Analytica Chimica Acta 1019 (2018) 1–13

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

PepsNMR for ^{1}H NMR metabolomic data pre-processing

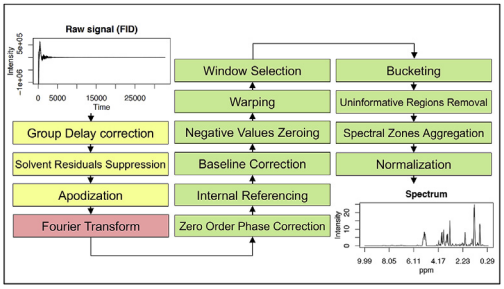
Manon Martin [a, *], Benoît Legat [b, 1], Justine Leenders [c], Julien Vanwinsberghe [d, 2], Réjane Rousseau [a, 3], Bruno Boulanger [e, 4], Paul H.C. Eilers [f], Pascal De Tullio [c], Bernadette Govaerts [a]

[a] Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA/IMMAQ), Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium
[b] Ecole Polytechnique de Louvain (EPL), Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium
[c] Laboratoire de Chimie Pharmaceutique, Université de Liège (ULg), Liège 1, Belgium
[d] Université Louis Pasteur, Strasbourg I, France
[e] Eli Lilly & Company, Statistical Department, Mont-St-Guibert, Belgium
[f] Department of Biostatistics, Erasmus University Medical Centre, Rotterdam, The Netherlands

HIGHLIGHTS

- A semi-automatic and complete pre-processing strategy is proposed as an R package called PepsNMR.
- The methodology of each pre-processing step is carefully described and applied to spectral matrices from human serum and urine.
- The spectral matrices repeatability is better with PepsNMR than with a gold standard pre-processing procedure.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 5 September 2017
Received in revised form
15 February 2018
Accepted 27 February 2018
Available online 12 March 2018

Keywords:
^{1}H nuclear magnetic resonance
Data pre-processing
R package
Metabolomics
Pre-processing quality evaluation

ABSTRACT

In the analysis of biological samples, control over experimental design and data acquisition procedures alone cannot ensure well-conditioned ^{1}H NMR spectra with maximal information recovery for data analysis. A third major element affects the accuracy and robustness of results: the data pre-processing/pre-treatment for which not enough attention is usually devoted, in particular in metabolomic studies. The usual approach is to use proprietary software provided by the analytical instruments' manufacturers to conduct the entire pre-processing strategy. This widespread practice has a number of advantages such as a user-friendly interface with graphical facilities, but it involves non-negligible drawbacks: a lack of methodological information and automation, a dependency of subjective human choices, only standard processing possibilities and an absence of objective quality criteria to evaluate pre-processing quality. This paper introduces PepsNMR to meet these needs, an R package dedicated to the whole processing chain prior to multivariate data analysis, including, among other tools, solvent signal suppression, internal calibration, phase, baseline and misalignment corrections, bucketing and normalisation. Methodological aspects are discussed and the package is compared to the gold standard procedure with two metabolomic case studies. The use of PepsNMR on these data shows better information recovery and

* Corresponding author.
E-mail address: manon.martin@uclouvain.be (M. Martin).
[1] Present address: Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Pôle en ingénierie mathématique (INMA), Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.
[2] Present address: ClinBAY SPRL, Genappe, Belgium.
[3] Present address: CMC Statistical Sciences, GlaxoSmithKline, Rixensart, Belgium.
[4] Present address: Arlenda, Mont-St-Guibert, Belgium.

https://doi.org/10.1016/j.aca.2018.02.067

predictive power based on objective and quantitative quality criteria. Other key assets of the package are workflow processing speed, reproducibility, reporting and flexibility, graphical outputs and documented routines.

1. Introduction and literature review

Among the Omics sciences, the field of untargeted metabolomics is prominent in the search for biomarkers in intricate biological systems. This approach studies the response of an organism to external stimuli, to determine the link between experimental conditions and phenotypic changes of interest [1]. Its main advantage is that it provides a metabolic snapshot of an organism, a tissue or cells, but its main challenge is related to the processing of large datasets, which requires powerful signal processing and statistical tools [2].

The most common analytical techniques used to generate metabolomic data are Gas Chromatography-Mass Spectrometry (GC-MS), Liquid Chromatography-MS and proton Nuclear Magnetic Resonance (^{1}H NMR) Spectroscopy. They work complementarily, each having its own purpose and limitations. Advantages of ^{1}H NMR spectroscopy are that it is low cost and requires minimal preparation. It is generally untargeted, thus unbiased, non-destructive, non-invasive, reproducible and fast [1,3,4].

Analytical and experimental variation that is unrelated to the biological perturbation under study deserve particular attention in metabolomic studies. This variation originates from sample collection, preparation and conservation, as well as data acquisition, and intra-individual (*e.g.* nutrients) or inter-individual (*e.g.* genetic) variations [5], especially in human studies [6]. When incomplete or inappropriate pre-processing methods are applied, this variation can lead to inconclusive results or the discovery of false biomarkers during data analysis. Pre-processing, here defined as in the field of chemometrics as all data editing done before data analysis [7], is used not only to enhance the Signal-to-Noise Ratio (SNR) but also to correct for or minimise instrumental artefacts and appropriately transform the data into interpretable spectral profiles that better represent the biological system under study.

Good data pre-processing is important. Commercial software such as TopspinTM, Chenomx or MestReNova remain the most used solutions for NMR data pre-processing but they do not provide advanced editing steps, lack modularity and require iterative, manual and arbitrary adjustments. These interventions will thus not necessarily improve the data quality and can impact the performance of the multivariate data analysis conducted afterwards. On the other hand, free software that are mostly open-source exist for ^{1}H NMR data processing (see Alonso et al. [2] for a review of the more important ones). However, their pre-processing tools remain basic and do not offer a global strategy dedicated to the chemometric approach[5] for further data analysis [8]. Besides, the multiplicity of pre-processing scenarios and methods [9], which may arbitrarily affect the final conclusions, would benefit from clear recommandations guiding the user's choice of when and how methods should be applied.

We introduce the R package PepsNMR (or Packaged Extensive Preprocessing Strategy for NMR data), for the (semi-automatic) pre-processing of 1D ^{1}H NMR Free Induction Decay (FID) datasets. The objective of this software is to provide an exhaustive and flexible workflow to deal with typical features of raw ^{1}H NMR data and cover the pre-processing (the cleaning of the raw instrumental data) and pre-treatment (the transformation of the cleaned data for data processing) steps as defined by Goodacre et al. [10]. After several steps the proposed pre-processing strategy transforms FID data into a binned spectral table that will enable data analysis using classic chemometric tools [11] that identify global trends in spectral profiles.

Innovative methods are combined with well known algorithms in order to correct for instrumental artefacts and irrelevant biological variability, to increase the SNR, appropriately transform the data domain/scale and reduce its dimensionality.

Section 2 presents the pre-processing methodology: its motivation, the description of each step, and the package features. Section 3 briefly describes the quantitative criteria used to assess the quality of the pre-processed data. Then Section 4 illustrates and discusses the use of PepsNMR on the basis of two case studies. Thereafter, the conclusion summarises and brings forward key points of the methodology (Section 5) and finally, Sections 6 and 7 describe software implementation and the datasets used in Section 4.

2. Methodology

2.1. PepsNMR workflow presentation

PepsNMR originates from an initially unpublished Matlab program called “Bubble” developed by Eli Lilly and one of the authors (Paul H.C. Eilers) [12,13]. The package works as an open system with full flexibility: it provides a complete series of pre-processing functions to be run sequentially or independently to transform the FIDs into the final spectra. Raw data in Bruker format can be imported directly, different methods and settings are available in most pre-processing steps and visualisation functions are available to monitor the process. The pre-processing package is depicted in a flowchart in Fig. 1 and the complete sequence of R functions is detailed below in the proposed order.

GroupDelayCorrection() This function corrects the frequency dependent — or first order — phase error due to the Group Delay induced by the Bruker's digital filter.
SolventSuppression() The variability of solvent residuals in the spectrum (especially for water) can mask informative signals from molecules of interest. This step removes the solvent signal (usually water) in the time domain using a Whittaker smoother.
Apodization() Apodization (mainly) aims to increase the SNR of the spectrum and/or its resolution. It multiplies the FID signal by a positive signal, often decaying. Several apodization functions are available in PepsNMR.
FourierTransform() The Fourier transform translates the complex FID signal in the time domain into a complex spectrum expressed in frequency (Hz). This function further calibrates the frequency scale and converts it into chemical shifts (ppm).

[5] Defined as the approach where the compounds of interest are identified after the statistical analysis of spectral intensities that leaded to the identification of significant spectral features.

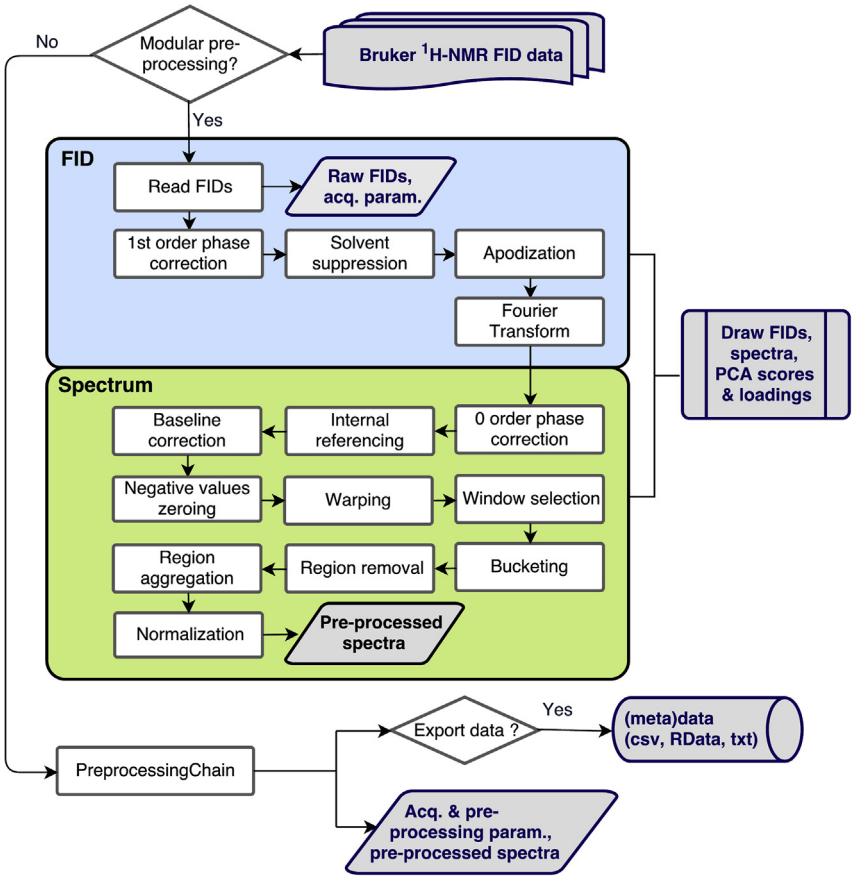


Fig. 1. PepsNMR flowchart with available functions, inputs and outputs.

ZeroOrderPhaseCorrection() In order to remove the frequency-independent phase shift, the spectrum is rotated to maximise the positiveness of its real part.
InternalReferencing() An internal compound (*e.g.* TMSP) is used for the chemical shift scale referencing.
BaselineCorrection() In order to remove the spectral baseline, this function uses asymmetric least squares with a roughness penalty.
NegativeValuesZeroing() Imperfect phase or baseline correction can lead to small residual negative values in the spectrum that cannot be interpreted. These negative values are thus set to zero.
Warping() The first global alignment solution provided in PepsNMR is a warping algorithm. It applies a warping function to spectra based on a reference spectrum. This warping function is a combination of polynomials and P-Splines that stretch/compress the chemical shift axis to increase the similarity between the transformed spectrum and the reference spectrum.
WindowSelection() This function selects the informative part of the spectrum to be kept for data analysis.
Bucketing() The second solution for minimising peak misalignment across spectra and for reducing the dimension of the data is soft bucketing. This function uses trapezoidal or rectangular interpolation to deliver equally-spaced bins with a specified length or number.
RegionRemoval() This function sets specific spectral regions of no interest in the study to zero.
ZoneAggregation() It is useful to apply targeted data reduction for molecules with an unstable chemical shift (*e.g.* citrate). This routine aggregates a pre-defined interval of descriptors into a single symmetrical peak.
Normalisation() This function normalises intensities in the spectrum to avoid unwanted inter-sample variation.

2.2. Details on the pre-processing steps

This section describes in a nutshell the methodological features of the pre-processing steps. For each R function, underlying mathematical and statistical concepts along with their algorithmic implementation are explained and discussed.

2.2.1. Raw signal

Each FID (Fig. 3) recorded with the ^{1}H NMR instrument is a complex decaying signal $S = S_x + iS_y$ of m data points (where S_x and S_y are respectively the real and imaginary parts of the signal). It represents the net sum of distinct signal components from different proton nuclei. The theoretical form of each single signal can be written as a function of its offset ν, amplitude s_0 and transverse relaxation time T with an exponential decay over time t: $s_0 \exp(i2\pi\nu t)\exp\left(\frac{-t}{T}\right)$. The function ReadFids () imports raw Bruker files into two matrices: the complex FID signal matrix and the Bruker acquisition parameters needed for the pre-processing steps.

2.2.2. Phase errors and group delay suppression

For a perfectly phased signal, its real part starts at a maximum value and its imaginary part starts at 0. The corresponding spectrum has a positive real part in absorptive mode and an imaginary part in dispersive mode with both positive and negative peaks. For technical reasons (Bruker digital filter, incorrect magnetisation, etc.) the real and imaginary parts of the signal result in a mixture of both absorptive and dispersive mode line shapes [14]. These lines

need to be phase corrected. The total phase error Φ is usually written as a sum of frequency-independent (φ_0) and frequency-dependent ($\varphi_1(\nu)$) phase errors, $\Phi = \varphi_0 + \varphi_1(\nu)$, that are corrected respectively with the zero and first order phase corrections. The initial non-phased signal is: $S = S_{\text{phased}}\exp(i\Phi)$.

This first pre-processing step aims to partly correct the first order phase error introduced by the Bruker digital filter. Indeed, this filter induces a pre-acquisition delay: the real part of the FID does not have a maximum value at $t = 0$. This delay, usually of tens of data points, is illustrated in Fig. 2.

In PepsNMR, the delay is removed from the FID based on properties of the Discrete Fourier Transform (DFT) [15]: a circular shift on the time domain signal by $\tau > 0$ samples, called the Group Delay, is identical to multiplying the corresponding spectrum F by the linear term $\exp(-i2\pi\nu\tau)$ *i.e.* applying a frequency-dependent linear phase shift in the frequency domain. The signal is first Fourier transformed, then multiplied by the phase shift, and finally back-transformed with the inverse Fourier transform. This double transformation is mandatory in the common case of a non-integer Group Delay. The function GroupDelayCorrection () removes the first order phase shift with τ recovered from the spectrometer acquisition parameters and returns the processed FIDs.

2.2.3. Solvent signal suppression

Water is the most common solvent for biofluid samples. It has a rather broad resonance, and its chemical shift varies according to solution conditions. Usual solvent suppression techniques aim to reduce the magnitude of the solvent resonance before the NMR signal reaches the receiver. Three different approaches are [16]: pre-saturation of the water resonance, production of zero net excitation of the water resonance and destruction of the water resonance with pulsed field gradients. Nevertheless, residual peaks of water are still measured and their intensity and instability can mask other interesting compounds. Moreover, and based on our experience, solvent suppression improves the efficiency of subsequent correction steps such as the baseline correction.

PepsNMR solves this common issue with non-parametric modelling and the subtraction of the residual water signal directly from the FID. Under the assumption that water is the main compound of the analysed samples, a discrete penalised least squares with second-order differences presented by Eilers [17] and based on the Witthaker's graduation method [18], is used to separately model the real and imaginary parts of the water signal.

This method estimates the smoothed series W that minimises the Q criterion:

$$Q = \sum_{j=1}^{m} (S_j - W_j)^2 + \lambda \sum_{j=3}^{m} \left(\Delta^2 W_j\right)^2 \quad (1)$$

This expression is a balanced combination of two conflicting objectives, *i.e.* fidelity to the data and smoothness. The first term is the sum of squares of differences measuring the lack of fit of W to the real or imaginary signal S on the m data points. The second term penalises the series' roughness with λ the penalty parameter multiplied by the sum of squared second order differences (defined as $\Delta^2 W_j = W_j - 2W_{j-1} + W_{j-2}$), which quantifies the roughness. The larger λ is, the smoother the water signal. An analytical solution is provided by Eilers [17] and the use of sparse matrices greatly reduces the computational time for large series such as ^{1}H NMR signals. The function SolventSuppression (Fid_data, lambda.ss = 1e6, …) estimates the water signal for both the real and the imaginary parts of the FID and returns the processed FIDs and the real and imaginary water signals. lambda.ss accounts for the parameter λ and its default value (1e6) is a practical choice, based on empirical results. Nevertheless, Eilers [17] and Frasso and Eilers [19] describe several ways to tune λ in order to optimise the smoothing: either visually, by cross-validation or using V-curves with R code available for the latter.

2.2.4. Apodization

Apodization multiplies the FID by a signal that changes its shape and emphasises distinct spectral portions [16] to enhance the sensitivity and/or the resolution of the spectrum. Due to its nature, NMR signal intensity decays over time whereas noise has random fluctuations with constant amplitude. Thus, the SNR is higher at the beginning of the FID and declines over time. An increase in sensitivity (higher SNR) can thus be achieved by multiplying the FID by a decaying signal. However, a faster decay of the FID (already dependent on the decay time T) implies broader spectral peaks line widths, lowering their height and thus their resolution since the area under the peaks remains unchanged. Therefore, optimal trade-off parameters for the apodization signal are needed to combine both of these objectives. A classic apodization step applies an exponentially decaying signal principally to enhance the SNR but

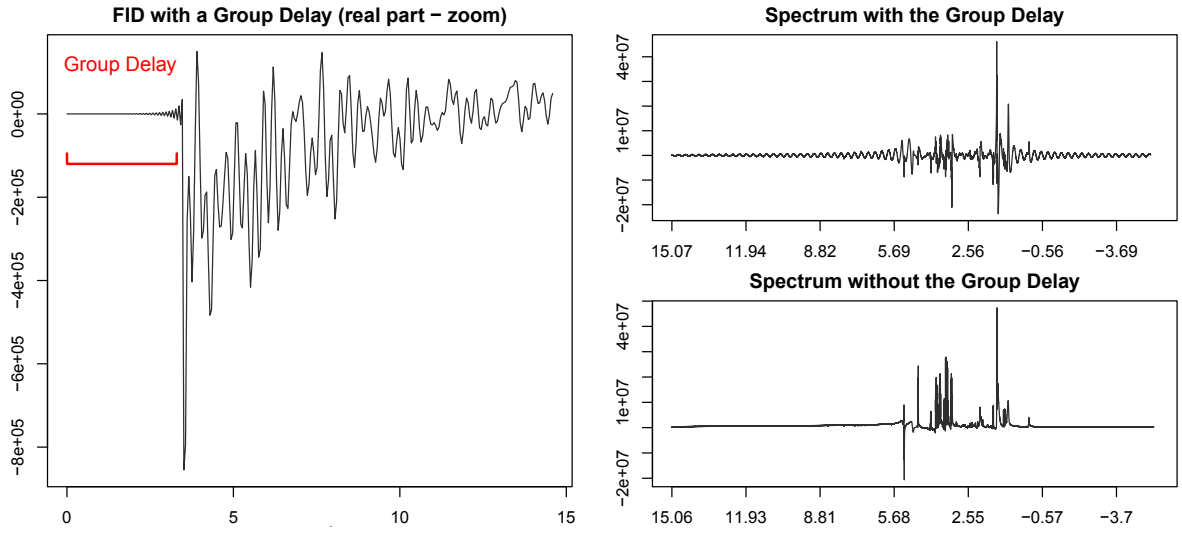


Fig. 2. The Group Delay recorded before the signal acquisition has to be removed from the FID.

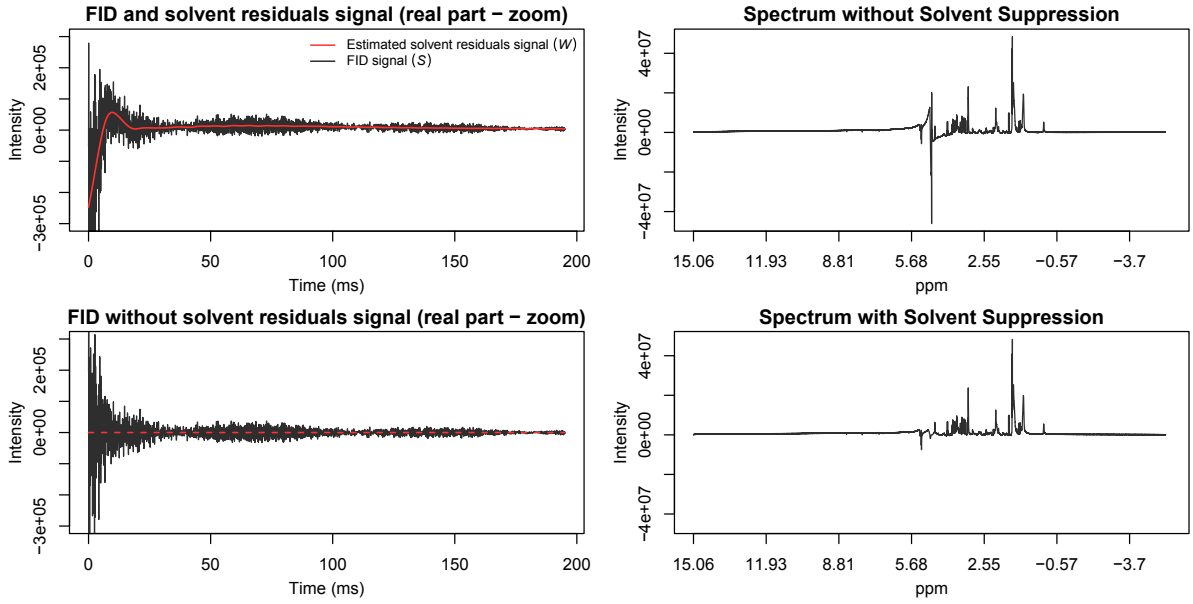


Fig. 3. The solvent residuals signal is first estimated with a smoothing function and then subtracted from the FID.

other apodization factors, such as a concave signal, can be used to enhance both sensitivity and resolution.

Several types of apodization functions are provided by PepsNMR with the option type.apod in Apodization (Fid_data, type.apod = "exp", …): decreasing exponential or gaussian functions, Hanning window, etc. By default PepsNMR multiplies the FID by a decaying exponential function: $h(t) = \exp(-LB{*}t)$ with *LB* the Line Broadening parameter. If *LB* increases, the SNR increases at the expense of the spectral resolution (broader line widths). Usual values in proton NMR for *LB* found in the literature are 0.3 for the NOESY presat pulse sequence and −0.01 for the CMPG presat pulse sequence. *LB* should not exceed the value of 1 to avoid information loss.

2.2.5. Fourier transform and chemical shift scale

At this stage, the recorded signal in the time domain *S* is ready to be converted into a spectrum in the frequency domain *F*. The Fourier transform (FT) provides a general solution to transforming the complex signal sequence into a complex spectral sequence. It extracts each signal and translates it into peaks in a spectrum with specific heights, positions and widths that depend on the amplitude of the signal, the frequency of the corresponding compound and the relaxation time *T* [14].

PepsNMR achieves this transformation with the function FourierTransform (), which applies the Discrete Fourier Transform (DFT) for uniformly spaced samples with the Fast Fourier Transform (FFT) algorithm.

The DFT can be written as:

$$F_j = \sum_{t=0}^{m-1} S_t \exp(-2i\pi jt/m) \tag{2}$$

for $j = 0, \ldots, m-1$.

The scale of the *F* vector has yet to be defined. PepsNMR first expresses it in Hertz on the basis of the spectrometer acquisition parameters with the following formula: $\nu_j = \nu_S + \Delta_\nu \times (j - (m-1)/2)$ where ν_S is the transmitter frequency (called SFO1 in the Bruker acquisition file) located at the center of the spectral window and Δ_ν represents the difference in frequency between 2 data points and is calculated from the spectral window width (or sweep width). Note that ν_S is a sum of ν_B, the basic transmitter frequency (set automatically depending on the spectrometer and the observed ^{1}H nucleus) and ν_O, the transmitter frequency offset, tuned by the user to shift the center of the spectral window from ν_B to ν_S.

The scale in Hertz is then converted into a chemical shift vector δ expressed in parts per million (ppm), a dimensionless scale that is unrelated to the magnetic field strength of the spectrometer [14]. A chemical shift represents the frequency ν_j relative to ν_B, where the 0 ppm is located at the ν_B frequency. The conversion equation is: $\delta_j = 10^6 \times \left(\frac{\nu_j - \nu_B}{\nu_B}\right)$.

A spectral matrix is then created where each row corresponds to a spectral profile and each column represents a spectral intensity at a specific chemical shift. By convention, the chemical shift along the horizontal axis is set as decreasing towards the right.

2.2.6. Zero order phase correction

For the technical reasons explained above, the spectrum can still exhibit a zero order phase shift error of a certain angle φ_0 that can be expressed as $F = F_{\text{phased}}\exp(i\varphi_0)$, where F_{phased} is the perfectly phased spectrum. This phase shift is constant for all signal vectors and thus independent of the spectral frequencies.

Based on the principle that the real part of the perfect spectrum F_{phased} has a pure absorptive mode, an optimal angle φ_0 can automatically be found in PepsNMR by maximising an adequate criterion of the positiveness of the real part of the spectrum.

PepsNMR proposes two positiveness criteria measured on a range *R* of descriptors: it can be quantified by the maximisation of the highest positive peak or by the *rms* criterion that represents the ratio between the sum of squares of positive intensities and the sum of squares of all intensities in the spectrum:

$$\varphi_0 = \underset{\phi \in [0,2\pi]}{\operatorname{argmax}} \frac{\sum_{j \in R} \left(F_j^\phi\right)^2 \times \text{ind}_j}{\sum_{j \in R} \left(F_j^\phi\right)^2} \tag{3}$$

where $F^\phi = F\exp(-i\phi)$ is the real part of the rotated spectrum with a correction angle $-\phi$ and $\text{ind}_j = 1$ if $F_j^\phi > 0$ and 0 otherwise. The

user can decide on which range R to maximise the positiveness criterion since particular areas that can be difficult to phase (especially the spectral water region) can be excluded. The function ZeroOrderPhaseCorrection (RawSpect_data, type.zopc = "rms", …) rotates each spectrum in the spectral matrix and returns the processed spectra with their rotation angles. In addition, a vector of angles can be directly provided by the user with the argument type.zopc = "manual".

Note that residuals of first order phase shift can still be present. They are currently not removed by PepsNMR but are negligible after the Group Delay removal.

2.2.7. Referencing with an internal standard

In order to be more reliable, the scale can be referenced to a known standard, *i.e.* an internal reference compound whose chemical shifts are minimally influenced by external factors (*e.g.* temperature or concentration) and ideally located outside the spectral region to be clearly identifiable. The reference compound used in NMR is usually a silane derivative such as Trimethylsilylpropionate (TMSP) or 4,4-Dimethyl-4-silapentane-1-sulfonic acid (DSS). The chemical shift of these standard peaks is referenced to 0 ppm by convention but other reference compounds may have a non-null ppm value.

In PepsNMR, the function InternalReferencing (RawSpect_data, shiftHandling = "cut", method = "max", range = "near0", …) proposes two ways to locate the reference compound peak in each spectrum within a range of intensities (range) with the method argument: it selects either the maximum intensity or the first peak in the search range higher than a predefined threshold.

2.2.8. Baseline correction

Ideal spectra are expected to have a flat baseline but, even after phase correction, baseline artefacts can still appear. These artefacts result from multiple sources, such as the presence of macromolecules, a not entirely linear electronic detection process or calibration errors from the 180° pulse [20]. It is critical to remove these artefacts with an appropriate methodology in order to enhance the warping efficiency and to perform accurate data analysis on meaningful and unaltered peak shapes. Several algorithms have been suggested that estimate the baseline to be removed from the spectra. They mainly involve iterative penalised least squares methods, robust baseline estimation, iterative polynomial fitting, local means or wavelets transforms [21,22].

The AsLS smoothing algorithm [23] in PepsNMR allows flexible baseline estimation with fast and reproducible computations. It is a nonparametric penalised method that combines asymmetric weighting of deviations from the trend with a penalty on any non-smooth behaviour of the estimated series.

This generalisation of the Whittaker smoother provides a robust baseline estimate with no artefacts at the end of the spectrum, where a polynomial fit [7].

The baseline estimator Z of a signal F, is found by minimising the objective function (4), similar to Equation (1), where the first term represents the lack of fit to the data, measured by asymmetric least squares, and the second term is the roughness penalty. The asymmetric weighting will favour positive corrected intensities as positive deviations are less weighted than negative ones.

$$Q = \sum_{j=1}^{m} \psi_j (F_j - Z_j)^2 + \lambda \sum_{j=3}^{m} \left(\Delta^2 Z_j\right)^2 \tag{4}$$

In Equation (4), ψ_j is a specific weight for the frequency j: if $F_j > Z_j$, $\psi_j = p$, otherwise $\psi_j = (1-p)$ where $p \in [0,1]$ is close to 0. The function BaselineCorrection (RawSpect_data, lambda.bc = 1e7, p.bc = 0.05, …) iteratively estimates the weights ψ_j and the smoother Z with a gradient algorithm [24] and subtracts the estimated baseline function from the spectrum: $F^* = F - Z$. An illustration of this process is given in Fig. 4. The larger λ is, the smoother the baseline will be and the smaller p is, the more the baseline will tend to be both below and close to the spectral line (containing random noise and signal). Based on experience, Eilers and Boelens [23] recommend the following ranges of parameters values: $0.001 \leq p \leq 0.1$ and $10^2 \leq \lambda \leq 10^9$. Often, a visual inspection is sufficient for tuning these parameters. Nonetheless, these authors provide ad hoc computation procedures in their article to optimise their values.

Since this baseline correction algorithm can create artificial positive intensities in the water region [12], a previous solvent suppression step is recommended and a subsequent step is dedicated to the removal of uninformative regions (*e.g.* water) to cope with this behaviour.

2.2.9. Negative value zeroing

Despite the application of baseline and phase corrections, spectra may still have negative intensities at specific frequency values. This step simply sets all these negative values to zero since they cannot be properly interpreted.

In PepsNMR, this operation is done with the function NegativeValuesZeroing ().

2.2.10. Warping

In metabolomic applications, due to the nature of biological samples and variation in experimental conditions (*e.g.* pH, temperature or concentration), peak shifts, or misalignment between identical features from different spectra, are observed. In the literature, peak alignment solutions commonly applied to NMR spectra are warping and bucketing,[6] and both methods are included in PepsNMR. Warping will indeed enhance the similarity between profiles by the way of shifts, stretches and/or compressions along their horizontal axis. Recent reviews by Bloemberg et al. [26] and Vu and Laukens [27] report the advantages and drawbacks of the most important warping procedures for signal alignment, such as parametric time warping (PTW), semi-parametric time warping (SPTW), dynamic time warping (DTW), correlation optimised warping (COW), icoshift, CluPA, etc.

The warping method implemented in PepsNMR is inspired by different elements of time warping as developed by Eilers [24] and Van Nederkassel et al. [28]. It applies a warping function $w(\nu)$ to a normalised spectrum F, *i.e.* a distortion function of the ppm axis that combines a polynomial term and a penalised B-splines (P-splines) term. The warping function $w(\nu)$ is defined as:

$$w(\nu) = \sum_{k=0}^{K} \beta_k \nu^k + \sum_{l=1}^{L} \alpha_l B_l(\nu) \tag{5}$$

The first term is a polynomial of order K with β_k the corresponding polynomial coefficients. The second term is a weighted sum of B-splines, with L the number of B-splines and α_l the coefficient for the l^{th} B-spline $B_l(\nu)$. The B-spline curves are constructed from polynomial pieces and smoothly joined together. The normalised and now distorted spectral profile $F(w(\nu))$ is interpolated from the discrete warping function.

The warping principle consists of building $w(\nu)$ such that the distance between a warped spectrum $F(w(\nu))$ and the reference spectrum G is minimised. The parameters are found by the iterative

[6] Data reduction technique where the values are integrated over discrete spectral areas (25).

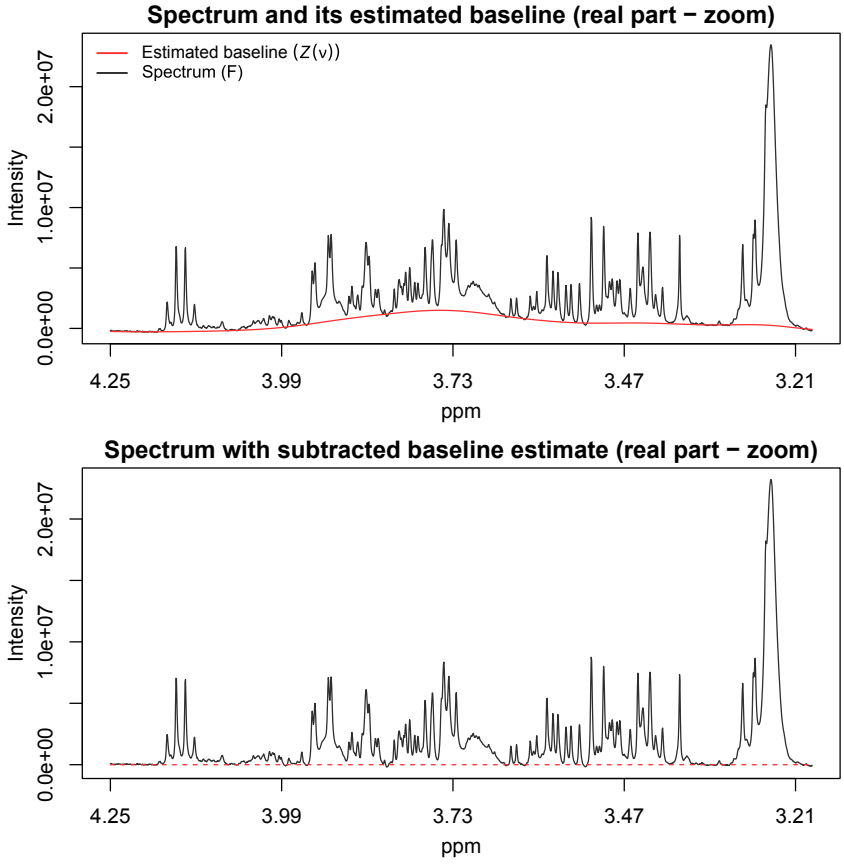


Fig. 4. The baseline is first estimated and then removed from the spectrum.

minimisation of the following penalised least squares objective function:

$$Q = \sum_{j=1}^{m} \left(G_j - F_j(w(\nu))\right)^2 + \lambda \sum_{l=3}^{L} \left(\Delta^2 \alpha_l\right)^2 + \kappa \sum_{l=1}^{L} \alpha_l^2 \qquad (6)$$

where the unknowns are β_k and α_l. Two penalty terms are applied to the B-splines coefficients [29]. The first one is based on second order differences $\Delta^2\alpha_l$ and tunes the B-splines smoothness with parameter λ whereas the second one is a Ridge penalty applied to regularise the coefficients (*i.e.* control their variance) with parameter κ. The function Warping () warps the spectra with the following default parameter values empirically chosen based on the authors' experiments: $K = 3$, $L = 40$, $\lambda = 0.01$ and $\kappa = 0.0001$. An illustration of warped spectra is given in Fig. 5.

Each spectrum must be first normalised and then denormalised after warping. Normalisation is defined as a row operation on the spectral matrix lines for making the samples more directly comparable with each other. The choice of a reference spectrum among all profiles is not straightforward. With no prior knowledge and no pooled samples, a robust reference selection process, available in PepsNMR, selects the spectrum that minimises the sum of squared distances with all other spectra before or after warping. Another possibility in PepsNMR is to refer to a manually defined sample, such as an artificial sample with known metabolic concentrations. Note that when the samples come from different classes, using one reference sample per class can be misleading and is likely to induce artificial between-class variability [26].

2.2.11. Spectral window selection

In order to focus the data analysis on particularly informative spectral areas, the function WindowSelection () trims the spectral window between right and left chemical shift bounds (by default: 0.2 ppm and 10 ppm), decreasing the number of descriptors m in the spectrum.

2.2.12. Bucketing

The high dimensionality of the data and small residual peak shifts can impede future multivariate data analyses [7]. Bucketing integrates the m original spectral intensities into m^b predefined intervals. Since the warping already corrects for chemical shift misalignments, a *soft* bucketing technique is sufficient to smooth and correct for small remaining deviations, provided that shifts of a same peak are small enough to be included in the same interval. Also called data reduction, bucketing therefore reduces the number of points, usually from $m > 15000$ to $m^b < 1000$.

In this step, there is a trade-off between keeping spectral information and removing peaks shifts as well as decreasing the total number of variables. Among possible binning methods, PepsNMR's bucketing proposes two integration options, either trapezoidal or rectangular, with equally sized buckets and generalised to cut the original x-axis at any chosen location. Fig. 6 illustrates the two possibilities. The function Bucketing (Spectrum_data, mb = 500, …) reduces the spectra to m^b buckets.

2.2.13. Region removal

Included by default in the selected spectral window, resonances without interest can eventually alter data analysis introducing unwanted variation. With biological samples, one major problem is

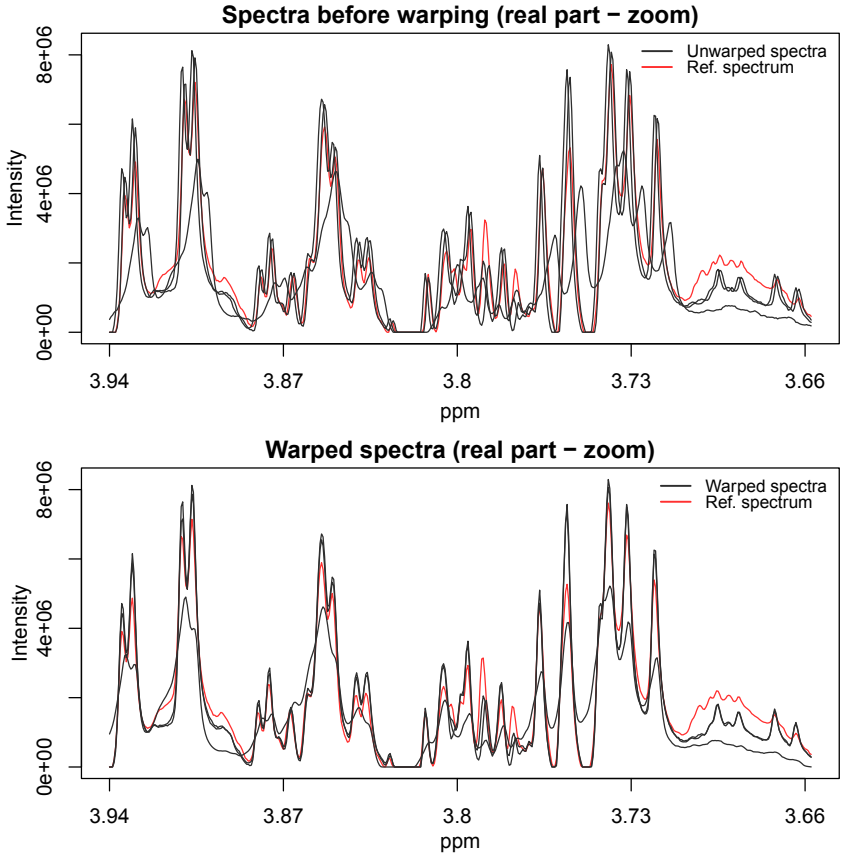


Fig. 5. Spectra are horizontally warped to remove peak shifts and to more accurately compare intensity differences.

the presence of water resonance residuals and positive values created by the baseline correction step, even with a previous application of pre-saturation and solvent signal suppression techniques [3]. By default, the function RegionRemoval () removes urea, maleic acid and water (4.5–6.1 ppm) peaks for urine spectra, whereas the water resonance area (4.5–5.1 ppm) is only removed by default for serum spectra.

2.2.14. Region aggregation

Strong variations in peak position are particularly present for ionised metabolites such as citrate, taurine, succinate or lactate [26] and cannot be totally avoided through sample preparation procedures (*e.g.* pH buffer), constrained warping methods or soft bucketing.

To avoid this unwanted source of variation, targeted data reduction can be applied with the function ZoneAggregation () in PepsNMR. Peaks are appropriately reshaped with the aggregation technique suggested in Rousseau [12]: the total intensity in an $[a, b]$ ppm interval $T = \sum_{j=a}^{j=b} F_j$ is distributed across this area so that the metabolite has a unique symmetrical triangular peak and null intensities at the border. A future version of the package will include the possibility of aggregating intensities into a Lorentzian curve.

2.2.15. Normalisation

Normalisation is defined as a row operation [25] $F_i \times c_i$ where each spectrum i is multiplied by a constant term c_i. It mostly addresses the dilution factor problem: variations in absolute concentrations can be observed between biofluid samples whereas data analysis focuses only on relative peak differences between them. This variation is especially present in urine samples that do not possess homeostasis properties such as serum samples do, but other technical issues can also create this unwanted variation. Various normalisation methods coexist in the literature (see for example Craig et al. [25] or Smolinska et al. [3], Marion [22]).

The constant sum normalisation (CSN) technique is advised for sera, cell media or biopsy samples where large differences between groups of patients are not expected. With this method, $c_i = \frac{m^b}{\sum_{j=1}^{m^b} F_{ij}}$. The mean intensity for each spectrum becomes m^b. The probabilistic quotient normalisation (PQN) [30] — specially designed for urine spectra and for distinguishing the effects of overall and local metabolite concentration changes on the spectrum — is another popular method included in PepsNMR. Other normalisation techniques available are: by the median, by the first quartile or by the intensity of an endogenous metabolite, usually creatinine for urine samples whose concentration is then assumed to be constant across samples under certain conditions [31]. Guidance on the choice of normalisation approach based on the type of sample analysed can be found in Wu and Li [32].

2.3. Other PepsNMR features and functions

For the sake of reproducibility, PepsNMR gives the possibility to modify, view and save spectra together with their acquisition parameters as well as pre-processing parameters during or after the whole pre-processing workflow. These different features are explained below.

The PreprocessingChain() function A wrapper for data import and pre-processing steps. When working with this function, the

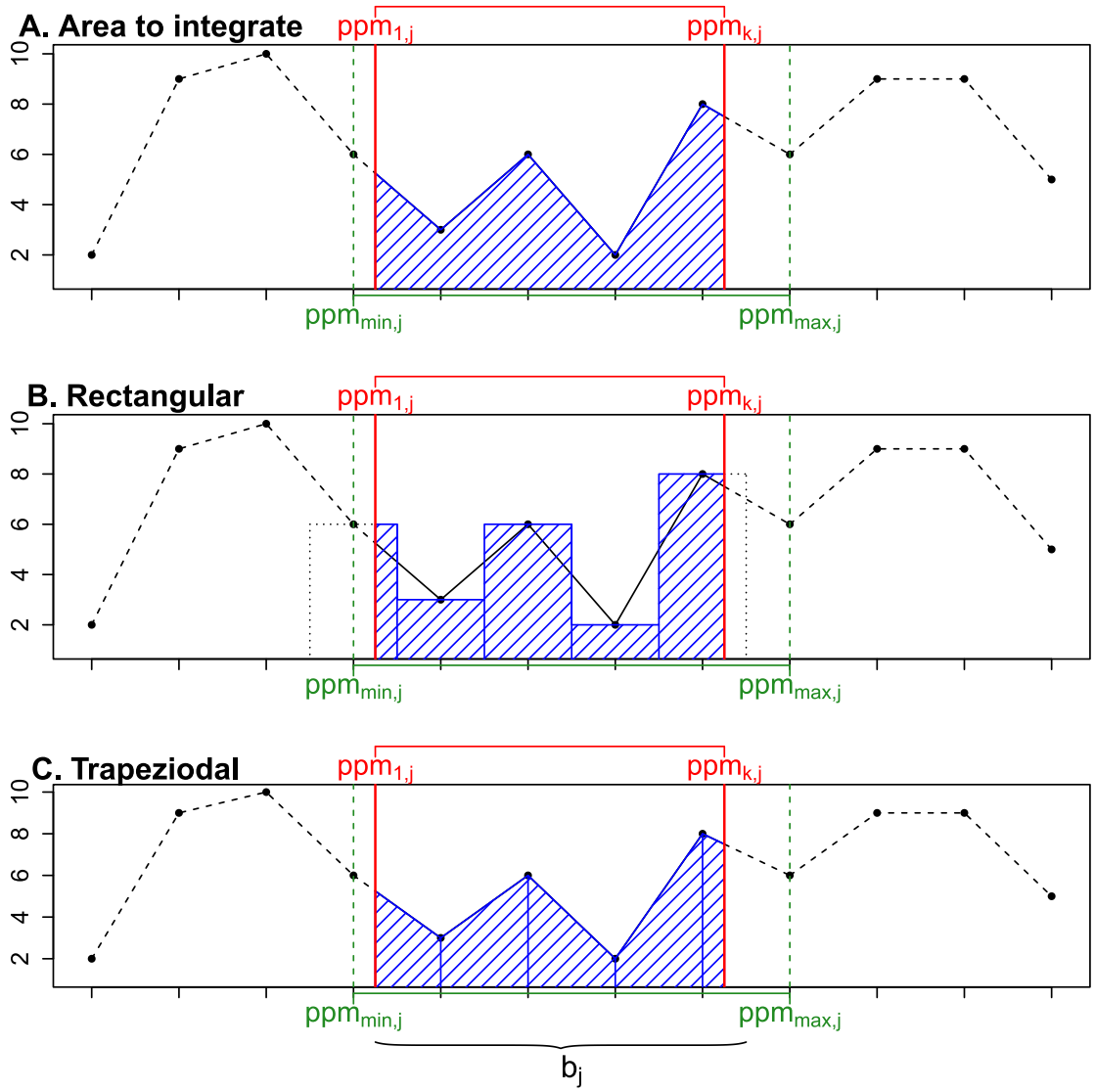


Fig. 6. Illustration of the target area to be integrated and the two bucketing methods (B. and C.) in PepsNMR.

pre-processed spectral intensities and acquisition parameters matrices can be exported (.RData;.txt or.csv) as well as a text file with all the pre-processing input parameters to ensure traceability.

Functions parameters For most pre-processing steps, different input parameters can be modified to adjust the methods to every dataset. Default values are in general assigned to most of those parameters in the package but can also be tuned by the user according to visual support, this article's guidelines and the related *help* support in R. They can also be optimised with an appropriate experimental design and/or objective quality criteria as explained below in Section 3.

Functions order Some pre-processing blocks can be switched or bypassed with respect to the advised order, while some corrections should precede other steps (*e.g.* solvent suppression before baseline correction or warping before bucketing) and several are compulsory (*e.g.* Fourier transform or phase corrections). As for their parameters, quality criteria can guide the choice of an appropriate function order.

Visualisation tools After each step, data can quickly be visualised to check the effects of the applied pre-processing method(s) and detect unexpected or abnormal results that could mislead data analysis, but also to spot outliers and global trends [7]. PepsNMR provides different ggplot2-themed plotting possibilities with the Draw () function to represent either the signal/spectral intensities or PCA scores and loadings. They can be directly displayed in the R interface or saved as png/pdf to be recovered later.

3. Quality criteria for PepsNMR evaluation

In the same perspective but beyond the scope of this article is the importance of relying on objective quality criteria to assess and quantify pre-processing efficiency leading to spectra with more informative content. The development and wide use of such quality parameters, as named by Engel et al. [8], are still lacking and should be investigated more intensively across the scientific literature. Indeed, sub-optimal pre-processing strategies can deteriorate relevant spectral information and affect study conclusions. However, currently there is no global consensus on which selection criteria other than visual inspection, should be used to test pre-processing efficiency. In this paper, the evaluation of PepsNMR

performance is done as suggested by Engel et al. [8]: the decision relies on a robust chemometric approach consisting of different quantitative quality parameters that aggregate into a super-parameter. This approach can either be used to compare the pre-processing efficiency of competing strategies or to appropriately calibrate the PepsNMR function parameters and order.

This approach requires a validation dataset to be collected for every dataset under study in order to properly compare and calibrate the pre-processing strategies based on quantitative quality criteria. Indeed these criteria measure the spectral repeatability from a designed validation dataset with technical and/or biological replicates (*e.g.* subsamples from the same biological unit analysed at different times or involving different dilutions). The spectral repeatability is defined as the ratio between the repeatable, structured group "signal" from the replicates and i.i.d. noise under the hypothesis that the sampled spectra are composed of both structured signal and random noise. Note that this validation dataset should be similar but independent of the dataset of interest in the study.

In the next Section, quality criteria are measured for distinct pre-processing strategies that are applied on to identical validation dataset with known replicates class to rank their efficiency. These criteria are derived from unsupervised as well as supervised chemometric tools. They are gathered under the Metabolomic Informative Content (MIC) concept [33] and include Total Variance Decomposition (TSV), PCA, clustering and PLS-DA related criteria. Applying a Sums of Squares decomposition on the spectral matrix breaks down the total variance into two complementary parts: the variance between the groups and the variance within the groups of observations. For clustering, the (adjusted) Rand indexes measure the true class recovery efficiency and should be maximised, while the Dunn and Davies-Bouldin indexes measure the clustering homogeneity, and they have to be respectively maximised and minimised (formula details can be found in Féraud et al. [33]). PLS-DA builds a latent structure with Latent Variables (LV) to extract the variance components from the data but instead of maximising the spectral data matrix variance as PCA does, it maximises the variance of both the spectral data matrix (the regressors) and the class labels (the response) and their correlation. Its cross-validation criterion Q^2 assesses model efficiency in terms of class prediction.

The incentive to use unsupervised methods (PCA and K-means clustering) is to contrast the amount of class information that we are able to recover in the data after implementing different pre-processing pipelines, provided that the between-group variation is the main source of variability in the data. This information recovery is a blind process: no prior information is provided about the class labels. The supervised classification algorithm (PLS-DA) completes the performance analysis since it constitutes standard multivariate data analysis that a researcher would use to discriminate between different classes of spectra.

4. Results and discussion

In this section, the advanced pre-processing methods in PepsNMR are compared to a manual data pre-processing with TopspinTM 3.1 and AMIX 3.9.14 on the grounds of two typical biofluid media: human serum (HSerum) and human urine (HUrine) datasets. To that end, the data tables are fully pre-processed either with PepsNMR or with Topspin/AMIX. The HSerum dataset comes from 4 different blood donors whereas the HUrine dataset is measured from 3 different urine donors. The both of them are designed in such a way that the biological information in between-donor variability should be the main factor affecting spectral shape. Noisy parameters that were intentionally added to the data to reproduce standard analytical conditions are: day of measurement, sample replicates and sample dilution. More details on these data, their acquisition and pre-processing are available in Section 7.

4.1. Variance structure

If the pre-processing steps manage to accurately transform the data and correct for technical sources of variation, the main source of variability in the two case studies would correspond to the biological differences between the donors.

As explained in Section 3, TSV and PCA are first applied to characterise and quantify the amount of biological information recovered. Variance decomposition results are given in Table 1. For the two datasets, the variance between the groups of samples is larger with PepsNMR than with manual pre-processing. This difference in values for HUrine is small (+16.7%), since the manual pre-processing is already effective, but a real improvement can be observed for the HSerum data (+67.8%): a larger amount of TSV is captured by the variability among the groups rather than between them. Since a larger between-group variance implies better group separation, PepsNMR performs better than the manual alternative.

Table 1 shows that the sum of Principal Components (PC) 1 and 2 is similar for the two pre-processing approaches for the HSerum dataset but the group structure is only captured by PC2 for the manual pre-processing, as can be seen in the scores plot in Fig. 7. The main source of variation (79.53%) has another origin: significant peak shifts are observed in the PC1 loadings plot (Fig. 8). They are hiding the information of interest. Scores plot of PC1 and PC2 for the HUrine dataset shares the same structure, regardless of the pre-processing mode. Spectral profiles are closely tied together and groups are clearly separated. PC1 discriminates between donor 3 and the two others, whereas PC2 discriminates patient 1 from donor 2. Moreover, the PepsNMR workflow for this dataset recovers more informative variability in the two first components (+16.1%) than the manual pre-processing.

4.2. Clustering and classification performance

In this Sub-Section, the unsupervised and supervised classification algorithms (K-means clustering and PLS-DA) explained in Section 2.5 are applied to both datasets. Parameter K for clustering is the number of donors. The number of LVs for PLS-DA is chosen based on the cross-validated Root-Mean-Square Error (RMSE): 4 for HSerum and 2 for HUrine. Accurate classification and homogeneity of groups are examined to evaluate the unsupervised tool while the predictive quality criterion Q^2 is used to rate the PLS-DA performance. In the case of a null model, $Q^2 = 0$, and with a perfectly discriminating model, $Q^2 = 1$. Results from these two approaches can be visualised in Table 2.

Clustering-related indexes quantify either the similarity between two partitions (the true class and the one obtained with blind clustering) or the cluster homogeneity. For the HSerum, higher Rand and Adjusted Rand indexes for PepsNMR clearly indicate a better recovery of the true classes of samples, as well as better group

Table 1
Repeatability is partly assessed by variance decomposition methods: total variance decomposition (% of between and within variance) and PCA (cumulated % of variance for the first two PCs).

Dataset	Pre-processing	Inertia		PCA variance	
		Between (%)	Within (%)	PC1 (%)	PC2 (%)
HSerum	manual	21.4	78.6	79.5	11.0
HSerum	PepsNMR	89.2	10.8	76.2	14.6
HUrine	manual	82.2	17.8	67.5	15.4
HUrine	PepsNMR	98.9	1.1	79.4	19.6

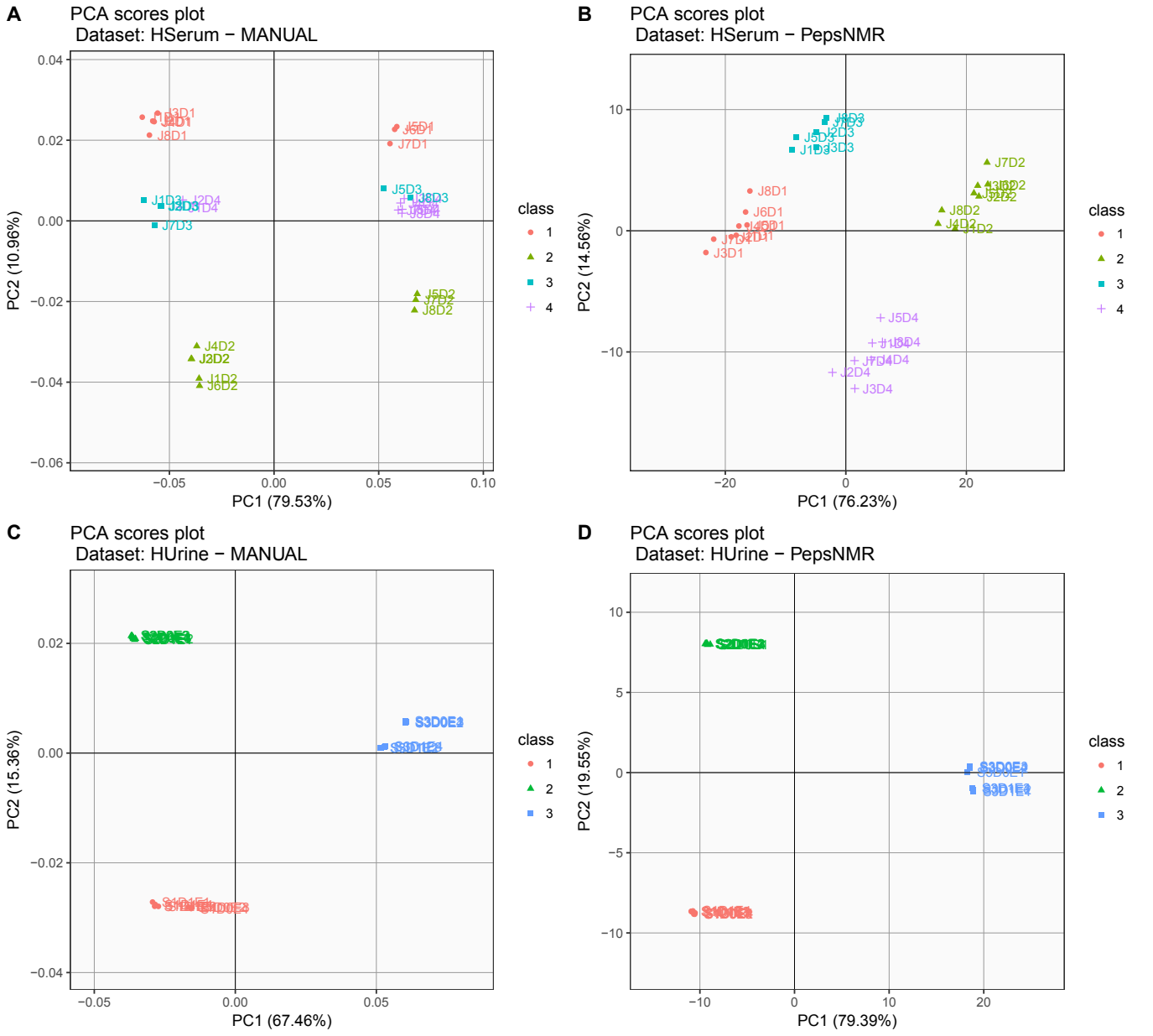


Fig. 7. PCA scores plot of PC1 and PC2 for the HSerum (A–B) and HUrine (C–D) datasets and the two pre-processing strategies (manual and PepsNMR). The colours represent the donor IDs. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

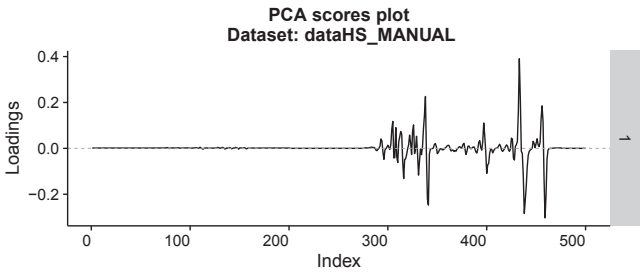


Fig. 8. First PCA loadings plot of the manually pre-processed HSerum dataset.

homogeneity according to the Dunn index. Less importantly, even though the Davies-Bouldin index shows a better group homogeneity for the manual pre-processing, the grouping does not represent correctly the classes. As for the variance structure analysis, the two pre-processing techniques for the HUrine dataset lead to analogous clustering results illustrated here by a perfect categorisation of samples. Nonetheless, clusters with the PepsNMR pre-processed datasets do have better homogeneity criteria.

Findings with the PLS-DA are consistent with the previous results: the classification algorithm has a very bad cross-validation performance for the manually pre-processed human serum dataset since the Q^2 is negative. With the PepsNMR strategy, the performance criterion is substantially improved. For the urine dataset, Q^2 is almost 1, reflecting the excellent class prediction for both pre-processing strategies.

5. Conclusions

Spectroscopic data pre-processing is a keystone procedure for

Table 2
Clustering/classification results for repeatability assessment: indexes derived from unsupervised K-means clustering ((Adjusted) Rand, Dunn and Davies-Bouldin (D-B) indexes) and a supervised PLS-DA validation criterion (the cross-validated coefficient of determination Q^2).

Dataset	Pre-processing	K-Means clustering				PLS-DA
		Rand	Adj. Rand	Dunn	D-B	Q^2
HSerum	manual	0.64	0.14	0.62	0.56	−0.47
HSerum	PepsNMR	0.88	0.72	0.66	0.69	0.73
HUrine	manual	1	1	0.83	0.67	0.99
HUrine	PepsNMR	1	1	3.71	0.17	1

enhanced information recovery in metabolomics, and PepsNMR is the only existing R package able to comprehensively address a very-detailed series of pre-processing steps designed for 1D ^{1}H NMR data. The two designed experiments comparing PepsNMR to a classic pre-processing strategy confirm its practical interest: its use can increase spectral repeatability, suggesting an overall better recovery of information, and it can enhance predictive power in a classification context. In addition, the package has built-in assets: it is open-source, it has routine flexibility, automation and objectivity, as well as structured documentation, and yields reproducible results and reporting. Although it is not possible to ensure that the package always performs better than classic pre-processing, its flexibility makes it possible to combine parts of it to other available tools for greater efficiency. Furthermore, this package should help to continuously improve the processing of metabolomic data characterised by a high degree of complexity, and in the future, complementary/alternative methods will be added to this first version. Methods from PepsNMR are already available in Workflow4Metabolomics [34], a web-interface dedicated to metabolomic data processing, analysis and annotation, where it works complementarily with other tools. Last but not least, the selection of optimal parameters as well as the set up of an optimal procedure for method selection is beyond the scope of this article. Methodological needs in this field [8] open up vast research possibilities.

6. Software availability

PepsNMR is available on GitHub (https://github.com/ManonMartin/PepsNMR). It has common R package characteristics: it can be easily installed via the following code lines: require (devtools); install_github ("ManonMartin/PepsNMR") in the R console, and has *help* pages for all user functions and internal datasets. A vignette with a practical pre-processing application is also available in the package.

Datasets used in this article are available in PepsNMR at various processing stages.

The additional R routines used to validate the pre-processing workflow can also be accessed in the MBXUCL R package (https://github.com/ManonMartin/MBXUCL).

7. Datasets

The package is applied to two real datasets from well-known human biofluid: serum (HSerum) and urine (HUrine), to demonstrate its performance. These datasets were both designed to collect multiple measures from the same experimental unit and capture other technical sources of variation, allowing the comparison of inter- and intra-unit variability. These datasets are available upon request.

7.1. Human serum

In the HSerum dataset, a blood sample was collected from 4 different donors. For each sample, 8 sub-samples were measured across 8 days with one sub-sample from each donor per day and permutations according to a latin hypercube sampling method. The total number of available FID signals is then $4 \times 8 = 32$. Data were acquired with a 500 MHz Bruker Avance spectrometer equipped with a TCI cryoprobe and using a CPMG relaxation-editing sequence with pre-saturation. Spectra are labelled as: JxDx where Jx is the day of measurement and Dx is the donor label. More details about this dataset can be found in Féraud et al. [33].

7.2. Human urine

In the second dataset, morning urine from 3 different donors was collected and directly aliquoted in eppendorfs with the following procedure: for each donor, 4 aliquotes containing 400 µl of urine each +300 µl of D2O buffer +10 µl of TMSP and 4 aliquotes containing 320 µl of urine +380 µl of D2O buffer and 10 µl of TMSP. Hence, each donor sample was divided into 2 sub-samples: one was kept pure and the other 25% diluted. The 4 aliquotes of each dilution were analysed on 4 different days. For each day, the order of measurement was held constant across the 6 sub-samples. The total number of collected FID signals is then: $3 \times 2 \times 4 = 24$. Spectra are labelled as: SxDx where x is the donor label and D is the dilution (0: no dilution; 1: 25% diluted).

7.3. Spectral pre-processing and outlier removal

Both datasets were pre-processed with PepsNMR and a combination of TopspinTM and Amix for manual pre-processing. With PepsNMR, the full advised workflow was applied with default parameters. The region removal and zone aggregation steps were adapted given the type of medium. For region removal, the water (4.5–5.1 ppm) area was removed for the serum while the water, urea and maleic acid (4.5–6.1 ppm) areas were removed for the urine medium. The zone aggregation step was only applied to the urine dataset for the citrate peak (2.5–2.7 ppm). The manual pre-processing included the following sequence: exponential apodization, Fourier transform, first and zero order phase corrections, internal calibration, polynomial baseline correction, bucketing and spectral region removal.

Outliers were detected with the ROBPCA [35] approach, a robust PCA used to detect spectra with a high score distance and/or with a high orthogonal distance to the PCA space. Two outlying observations were removed from the HSerum dataset (J6D4 and J6D3) and one from the HUrine dataset (S1D0E1). Note that a missing observation in the manual pre-processing dataset (J4D3) was completely removed for data analysis.

Acknowledgements

The first author gratefully acknowledges funding from the Belgian Fund for Scientific Research (F.R.S.-FNRS) with a FRIA grant. The F.R.S.-FNRS further supported this work with P. de Tullio as a

senior research associate. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) to the Institute of Statistics, Biostatistics and Actuarial Sciences is also gratefully acknowledged.

References

[1] A.K. Kosmides, K. Kamisoglu, S.E. Calvano, S.A. Corbett, I.P. Androulakis, Metabolomic fingerprinting: challenges and opportunities, Crit. Rev. Biomed. Eng. 41 (3) (2013) 205–221, https://doi.org/10.1615/CritRevBiomedEng.2013007736.

[2] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015, Frontiers in bioengineering and biotechnology 3 (2015) 23, https://doi.org/10.3389/fbioe.2015.00023.

[3] A. Smolinska, L. Blanchet, L.M. Buydens, S.S. Wijmenga, Nmr and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review, Anal. Chim. Acta 750 (2012) 82–97, https://doi.org/10.1016/j.aca.2012.05.049, 750th Anniversary Volume, http://www.sciencedirect.com/science/article/pii/S000326701200815X.

[4] T. Gebregiworgis, R. Powers, Application of nmr metabolomics to search for human disease biomarkers, Comb. Chem. High Throughput Screen. 15 (8) (2012) 595–610, https://doi.org/10.2174/138620712802650522. http://www.ingentaconnect.com/content/ben/cchts/2012/00000015/00000008/art00002.

[5] A.D. Maher, S.F. Zirah, E. Holmes, J.K. Nicholson, Experimental and analytical variation in human urine in 1h nmr spectroscopy-based metabolic phenotyping studies, Anal. Chem. 79 (14) (2007) 5204–5211, https://doi.org/10.1021/ac070212f pMID: 17555297.

[6] E.M. Lenz, I.D. Wilson, Analytical strategies in metabonomics, J. Proteome Res. 6 (2) (2007) 443–458, https://doi.org/10.1021/pr0605217 pMID: 17269702.

[7] K.H. Liland, Multivariate methods in metabolomics — from pre-processing to dimension reduction and statistical analysis, Trac. Trends Anal. Chem. 30 (6) (2011) 827–841, https://doi.org/10.1016/j.trac.2011.02.007. http://www.sciencedirect.com/science/article/pii/S0165993611000914.

[8] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M. Buydens, Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50 (2013) 96–106, https://doi.org/10.1016/j.trac.2013.04.015. http://www.sciencedirect.com/science/article/pii/S0165993613001465.

[9] L. Buydens, Towards tsunami-resistant chemometrics, The analytical Scientist 813 (2013) 24–30.

[10] R. Goodacre, D. Broadhurst, A.K. Smilde, B.S. Kristal, J.D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D.B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics 3 (3) (2007) 231–241, https://doi.org/10.1007/s11306-007-0081-3.

[11] S. Ren, A.A. Hinzman, E.L. Kang, R.D. Szczesniak, L.J. Lu, Computational and statistical analysis of metabolomics data, Metabolomics 11 (6) (2015) 1492–1513, https://doi.org/10.1007/s11306-015-0823-6.

[12] R. Rousseau, Statistical Contribution to the Analysis of Metabonomics Data in 1h Nmr Spectroscopy, Ph.D. thesis, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, 2011.

[13] J. Vanwinsberghe, Bubble: Development of a Matlab Tool for Automated 1h-nmr Data Processing in Metabonomics, Traineeship report (unpublished results), Strasbourg University, 2005.

[14] J. Keeler, 4 fourier transformation and data processing, in: Understanding NMR Spectroscopy, John Wiley & Sons, 2002, pp. 48–65, https://doi.org/10.17863/CAM.968.

[15] W.M. Siebert, Circuits, Signals, and Systems, vol 2, MIT press, 1986.

[16] T.D. Claridge, Chapter 3-practical aspects of high-resolution {NMR}, in: T.D. Claridge (Ed.), High-resolution {NMR} Techniques in Organic Chemistry, third ed., Elsevier, Boston, 2016, pp. 61–132, https://doi.org/10.1016/B978-0-08-099986-9.00003-8.

[17] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (14) (2003) 3631–3636, https://doi.org/10.1021/ac034173t pMID: 14570219.

[18] E.T. Whittaker, On a new method of graduation, Proc. Edinb. Math. Soc. 41 (1922) 63–75, https://doi.org/10.1017/S0013091500077853.

[19] G. Frasso, P.H. Eilers, L- and v-curves for optimal smoothing, Stat. Model. Int. J. 15 (1) (2015) 91–111. https://doi.org/10.1177/1471082X14549288.

[20] L.R. Euceda, G.F. Giskeødegård, T.F. Bathen, Preprocessing of nmr metabolomics data, Scand. J. Clin. Lab. Investig. 75 (3) (2015) 193–203, https://doi.org/10.3109/00365513.2014.1003593 pMID: 25738209.

[21] K.H. Liland, T. Almøy, B.-H. Mevik, Optimal choice of baseline correction for multivariate calibration of spectra, Appl. Spectrosc. 64 (9) (2010) 1007–1016, https://doi.org/10.1366/000370210792434350 pMID: 20828437.

[22] R. Marion, Pre-processing of Nmr Spectra: Review and Evaluation of Baseline Correction, Normalization, Scaling and Transformation Methods, Master thesis (unpublished results), Ecole de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, 2016.

[23] P.H.C. Eilers, H.F. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, Medical Centre Report (unpublished results), Leiden University, 2005.

[24] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2) (2004) 404–411, https://doi.org/10.1021/ac034800e pMID: 14719890.

[25] A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson, J.C. Lindon, Scaling and normalization effects in nmr spectroscopic metabonomic data sets, Anal. Chem. 78 (7) (2006) 2262–2267, https://doi.org/10.1021/ac0519312 pMID: 16579606.

[26] T.G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, L.M. Buydens, Warping methods for spectroscopic and chromatographic signal alignment: a tutorial, Anal. Chim. Acta 781 (2013) 14–32, https://doi.org/10.1016/j.aca.2013.03.048. http://www.sciencedirect.com/science/article/pii/S0003267013004224.

[27] T.N. Vu, K. Laukens, Getting your peaks in line: a review of alignment methods for nmr spectral data, Metabolites 3 (2) (2013) 259, https://doi.org/10.3390/metabo3020259. http://www.mdpi.com/2218-1989/3/2/259.

[28] A. van Nederkassel, M. Daszykowski, P. Eilers, Y.V. Heyden, A comparison of three algorithms for chromatograms alignment, J. Chromatogr. A 1118 (2) (2006) 199–210, https://doi.org/10.1016/j.chroma.2006.03.114. http://www.sciencedirect.com/science/article/pii/S0021967306007059.

[29] P.H.C. Eilers, B.D. Marx, Flexible smoothing with *b*-splines and penalties, Stat. Sci. 11 (2) (1996) 89–102. http://www.jstor.org/stable/2246049.

[30] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics, Anal. Chem. 78 (13) (2006) 4281–4290, https://doi.org/10.1021/ac051632c pMID: 16808434.

[31] K.W.A. Tang, Q.C. Toh, B.W. Teo, Normalisation of urinary biomarkers to creatinine for clinical practice and research—when and why, Singap. Med. J. 56 (1) (2015) 7.

[32] Y. Wu, L. Li, Sample normalization methods in quantitative metabolomics, J. Chromatogr. A 1430 (2016) 80–95.

[33] B. Féraud, B. Govaerts, M. Verleysen, P. de Tullio, Statistical treatment of 2d nmr cosy spectra in metabolomics: data preparation, clustering-based evaluation of the metabolomic informative content and comparison with 1h-nmr, Metabolomics 11 (6) (2015) 1756–1768, https://doi.org/10.1007/s11306-015-0830-7.

[34] F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thévenot, C. Caron, Workflow4metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics 31 (9) (2015) 1493, https://doi.org/10.1093/bioinformatics/btu813.

[35] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, Robpca: a new approach to robust principal component analysis, Technometrics 47 (1) (2005) 64–79, https://doi.org/10.1198/004017004000000563.