2014

Simultaneous Phase and Scatter Correction for NMR Datasets

Bradley Worley
University of Nebraska-Lincoln, bradley.worley@huskers.unl.edu

Robert Powers
University of Nebraska-Lincoln, rpowers3@unl.edu

Published in final edited form as:
Chemometr Intell Lab Syst. 2014 February 15; 131: 1–6. doi:10.1016/j.chemolab.2013.11.005.

Simultaneous Phase and Scatter Correction for NMR Datasets

Bradley Worley and **Robert Powers***
Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304

Abstract

Nuclear magnetic resonance (NMR) spectroscopy has proven invaluable in the diverse field of chemometrics due to its ability to deliver information-rich spectral datasets of complex mixtures for analysis by techniques such as principal component analysis (PCA). However, NMR datasets present a unique challenge during preprocessing due to differences in phase offsets between individual spectra, thus complicating the correction of random dilution factors that may also occur. We show that simultaneously correcting phase and dilution errors in NMR datasets representative of metabolomics data yields improved cluster quality in PCA scores space, even with significant initial phase errors in the data.

Keywords

Phase-scatter correction; MSC; SNV; NMR; PCA; PLS

1. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a ubiquitous instrumental method in chemistry, owing to its ability to reveal information on the structure, dynamics and environment of molecules or mixtures of molecules containing NMR-active nuclei. This capability and the near universality of NMR-active half-integer spin protons in organic and bio-organic molecules make NMR an ideal platform for chemometric analyses of chemical and biological systems [1–4]. More often than not, the chemometric analysis of NMR spectra involve dimensionality reduction procedures such as principal component analysis (PCA) [5] and partial least squares projections to latent structures (PLS) [6] to reveal differences between spectra in a dataset. Whereas the unsupervised PCA algorithm will reveal differences between experimental groups only when those differences account for the majority of the gross data variation, PLS is capable of forcing separation between statistically indistinguishable groups. Irrespective of the multivariate classification method used, greater statistical significance and increased biological relevance may be attributed to separations between experimental groups having greater variation between groups than within them [7].

*To whom correspondence should be addressed: Robert Powers, University of Nebraska-Lincoln, Department of Chemistry, 722 Hamilton Hall, Lincoln, NE 68588-0304, rpowers3@unl.edu, Phone: (402) 472-3039, Fax: (402) 472-9402.

Notes
The authors declare no competing financial interest.

As a consequence of the need to reduce within-group variation far below between-group variation, chemometric NMR spectral data is characteristically preprocessed to correct for errors in phase and baseline [8]. Subsequent pretreatment through variable scaling and spectral normalization prior to multivariate analysis is then performed to correct for large disparities in signal intensities and random 'dilution' errors, respectively [8, 9].

1.1 Phase correction

Modern FT-NMR spectrometers effectively acquire a rotating-frame free induction decay (FID) signal through the use of quadrature phase detection of the incoming signal [10]. This detection method imparts phase information to the FID by the creation of both an in-phase component $i(t)$ and a quadrature component $q(t)$, phased 90 degrees from $i(t)$. Ideally, the detected FID would arrive in-phase with respect to the receiver, and fine tuning of acquisition parameters can often accomplish this [11]. However, variations in receiver phase, dead time between the transmit and receive gating circuits, and delays arising from analog and digital filtering can all produce phase errors. After Fourier transformation, these phase errors result in a mixture of desirable absorptive spectral lines and broad dispersive lines between the $I(\omega)$ and $Q(\omega)$ frequency-domain signals, which are reversed through a process of phase correction as follows:

$$A(\omega) = I(\omega)\cos(\phi) - Q(\omega)\sin(\phi) \quad (1)$$

$$D(\omega) = I(\omega)\sin(\phi) + Q(\omega)\cos(\phi) \quad (2)$$

Phase correction ideally results in a purely absorptive spectrum in $A(\omega)$ and a purely dispersive spectrum in $D(\omega)$, and relies on the accurate determination of the phase error $\varphi(\omega)$, an expansion of phase error terms as powers of ω:

$$\phi(\omega) = \phi_0 + \phi_1\omega + \phi_2\omega^2 + \ldots \quad (3)$$

Realistically, phase errors higher than first-order are not observed, and phase correction rests on the determination of a zero-order phase error φ_0 and a first-order phase error φ_1. This determination may be performed manually, through software-interactive adjustment of zero- and first-order corrections by the spectroscopist. However, manual phase correction is generally too time-consuming in the case of chemometric datasets, in which there are tens or hundreds of spectra to correct. In that case, the task of phase correction is handed to any number of automated routines that correct each spectrum individually. Spectra may be automatically phase-corrected by maximization of the most negative absorptive data point [12], analysis of the absorption-versus-dispersion [13] or symmetry [14] characteristics of spectral lines, baseline optimization [15] or entropy minimization [16], to name a few. It is worthy of mention that, when the ultimate fate of the spectra is multivariate analysis, the optimization of each spectrum in isolation is wasteful of information that is available from treating the dataset as an ensemble. In fact, phase differences between spectra non-linearly affect both line shapes and baseline, possibly emphasizing spectral details that imply no experimentally relevant conclusions.

1.2 Normalization

Despite the quantitative nature of ^{1}H NMR experiments, chemometric samples exhibit variable total analyte concentrations due to variations in sample preparation, instrument stability, or even the samples themselves. These dilution errors are especially common in metabolomic analyses of biofluids such as urine, where total concentrations may vary

several orders of magnitude. To ensure spectral intensities in a dataset are directly comparable and related to concentrations, normalization is applied to the spectra. The most common normalization method used in chemometrics is unit-integral or constant-sum (CS) normalization, where each spectrum is scaled such that its total integral is unity [9]. CS normalization does more harm than good, however, as it introduces false correlations between spectral peaks and poorly tolerates large disparities in peak intensities.

In an attempt to overcome the drawbacks of CS normalization, Dieterle *et al.* introduced probabilistic quotient (PQ) normalization, in which the median quotient between all corresponding spectral data points is used as an estimate of the true dilution factor [17]. Shortly after, a method of normalization based on histogram matching (HM) was proposed as an alternative to PQ normalization, taking cues from image processing algorithms [18]. Based on their ability to more accurately recover true dilution factors, both PQ and HM normalization were reported to outperform CS normalization on real and simulated ^{1}H NMR metabolomics datasets. Quantitative evidence of improved PCA or PLS cluster quality was not provided using these new normalization methods. Finally, while more commonly applied to infrared spectroscopic data, standard normal variate (SNV) normalization and its mathematical cousin, multiplicative scatter correction (MSC), are candidate methods for ^{1}H NMR spectra [19].

Normalization applied directly to NMR data is sub-optimal, as even small phase differences between spectra can frustrate the estimation of dilution factors. Possibly worse, blind normalization of poorly phased spectra can accentuate experimentally irrelevant spectral features during dimensionality reduction, leading to erroneous conclusions. These difficulties motivated our development of phase-scatter correction (PSC) as a means of simultaneously correcting these coupled phase and dilution errors.

2. METHODS

2.1 NMR data processing

Previously collected one-dimensional (1D) ^{1}H NMR spectral data from published work [20] was leveraged as a typical metabolomics dataset for performance analysis of PSC versus other normalization methods. FIDs were extracted from Bruker-format files using the NMRPipe software package [21] and loaded into the GNU Octave environment [22] for processing. Time-domain signals were zero-filled to 32k real points and Fourier transformed, resulting in a complex data matrix of 177 spectra divided amongst 16 classes (N=177, K=32768, M=16). Spectra were both automatically phase corrected by simplex entropy minimization [16] and manually phase corrected by applying a constant phase correction value to all spectra. Both automatically and manually phase corrected datasets were then normalized using the CS, PQ, HM, SNV, MSC and PSC methods. Each normalized dataset was binned using a uniform 0.04 ppm bin width, scaled per-variable to unit variance, and subjected to PCA. The J_2 statistic [23] was calculated for each class to provide a measure of cluster quality for the scores from each normalization method, as follows:

$$J_{2,k}=\frac{|\boldsymbol{C}|}{|\boldsymbol{C_k}|} \quad (4)$$

where $\boldsymbol{C_k}$ is the covariance matrix of the scores in class k, $\boldsymbol{C}$ is the covariance matrix of all scores, and the vertical bars represent the determinant. Thus, as a cluster shrinks relative to the entirety of the scores-space data, its J_2 statistic will increase. While J_2 provides a measure of individual cluster tightness, it does not capture the degree of cluster overlap

within a dataset. Figure 1 shows the results of the J_2 calculation for normalization methods applied to real ^{1}H NMR metabolomics data.

To quantify differences between extracted PCA models of automatically and manually phase corrected datasets, the angle between the first principal component loadings of each pair of models (θ) was calculated as follows:

$$\theta = \cos^{-1}(\boldsymbol{p}_{auto} \bullet \boldsymbol{p}_{man}) \quad (5)$$

where $\boldsymbol{p}_{auto}$ and $\boldsymbol{p}_{man}$ are the first-component loadings resulting from a given normalization method's data after automatic and manual phase correction, respectively. The loading angle θ for a given normalization method is a reflection on that method's ability to properly normalize data and produce consistent PCA models from two different initial phase error conditions.

2.2 Simulated spectral datasets

The ^{1}H NMR spectra of 100 mM samples of 32 metabolites (Table 1) at pH 7.4 were downloaded from the Biological Magnetic Resonance Bank (BMRB, [24]) and fit to mixtures of complex Lorentzian functions using ACD/1D NMR Processor (Advanced Chemistry Development). Peak amplitudes (A), shifts (ω_0), and widths (λ) were loaded into Octave to generate simulated spectra having 64k real data points and a spectral width of 11 ppm, centered at 4.7 ppm, based on the following model function:

$$s(\omega_j) = \sum_{k=1}^{N} \frac{A_k \lambda_k}{\lambda_k + i(\omega_j - \omega_{0,k})} \quad (6)$$

where $s(\omega_j)$ is the j-th data point of the spectrum, N equals the number of peaks, and i equals the imaginary unit. Spectra were referenced and normalized to the DSS peak, and peaks corresponding to HOD and DSS were subsequently removed, resulting in a basis set of 32 perfectly-phased, noise-free metabolite spectra. Finally, the basis metabolite spectra were stored row-wise in a matrix $\boldsymbol{S}$ for later use in Monte Carlo calculations.

2.3 Monte Carlo experiments

Using the basis metabolite spectra, a dataset of 48 simulated metabolomics spectra ($\boldsymbol{X}$) was generated according to the following equation:

$$\boldsymbol{X} = \boldsymbol{A}(\boldsymbol{C}\boldsymbol{S} + \boldsymbol{R}) + \boldsymbol{E} \quad (7)$$

where $\boldsymbol{A}$ is a diagonal matrix of dilution factors a_i, $\boldsymbol{C}$ is a matrix of metabolite concentrations, $\boldsymbol{S}$ is the previously created metabolite basis set, $\boldsymbol{R}$ is a matrix of identical DSS reference peaks, and $\boldsymbol{E}$ is a matrix of Gaussian white noise. Dilution factors were generated from a log-normal distribution having zero mean and $\sigma = 0.25$. Concentrations in $\boldsymbol{C}$ where generated from normal distributions with parameters chosen to mimic those in Torgrip *et al.* (Table 2) [18]. The resultant data in $\boldsymbol{X}$ is a simulated set of 48 metabolite extracts, spiked with 100 μM DSS, where six distinct classes arise from differences in the concentrations of alanine, asparagine, glutamine, malate, proline, sucrose and valine. All other metabolites were assigned concentrations from a normal distribution having $\mu = 5$ μM and $\sigma = 0.5$ μM.

Monte Carlo simulations were run to assess the performance of all discussed normalization methods over various amounts of phase error added to $\boldsymbol{X}$. Forty-six phase error points were calculated, in which the standard deviation of φ_0 was linearly increased from 0° to 5°. The

standard deviation of φ_1 at each point was equal to one tenth that of φ_0. Both φ_0 and φ_1 were assigned zero mean. For each phase error point, 100 Monte Carlo iterations were performed with different sets of random dilution factors. Spectra in the de-phased $\boldsymbol{X}$ matrix were automatically phase corrected using simplex entropy minimization and normalized each time using CS, PQ, HM, SNV, MSC and PSC methods. Normalization to unit DSS integral was also performed for reference. An identical set of normalization calculations was performed on the unphased data. Estimated dilution factors were compared to the true value to produce a root-mean-square dilution error, RMSE(a), for each method. Figure 2 shows the RMSE(a) result of Monte Carlo simulation at 0.2° phase error. To assess normalization effects on multivariate model quality, spectra from each method were uniformly binned with 0.04 ppm bin widths, each bin scaled to unit variance, and subjected to PCA. Values of J_2 for each of the six classes were then calculated, and the median of the values was reported for each Monte Carlo iteration. The θ values between automatically phased and unphased PCA model loadings were also calculated at each iteration to assess each normalization method's ability to produce consistent models in the presence of phase errors. Figure 3 summarizes the results of Monte Carlo simulation over all phase errors based on RMSE(a), J_2 and θ.

3. CALCULATION

Phase-scatter correction (PSC) is effectively an extension of multiplicative scatter correction (MSC) to handle phase errors during normalization. In MSC, each spectrum is scaled around its mean intensity and shifted to match a reference spectrum, typically the mean of the dataset [19]. Optimal values of scale ($\boldsymbol{b}$) are determined by linearly regressing the mean-centered reference onto the mean-centered data matrix:

$$\left(\boldsymbol{X}-\overline{\boldsymbol{X}}\right)^{\mathrm{T}}\boldsymbol{b}-(\boldsymbol{r}-\overline{r})^{\mathrm{T}} \quad (8)$$

where spectra are arranged as rows in $\boldsymbol{X}$ and $\boldsymbol{r}$. The solution of the above equation for $\boldsymbol{b}$ has a closed-form expression, and thus MSC is rather computationally efficient. PSC additionally corrects zero- and first-order phase errors during normalization, requiring a nonlinear optimization of the form:

$$\left\{\widehat{b},\widehat{\phi}_0,\widehat{\phi}_1\right\}=\underset{b,\phi_0,\phi_1}{\operatorname{argmin}}\sum_{j=1}^{N}\left|b\cdot s\left(\omega_j\right)e^{i\left(\phi_0+\phi_1\omega_j\right)}-r\left(\omega_j\right)\right|^2 \quad (9)$$

where $s(\omega_j)$ is the j-th point of a given mean-centered row in $\boldsymbol{X}$ and $r(\omega_j)$ is the corresponding point in a suitably chosen mean-centered reference spectrum. Minimization is carried out for every spectrum in the dataset using Levenberg-Marquardt nonlinear least squares [25] as implemented by the *leasqr* function in Octave, a function similar to MATLAB's *nlinfit*. The corrected spectrum is then returned from minimization as follows:

$$s^*\left(\omega_j\right)=\widehat{b}\cdot s\left(\omega_j\right)e^{i\left(\widehat{\phi}_0+\widehat{\phi}_1\omega_j\right)}+\overline{r} \quad (10)$$

Phase-scatter correction of 50 spectra having 32k real points each requires approximately 30 seconds (Supplementary Figure S-2) on a single-core 3.2 GHz Intel workstation running GNU Octave 3.6.

4. RESULTS

On the real metabolomics spectra, PSC normalization resulted in the highest quality clusters (Figure 4) according to the lower bound of the J_2 statistic shown in Figure 1. Given the fact that the spectra were each automatically phase corrected before any normalization was applied, this observed increase in J_2 must be due to the correction of subtle phase differences *between* spectra not detectable by correcting each spectrum individually. It is important to note that, while PQ and HM produce higher median J_2 values, this is an artifact of large distortions of their respective PCA loadings, and not always reflective of higher quality clusters (See Supplementary Figures S-1, S-2 and S-3). Because J_2 is a per-cluster statistic, it is only an ideal measure of overall scores-space model quality when all clusters are nearly identically distributed. Models containing highly distorted components may contain several high-quality clusters and a few extremely low-quality clusters, resulting in a high mean or median J_2 value. For that reason, the lower bound of J_2 for each method – effectively the worst cluster quality – was chosen as a better indicator of overall model quality than the median. In fact, PSC produced the most consistent model loadings between automatically and manually phase corrected data, with a θ value of 14.5°. This can be compared to θ values of 89.6° and 20.2° for PQ and HM, respectively.

Moreover, Monte Carlo analyses of PSC versus contemporary normalization methods show that PSC offers a unique advantage during multivariate analysis. Results of Monte Carlo normalization after automatic phase correction are summarized in Figure 3, and scatter plots of recovered dilution factors are shown in Figure 2. While PSC fails to recover true dilution factors as accurately as DSS, CS or HM normalization, it does remain competitive with MSC at all phase errors (Figure 3(a)). PSC normalization yields tighter clusters than all other methods, as is apparent from Figure 3(b) and further supported by Supplementary Figure S-7. Furthermore, PSC results in dramatically lower values of θ than all other methods, indicating that residual phase errors left uncorrected by automatic phase correction are significant enough to distort principal component loadings when normalized by any method other than PSC (Figure 3(c)).

5. DISCUSSION

As evident from visual inspection of both the real metabolomics dataset and the Monte Carlo simulated datasets, correction of minute phase differences between spectra yields a substantial improvement in cluster quality in multivariate analysis. In general, phase differences contribute significantly to spectral lineshape differences in ^{1}H NMR data. This effect is especially pronounced in the case of PSC correction of spectra containing significant and consistent broad background signals, where normalization alone cannot comparably standardize baselines.

One particularly striking result of the Monte Carlo simulations is the difference between automatically phase corrected and unphased dilution factor estimates (Figure 2). In fact, examination of dilution factors estimated by DSS integration clearly shows that automatic phase correction introduces variation into the dataset through minute differences in φ_0 and φ_1 between spectra. This artificial variation is then amplified through normalization, as is especially apparent in the case of PQ normalization.

In their report on HM normalization, Torgrip *et al.* noticed the potential unsuitability of explained sum of squares (R_2) for assessing model quality differences due to normalization methods [18]. As a percentage measure, explained sum of squares is not suitable for comparing the qualities of PCA models, or any preprocessing done prior to building the models [26]. Therefore, the J_2 statistic was chosen as an alternative means of comparing

cluster quality during Monte Carlo simulation. Effectively, J_2 measures the ratio of the area of a cluster in scores space relative to the total scores-space area, regardless of how much variation the model captures. Even still, because J_2 is a per-cluster statistic, it not an ideal measure of overall scores-space model quality, especially for models containing highly distorted components. Mean or median J_2 values of a model may be high in this case, despite the fact that the model scores are useless from the perspective of class discrimination. Thus, the minimum J_2 was chosen as a more effective indicator of overall cluster quality.

It is important to note that phase-scatter correction is generally applicable when the Discrete Fourier Transform (DFT) is used to yield phase-sensitive spectra from time-domain data. While some newer parametric methods of NMR time-frequency transformation render phase correction unnecessary [11, 27], they complicate chemometric analyses in other ways and do not detract from the utility of PSC in DFT-processed datasets. Lastly, uniform binning was utilized during Monte Carlo simulation immediately prior to PCA modeling merely to accelerate the tens of thousands of iterations performed. In fact, binning is by no means a requisite operation of the algorithm and PSC is designed to be applied directly to full-resolution NMR spectra.

Finally, use of PSC requires an initially phased dataset before performing normalization and further analysis. In other words, PSC does not replace general phase correction routines for producing pure-absorptive NMR spectra: it can only correct phase differences *between* spectra. However, the required initial phase correction may be performed by any of the aforementioned automatic phase correction algorithms, making PSC an attractive normalization method when highly automated spectral processing is required.

6. CONCLUSIONS

Phase-scatter correction is a novel algorithm for simultaneously correcting zero- and first-order phase errors and random dilution factors in ^{1}H NMR chemometric data. While PSC only performs comparably to MSC in dilution factor estimation, it more consistently yields high-quality clusters and reliable models than all other methods when given imperfectly phased data. PSC can be fully automated through prior automatic phase correction of the dataset, has no tunable parameters, and makes no assumptions regarding line shape, baseline flatness, or intensity distributions in the data. These qualities lend PSC to use in chemometrics as a new method of normalizing NMR data entering into multivariate analyses such as PCA or PLS. An implementation of the PSC algorithm is available in open-source GNU Octave code as part of a toolbox for processing and analyzing NMR chemometric data, downloadable at http://bionmr.unl.edu/mvapack.php.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge Ralf Torgrip at Stockholm University for his contribution of example source code for the histogram matching normalization method. This project was supported by National Institutes of Health Grants P20 RR-17675, P30 GM103335, R01 CA163649-01A1, and R01 AI087668-01A1.

References

1. Worley B, Powers R. Multivariate Analysis in Metabolomics. Current Metabolomics. 2013; 1:92–107.

2. Zhang B, Powers R. Using NMR-based metabolomics to study the regulation of biofilm formation. Future Med Chem. 2012; 4:1273–1306. [PubMed: 22800371]
3. Gebregiworgis T, Powers R. Application of NMR Metabolomics to Search for Human Disease Biomarkers. Comb Chem High Throughput Screening. 2012; 15:595–610.
4. Powers R. NMR metabolomics and drug discovery. Magnetic Resonance in Chemistry. 2009; 47:S2–S11. [PubMed: 19504464]
5. Jolliffe, IT. Principal Component Analysis. 2. Springer; New York: 2002.
6. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab. 2001; 58:109–130.
7. Worley B, Halouska S, Powers R. Utilities for Quantifying Separation in PCA/PLSA-DA Scores. Anal Biochem. 2013; 433:102–104. [PubMed: 23079505]
8. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Calmani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjostrom M, Trygg J, Wulfert F. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics. 2007; 3:231–241.
9. Craig A, Cloareo O, Holmes E, Nicholson JK, Lindon JC. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal Chem. 2006; 78:2262–2267. [PubMed: 16579606]
10. Levitt, MH. Spin dynamics : basics of nuclear magnetic resonance. 2. John Wiley & Sons; Chichester, England; Hoboken, NJ: 2008.
11. Chylla RA, Volkman BF, Markley JL. Practical model fitting approaches to the direct extraction of NMR parameters simultaneously from all dimensions of multidimensional NMR spectra. J Biomol Nmr. 1998; 12:277–297. [PubMed: 9751999]
12. Siegel MM. The Use of the Modified Simplex-Method for Automatic Phase Correction in Fourier-Transform Nuclear Magnetic-Resonance Spectroscopy. Anal Chim Acta-Comp. 1981; 5:103–108.
13. Craig EC, Marshall AG. Automated Phase Correction of Ft Nmr-Spectra by Means of Phase Measurement Based on Dispersion Versus Absorption Relation (Dispa). J Magn Reson. 1988; 76:458–475.
14. Heuer A. A New Algorithm for Automatic Phase Correction by Symmetrizing Lines. J Magn Reson. 1991; 91:241–253.
15. Brown DE, Campbell TW, Moore RN. Automated Phase Correction of Ft Nmr-Spectra by Baseline Optimization. J Magn Reson. 1989; 85:15–23.
16. Chen L, Weng ZQ, Goh LY, Garland M. An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. J Magn Reson. 2002; 158:164–168.
17. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Anal Chem. 2006; 78:4281–4290. [PubMed: 16808434]
18. Torgrip RJO, Aberg KM, Alm E, Schuppe-Koistinen I, Lindberg J. A note on normalization of biofluid 1D H-1-NMR data. Metabolomics. 2008; 4:114–121.
19. Fearn T, Riccioli C, Garrido-Varo A, Guerrero-Ginel JE. On the geometry of SNV and MSC. Chemometr Intell Lab. 2009; 96:22–26.
20. Halouska S, Fenton RJ, Barletta RG, Powers R. Predicting the in Vivo Mechanism of Action for Drug Leads Using NMR Metabolomics. Acs Chem Biol. 2012; 7:166–171. [PubMed: 22007661]
21. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. J Biomol Nmr. 1995; 6:277–293. [PubMed: 8520220]
22. Eaton, JW.; Bateman, D.; Hauberg, S. GNU Octave Manual Version 3. Network Theory Limited; 2008.
23. Koutroumbas, K.; Theodoridis, S. Pattern Recognition. 3. Elsevier/Academic Press; Amsterdam, Boston: 2006.
24. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL. BioMagResBank. Nucleic Acids Res. 2008; 36:D402–D408. [PubMed: 17984079]

25. Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. J Soc Ind Appl Math. 1963; 11:431–441.

26. Kjeldahl K, Bro R. Some common misunderstandings in chemometrics. J Chemometr. 2010; 24:558–564.

27. Hu HT, Van QN, Mandelshtam VA, Shaka AJ. Reference deconvolution, phase correction, and line listing of NMR spectra by the 1D filter diagonalization method. J Magn Reson. 1998; 134:76–87. [PubMed: 9740734]

Highlights

- Protocol simultaneously corrects coupled phase and dilution errors in NMR metabolomics datasets
- Improves cluster quality in PCA scores space and can be fully automated
- Outperforms other common normalization techniques when normal phase errors exist

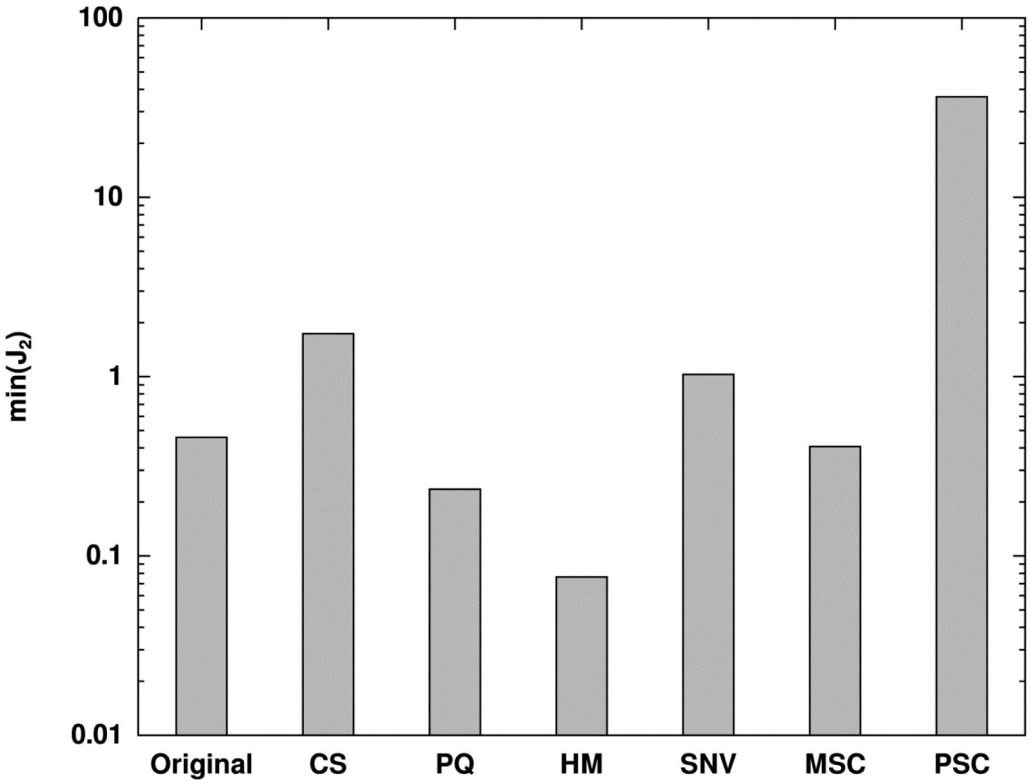

Figure 1.
Comparison of PCA cluster quality for ^{1}H NMR metabolomics data normalized using different algorithms. The minimum J_2 value (worst cluster quality) for each model is reported here, as it is a more effective indicator of overall model and cluster quality than the mean or median. See Supplementary Figure S-10 for complete five-number summaries of the J_2 values obtained from normalization of this example dataset.

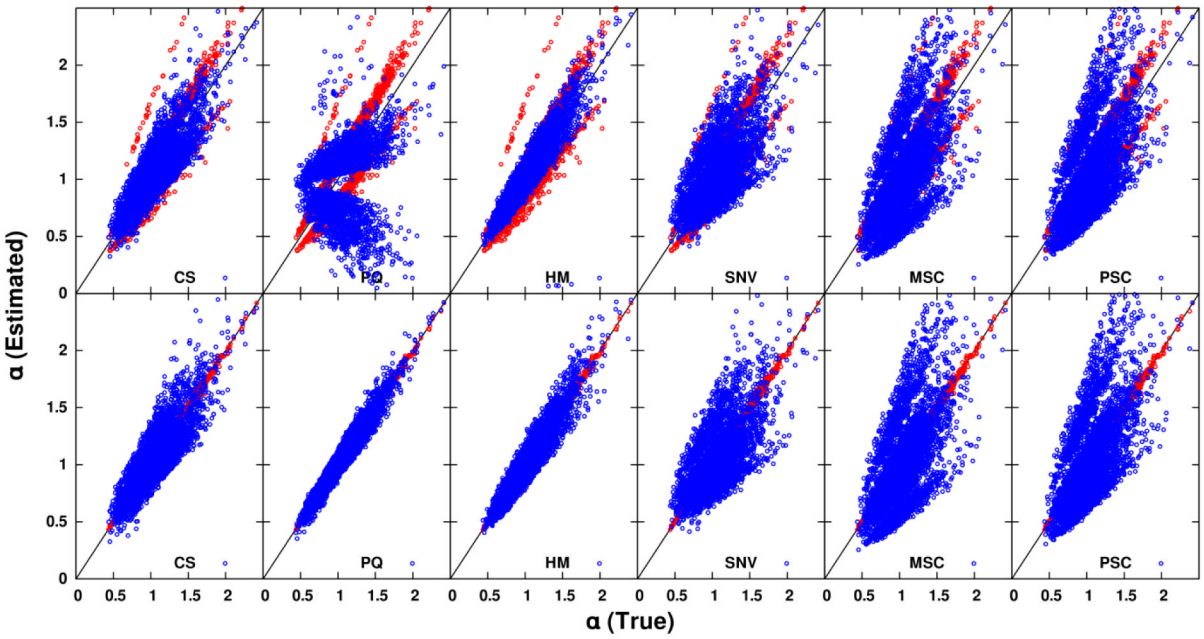

Figure 2.

Results of 100 Monte Carlo iterations at 0.2° zero-order phase error, indicating the ability of all compared normalization methods to recover the true dilution factor of a nearly perfectly phased dataset. Red points reflect the dilution factors calculated by integrating the DSS peak and blue points reflect the dilution factor estimates from normalization. Upper panels show the dilution factors recovered from automatically phased data after normalization, and lower panels show dilution factors recovered from unphased data after normalization.

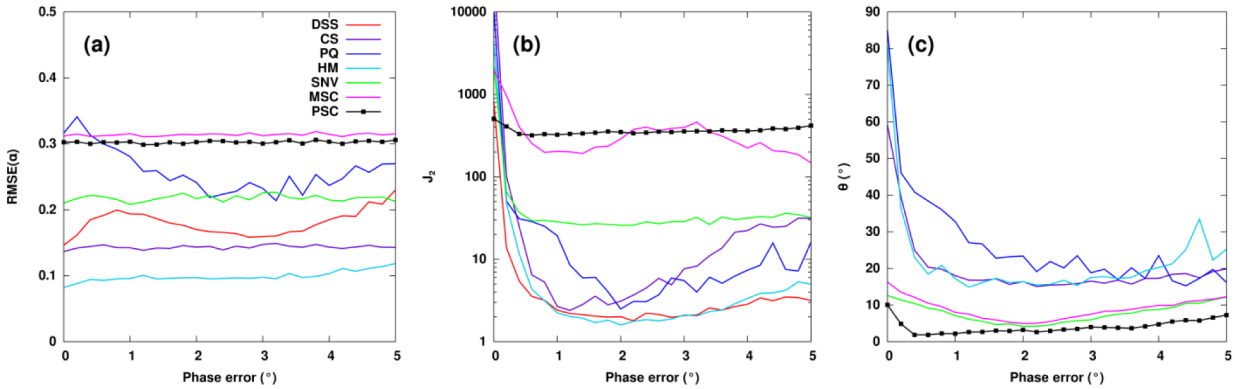

Figure 3.

Results of the Monte Carlo simulation over all phase error points. **(a)** As phase error increases, dilution factor estimates from all methods except remain effectively stable. Estimates from PSC compete with MSC, but suffer in comparison with HM. **(b)** However, J_2 values indicate that PSC outperforms all other normalization methods at producing tight clusters at any realistic phase error. **(c)** Finally, values of θ. calculated from PCA loadings indicate that PSC maintains the highest model consistency in the face of imperfectly phased data. Phase error on the x-axis refers to zero-order error; it should be noted that each point also contains first-order phase error as discussed in Methods. See Supplementary Figures S-4 and S-6 for versions of these figures calculated from normalization of unphased data, and Figures S, S-7 and S-9 for versions with confidence regions applied.

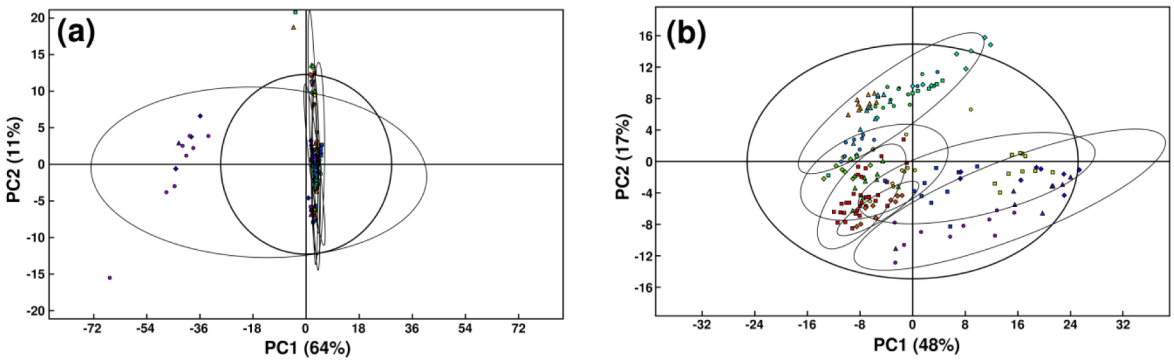

Figure 4.

PCA scores plots of a typical metabolomics dataset after automatic phasing following with either PQ or PSC normalization. In both plots, ellipses denote different classes of antibiotic treatment of *Mycobacterium smegmatis* and differing symbols within each ellipse represent differing antibiotic subclasses. **(a)** PQ normalization amplifies residual phase differences left behind after automatic phasing. **(b)** PSC normalization produces a more valid PCA model by correcting residual phase differences.

Table 1

Metabolite spectra used in Monte Carlo simulations.

Aminobutyrate	Adenosine	Alanine	Arginine
Asparagine	Aspartate	Choline	Citrulline
Ethanolamine	Fructose	Galactose	Glucose
Glutamate	Glutamine	Glycine	Histidine
Isoleucine	Lactate	Leucine	Lysine
Malate	Maltose	Myoinositol	Ornithine
Phenylalanine	Proline	Putrescine	Serine
Succinate	Sucrose	Threonine	Valine

Table 2

Metabolite concentrations used in Monte Carlo simulations.

Metabolite	C_A (μM)	C_B (μM)	C_C (μM)	C_D (μM)	C_E (μM)	C_F (μM)
Alanine	9.2 ± 1.4	19.6 ± 1.6	16.9 ± 1.2	6.5 ± 0.66	26.2 ± 3.6	13.5 ± 1.1
Asparagine	6.8 ± 0.86	11.7 ± 1.8	19.0 ± 1.9	14.7 ± 1.2	24.8 ± 2.6	17.4 ± 1.0
Glutamate	13.3 ± 1.7	9.2 ± 1.5	18.8 ± 1.9	16.9 ± 2.1	25.0 ± 3.5	6.9 ± 1.0
Malate	14.2 ± 1.2	11.9 ± 1.4	22.0 ± 5.1	6.7 ± 0.68	9.4 ± 0.72	18.0 ± 2.4
Proline	11.4 ± 1.5	18.4 ± 3.1	14.7 ± 2.4	6.9 ± 0.62	9.8 ± 1.5	23.7 ± 2.9
Sucrose	7.1 ± 0.90	17.2 ± 2.1	19.3 ± 2.0	13.2 ± 1.9	9.3 ± 0.56	23.3 ± 2.7
Valine	9.0 ± 0.85	26.3 ± 2.3	13.4 ± 1.2	20.4 ± 1.7	6.7 ± 0.90	17.0 ± 1.5