets were extracted from the literature (wine and brain tumor data) and one was acquired in our lab (biodiesel–diesel data).

2.2. Wine data

Wine ^{1}H NMR spectra have been studied by Larsen et al. [14], where the methodology for data acquisition has been described. The data matrix $\mathbf{X}$(40x8712), downloaded from http://www.models.kvl.

dk, is composed of 40 table wine samples (distributed in white, red and rosé) covering the region from 0.50 to 6.00 ppm, and the lactic acid content (one important organic acid for the taste profile of the wine [14]) from each sample. In this region, one can observe several peaks attributed to ethanol, organic acids, carbohydrates and, in smaller quantities, polyphenols, other aromatic compounds and colorants. All these peaks present misalignments that occur mainly due to the differences in the sample pH, which were not adjusted before the analysis. OBA was applied to the NMR spectra using 0.05 ppm as the initial width of the buckets and slackness of 50%. For this data set the sampling interval was about 0.00063 ppm leading to $N = 0.05/0.00063 = 79$ points and $s = 50 * 0.01 * 79 = 39$ points (these parameters were rounded). As a result, the matrix of buckets **Z** with dimensions (40 × 109) was obtained. Buckets with 0.05 ppm were used for the conventional bucketing procedure and in this case the **Z** matrix has dimensions (40 × 110). Since the lactic acid content was available, pretreated (by conventional and optimized bucketing) and raw data, using the entire spectra and a selected region, were used to build partial least squares (PLS) and multiple linear regression (MLR) models on mean centered data matrices. All the models had their predictive performance evaluated by the leave-one-out cross-validation approach where the coefficient of determination (R^2) was calculated. For construction of the models, in-house routines built in Matlab code were used (The MathWorks, Natick, MA, USA).

2.3. Biodiesel–diesel data

One hundred samples of biodiesel–diesel blends collected in gas stations in the state of São Paulo were supplied by the Analytical Center of the Institute of Chemistry, University of Campinas. The samples were classified as metropolitan (the city of São Paulo and metropolitan region of Campinas) and non-metropolitan (other smaller cities of the state), according to their regions of commercialization. The NMR analysis was carried out on a Bruker Avance DRX400 spectrometer 400.13 MHz to ^{1}H at room temperature, using 550 μL of neat biodiesel samples in a 5 mm Bruker BBO probehead without spinning using the standard 90° pulse sequence for ^{1}H. Homogeneity of the field was obtained by inspections of the spectrum of the standard lineshape sample (0.3% chloroform in acetone d6). This field condition was used for all samples during all days of the analyses. All spectra were acquired with 32 K points in the time domain, 20 ppm (^{1}H) and 16 scans. FIDs were processed with TOPSPIN 2.1 with 64 K points, multiplied by an exponential window function with line broadening constant of 0.3 Hz (^{1}H) and normal Fourier transformation. The phase of the final spectra was adjusted one by one by direct inspection; the baseline was made using an automatic linear function. All the spectra were referenced using a digital lock field position obtained using TMS in acetone at 0 ppm. The ^{1}H NMR spectra were organized in a data matrix **X** with dimensions (100 × 15,850) relative to the region from 0.02 to 10.00 ppm, which was reduced into buckets by the conventional way, using buckets with widths of 0.05 ppm and OBA with slackness of 50% (sampling interval = 0.00063 ppm, $N = 79$ points, $s = 39$ points), followed by normalization to unit area. The bucketing procedures provided **Z** matrices with dimensions (100 × 200) and (100 × 199) for the conventional and optimized ones, respectively. The raw data and the bucketed normalized matrices were mean centered and submitted to exploratory analysis by principal component analysis (PCA) using the software Pirouette 3.11 (Infometrix, Seattle, WA, USA).

2.4. Brain tumor data

The ^{1}H NMR spectra from human brain tumor extracts have been studied by Faria et al. [23] where the methodology for data acquisition has been described. From the ^{1}H NMR spectra reported, 16 and 13 spectra corresponding to non-neuroglial (NN) and high-grade neuroglial (Hg) tumors, respectively, were selected for the present study. The spectra corresponding to the region between 1.22 and 4.25 ppm were organized in a data matrix **X** with dimensions (29 × 4964). OBA was applied to the NMR spectra using 0.002 ppm as the width of buckets and *slackness* of 50%. For this data set the sampling interval was about 0.00060 ppm leading to $N = 0.002/0.00060 = 3$ points and $s = 50 * 0.01 * 3 = 2$ points (these parameters were rounded). As a result, a matrix **Z** of buckets with dimensions (29 × 1416), was obtained. Partial least squares discriminant analyses (PLS-DA) with the leave-one-out cross-validation approach were used to build classification models (two types of tumors, NN and Hg) from the raw data set and the data set after pretreatment with OBA using a **y**-vector of classes where the value 1 was set to NN tumors and the value −1 was set to Hg tumors. The analyses were carried out on mean centered matrices. The performance of the PLS-DA model of each data set was evaluated through the number of misclassifications (NMC) as diagnostic statistics and related to the diagnostic statistics obtained from 10,000 permutation tests computed using the permuted vector of classes (**y**-randomization). The NMC values were calculated as the sum of false positives and false negatives in the models, obtained by relating the predicted class labels to a discriminative threshold defined using estimated distributions for the predicted values in each class. The threshold is selected at the point where the two estimated distributions are equal, these distributions being approximately normal (Gaussian distributions with the mean and the standard deviation of all the predictions for each class). For the models from the permutations a null hypothesis H_0 assumes that there is no difference between the groups. Thus, the statistical significance of the number of misclassifications of the models is assessed by comparing them to values of their null hypothesis distributions H_0 [24,25]. From these comparisons each *p*-value (one plus the number of elements in the null distribution that are smaller or equal to the NMC for the unpermuted model divided by the number of permutation tests, in this case, 10,000) was calculated [24] and associated to the significance threshold $\alpha = 0.05$. Additional details about this performance evaluation may be found elsewhere [24,25]. All analyses were performed using in-house routines built in Matlab code (The MathWorks, Natick, MA, USA).

3. Results and discussion

3.1. Case study: wine data

In the wine data set a broad range of peak shifts is observed strongly dependent on the sample pH. This can be seen in Fig. 2 in the NMR spectra for all samples, in the zoomed regions related to the signals from ethanol (quartet — methylene group) and lactic acid (doublet — terminal methyl group). The alignment of these spectra using correlation optimized warping (COW) [14] and icoshift [15] approaches, has already been reported in the literature, besides the results for PLS regression models.

Table 1 summarizes the PLS regression results (on mean-centered data) for lactic acid content using the raw data and pretreated data by conventional and optimized bucketing. For comparison, the results from icoshift [15] and COW [14] aligned matrices are also included. The MLR models for bucketed regions associated with chemical shifts in the region of 1.35 and 1.45 ppm corresponding to two buckets (#92 and #93 from conventional bucketing and #91 and #92 from optimized bucketing) are also shown in this table. The two buckets in each situation present little correlation (correlation coefficients 0.1363 and −0.0649 from the conventional and optimized bucketing, respectively), thus there is no redundancy in the MLR models.

Based on the R^2 values, the worst results were those relative to the whole spectral range (smallest R^2 values). Possibly, the non-linearity introduced by the misalignments imposes to PLS model, the use of a higher number of factors (four latent variables) to capture the correlation between the spectral data and the lactic acid content. Moreover, it can be seen that the PLS models built after the COW and icoshift correction are worse than the ones for the raw data, leading to a lower value of

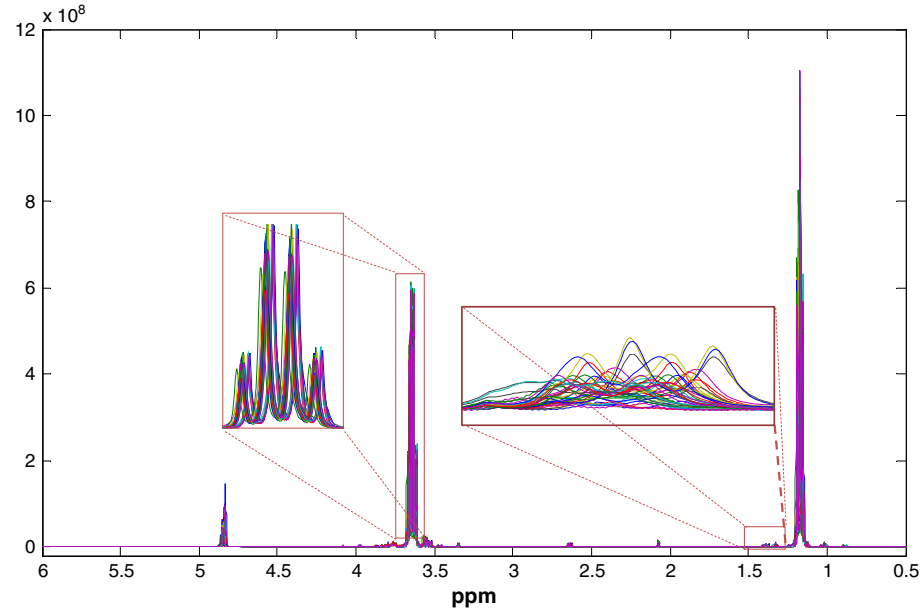

Fig. 2. NMR spectra for wine samples and the zoomed regions related to the signals from ethanol (quartet — methylene group) and lactic acid (doublet — terminal methyl group).

R^2, with the same number of latent variables, which suggests that these corrections were ineffective to solve the misalignments associated to the signals of the lactic acid. Indeed, in the original publication [14] relative to the COW procedure, the authors used an additional tool based on multistep interval procedure utilizing coshift, for improving the regression model of the lactic acid (second line for COW in Table 1), since the COW method was only able to correct for the shift of the dominant ethanol peaks. Also, in the original publication [15] about icoshift alignment, the authors improved the lactic acid regression model (second line for icoshift in Table 1) using custom intervals, defined through the prior knowledge of the NMR peak assignments. Thus, both COW and icoshift did not work well for the minor constituent lactic acid, requiring further optimizations.

The bucketing procedures presented superior performance for the PLS regression models, when the whole spectral range was used (Table 1), without the necessity of subsequent manipulations. The conventional bucketing pretreatment provided a PLS regression model with 4 latent variables, $R^2 = 0.63$ and RMSECV = 0.310 g L^{-1} (Table 1), while the optimized bucketing pretreatment reached $R^2 = 0.84$ with the same number of factors (4 LV) and with a lower error in the cross-validation (RMSECV = 0.200 g L^{-1}). To avoid overfitting, the number of latent variables was chosen by observing the plots of RMSECV *versus* the number of factors, as shown in Fig. 3. As can be seen in this figure, for example, in the curve relative to PLS models after OBA, the models with more than five latent variables are overfitted. The better results from the optimized bucketing for the whole spectra may be associated to the advantage of this methodology to concentrate the signals in a few buckets, avoiding peak splitting. Indeed, the conventional bucketing is also able to allocate the signals in a few buckets, but in this case peak splitting is not completely avoided. The results suggest that for this data set the bucketing procedures are advantageous, considering the serious misalignments of the lactic acid signals.

Using only the signal from terminal methyl of lactic acid, the icoshift method illustrated the best PLS model with 2 LVs (RMSECV = 0.104 g L^{-1} and $R^2 = 0.96$), while the PLS model obtained after the

Table 1
Results from PLS regression and MLR models for lactic acid content (reference values, mean = 1.03 g L^{-1} and standard deviation 0.51 g L^{-1}).

Pretreatment	Spectral region (ppm)	#LV[a]	RMSECV[b] (g L^{-1})	R^{2c}
None (raw data)	0.5–6.0	4	0.369	0.48
	1.35–1.45	3	0.113	0.95
	1.35–1.45	2	0.136	0.93
icoshift[d]	0.5–6.0	4	0.400	0.39
	1.35–1.45	2	0.104	0.96
COW[e]	0.5–6.0	4	0.440	0.27
	1.3–1.6	3	0.100	0.96
Conventional bucketing	0.5–6.0 (bucketed)	4	0.310	0.63
	1.35–1.45 (bucketed)[f,g]	–	0.114	0.95
Optimized bucketing	0.5–6.0 (bucketed)	4	0.200	0.84
	1.35–1.45 (bucketed)[f,h]	–	0.124	0.94

Parameters are based on leave-one-out cross-validation.
[a] #LV = number of latent variables.
[b] RMSECV = root mean squared error of cross validation.
[c] R^2 = coefficient of determination.
[d] From reference [15].
[e] From reference [14].
[f] MLR models.
[g] Buckets #92 and #93 (from 1.4608 to 1.4110 ppm = bucket #92 and from 1.4110 to 1.3611 ppm = bucket #93).
[h] Buckets #91 and #92 (from 1.4804 to 1.4261 ppm = bucket #91 and from 1.4261 to 1.3573 ppm = bucket #92).

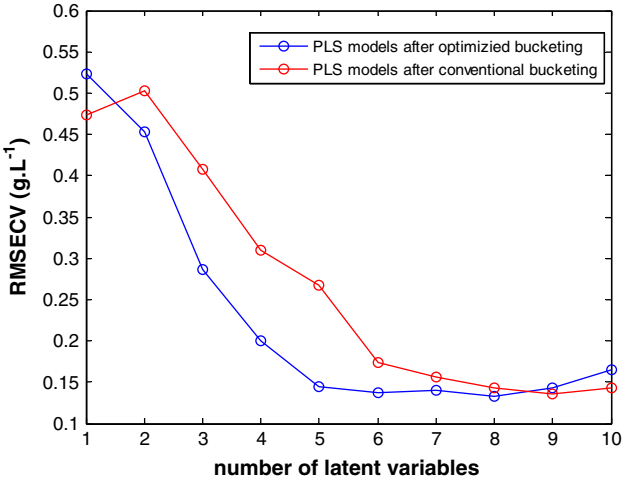

Fig. 3. RMSECV values *versus* the number of latent variables obtained in the leave-one-out cross-validation approach for PLS models after optimized (blue line) and conventional (red line) bucketing.

correction provided by the COW approach yielded a regression model with similar statistics (RMSECV = 0.100 g L^{-1} and $R^2 = 0.96$), but was more complex with 3 LVs. The MLR models obtained from each bucketing procedure showed similar performance to that one obtained from the icoshift procedure. This fact constitutes a great advantage for the bucketing approaches, because the MLR models are unbiased and simpler than PLS models and they do not require the optimization of latent variable numbers.

The MLR model achieved after the use of the conventional bucketing pretreatment (RMSECV = 0.114 g L^{-1} and $R^2 = 0.95$) was slightly better than the one obtained after OBA (RMSECV = 0.124 g L^{-1} and $R^2 = 0.94$), but they are still comparable, as verified by an F test on the predicted lactic acid contents obtained in the leave-one-out cross validation procedure for both MLR models, at the $\alpha = 0.05$ significance level with the number of degrees of freedom of numerator and denominator equal to 39 (p-value = 0.9926). Also, a t test was performed on the means of the predicted lactic acid contents determined by the two MLR models, where it was found that the means were not significantly different at the $\alpha = 0.05$ significance level (p-value = 1.00).

For the specific region studied, the signals from the ethanol ^{13}C satellites can be found (between 1.30 and 1.35 ppm) very close to the lactic acid peaks, which, in part, may have hampered the search for local minima and consequently affected the regression models. Actually, the two buckets (Table 1) used for both MLR models (after conventional and optimized bucketing) do not cover exactly the same region. Considering this problem, one possible solution is to modify the parameters in the OBA (*slackness* and initial bucket width), but this was not performed in the present work. Despite this, OBA proposed herein worked very well for obtaining the lactic acid content from the wine NMR spectra.

3.2. Case study: biodiesel–diesel data

Fig. 4 shows the comparison between conventional and optimized bucketing applied to the ^1H NMR spectra of the 100 biodiesel–diesel blends. It is possible to notice in Fig. 4b and c the superior performance of OBA for solving the alignment problem (see the enlarged regions). The proton signals in the saturated chains in biodiesel and diesel hydrocarbons tend to suffer more peak shifts due to their greater conformational freedom, which is strongly temperature-dependent and which leads to the misalignments seen in Fig. 4a. A simple alignment of the spectra according to the reference signal cannot correct for such shifts.

The blends are divided into two classes according to the location where the biodiesel were commercialized, as metropolitan and non-metropolitan. Fig. 5a, b, and c compare the clustering obtained through PCA (PC1 *versus* PC2) for the mean centered spectra before the bucketing

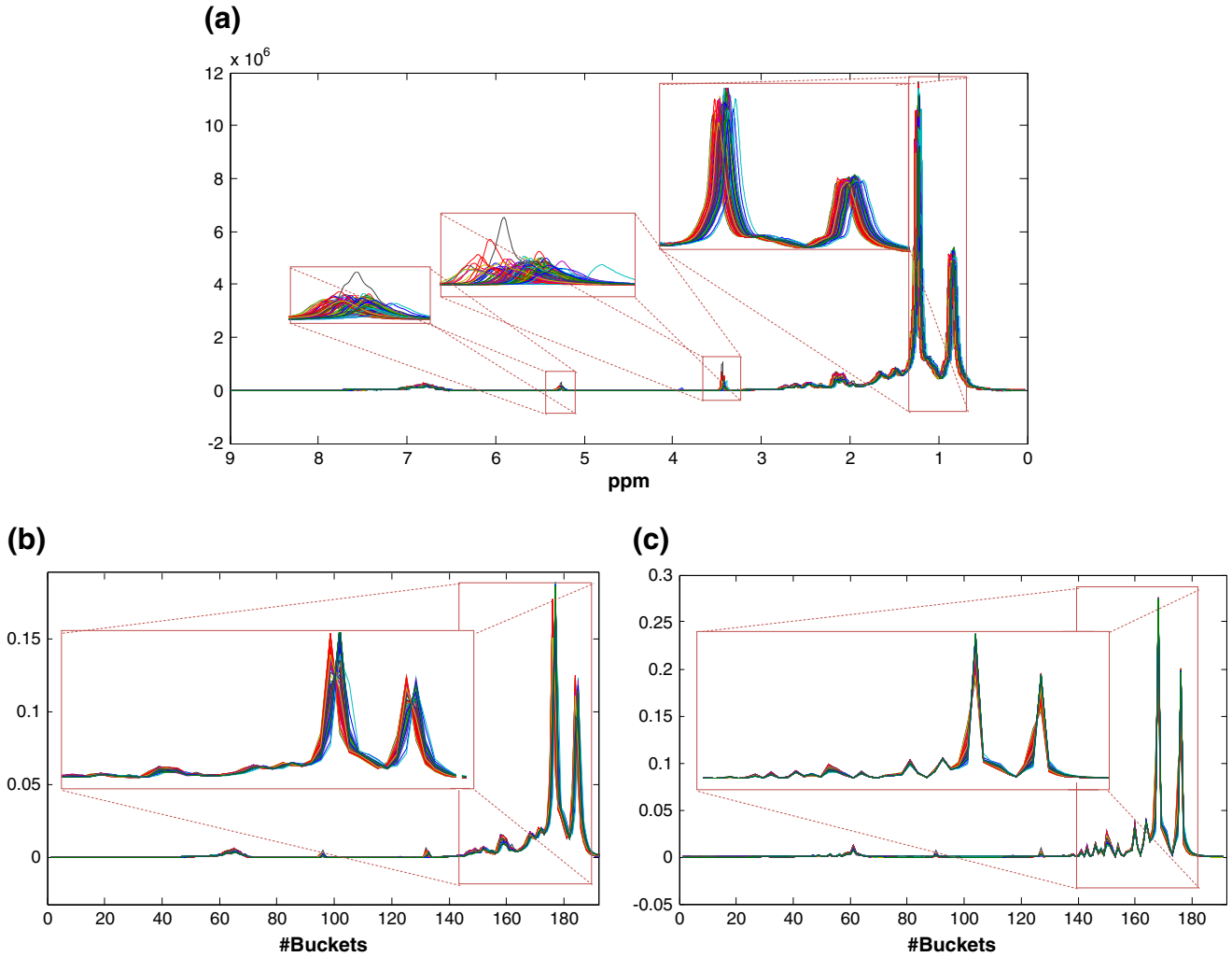

Fig. 4. (a) The raw NMR spectra with some highlighted misalignments in enlarged regions. Bucketed ^1H NMR spectra by (b) conventional and (c) optimized bucketing. The differences between the values in the y-axis occur because the bucketed spectra are normalized to unit area. The bucketed matrix in (b) is **Z** (100×200), while in (c) is **Z** (100×199).

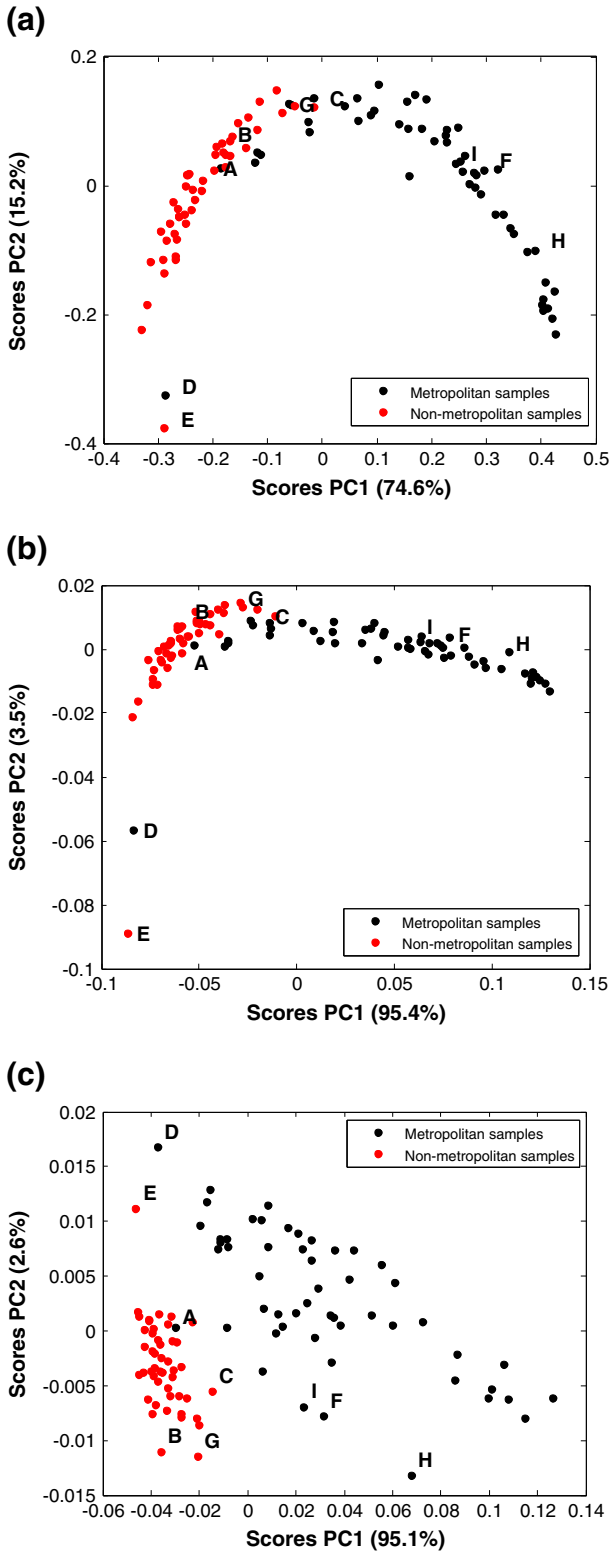

Fig. 5. PCA score plots obtained for: (a) raw data matrix; (b) data matrix after conventional bucketing; and (c) data matrix after optimized bucketing. The explained variances are shown inside parentheses in each PC.

pretreatment and after using the conventional and optimized bucketing approaches, respectively.

The analysis of the loading plots (data not shown) relative to scores plots presented in Fig. 5a, b, and c indicated that the distribution of the samples is determined by the two largest peaks, between 0.7 and 1.4 ppm, which are precisely those relative to the protons in hydrocarbons with saturated chains more affected by the misalignment issue. In Fig. 5a, one can see that PC1 captures the difference between the two kinds of samples (metropolitan and non-metropolitan), while PC2 basically captures the misalignment and the information about two samples labeled as D and E with more negative score values. The clustering in Fig. 5a is similar to that obtained after the conventional bucketing (Fig. 5b) and is very different from that one obtained when using the optimized bucketing pretreatment (Fig. 5c), because in Fig. 5b, the peak alignments are still not fully achieved, as can be seen in the enlarged region from Fig. 4b.

From a prior knowledge obtained by standard chemical analysis in the field of biofuels (not shown), for this data set, all labeled samples (A, B, C, D, E, F, G, H and I) are anomalous (outliers) blends, that is, they are out of specifications. Therefore, as can be seen in Fig. 5a and b, the variance associated to the misalignment (captured in PC2), in this case, seriously hampers the identification of the outliers, not being interesting from the viewpoint of exploratory analysis. Otherwise, it is clear in Fig. 5c that the outliers are evidenced from their own groups (metropolitan and non-metropolitan samples) with extreme score values, making easier the identification and interpretation. After the misalignments are corrected by OBA, the clustering obtained (Fig. 5c) describes 95.1% of the total variance along PC1, practically all variance contained in three components referring to the clustering without any bucketing (Fig. 5a — PC1 74.6%, PC2 15.1% and PC3 5.7%) and 2.6% of the total variance along PC2. In this example, the unnecessary complexity of the spectral profiles, provided by the misalignments was properly corrected.

Besides the analysis described above, PCA was performed for the intervals from 6.4 to 8.5 ppm and from 0.4 to 1.4 ppm, independently. The first interval corresponds to the regions with signals related to aromatic protons that differ from those in hydrocarbons and suffer less with the misalignment (there is a conformational hindrance for the protons in aromatic rings). In fact, the three clusters obtained for the mean centered data before and after applying both bucketing pretreatments were almost identical. For the second interval (larger peaks), the observations were very similar to those cited above for the whole spectral range (data not shown).

OBA also allowed a data reduction in each ^{1}H NMR spectrum from 32,768 frequency domain points to 191 buckets. This fact can be important from the computational viewpoint, since the reduced number of variables can decrease computational time. However, the data reduction is accompanied by a possible decrease in spectral resolution, which may lead to loss of information, especially when subtle differences are expressed by the samples and high resolution is required. Other solutions for the alignment, such as dynamic time warping (DTW), correlation optimized warping (COW), and the icoshift method may be used without a decrease in spectral resolution. Nevertheless, these alternatives are theoretically less simple than the bucketing approach and commonly involve expensive computational operations and user expertise. Finally, it is noteworthy that OBA provides flexibility of the definition of the input parameters, which may be adjusted in order to avoid serious loss of spectral resolution.

3.3. Case study: brain tumor data

Fig. 6a and b present the NMR spectra for the two classes of brain tumors, non-neuroglial (NN) and high-grade neuroglial (Hg). Fig. 6a shows the raw data, where it is possible to visualize the small extent of misalignments through the enlarged region. By contrast, Fig. 6b shows the corrected NMR spectra after the pretreatment by OBA, which resulted in a bucketed matrix **Z** with dimensions (29×1416). The small initial bucket width of 0.002 ppm was successful to overcome the misalignments, as can be observed in the enlarged region from Fig. 6b, showing that the peaks in this region become sharper.

At this point, it is worth to cite the correspondence problem usual for complex regions in ^{1}H NMR spectra, where owing to inhomogeneous magnetic field or incomplete phase correction, the peak shapes

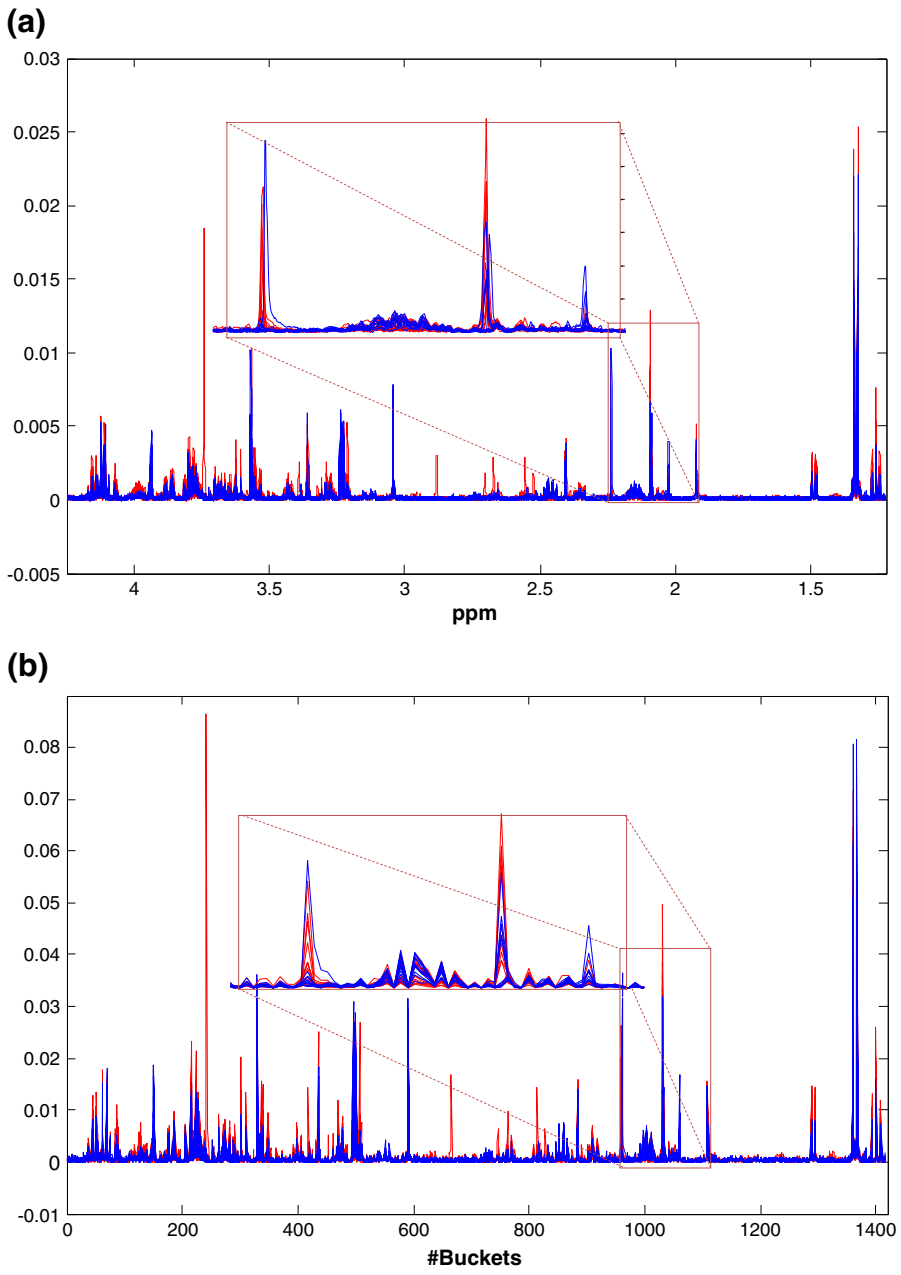

Fig. 6. Brain tumor data, (a) raw data; (b) pretreated by OBA (width of buckets of 0.002 ppm and slackness of 50%). NMR spectra for NN tumor samples are in red and for Hg tumor samples in blue.

may be distorted from the ideal symmetrical shape, and besides that, the peak positions can change due to temperature, pH, and ionic strength, even leading to an inversion in the order of the signals [26]. OBA proposed here does not deal with these issues and relies that for the data set, extreme correspondence problems are not present. Indeed, this is an inherent weak point for all bucketing procedures, being addressed by some works in literature [27,28] proposing alternatives, such as, the use of the generalized fuzzy Hough transform (GFHT) in order to establish the objectively true correspondence. Despite this, the bucketing is still widely used because no method has proven to be sufficiently easy to use and sufficiently successful in producing good results [26]. For the brain tumor data set the correspondence problem does not exist, which was properly observed through a heat map created after sorting the data using the creatine signal as reference.

In order to show the advantages of OBA, PLS-DA models were built to assess the discrimination of the two kinds of tumors. The number of latent variables used in each model was defined by choosing the one with the smallest number of misclassifications (NMC), avoiding overfitting in a similar way to that presented in the first case study (Fig. 3). Therefore, based on this diagnostic statistic, 2 latent variables were determined in the optimization of the PLS-DA model relative to the mean centered pretreated (bucketed) data set, where the model reached four misclassifications between the kinds of tumors. For the PLS-DA model using the mean centered raw data set, 4 latent variables with five misclassifications were selected. Both models were significant at a level $\alpha = 0.05$ in the permutation tests. Fig. 7a and b show the distribution of 10,000 permutation tests for NMC of the PLS-DA models from the raw and pretreated data sets. From Fig. 7, it is clear that the number of permutations (10,000) used was enough to sample the tails of the distributions, resulting in distributions with Gaussian shapes. The PLS-DA model for the pretreated data set reached a p-value = 0.0032 against the p-value = 0.0005 for the PLS-DA model

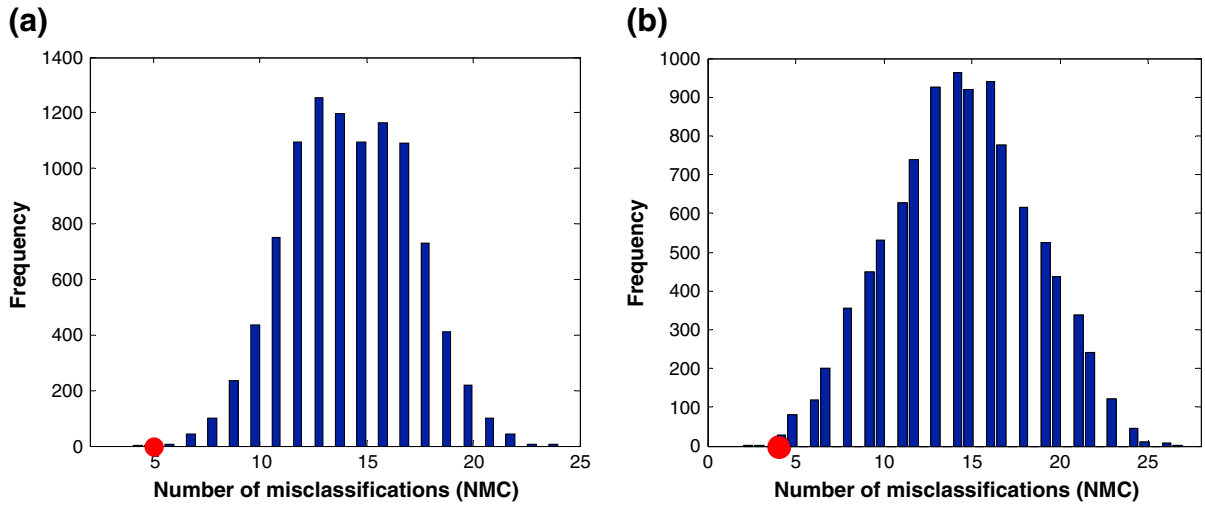

Fig. 7. Distribution of 10,000 permutation tests for NMC of the PLS-DA models from (a) the raw data set and (b) the pretreated (bucketed) data set. The red balls indicate the NMC for each PLS-DA model from unpermuted data sets.

relative to the raw data set. A *p*-value smaller than the significance threshold $\alpha = 0.05$ indicates that the null hypothesis H_0 (no difference between the two classes of tumors) may be rejected and, at this level of significance, differences between the classes are observed. For the two models, the classifications were significant, however, the PLS-DA model built using the bucketed data set was more parsimonious in determining the latent variable subspace that enables discrimination, besides achieving an improvement in NMC. Moreover, by analyzing the regression vector (data not shown) obtained for this model and the raw data, the regions more discriminants between the two types of tumors, as pointed out by the literature [23], became more evident owing to the reduction in the complexity of the data set provided by OBA, which may be important in the context of the search for biomarkers.

Finally, besides the superiority shown due to the "alignment" provided by OBA, there is the advantage of the reduction in the number of variables, which may be quite important from a computational viewpoint, especially when dealing with large data sets. This example showed the great applicability of the new algorithm and its flexibility in solving issues in a complex data set.

4. Conclusions

It has been shown that OBA has superior performance compared to conventional bucketing, widely used in the literature. For the wine data set, the results demonstrated that the optimized bucketing strategy can be useful for building less complex models (MLR models) with good predictive abilities, even comparable to PLS models obtained when sophisticated alignment methods are used. For the biodiesel–diesel blend data set, the good performance of OBA in the exploratory analysis was shown, which can be of great significance for pattern recognition purposes. The main point resides in the improvement of the explained variance by the principal components with consequent increase in interpretabilities. In this example, OBA provided good results, even in a data set with a large number of misalignments. In the brain tumor data set, OBA allowed obtaining significant PLS-DA models for discrimination of the tumors with a lower number of misclassifications by correcting a small number of misalignments, especially important to the more discriminatory signals. The proposed algorithm in this paper is easy to use, where the users just need to know about the extent of the misalignments at the baseline to set a suitable initial width of buckets and slackness. This point is very important, because the bucketing methodology has an inherent decrease of resolution, which can be minimized by the use of suitable input parameters. In our applications a slackness of 50% has been suitable, but it is not possible to generalize this parameter, since different data sets may require different values. The algorithm consists of an in-house Matlab routine available for free download at http://lqta.iqm.unicamp.br.

Acknowledgments

The authors acknowledge CNPq, CAPES, and FAPESP for the financial support and Prof. Carol H. Collins for revision of the English text. The authors also thank Prof. José Dias de Souza Filho for the NMR analysis of the biodiesel–diesel blends.

References

[1] E.M. Lenz, J. Bright, I.D. Wilson, S.R. Morgan, A.F.P. Nash, A ^{1}H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects, Journal of Pharmaceutical and Biomedical Analysis 33 (2003) 1103–1115.
[2] S. Agnolet, J.W. Jaroszewski, R. Verpoorte, D. Staerk, ^{1}H NMR-based metabolomics combined with HPLC-PDA-MS-SPE-NMR for investigation of standardized *Ginkgo biloba* preparations, Metabolomics 6 (2010) 292–302.
[3] E.F. Boffo, L.A. Tavares, M.M.C. Ferreira, A.G. Ferreira, Classification of Brazilian vinegars according to their ^{1}H NMR spectra by pattern recognition analysis, LWT- Food Science and Technology 42 (2009) 1455–1460.
[4] M.R. Monteiro, A.R.P. Ambrozin, L.M. Lião, E.F. Boffo, L.A. Tavares, M.M.C. Ferreira, A.G. Ferreira, Study of Brazilian gasoline quality using hydrogen nuclear magnetic resonance (^{1}H NMR) spectroscopy and chemometrics, Energy & Fuels 23 (2009) 272–279.
[5] C. Daolio, F.L. Beltrame, A.G. Ferreira, Q.B. Cass, D.A.G. Cortez, M.M.C. Ferreira, Classification of commercial catuaba samples by NMR, HPLC and chemometrics, Phytochemical Analysis 19 (2008) 218–228.
[6] R.H. Jellema, Variable shift and alignment, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Chemical and Biochemical Data Analysis, vol. 2, Elsevier, Oxford, 2009, pp. 85–108.
[7] H. Winning, Quantitative multivariate NMR spectroscopy in Food Science and Nutrition, Thesis, University of Copenhagen, Frederiksberg, 2009.
[8] N. Trbovic, F. Dancea, T. Langer, U. Günther, Using wavelet de-noised spectra in NMR screening, Journal of Magnetic Resonance 173 (2005) 280–287.
[9] J. Forshed, R.J.O. Torgrip, K.M. Åberg, B. Karlberg, J. Lindberg, S.P. Jacobsson, A comparison of methods for alignment of NMR peaks in the context of cluster analysis, Journal of Pharmaceutical and Biomedical Analysis 38 (2005) 824–832.
[10] V. Pravdova, B. Walczak, D.L. Massart, A comparison of two algorithms for warping of analytical signals, Analytica Chimica Acta 456 (2002) 77–92.
[11] T. Skov, F. van der Berg, G. Tomasi, R. Bro, Automated alignment of chromatographic data, Journal of Chemometrics 20 (2006) 484–497.
[12] G. Tomasi, F. van der Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, Journal of Chemometrics 18 (2004) 231–241.
[13] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping, Journal of Chromatography. A 805 (1998) 17–35.
[14] F.H. Larsen, F. van der Berg, S.B. Engelsen, An exploratory chemometric study of ^{1}H NMR spectra of table wines, Journal of Chemometrics 20 (2006) 198–208.

[15] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: a versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance 202 (2010) 190–202.
[16] B. Lefebvre, S. Golotvin, L. Schoenbachler, R. Beger, P. Pryce, J. Megyesi, R. Safirstein, Intelligent bucketing for metabonomics — Part 1, http://www.acdlabs.com/download/publ/2004/enc04/intelbucket.pdf/, (20 january 2010).
[17] B. Lefebvre, R. Sasaki, S. Golotvin, A.W. Nicholls, Intelligent bucketing for metabonomics — Part 2, http://www.acdlabs.co.uk/download/publ/2004/intelbucket2.pdf/, (20 January 2010).
[18] A.W. Nicholls, R.J. MortiShire-Smith, J.K. Nicholson, NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats, Chemical Research in Toxicology 16 (2003) 1395–1404.
[19] L. Jang-Eun, H. Geum-Sook, V.D.B. Frans, L. Cherl-Ho, H. Young-Shick, Evidence of vintage effects on grape wines using ^{1}H NMR-based metabolomic study, Analytica Chimica Acta 648 (2009) 71–76.
[20] R.A. Davis, A.J. Charlton, J. Godward, A.J. Stephen, M. Harrison, J.C. Wilson, Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform, Chemometrics and Intelligent Laboratory Systems 85 (2007) 144–154.
[21] P.E. Anderson, N.V. Reo, N.J. DelRaso, T.E. Doom, M.L. Raymer, Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics, Metabolomics 4 (2008) 261–272.
[22] P.E. Anderson, D.A. Mahle, T.E. Doom, N.V. Reo, N.J. DelRaso, M.L. Raymer, Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data, Metabolomics 7 (2011) 179–190.
[23] A.V. Faria, F.C. Macedo Jr., A.J. Marsaioli, M.M.C. Ferreira, F. Cendes, Classification of brain tumor extracts by high resolution ^{1}H MRS using partial least squares discriminant analysis, Brazilian Journal of Medical and Biological Research 44 (2011) 149–164.
[24] E. Szymanska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistic for PLS-DA models in metabolomic studies, Metabolomics 8 (2012) S3–S16.
[25] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89.
[26] K.M. Åberg, E. Alm, R.J.O. Torgrip, The correspondence problem for metabonomics datasets, Analytical and Bioanalytical Chemistry 394 (2009) 151–162.
[27] E. Alm, R.J.O. Torgrip, K.M. Åberg, I. Schuppe-Koistinen, J. Lindberg, A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support, Analytical and Bioanalytical Chemistry 395 (2009) 213–223.
[28] E. Alm, T. Slagbrand, K.M. Åberg, E. Wahlström, I. Gustafsson, J. Lindberg, Automated annotation and quantification of metabolites in ^{1}H NMR data of biological origin, Analytical and Bioanalytical Chemistry 403 (2012) 443–455.