Table of Content Abstract:

Cleaning up NMR spectra with reference deconvolution for improving multivariate analysis of complex mixture spectra

Parvaneh Ebrahimi[1*], Mathias Nilsson[1, 2], Gareth A. Morris[2], Henrik M. Jensen[3], and Søren B. Engelsen[1]

[1] Department of Food Sciences, Copenhagen University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

[2] School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

[3] DuPont Nutrition Biosciences ApS (Edwin Rahrs Vej 38, Brabrand)

[*] Corresponding author:

Parvaneh Ebrahimi, Department of Food Sciences, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark, E-mail: parvaneh@food.ku.dk, Phone: +45 353 32971

Short Abstract:

The aim of this study is to investigate how reference deconvolution can improve the results obtained by multivariate analysis of NMR data. ^{1}H NMR data was recorded for a set of samples and spectra were then produced with and without reference deconvolution, and subsequently, analyzed by PCA and PLS. The results confirmed that reference

deconvolution resulted in simpler and improved models, requiring fewer latent variables to explain the same or higher percentage of the variance.

Abstract

NMR spectroscopy provides valuable data for metabolomics, but the information sought can be partly obscured by errors from hardware imperfection, causing frequency, phase and spectral lineshape to change significantly between measurements. Clearly, this is a highly undesirable source of variation in multivariate quantitative studies such as metabolomics. Fortunately, many hardware imperfections affect all resonances in the same way. They can therefore be corrected for by comparing an experimental reference peak with the known correct peak shape, in a procedure known as reference deconvolution. This post-measurement processing method can correct many systematic errors in data. The aim of this study is to investigate how reference deconvolution can improve the results obtained by multivariate analysis of NMR data. For this purpose, ^{1}H NMR data were recorded for a set of 136 mixture samples. Spectra were then produced with and without reference deconvolution, and analyzed by Principal Component Analysis (PCA) and Partial Least Squares (PLS). The results showed that reference deconvolution resulted in simpler and improved models, requiring fewer latent variables to explain the same or higher percentage of the variance. It was also evident that the recovery of the design concentrations was significantly enhanced. This confirms that reference deconvolution can significantly improve multivariate data analysis and should be considered as a standard tool in high throughput quantitative NMR spectroscopy.

Keywords: reference deconvolution, NMR, multivariate data analysis, PCA, PLS

1. Introduction

The quality of NMR spectra has improved substantially with recent improvements in spectrometer design and manufacture. Despite these improvements, significant instrumental imperfections remain, and these are often the limiting factor in determining the amount and quality of information obtainable from NMR experiments. This is particularly true for experiments involving multiple data acquisitions, such as multidimensional NMR methods and chemometric studies, for which instrumental reproducibility is vital. Most instrumental imperfections affect all the signals in a spectrum in the same way (1). For example, magnetic field inhomogeneity broadens all lines to the same extent, radiofrequency pulse phase error imposes the same phase shift on all signals, and receiver gain variation changes all signal amplitudes equally. Other imperfections in NMR data vary from signal to signal, for example shifts in peak position due to changes in temperature, concentration and pH. All of these types of imperfection in NMR data represent sources of unwanted, non-relevant variance, and can blur the picture obtained from the biological variance in a metabolomics study, or any investigation relying on quantitative pattern recognition. Any method that can reduce or suppress such unwanted spectral distortion is a welcome addition to the arsenal of methods available to the data analyst.

One data processing method that is highly effective at correcting systematic errors in NMR data is reference deconvolution (1-5). This extracts the signal of a known reference material from the experimental data, compares it to the theoretically expected form, and constructs the correction function needed to convert the full experimental dataset into the form it would have had if the unwanted perturbations experienced by the reference signal had not been present. The reference signal should ideally be a well resolved singlet, of high signal-to-noise ratio (S/N), for which the theoretical lineshape is known (1). Typical examples of suitable signals are those from 3-(trimethylsilyl)propanoic acid (TSP) and tetramethylsilane (TMS), compounds that are commonly added to NMR samples to provide an internal standard for quantification and calibration of the chemical shift axis. Reference deconvolution is fast, linear (to a good approximation –the noise structure is changed slightly because the experimental noise in the reference region is convoluted onto the full spectrum), and robust; it has been known to NMR spectroscopists for many years, but it appears to have been neglected by the NMR-based metabolomics/chemometrics community. The algorithm used in this article, Free Induction Decay Deconvolution for Lineshape Enhancement (FIDDLE), has been used in a wide variety of different contexts (1-9), but has yet to be applied to chemometrics.

This study investigates how reference deconvolution can help in multivariate data analysis of NMR data. For this purpose, a ternary experimental design was prepared of 136 mixture samples with different concentrations of lactic acid, propionic acid, and lactose, and a constant artificial 'metabolic' background consisting of eight different amino acids and carbohydrates. ^{1}H NMR spectra were acquired using a standard metabolomics protocol (10), except that a higher than usual concentration of the reference material (TSP) was used. The effect of reference deconvolution was then investigated by subjecting the corrected and uncorrected experimental data to two of the most common data mining methods, principal component analysis (PCA) and partial least squares (PLS). PCA and PLS models from the corrected data were superior to those from the uncorrected spectra, demonstrating the ability of reference deconvolution to reduce systematic imperfections in NMR data and, in turn, to improve the consistency of a spectral dataset.

2. Theory

A number of different algorithms have been proposed for reference deconvolution (2-4), but they are all based on the same foundations. The FIDDLE (Free Induction Decay Deconvolution for Lineshape Enhancement) algorithm is effective and simple; the theoretical basis has been discussed extensively in the literature (1,5,7,8,11), but a graphical illustration of the key elements is shown in Figure 1 and explained in the following.

The NMR time-domain data, the free induction decay or FID (Fig. 1a), are zero-filled (to retain all the spectral information), Fourier transformed (FT) and phase corrected, to yield the raw NMR spectrum (Fig. 1b). A suitable reference signal in the spectrum is then chosen and the rest of the spectrum set to zero. The real part (the absorption mode) of this filtered spectrum is subjected to inverse Fourier transformation (IFT) to give a complex FID that contains only the reference signal (Fig. 1c). Choosing to retain only the real part of the reference spectrum excludes dispersion mode signals, making clean extraction of the reference signal much easier; no information is lost if the initial FID is zero-filled (6). In parallel, a synthetic FID (Fig 1e) is calculated for the reference signal, using the known frequency (or frequencies; in the case of a reference such as TSP, ^{29}Si and ^{13}C satellite signals are included) and a specified line shape. The latter is chosen by the user, according to need; while the true theoretical line shape is typically a Lorentzian, it can often be advantageous to use a Gaussian shape as this has a narrower base. This choice of target lineshape is analogous to the choice of window function (apodization) in normal FT processing, and the same considerations for resolution or sensitivity enhancement apply (8). The most conservative choice is a Lorentzian lineshape of approximately the same width as the experimental reference signal (Fig. 1d); this regularizes the lineshape (and phase and frequency) with minimum change in resolution and signal-to-noise ratio. A complex correction function is then constructed by dividing the ideal reference FID (Fig. 1e) by the experimental reference FID (Fig. 1c). The cumulative effect of instrumental imperfections such as field inhomogeneity, pulse phase error, modulation sidebands etc. is to multiply the FID that would have been recorded if the instrument had behaved ideally by a complex time-domain error function. The correction function calculated here is the inverse of that function, so when the original (full) experimental FID (Fig. 1a) is multiplied by it, the result is a corrected FID (Fig. 1f) in which all the multiplicative errors seen in the reference FID have been corrected. The corrected FID can then be Fourier transformed to yield the reference deconvoluted spectrum (Fig. 1g), in which such imperfections as lineshape distortions, signal amplitude errors and signal phase changes have been corrected (5,7-9).

For best results, the reference peak should be a well-resolved singlet which is present with high amplitude in all the spectra being deconvoluted (1). The noise in the vicinity of the reference signal will be convoluted onto the entire spectrum, so if the signal-to-noise ratio of the reference signal is too low, it can significantly degrade the quality of the data (5). Multiplets are a much poorer choice for reference signals as they have FIDs that have zero amplitude at regular intervals, which results in singularity problems that are mathematically challenging (9). The zeroes make interpolation necessary, introducing an element of non-linearity into the algorithm. While the use of a doublet as the reference signal has been reported (12), most software for reference deconvolution does not cater for multiplet reference signals. In this study only singlet reference signals have been used.

The choice of the ideal peak lineshape and linewidth (the “target lineshape”) is important, and warrants further discussion. The lineshape chosen for the ideal reference signal is typically Lorentzian, Gaussian or a mixture of the two (1), although there are many other possibilities. As noted above, there is a close analogy between the choice of target lineshape and the apodization procedure used in conventional Fourier transform processing. Since most reference signals have a Lorentzian natural shape, and the effects of static field inhomogeneity also often approximate to a Lorentzian distribution of signals as a function of frequency, the choice of a Lorentzian target lineshape with a width close to that of the experimental reference line will produce a spectrum similar in appearance to the original, but with errors in lineshape, phase, frequency etc. corrected. However, it is often useful to change the target lineshape to aid the extraction of the features of interest from the data under analysis. If a Lorentzian target lineshape narrower than the experimental reference line is chosen, resolution will be increased, but at a severe cost in signal-to-noise ratio; if too narrow a lineshape is used, numerical instabilities in the correction will cause severe spectral distortions. Choosing a target lineshape wider than the experimental reference line will increase the signal-to-noise ratio at a cost in resolution, with a maximum S/N improvement at twice the experimental linewidth (so-called matched filtration). The choice of a Gaussian or mixed lineshape is often a good alternative, as the narrow base of a Gaussian improves resolution, but at a moderate cost in sensitivity. The optimum target lineshape naturally depends on the objective of the analysis, and comparison between spectra corrected with different target lineshapes is often worthwhile.

3. Materials and Methods

3.1.Experimental Design

A ternary mixture of lactic acid, propionic acid, and lactose was designed using JMP software, Version 9 (SAS Institute Inc., Cary, NC, USA) with 16 increments from 0-15 mM for each component, which yielded a total of 136 mixtures (see experimental design in Figure 2). Each ternary mixture was prepared in distilled water and added to a 'metabolic background' consisting of a mixture of amino acids and carbohydrates (L-alanine, L-asparagine, L-glutamate, L-leucine, L-phenylalanine, sucrose, glucose, and galactose) at 15 mM each in distilled water. Sodium azide was added to prevent the growth of bacteria and fungi (20 mg per 100 mL of the metabolic background solution). Phosphate buffer with pH 7.4 was also prepared with deuterated water according to a protocol for biological samples (10) which includes TSP as a chemical shift reference. However, concentration of TSP was increased by a factor of 10, relative to the concentration in the original protocol, to 10 mM, in order to ensure high signal-to-noise-ratio for the TSP singlet to be used as the reference signal in reference deconvolution. The 10-fold increase in the concentration of TSP did not affect the pH of the buffer. To prepare samples for NMR measurement, 200 µL of the artificial ’metabolic’ background and 200 µL of the phosphate buffer were added to 200 µL of each ternary design mixture. In the final samples, the concentrations of the ternary design components varied between 0 and 5 mM.

3.2. NMR Data Acquisition and Processing Methods

^{1}H NMR spectra of the samples were recorded on a Bruker DRX 500 spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at a proton frequency of 500.13 MHz. For each spectrum, 32 768 complex points were acquired in 64 scans with a recycle delay of 2 seconds at a nominal temperature of 298 K. The spectrometer was equipped with a 5 mm BBI probe and spectra were recorded using the one-dimensional (1D) NOESY for suppression of the solvent (water) signal. All processing of the data, including phase correction, apodization, Fourier transformation, baseline correction, referencing to TSP signal, and reference deconvolution, was performed using the DOSY Toolbox (13). Spectra were processed with and without reference deconvolution. Linewidths are expressed as full widths at half-height throughout this paper. Reference deconvolution was performed using the TSP methyl signal as reference, using Gaussian or Lorentzian lineshapes with linewidths ranging from 1 to 5 Hz in 0.25 Hz increments. In order to ensure comparability, FIDs that were not reference deconvoluted were weighted with Gaussian and Lorentzian apodization functions adjusted to give reference linewidths corresponding as closely as possible to those obtained using reference deconvolution. For example, to make models of conventional data and reference deconvoluted with 3 Hz Lorentzian data comparable, line broadening was added to the FID in conventional data to make the width of the reference peak equal to that of the reference signal. The resultant spectra from the DOSY Toolbox were imported into MATLAB 2012b (MathWorks, Inc., Natick, MA, USA) and further processed by normalizing the spectra relative to the TSP signal area. The Matlab code for the DOSY toolbox is freely available from www.models.life.ku.dk.

3.3. Multivariate Analysis

Prior to the multivariate analysis, spectral regions containing only noise, water or TSP signals were removed from the data. The PLS Toolbox, Version 7.0 (Eigenvector Research, Inc., WA, USA) was used for the multivariate analysis. Principal Component Analysis (PCA) models (14) were calculated for mean-centered datasets. Partial Least Squares (PLS) models (15) between the mean-centered data and concentration of lactic acid in the samples were also calculated, and cross-validated by the leave-one-out method. Two of the samples were in all cases identified as score outliers (outside the limit of confidence in the primary scores plot) and removed from the datasets.

4. Results and Discussion

Selected regions of the conventional and reference deconvoluted spectra from the 136 samples are shown in Figure 3. The spectra were reference deconvoluted with a 1.5 Hz

Lorentzian target lineshape. The experimental linewidths for the reference (TSP) signal in the spectra measured were around 1.5 Hz; the aim here was to correct spectral errors while minimizing any change in linewidth between uncorrected and corrected spectra, in order to facilitate comparison.

Comparing the conventional and reference deconvoluted spectra in Figure 3, it can be seen that the signals from the constant ‘metabolic’ background in the samples are much more consistent in the reference deconvoluted spectra. For these signals, reference deconvolution has significantly reduced the effects of experimental and instrumental irreproducibilities − which do not have a chemical/biological source − between the spectra. Inspecting the lactic acid doublet, it is also clear that in the reference deconvoluted spectra the lineshapes are much more consistent, and the 16 increments in concentration in the design can be easily observed. Depending on the nature and extent of the lineshape errors in the experimental data, reference deconvolution with a target linewidth equal to the experimental width can increase or decrease signal-to-noise ratio. The effect on signal-to-noise ratio here was, as expected for good quality data, marginal, the S/N ratio of the lactic acid doublet for the average spectrum in Figure 3 decreasing from 3.0×10^4 in the normal spectrum to 2.9×10^4 in the reference deconvoluted spectrum. Just as in conventional processing of NMR data, the target lineshape in reference deconvolution can be chosen to enhance either the sensitivity or the resolution of the spectrum. Typical choices are a Lorentzian target lineshape broader than the experimental reference line for the former, and a Gaussian lineshape narrower than the experimental reference line for the latter. If necessary, the target lineshape can be varied between datasets to maintain the desired balance between resolution and signal-to-noise ratio. Where the spectral lines of interest are naturally broader than that of the reference material, resolution enhancement is best achieved by choosing a target lineshape for the reference that contains a negative Lorentzian width contribution and a positive Gaussian (i.e. the corresponding time-domain function corresponds to a rising exponential multiplied by a decaying Gaussian). The negative Lorentzian contribution should correspond to the difference in natural linewidth between the signals of interest and the reference.

In order to optimize the target linewidth, reference deconvolution with a Gaussian linewidth varying from 1 to 5 Hz in 0.25 Hz increments was performed on all the spectra. The Gaussian lineshape was chosen because it represents a good compromise between resolution and S/N. For each increment, a PLS model was calculated between the spectral data and the concentration of lactic acid as the response variable. The resulted RMSECV and R^2_{CV} values as a function of target linewidth are plotted in Figure 4. It can be seen that linewidths between 2 and 3 Hz resulted in the lowest RMECV's and the highest R^2 values. As the optimum region forms a plateau, a linewidth value of 2.5 Hz can safely be chosen as the optimum. The optimum value will depend strongly on the data: where peaks in the raw data are well-resolved, an increase in signal-to-noise ratio is beneficial, while for crowded spectra resolution enhancement may be the better option.

In order to investigate further the spectral variance in the ternary design, and to demonstrate how reference deconvolution can improve component modeling of the data, a PCA model was calculated (16,17). Figure 5 shows the PCA scores plot of the normal spectra and that of the reference deconvoluted spectra. In this case, the reference deconvoluted spectra were calculated using the optimal 2.5 Hz Gaussian lineshape and the normal spectra were weighted with a -1.5 Hz Lorentzian and a +2.5 Hz Gaussian apodization function in order to achieve similar lineshapes and facilitate comparison. From Figure 5, it is clear that the triangular design is much better recovered in the scores plot from reference deconvoluted data. Moreover, the percentage of the explained variance for the first two principal components is higher for the reference deconvoluted data. These are both strong and credible indicators that systematic irregularities have been removed from the data by reference deconvolution, and that as a result, simpler PCA models are required to explain the data.

In order to obtain a quantitative measure of the regularity of the PCA scores plots shown in Figure 5, the distances between the scores in normalized scores plots were calculated. This allows numerical confirmation of the higher regularity observed for reference deconvoluted data. The density plot of the resulted distance distributions (in PC1 and PC2 scores) is shown in Figure 6. The average distance between the sample scores in a normalized plot, considering the span of normalized plots and the 16 increments in the ternary design, should be approximately 0.13 (dividing 2 by the 15 gaps between the scores in the base of the triangle). As evidenced by the plot, for the reference deconvoluted data, a clear and well-defined peak is observed around 0.13, as compared to the uncorrected data which only shows a broad shoulder. This implies that in the scores space, the samples appear closer to the correct positions expected for the ternary design. In addition, the distribution is more regular for reference deconvoluted data, and the density of distances below 0.08 is zero.

Subsequently, PLS models were calculated between the spectral data and the lactic acid concentration of the samples. Models were calculated for a number of different sets, including uncorrected data, reference deconvoluted data with 1.5 Hz Lorentzian target linewidth, uncorrected data with 1 Hz Lorentzian apodization, reference deconvoluted data with 2.5 Hz Lorentzian linewidth, uncorrected data with -1.5 Hz Lorentzian and 2.5 Hz Gaussian apodization, and reference deconvoluted data with 2.5 Hz Gaussian target linewidth. To test the predictive ability of the PLS models, the central part of the triangular design − shown with dashed lines in Figure 2 − was used as the calibration set (28 samples) and all the other samples in the design as the test set (106 samples). The statistics for all the PLS models are summarized in Table I. The most noticeable result is that for the PLS models built on the reference deconvoluted data, each latent variable explains more variance compared to the uncorrected data, and fewer latent variables are needed to describe the data adequately. For the uncorrected data, both with and without apodization, PLS models comprised of 4 latent variables are appropriate, whereas for reference deconvoluted data, only 3 latent variables are needed. This is mainly because in reference deconvoluted data, variations in peaks shape and amplitude − as were observed for lactic acid doublet in Figure 3 - due to instrumental inconsistencies have been corrected. As a result, the data become more bilinear, and simpler multivariate models can be constructed to explain the data and focus on the interesting variance. The Root Mean Square Error of Calibration (RMSEC) and Root Mean Square Error of Prediction (RMSEP) values decrease when window functions are applied; this is attributable to the smoothing effect of apodization and broadening of the lines. However, both RMSEC and RMSEP values are further improved in the reference deconvoluted data when compared to uncorrected data with corresponding apodization (linewidth); this improvement is not attributable to the smoothing effect. Consistent with the prediction errors, the squared Pearson correlation coefficients of the calibration (R^2_{cal}) and the prediction (R^2_{pred}) are higher in the PLS models of the reference deconvoluted data.

The percentages of the cumulative variances captured for the **X** and **y** blocks are given in Table I. The cumulative variance captured for the **X** block shows an increase with apodization of the raw data, and increases by approximately 10% when reference deconvolution is used. Plots of the variance captured for the **X** and **y** blocks versus number of latent variables are shown in Figure 7; uncorrected data, uncorrected data with 2.5 Hz Gaussian apodization, and reference deconvoluted data with 2.5 Hz Gaussian target lineshape are included. Inspection of the **X** variance captured for each latent variable in Figure 7a shows that for the models built on the reference deconvoluted data, the latent variables explain more of the variance in the **X** block, and that with only 3 latent variables almost all the **X** variance is explained. In contrast, for the uncorrected datasets, lower **X** variance is explained for each latent variable. The cumulative variance captured for the response variable (**y** block in Table I) is also higher in the reference deconvoluted data than for uncorrected data. Figure 7b shows the **y** variance captured for each latent variable; for the reference deconvoluted data only 2 latent variables explain almost 100% of the variance, whereas for the uncorrected spectra, at least 4 latent variables are required to explain a comparable amount of variance in the **y** block.

5. Conclusions

For a designed set of 136 samples, ^{1}H NMR spectra were recorded and processed with and without reference deconvolution. Then, PCA and PLS models were calculated and a comparison was made between the models of the data with and without reference deconvolution. The results clearly demonstrate that reference deconvolution substantially improves PCA and PLS models of the NMR data. This is mainly because reference deconvolution corrects systematic artifacts such as lineshape errors, and as a result, data become more bilinear. The resultant multivariate models become simpler, as they can capture more of the relevant variance, and fewer latent variables are needed to explain the data. Reference deconvolution can be particularly helpful in quantitative NMR spectroscopy, and where quantitative pattern recognition of NMR data is of interest, e.g. in NMR-based metabolomics. Investigations are in progress to study the extent to which multivariate analysis of data from real NMR metabolomics studies can benefit from reference deconvolution.

6. Acknowledgements

The strategic research council is acknowledged for generous support to the MicroPAT project under the inSPIRe (Danish Industry-Science Partnership for Innovation and Research in Food Science) consortium. The Faculty of Science and The Ministry of Science and Technology are acknowledged for a grant to the NMR Metabolomics infrastructure.

References

1. Morris GA. Compensation of instrumental imperfections by deconvolution using an internal reference signal. *J Magn Reson* 1988; **80**: 547-552.
2. De Graaf AA, Van Dijk JE, Bovée WMMJ. QUALITY: Quantification improvement by converting lineshapes to the lorentzian type. *Magn Reson Med* 1990; **13**: 343-357.
3. Wouters JM, Petersson GA. Reference lineshape adjusted difference NMR spectroscopy. I. Theory. *J Magn Reson* 1977; **28**: 81-91.
4. Wouters JM, Petersson GA, Agosta WC, Field FH, Gibbons WA, Wyssbrod H, Cowburn D. Reference lineshape adjusted difference NMR spectroscopy. II. experimental verification. *J Magn Reson* 1977; **28**: 93-104.
5. Morris GA. *eMagRes,* 2002; 9: 125-131; DOI: 10.1002/9780470034590.emrstm0449.
6. Bartholdi E, Ernst RR. Fourier spectroscopy and the causality principle. *J Magn Reson* 1973; **11**: 9-19.
7. Metz KR, Lam MM, Webb AG. Reference deconvolution: A simple and effective method for resolution enhancement in nuclear magnetic resonance spectroscopy. *Concepts Magn Reson* 2000; **12**: 21-42.
8. Morris GA, Barjat H, Horne TJ. Reference deconvolution methods. *Prog Nucl Magn Reson Spectrosc* 1997; **31**: 197-257.
9. Morris GA. *Data Handling in Science and Technology* (Chapter 16: Reference deconvolution in NMR)*,* vol 18 Elsevier, 1996, 346-361.
10. Beckonert O, Keun HC, Ebbels TMD, Bundy J, Holmes E, Lindon JC, Nicholson JK. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protocols* 2007; **2**: 2692-2703.
11. Morris GA. *eMagRes*, 2007; DOI: 101002/9780470034590emrstm0449.
12. Barjat H, Morris GA, Swanson AG, Smart S, Williams SC. Reference deconvolution using multiplet reference signals. *J Magn Reson, Ser A* 1995; **116**: 206-214.
13. Nilsson M. The DOSY Toolbox: A new tool for processing PFG NMR diffusion data. *J Magn Reson* 2009; **200**: 296-302.
14. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab* 1987; **2**: 37-52.
15. Wold S, Martens H, Wold H. Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad 1983; 286-293.
16. Winning H, Larsen FH, Bro R, Engelsen SB. Quantitative analysis of NMR spectra with chemometrics. *J Magn Reson* 2008; **190**: 26-32.
17. Engelsen SB, Savorani F, Rasmussen MA. *eMagRes*, 2013; 2**:** 267–278 DOI: 101002/9780470034590emrstm1304.

Table I. Statistics of the PLS models between the spectra and lactic acid concentration as the response variable. Samples from the central part of the triangular design were used as the calibration set and all the other samples as the test set (the two outliers were removed-see section 3.3). ‘FT’ denotes uncorrected spectral data and ‘RD’ reference deconvoluted spectral data.

Datasets	**Number of LVs**	**RMSEC**	**$R^2_{cal.}$**	**RMSEP**	**$R^2_{pred.}$**	**X Cum. Var. (%)**	**y Cum. Var. (%)**
FT	4	0.0080	0.995	0.0213	0.994	83.06	99.52
RD-1.5 Hz Lorentzian	3	0.0022	0.999	0.0096	0.999	97.72	99.96
FT-1.5 Hz Lorentzian Apodization	4	0.0055	0.998	0.0132	0.997	83.77	99.78
RD-2.5 Hz Lorentzian	3	0.0034	0.999	0.0055	0.999	95.79	99.91
FT-2.5 Hz Gaussian Apodization*	4	0.0073	0.996	0.0220	0.994	86.16	99.60
RD-2.5 Hz Gaussian	3	0.0022	0.999	0.0058	0.999	98.40	99.96

***** Besides +2.5 Hz Gaussian apodization, -1.5 Hz Lorentzian apodization was also used to eliminate the natural linewidth.

Figures Captions:

Figure 1. Schematic illustration of the FIDDLE algorithm for reference deconvolution. The reference peak is extracted from the experimental spectrum (b) and its inverse Fourier transform (c) is compared to that of 'perfect' FID (e) to yield a correction function (e/c). The correction is then applied in the time domain to the entire experimental FID (a) to produce the corrected FID (f).

Figure 2. A schematic illustration of the ternary experimental design. A total of 136 mixture samples of lactic acid, propionic acid, and lactose were designed by JMP software. To validate the PLS models (Section 4), mixtures in the center of the design (shown by the dashed triangle) were used as the calibration set, and the remainder of the samples as the test set.

Figure 3. NMR spectra with and without reference deconvolution with a 1.5 Hz Lorentzian target lineshape: a) signals from the constant 'metabolic' background without reference deconvolution; b) signals from the constant background with reference deconvolution; c) the doublet originating from lactic acid without reference deconvolution; and, d) the doublet originating from lactic acid with reference deconvolution.

Figure 4. Root Mean Square Error of Cross Validation (RMSECV)/R^2 and cross validation (R^2_{cv}) of the PLS models calculated for the experimental NMR data using reference deconvolution with different Gaussian linewidths. Lactic acid concentration was used as the response variable. All the samples were included in the models with the exception of the two outliers.

Figure 5. PCA scores plots of: a) raw data weighted with -1.5 Hz Lorentzian and +2.5 Hz Gaussian apodization functions, and b) data reference deconvoluted using a 2.5 Hz Gaussian target lineshape.

Figure 6. Score distance density plot showing the regularity of the PCA scores. Plots show the density of the distances between the scores for the uncorrected spectral data, (**red** line), and reference deconvoluted spectral data generated using an optimal 2.5 Hz Gaussian linewidth, (**blue** line).

Figure 7. Captured variance in the **X** and **y** blocks versus number of latent variables. a) Variance in the **X** block, and b) Variance in the **y** block; Plots from normal Fourier transformed data, Fourier transformed data with 2.5 Hz Gaussian apodization, and reference deconvoluted data with a 2.5 Hz Gaussian target lineshape are shown in blue, green and red, accordingly.