Supplementary Information for:

MVAPACK: A Complete Data Handling Package for NMR Metabolomics

Bradley Worley and Robert Powers*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304

TABLE OF CONTENTS

Supplementary Methods

Coffee sample preparation

Four freshly brewed roasts of coffee (Light, Dark, Medium Regular, and Medium Decaffeinated) were purchased from a local coffee shop. From each roast, sixteen 1.2 mL samples were drawn and stored at -80°C for 24 hours. The samples were then lyophilized at -50°C and 0.1 mBar for 24 hours and subsequently redissolved in 1.0 mL of 99.8% D_2O (Isotec, St. Louis, MO; ChEBI:41981) without pH adjustment. Following redissolution, the samples were centrifuged at 12k RPM and 25°C for 5 minutes and 800 µL of the supernatant was collected into NMR tubes. The samples were stored in their NMR tubes at 4°C for 36 hours prior to data collection.

Caffeine extraction

Measurement of the caffeine concentration in each coffee roast was performed based on previously outlined procedures[1]. Triplicate standards of caffeine were made by dissolving 2.9 mg of caffeine (Sigma-Aldrich, St. Louis, MO; ChEBI:27732) into 100 mL of 99.5% CH_2Cl_2 (Sigma-Aldrich, St. Louis, MO; ChEBI:15767) for a final concentration of 149 uM. From each purchased coffee roast, 25 mL of brewed coffee was combined with 25 mL of CH_2Cl_2 in a separatory funnel in a two-step liquid-liquid extraction. Extracted caffeine in CH_2Cl_2 was diluted 20-fold into 1.0 mL and subjected to UV/Vis absorption spectroscopy for caffeine quantitation.

UV/Vis spectroscopy

Absorption spectra of caffeine standards and extracts were collected on a Shimadzu UV-2501PC with a 1.0 nm slit width and 1.0 cm quartz cuvettes. Spectra were collected between the wavelengths of 500 nm and 230 nm.

NMR spectroscopy

All NMR experiments were collected on a Bruker Avance DRX 500.13 MHz spectrometer equipped with a 5 mm inverse triple-resonance (^{1}H, ^{13}C, ^{15}N) cryoprobe with a *z*-axis gradient. A Bruker BACS-120 sample changer and ICON-NMR software was used to automate NMR data collection. A standard 1D ^{1}H NMR spectrum using a SOGGY pulse sequence[2, 3] and a T_2-filtered 1D ^{1}H NMR spectrum *z*-filtered CPMG sequence[4] with an identical SOGGY water suppression element were acquired for each sample. All experiments were performed at 20°C with 128 scans, 32 dummy scans, a carrier frequency offset of 4.7 ppm, a 6009.6 Hz spectral width, and a 1.0 s relaxation delay. For T_2 filtering, 20 repetitions of a CPMG-*z* element having a delay (τ) of 5.0 ms were performed, for a total filter time ($2n\tau$) of 200 ms. Free induction decays were collected with 32k total data points resulting in a total acquisition time of 10 minutes per experiment.

Caffeine quantitation

A reference spectrum of caffeine in CH_2Cl_2 was generated from the three standard UV/Vis absorption spectra by taking the mean of the spectra after multiplicative scatter correction. To quantify caffeine in the extracts, the absorption spectrum of each extract was fit by nonlinear least squares to the sum of a scaled caffeine reference spectrum and no more than two extra

'background' Gaussian bands. The ratio of the fit caffeine reference spectrum in each extract to that of the known samples was used as an estimate of caffeine concentration in the extracts. Concentrations of the medium regular, medium decaffeinated, dark and light roasts were 1.526 mM, 0.217 mM, 1.979 mM and 4.993 mM, respectively.

Multivariate analysis

All NMR spectra were loaded, pre-processed, pre-treated and modeled inside the GNU Octave 3.6 programming environment[5] using functions available in MVAPACK. Spectra were loaded in from Bruker DMX binary format and corrected for group delay errors by a circular shift of their time-domain data points. All spectra were Fourier transformed, automatically phase-corrected and referenced to match the chemical shifts of caffeine with known database values. Spectral regions upfield of 0.44 ppm and downfield of 9.16 ppm were removed from the dataset, as they contained no informative signals. As solvent resonances were adequately suppressed by the excitation sculpting pulse sequence, no spectral regions were removed around the water resonance. Figure 2 shows representative processed spectra of each coffee roast analyzed by NMR.

For Principal Component Analysis (PCA), the dataset was normalized by the Probabilistic Quotient (PQ) method[6] and subjected to optimized bucketing[7]. Low-variation "noise" bins were automatically removed from the dataset[8] resulting in a final dataset having 64 observations and 371 variables. The dataset was scaled to unit variance and PCA modeling produced six

significant components having cumulative R^2 and Q^2 statistics of 0.9643 and 0.9561, respectively.

Linear Discriminant Analysis (LDA) was performed on the resulting six-dimensional PCA scores to yield a three component model that best captured the between-class variation present in the six orthonormal input score vectors. LDA modeling yielded a model having a cumulative R^2_X statistic of 0.9959 and cumulative R^2_Y and Q^2 statistics of 1.0. Figure 3 illustrates the scores-space data resulting from both PCA before and after LDA transformation.

For Orthogonal Projections to Latent Structures Regression (OPLS-R), the dataset was aligned using a per-class application of interval correlation-optimized shifting[9] and PQ normalization, resulting in a final dataset having 64 observations and 11,888 variables. The Pareto-scaled dataset was regressed by OPLS against a response vector containing caffeine concentrations estimated by UV/Vis analysis of the four coffee roasts, yielding a model with one predictive component and one orthogonal component ($R^2_{X(pred)}$ = 0.5270, $R^2_{X(orth)}$ = 0.1306, R^2_Y = 0.9852, Q^2 = 0.9177). CV-ANOVA significance testing returned a p value equal to zero (F = 166.42) to within double-precision floating point error[10], indicating a reliable model. Figures 4 and 5 summarize the results of OPLS regression.

The OPLS-R and LDA models were further validated using response permutation tests having 1000 iterations each. The permutation tests of both models resulted in p values equal to zero to

within double-precision floating point error for both the R^2 and Q^2 statistics, further indicating high model reliability. See the Supporting Information for complete graphical results of internal cross-validation and permutation testing.

Validation against SIMCA-P+

Correctness of the PCA and OPLS-R models generated by MVAPACK was verified by exporting the final processed and treated data matrices from Octave and modeling them in SIMCA-P+ 13.0 (Umetrics AB, Umea, Sweden). The scores extracted from SIMCA and MVAPACK were found to have coefficients of determination (R^2) of 0.999976 and 0.999989 for the PCA and OPLS models, respectively. The 'imperfect' non-unity values of R^2 reflect the fact that SIMCA-P+ 13.0 only permits the export of scores with no more than four decimal places. Graphical comparisons of the model scores from SIMCA and MVAPACK are available in the Supporting Information.

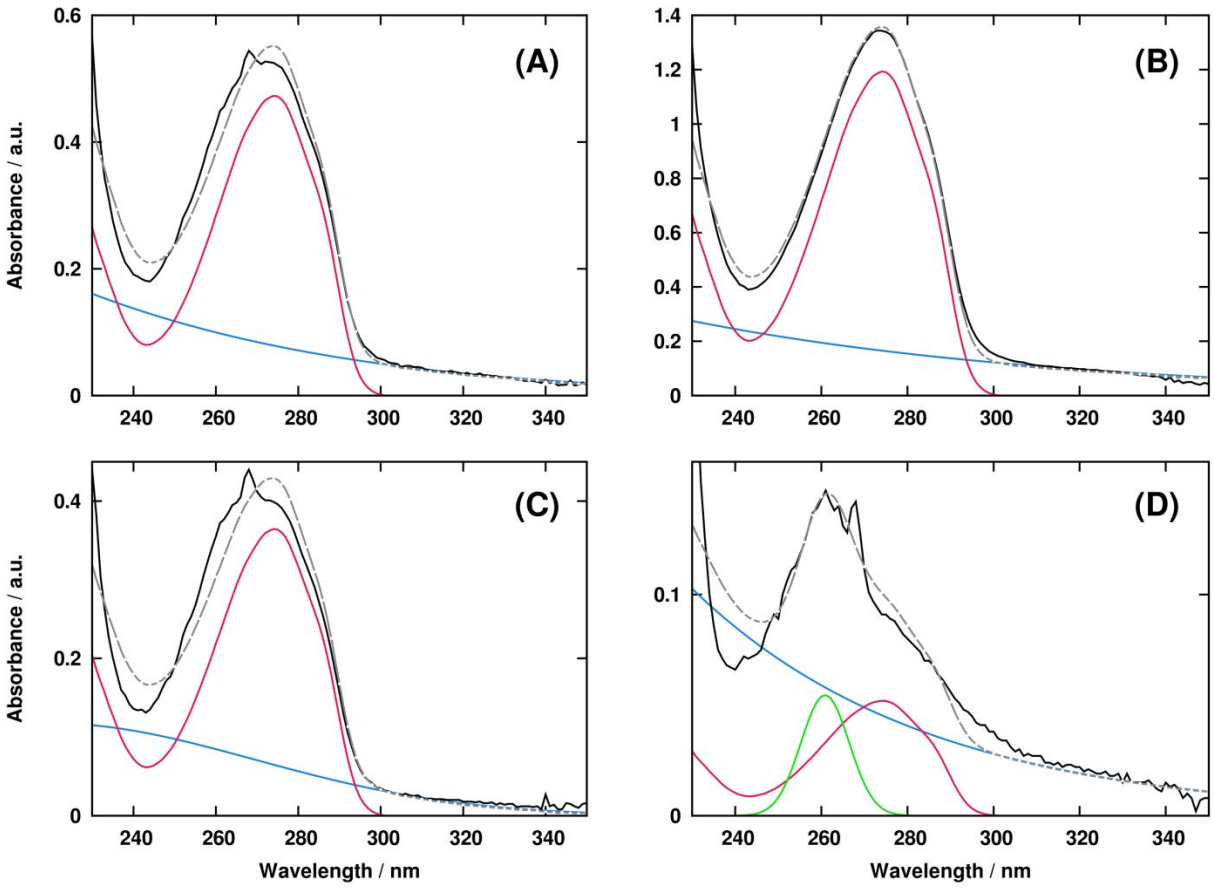

Figure 1S. UV/Vis absorbance band-fitting results for caffeine concentration estimation of dark roast **(A)**, light roast **(B)**, medium regular roast **(C)** and medium decaffeinated roast **(D)**. Black lines represent observed spectra, dashed grey lines represent fitted spectra, red lines represent fitted caffeine, and blue and green lines represent additional Gaussian bands required for fitting.

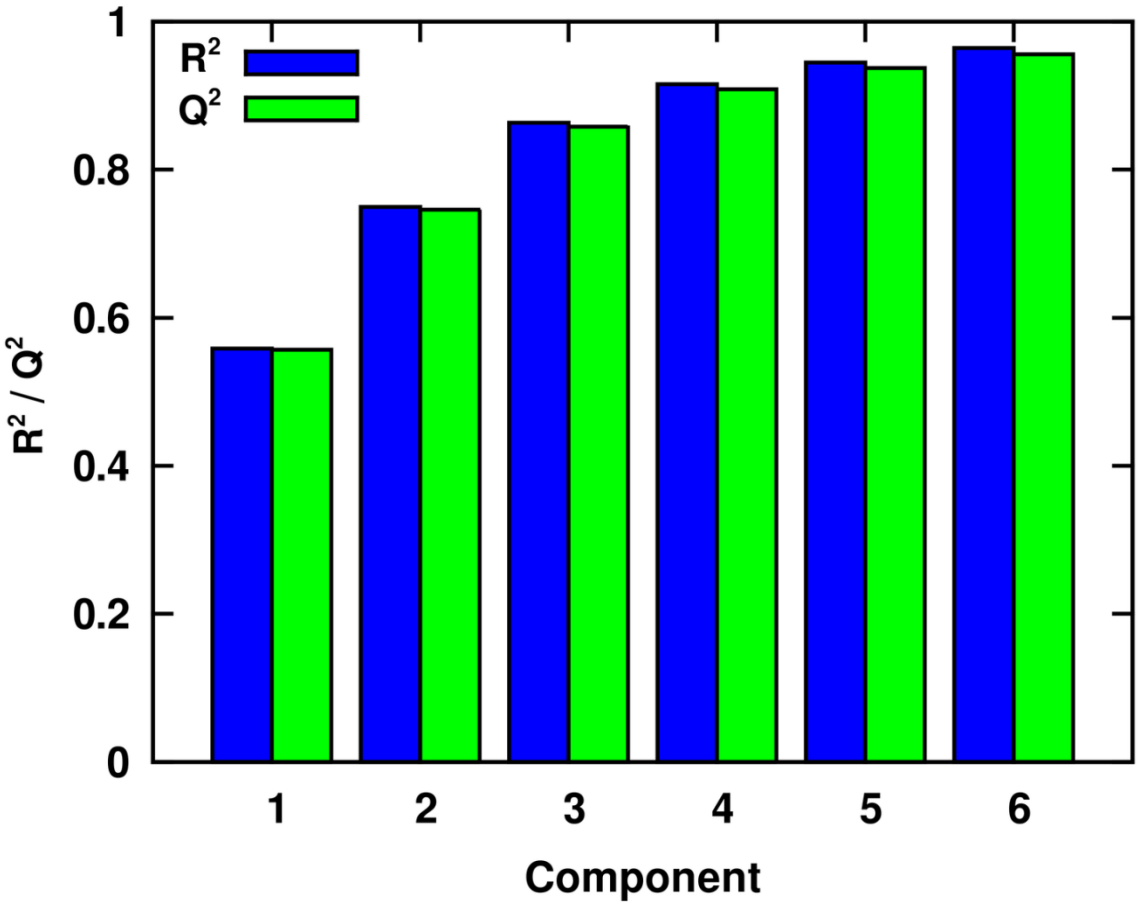

Figure 2S. Internal cross-validation results of PCA-modeled coffee data. R^2 represents the explained variation in X, and Q^2 represents its cross-validated equivalent obtained by recalculation of model scores and loadings after a zeroing of each value in X.[11] In the case of a perfect PCA model, X is completely explained by the set of model components and Q^2 will equal R^2. In all other cases, X is approximated by the model and $Q^2 < R^2$.

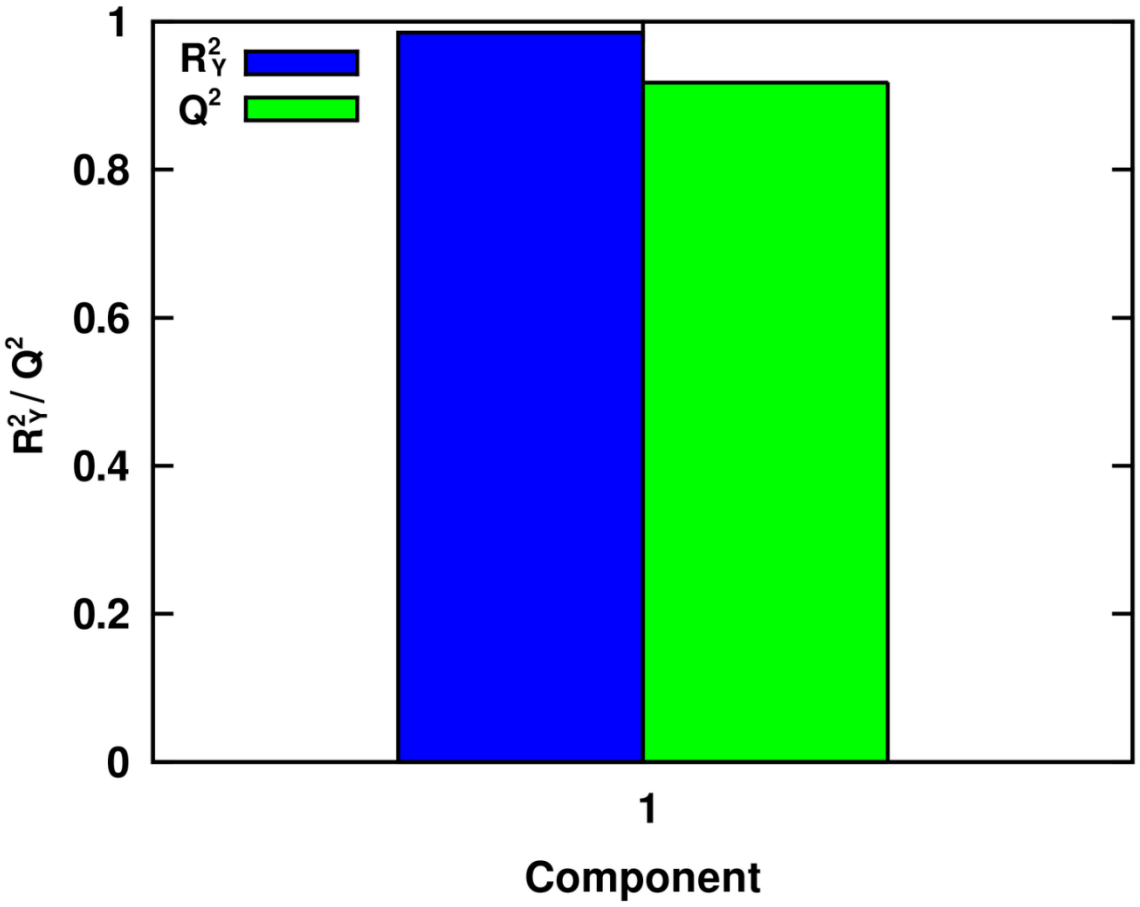

Figure 3S. Internal cross-validation results of OPLS-R modeled coffee data. R^2_Y represents the explained variation in *Y*, and Q^2 represents its cross-validated equivalent obtained by 10 iterations of 7-fold internal cross-validation. Similar to R^2 and Q^2 in the case of PCA, Q^2 will approach R^2_Y as the OPLS model approaches perfect prediction ability for *Y*.

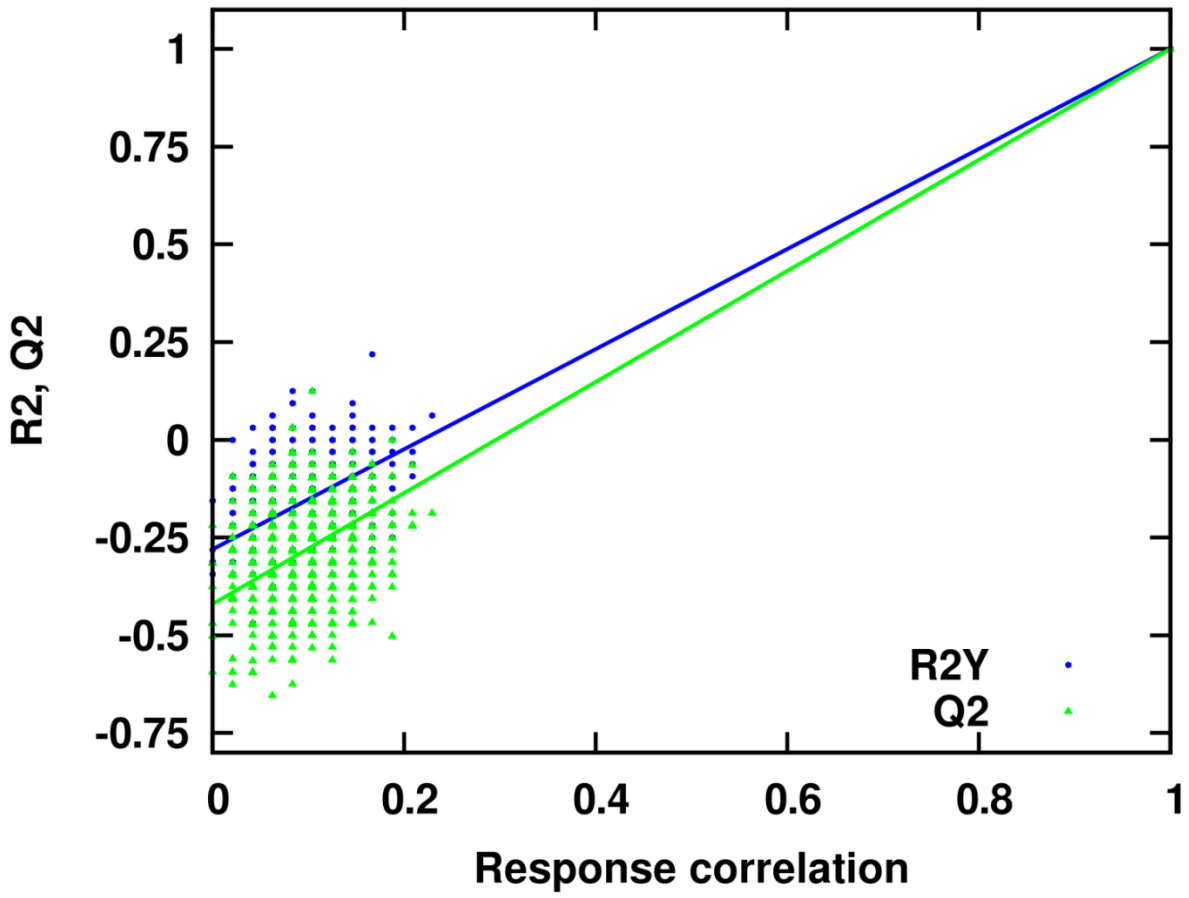

Figure 4S. Response permutation testing results for LDA-modeled PCA scores after 1000 random permutations of *Y*. Model significance is inferred from the degree of vertical separation between the null distribution (leftmost) and the true R^2_Y and Q^2 values (rightmost). The apparent discretization along the correlation axis is a result of using binary class labels in *Y*.

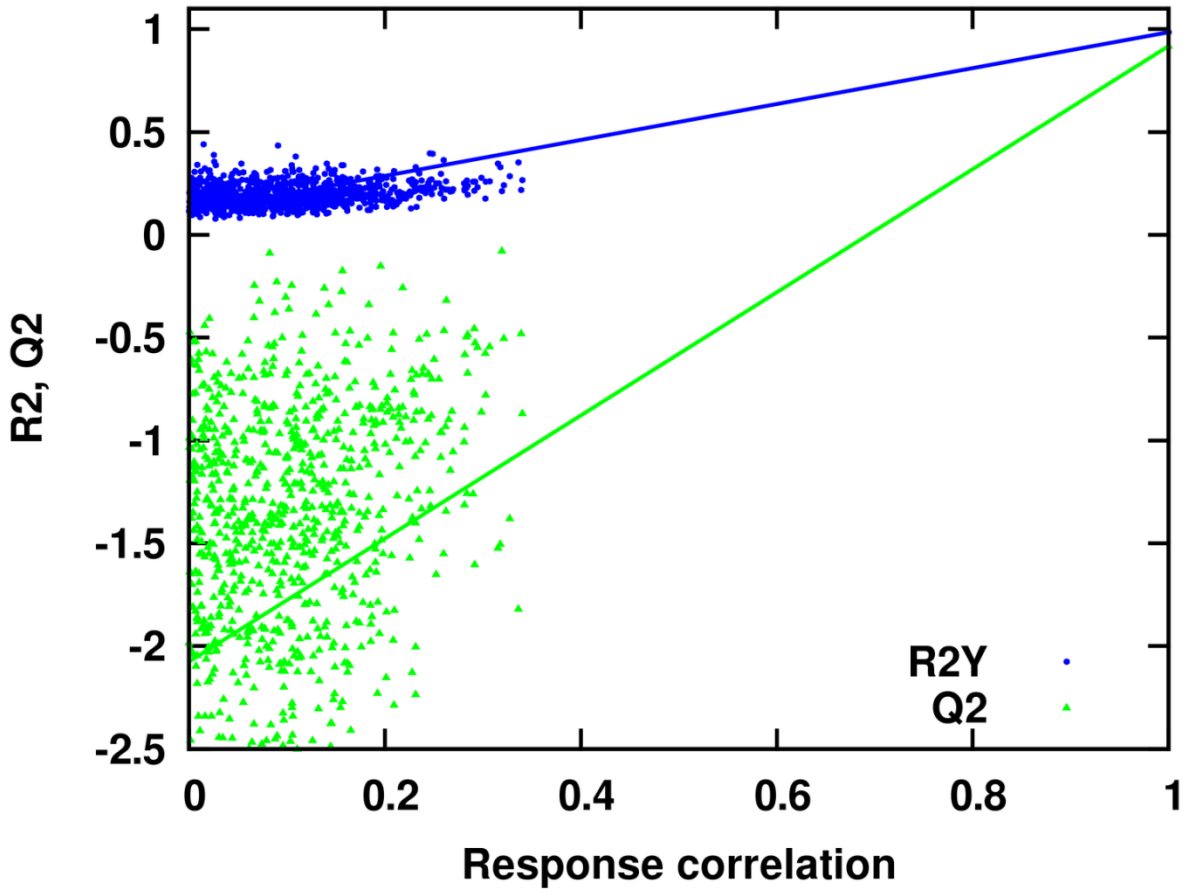

Figure 5S. Response permutation testing results for the OPLS-R model after 1000 random permutations of *Y*. Again, from the rules described in Figure 4S, the model is deemed significant (*p* values equal to zero to within 64-bit floating point precision).

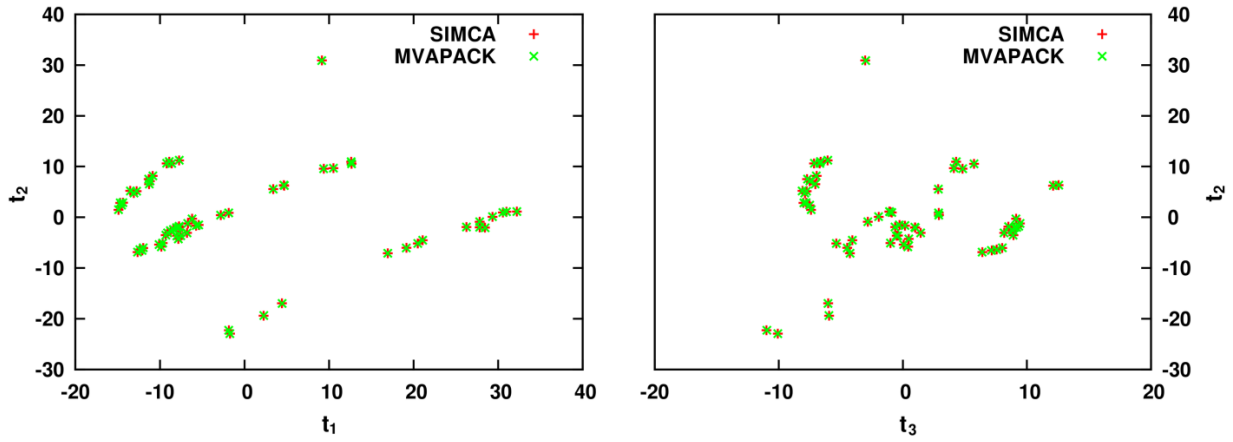

Figure 6S. Comparison of calculated PCA scores from SIMCA-P+ and MVAPACK along the first three principal components. The results are visually identical and exact to within the tolerance of the NIPALS algorithm (relative error < 10^{-6}).

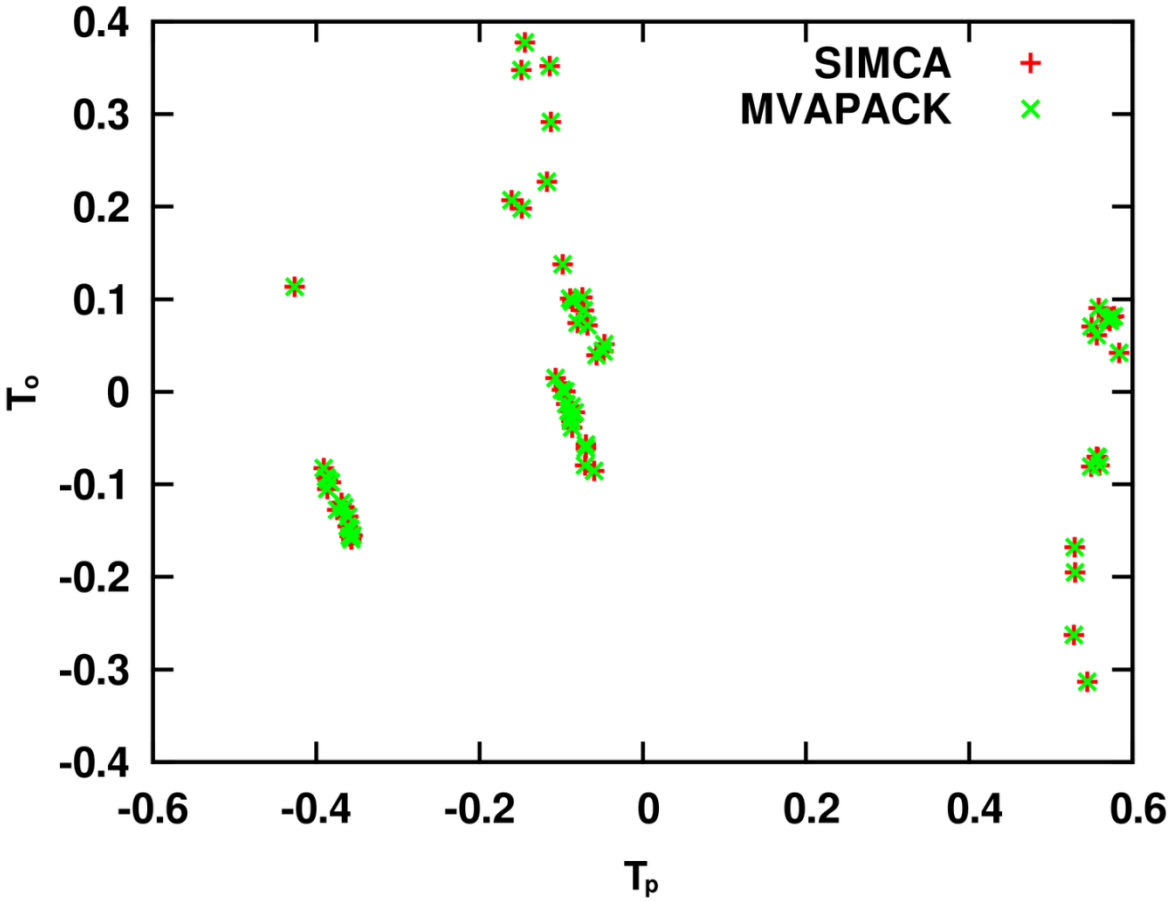

Figure 7S. Comparison of calculated OPLS-R scores from SIMCA-P+ and MVAPACK along the predictive and orthogonal component, respectively. The results are visually identical and exact to within the tolerance of the OPLS algorithm (relative error < 10^{-6}).

Supplementary References

1. Belay, A., Ture, K., Redi, M., and Asfaw, A. (2008) Measurement of caffeine in coffee beans with UV/vis spectrometer, *Food Chem 108*, 310-315.
2. Hwang, T. L., and Shaka, A. J. (1995) Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients, *J Magn Reson Ser A 112*, 275-279.
3. Nguyen, B. D., Meng, X., Donovan, K. J., and Shaka, A. J. (2007) SOGGY: Solvent-optimized double gradient spectroscopy for water suppression. A comparison with some existing techniques, *J Magn Reson 184*, 263-274.
4. Rastrelli, F., Jha, S., and Mancin, F. (2009) Seeing through Macromolecules: T-2-Filtered NMR for the Purity Assay of Functionalized Nanosystems and the Screening of Biofluids, *J Am Chem Soc 131*, 14222-+.
5. Eaton, J. W., Bateman, D., and Hauberg, S. (2008) *GNU Octave Manual Version 3*, Network Theory Limited.
6. Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics, *Anal Chem 78*, 4281-4290.
7. Sousa, S. A. A., Magalhaes, A., and Ferreira, M. M. C. (2013) Optimized bucketing for NMR spectra: Three case studies, *Chemometr Intell Lab 122*, 93-102.
8. Zhang, B., Halouska, S., Gaupp, R., Lei, S., Snell, E., Fenton, R. J., Barletta, R. G., Somerville, G. A., and Powers, R. (2013) Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes, *Journal of Integrated OMICS 3*, 120-137.

9. Savorani, F., Tomasi, G., and Engelsen, S. B. (2010) icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, *J Magn Reson 202*, 190-202.

10. (2008) IEEE Standard for Floating-Point Arithmetic, *IEEE Std 754-2008*, 1-58.

11. Eshghi, P. (2014) Dimensionality choice in principal components analysis via cross-validatory methods, *Chemometr Intell Lab 130*, 6-13.